CAHIERS

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

DOCUMENTS DE TRAVAIL DE LA DG TRÉSOR

Numéro 2013/02 –Juillet 2013

Prévision de court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques : impact de la sélection des variables

Stéphanie COMBES
Catherine DOZ
Jean-Marie FOURNIER

TRÉSOR
DIRECTION GÉNÉRALE

PRÉVISION DE COURT TERME DE LA CROISSANCE DU PIB FRANÇAIS À L'AIDE DE MODÈLES À FACTEURS DYNAMIQUES : IMPACT DE LA SÉLECTION DES VARIABLES

Stéphanie COMBES
Catherine DOZ
Jean-Marie FOURNIER

* **Stéphanie COMBES** est en poste à la Direction Générale du Trésor du Ministère de l'Économie et des Finances et du Ministère du Commerce Extérieur (France)
stephanie.combes@dgtresor.gouv.fr

* **Catherine DOZ** est en poste à l'Université Paris 1 Panthéon-Sorbonne et Paris School of Economics
catherine.doz@univ-paris1.fr

* **Jean-Marie FOURNIER** est en poste à l'Institut National de la Statistique et des Études Économiques (France)
jean-marie.fournier@insee.fr

TRÉSOR
DIRECTION GÉNÉRALE

Table des matières

Résumé 3
Introduction 4
1. Utilisation des modèles à facteurs dynamiques pour la prévision de court terme : rappel de l'étude menée par M.Bessec et C.Doz en 2011 6
1.1. Les horizons de prévision 6
1.2. Utilisation des modèles à facteurs dynamiques pour la prévision 7
1.2.1. Les modèