Forecasting Regional GDP with Factor Models: How Useful are National and International Data?

Alexandre Kopoin
Kevin Moran
Jean-Pierre Paré

Corresponding author: Kevin Moran
Department of Economics, Université Laval
1026 avenue des sciences humaines, Quebec, QC Canada G1V 0A6
(418) 656-2131x5870
kmoran@ecn.ulaval.ca

June 2013

Abstract

We assess the contribution of national (country-wide) and international data to the task of forecasting the real GDP of Canadian provinces. Using the targeting predictors approach of Bai and Ng (2008) [*Journal of Econometrics 146:304-317*], we find that larger datasets containing regional, national and international data help improve forecasting accuracy for horizons below the one-year-ahead mark, but that beyond that horizon, relying on provincial data alone produces the best forecasts. These results suggest that shocks originating at the national and international levels are transmitted to Canadian regions, and thus reflected in the regional timeseries, fairly rapidly.

JEL Classification: C33, C53, C83.

Keywords: Dynamic factor models, targeted predictors, regional forecasts

1 Introduction

Forecasting regional economic timeseries is an important policy goal for provincial and state agencies, or in regional branches of central banks. To pursue that goal, data from the same region as the variables to be predicted constitute a natural first source of information, but national (country-wide) and international data could also prove important.

Does regional forecasting require all these data? This question is relevant because incorporating the information contained in several hundred regional, national, and international timeseries may prove computationally intensive. Further, even factor models, specifically designed to manage large numbers of potential predictors, have been shown to lose forecasting ability in some settings where additional timeseries are incorporated to already large datasets (Boivin and Ng, 2006).

To analyze this question, this paper uses a targeted factor modeling approach to assess how much national and international data help forecast provincial GDPs in Canada. The targeted factor approach, introduced by Bai and Ng (2008), pre-selects timeseries before estimating the common factors used to forecast. It is less susceptible to the problems documented in Boivin and Ng (2006) because it focuses on timeseries likely to contain relevant information for the variable one wishes to forecast.

We conduct several forecasting exercises for the real GDP of two Canadian provinces.[1] We start by using only provincial timeseries to forecast GDP. Next, we add data from other provinces, Canada as a whole, and from the United States, and compare forecasting ability to the one achieved using only provincial data at each step.

Our results indicate that national and US series can significantly improve the forecasting ability of targeted factor models for Canadian provincial GDPs. This effect is present only at short-term horizons, however: beyond the one-year-ahead mark, relying on provincial data alone produces the best forecasts. These results suggest that shocks originating at the national and international levels are transmitted to Canadian regions fairly rapidly, and thus regional timeseries encompass all relevant information within a short period following a shock.

Although forecasting research typically emphasizes national variables, a small

[1] The two provinces, Ontario and Quebec, are the most important economically in Canada. We limit our analysis to these provinces because of the limited availability of quarterly real GDP measures for other provinces.

but growing literature studies how best to predict important regional data. Among those, Rapach and Strauss (2010, 2012), for the United States, examine employment forecasting at the state level; Lehman and Wohlrabe (2012), for Germany, study regional GDPs; and Kwan and Cotsomitis (2006), for Canada, analyze provincial household spending. Our paper provides an important contribution to this growing literature and our results about the significance of national and international data may guide future research on this theme.

Our analysis is also related to contributions that analyze the role of international data for forecasting country-level data (Cheung and Demers, 2007; Schumacher, 2010; Eickmeier and Ng, 2010). These papers find that using larger datasets that include international data can help improve factor models' ability to forecast aggregates like national GDP. Our work extends the scope of these results, by showing that regional forecasting can also benefit from the use of larger datasets that include countrywide and international data.[2]

The remainder of this paper is organized as follows. Section 2 describes the targeted factor approach. Section 3 presents the data used and our forecasting experiments. Section 4 presents our mains findings, while Section 5 concludes.

2 Factor models with targeted predictors

Influential papers by Stock and Watson (2002a,b) have helped popularize the use of factor models for forecasting. Thee models synthesize the information contained in a large number of variables into a few factors, which are then used to help forecast the variable of interest. They have been shown to possess significant forecasting ability in a wide variety of settings and are now part of the forecaster's standard toolkit.

Let $\mathbf{X}_t = (X_{1t}, X_{2t}, ...X_{Nt})$, $i = 1, \ldots, N, t = 1, \ldots T$ represent a large number of potential predictors for y_{t+h}, the value at $t+h$ of the variable of interest (here the growth rate of regional GDP). The first step of a factor-model forecasting approach for y_{t+h} is to express each variable X_{it} as the sum of two components, as in

$$X_{it} = \lambda_i' \mathbf{F}_t + e_{it}, \tag{1}$$

where $\mathbf{F}_t$ is a $(r, 1)$ vector of factors common to all X_{it} (λ_i' collects the influence of each factors on X_{it}) and e_{it} is an *i.i.d.* idiosyncratic component. The r factors are

[2] Conversely, an interesting literature exemplified by Hernandez-Murillo and Owyang (2006) analyzes how the information contained in regional data can help forecast countrywide aggregates.

then estimated by applying the method of principal components to the covariance matrix of $\mathbf{X}_t$.[3] The second step uses the estimated factors $\mathbf{F}_t$ to help forecast y_{t+h} using

$$y_{t+h} = \alpha'\mathbf{W}_t + \beta'\mathbf{F}_t + \nu_{t+h}, \tag{2}$$

where $\mathbf{W}_t$ is a vector of predetermined variables possibly including lagged values of y_t and ν_{t+h} is the prediction error.

Under this approach, the dataset with the largest possible number of potential predictors could be, at first, expected to produce the best estimates of the common factors $\mathbf{F}_t$ and thus the superior forecasts. However, Boivin and Ng (2006) have shown that if the additional data is noisy and contains limited information about $\mathbf{F}_t$, its addition can decrease the model's forecasting ability. In addition, while the principal components are meant to explain the overall covariance structure of $\mathbf{X}_t$, the ultimate objective here is to forecast one single variable, namely y_{t+h}. It might occur that information contained in variables particularly useful for predicting y_{t+h} is deemphasized because its marginal significance is weaker for the common factors $\mathbf{F}_t$.

In this context, Bai and Ng (2008) have proposed preselecting timeseries likely to contain predictive power for y_{t+h} before including them in the data pool used to identify the common factors. To implement this *targeted* approach, they propose two classes of preliminary tests to identify the relevant variables.

The first class is labeled *hard thresholding* and consists of estimating the following regression for each candidate variable X_{it}

$$y_{t+h} = \alpha'\mathbf{W}_t + \gamma_i X_{it} + \upsilon_{t+h}, \tag{3}$$

and then sort the variables X_{it} by descending order of marginal predictive power for y_{t+h} (as measured by their t-statistic). The factor analysis then proceeds using only the variables whose t-statistic exceed a given threshold significance level t^*.

One drawback of *hard thresholding* is that by testing each variable X_i separately, the approach can end up keeping variables too similar to each other and discarding others less strongly connected to y_{t+h} but with unique information. To help identify valuable predictors for y_t while taking other predictors into account, Bai and Ng (2008) propose a *soft thresholding* approach that analyzes the system-wide regression

$$y_{t+h} = \alpha'\mathbf{W}_t + \gamma\mathbf{X}_t + \upsilon_{t+h}, \tag{4}$$

[3] Practical estimation issues include deciding how many factors to include (the value of r) and whether to include a lag structure for the factors in (1).

and estimates γ using an "elastic net" penalty that seeks to shrink parameter values and discard uninformative variables. Specifically, this is operationalized by solving the optimization problem

$$\min_{\gamma} RSS + \kappa_1 \sum_{i=1}^{N} |\gamma_i| + \kappa_2 \sum_{i=1}^{N} \gamma_i^2, \tag{5}$$

where RSS is the residual sum of squares from (4), while κ_1 and κ_2 are parameter values specified by the user. In practice, the calibration of κ_1 is recast as a rule for the maximum number of variables with non-zero γ_i included in the analysis.[4]

3 Data description and experiments

We study regional data from two economically important Canadian provinces, Quebec and Ontario. For these provinces, the datasets contain 373 and 70 times series, respectively, covering the sample 1983Q1 to 2011Q1 at a quarterly frequency. In each dataset, the available timeseries include the real GDP for each province, as well as various real activity indicators, monetary and financial indicators, GDP components on the expenditures side, and retail trade and price indicators. We also use aggregate (country-wide) data for Canada and international (US) data, arranged in two additional datasets containing 480 and 200 timeseries, respectively. The aggregate datasets also include various GDP components, price, employment, and financial data. The four separate datasets are labeled Qc, On, CA, and US. Following standard procedures in factor analysis, all series are seasonally adjusted, screened for outliers and appropriately differenced to obtain stationarity. In addition all series are standardized to $\hat{x}_t = (x_t - \overline{x})/\sigma_x$ for any variable x_t, with $\overline{x}$ and σ_x the sample mean and standard deviation of the variable, respectively.[5]

We construct forecasts for $y_{t+h} \equiv \sum_{i=1}^{h} \Delta log(GDP_{t+i})$, the cumulative growth between t and $t+h$ for each of the two provincial GDPs, using data up to and including time-t. We cover the range between $h = 1$ (one-quarter-ahead forecasts) to $h = 12$ (three-year-ahead forecasts). The goal is to determine whether national and international data contains information relevant for forecasting regional GDPs over and above that already present in the regional data.

[4] As in Bai and Ng (2008) and Schumacher (2010), we solve (5) using the *least angle regression* (LARS) algorithm, a computationally-efficient method for conducting forward selection regressions.

[5] A description of all timeseries used is available in a data appendix available from the authors.

Towards that objective, we first implement the targeted factor models in Bai and Ng (2008) using only the relevant provincial data (Qc in the case of Quebec for example). Forecasts obtained in this step serve as benchmark. We then repeat the targeted procedure by computing forecasts with the datasets $Qc + CA$, $Qc + On$, $Qc + US$, and $Qc + On + CA$. The entire procedure is repeated using data from Ontario as the benchmark, so that the datasets sequence is On, $On+CA$, $On+QC$, $On + US$ and $On + CA + Qc$.

4 Results

Table 1 to 5 present our results. Each table illustrates the implications of a specific implementation of the targeting approach in Bai and Ng (2008). In all tables, the entries report the mean squared error (MSE) of an individual forecasting experiment, *relative* to the MSE achieved using only the provincial data. Entries lower than 1 thus suggest that adding other regional, country-wide, or international data helps produce better forecasts for the provincial GDP. For example, the first element of Table 1, 0.823 reports that forecasting Quebec's GDP one-quarter ahead when data from Quebec and from Canada are used lowers the MSE by almost 18% relative to the case using only information from Quebec. The symbols *, ** and *** indicate statistically significant differences in forecasting ability according to the Giacomini and White (2006) test.

4.1 Hard Thresholding

The first targeting method advocated by Bai and Ng (2008) constructs the factors using only timeseries whose individual significance for the targeted variable is higher than a threshold t-statistic t^*. Tables 1, 2, and 3 depict results obtained using $t^* = 1.28$, 1.65, and 2.58, respectively (these correspond to 10%, 5% and 1% critical values for two-tailed t-tests).

Table 1 illustrates the case when $t^* = 1.28$. The main result of the table is that for one-quarter-ahead and two-quarter-ahead forecasts ($h = 1$ and $h = 2$) larger datasets, containing data from other provinces, country-wide aggregates or from the US, can improve the performance of our regional forecasting model in a statistically significant way. In addition, improvements are economically meaningful, reaching 19% (a relative MSE of 0.81) when forecasting Quebec's GDP and up to 38% for Ontario's GDP.[6] By contrast, additional information from the other province, from Canada-wide variables or from the United States does not improve forecasting performance for forecasting horizons of four-quarter ahead or above: instead, relying on provincial data alone is sufficient.

Table 2 (for $t^* = 1.65$) and Table 3 ($t^* = 2.58$) confirm these results. Forecasts for Quebec and Ontario's GDP can be improved by the use of extra-regional information for $h = 1$ and $h = 2$, but for longer-term horizons, regional data alone obtains the best results. For Quebec's GDP, some of the improvements in short-term forecasting accuracy are important and reach 25% in some cases.

Taken together, the results in Table 1-3 are coherent with the message from Boivin and Ng (2006): larger datasets can, but do not necessarily improve the forecasting performance of factor models. In our setting, national and international data appear to contain relevant information for forecasting at short-term horizons but not at longer-term ones. One interpretation of this general result is that shocks originating at the national and international levels are reflected relatively rapidly into regional indicators, because of close economic integration (intranational and international) in Canada.[7]

4.2 Soft Thresholding

Table 4 and 5 report results arrived at by the $LARS - EN$ soft thresholding approach. Recall that this selects variables to be included in the pool from which factors are estimated by pursuing the system-wide approach represented by (4). Two parameters, κ_1 and κ_2, need to be calibrated. Since the choice of κ_1 can be relabeled as the maximum number of variables include in the pool, we first pursue a conservative strategy where $N_{\mathcal{D}} = 30$ variables are selected from each provincial dataset. Meanwhile, κ_2 is set to 0.25 following Bai and Ng (2008). The results of this experiment are reported in Table 4. Next, Table 5 reports the results of allowing considerably more variables to enter the pool: 110 for the Quebec dataset and 75 for Ontario.

Overall, the results in Table 4 and 5 are consistent with those from Table 1-3: at horizons shorter than four-quarters-ahead, adding national and international variables can improve the forecasting performance of factor models with targeted

[6]The relatively small size of our *Ontario*-only dataset might explain the larger improvements.

[7]Similar results are obtained by Schumacher (2010), who shows that international data can help forecast national (German) GDP better, but only at horizons of less than a year. Beyond that horizon, the information contained in national data is enough to produce the best forecasts.

predictors; for longer horizons by constrast, no improvements is gained from larger datasets. However, even for horizons at which larger datasets help, Table 4 and 5 report more modest improvements in forecasting ability than those depicted in Tables 1 to 3. This may seem puzzling, as the soft thresholding approach is designed to offer a more flexible and thus efficient approach. We intrepret this finding as suggesting that soft thresholding does indeed improve noticeably forecasting ability, but by helping the provincial-only dataset, it lessens the need add information from national and international data. Said otherwise, absolute forecasting accuracy is always best with soft thresholding, but because it improves the performance of provincial datasets more, the relative MSEs reported in the tables are smaller.

5 Conclusion

Regional economic performance has important implications for state, provincial or regional policy makers. However, the development of forecasting tools aimed at predicting regional economic variables has lagged the rapid advancements that have taken place in the literature on national-level forecasting methods. The present paper contributes to bridge that gap and reports that a factor modeling approach with targeted predictors can significantly help forecasting accuracy for regional GDPs in Canada. This improved performance is present only at short horizons however, suggesting that a longer-term, regional data forecasting can be conducted with regional data alone with a careful approach and a relatively large regional dataset.

References

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146, pages 304–317, 2008.

J. Boivin and S. Ng. Are more data always better for factor analysis? *Journal of Econometrics* 132, pages 169–194, 2006.

C. Cheung and F. Demers. Evaluating forecasts from factor models for Canadian GDP growth and core inflation. *Bank of Canada Working Paper* 2007-8, 2007.

S. Eickmeier and T. Ng. Forecasting national activity using lots of international predictors: An application to New Zealand. *International Journal of Forecasting* 27, pages 496–511., 2010.

R. Giacomini and H. White. Tests of conditional forecast accuracy. *Econometrica* 77, pages 1545–1578., 2006.

R. Hernandez-Murillo and M. T. Owyang. An empirical comparison of methods for forecasting using many predictors. *Economic Letters 90*, 2006.

A. C. C. Kwan and J. A. Cotsomitis. The usefulness of consumer confidence in forecasting household spending in CANADA: A national and regional analysis. *Economic Enquiry* 12, page 185197, 2006.

R. Lehman and K. Wohlrabe. Forecasting GDP at the regional level with many predictors. *CESIFO Working Paper* 3956, 2012.

D. E. Rapach and J. K. Strauss. Bagging or combining (or both)? an analysis based on forecasting U.S. employment growth. *Econometric Reviews* 29, page 511533, 2010.

D. E. Rapach and J. K. Strauss. Forecasting US state level employment growth: An amalgamation approach. *International Journal of Forecasting* 28, pages 315–327, 2012.

C. Schumacher. Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters* 107, pages 95–98, 2010.

J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, pages 1167–1179., 2002a.

J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, pages 147–162., 2002b.

Table 1: **Forecasting Performance with Targeted Predictors**

Relative to forecasts using regional data only

Targeting Method :	**Hard Threshold with** $t^* = 1.28$				
	Forecasting horizon				
Dataset	$h=1$	$h=2$	$h=4$	$h=8$	$h=12$
Panel A: Forecasting Quebec's GDP					
$Qc+Ca$	0.823***	0.892***	1.067	1.058	0.994
$Qc+On$	0.911***	1.001**	1.039	1.049	0.992
$Qc+Us$	0.871***	0.982	1.072	1.037	0.992
$Qc+Ca+On$	0.814***	1.004	1.073*	1.072	0.992
Panel B: Forecasting Ontario's GDP					
$On+Ca$	0.660***	0.874***	1.055*	1.021	1.001
$On+Qc$	0.858***	0.847***	1.036*	1.934	1.010
$On+Us$	0.691***	0.798***	1.009	1.033	1.002
$On+Ca+Qc$	0.819***	0.822***	0.930	1.032	0.994

Note : Each entry is the ratio of the mean-squared error of the forecasts obtained with a targeted factor model using the larger dataset to that obtained when the forecasts are obtained with data from Quebec only (first panel) or Ontario only (second panel). Entries under 1 suggest the larger dataset has superior forecasting performance. The symbols *,** and *** indicate rejection of the null of equal predictive accuracy at the 10%, 5% and 1% level, respectively, according Giacomini and White (2006). The model is estimated and forecasts are computed separately for each horizon, using a rolling window of 21 periods.

Table 2: **Forecasting Performance with Targeted Predictors**

Relative to forecasts using regional data only

Targeting Method :	**Hard Threshold with** $t^* = 1.65$				
	Forecasting horizon				
Dataset	$h = 1$	$h = 2$	$h = 4$	$h = 8$	$h = 12$
Panel A: Forecasting Quebec's GDP					
$Qc + Ca$	0.748***	0.849***	1.016	1.110*	0.993
$Qc + On$	0.909***	0.956***	1.034	1.074	0.992*
$Qc + Us$	0.889***	1.015	1.076**	1.058	0.993
$Qc + Ca + On$	0.743***	0.993	1.090*	1.103*	0.991
Panel B: Forecasting Ontario's GDP					
$On + Ca$	0.739***	0.918***	1.049	1.019	1.000
$On + Qc$	0.906***	0.909***	1.068***	1.009	0.999
$On + Us$	0.730***	0.854***	1.166***	1.081*	0.997
$On + Ca + Qc$	0.762***	0.904***	1.030	1.015	1.000

Table 3: **Forecasting Performance with Targeted Predictors**

Relative to forecasts using regional data only

Targeting Method :	**Hard Threshold with** $t^* = 2.58$				
	Forecasting horizon				
Dataset	$h = 1$	$h = 2$	$h = 4$	$h = 8$	$h = 12$
Panel A: Forecasting Quebec's GDP					
$Qc + Ca$	0.810***	0.954***	1.054***	1.073	1.001
$Qc + On$	0.900***	0.995***	1.026***	1.042	1.002
$Qc + Us$	0.774***	0.983***	1.146***	1.105	1.007
$Qc + Ca + On$	0.978***	1.007***	1.059***	1.085	1.003*
Panel B: Forecasting Ontario's GDP					
$On + Ca$	0.739***	0.918***	1.049	1.019	1.000
$On + Qc$	0.906***	0.909***	1.068***	1.009	0.999
$On + Us$	0.730***	0.854***	1.166***	1.081*	0.997
$On + Ca + Qc$	0.762***	0.904***	1.030	1.015	1.000

Table 4: **Forecasting Performance with Targeted Predictors**

Relative to forecasts using regional data only

Targeting Method :	**soft threshold with** $N_{\mathcal{D}} = 30$ **(Lars-EN)**				
	Forecasting horizon				
Dataset	$h = 1$	$h = 2$	$h = 4$	$h = 8$	$h = 12$
Panel A: Forecasting Quebec's GDP					
$Qc + Ca$	0.972***	0.996***	1.005	1.026***	0.991
$Qc + On$	0.934***	0.944***	1.028*	.994	0.992*
$Qc + Us$	0.929***	1.004**	1.147***	1.065*	0.982
$Qc + Ca + On$	0.893***	1.085	1.014*	1.083	1.007
Panel B: Forecasting Ontario's GDP					
$On + Ca$	0.841***	0.898***	1.086***	1.081**	1.002
$On + Qc$	0.981***	0.975***	0.922	1.128***	0.969
$On + Us$	0.950***	0.905**	1.331***	1.432***	1.047
$On + Ca + Qc$	0.884***	0.881***	1.066***	1.052	0.999

Table 5: **Forecasting Performance with Targeted Predictors**

Relative to forecasts using regional data only

Targeting Method :	**Soft threshold with** $N_{\mathcal{Q}} = 110$ **and** $N_{\mathcal{O}} = 75$				
	Forecasting horizon				
Dataset	$h = 1$	$h = 2$	$h = 4$	$h = 8$	$h = 12$
Panel A: Forecasting Quebec's GDP					
$Qc + Ca$	0.936***	1.014***	0.918	1.011	1.033***
$Qc + On$	1.077	1.076***	0.864***	1.016	0.984
$Qc + Us$	0.881***	0.912**	0.912	1.012	0.988
$Qc + Ca + On$	0.945***	1.035***	0.928	1.009	1.031***
Panel B: Forecasting Ontario's GDP					
$On + Ca$	0.977***	1.031***	1.054***	1.063***	0.982
$On + Qc$	0.981***	0.975***	0.922	1.004	0.969
$On + Us$	0.919***	0.841***	1.163***	1.317***	1.068
$On + Ca + Qc$	0.951***	1.028***	1.021***	1.025*	0.993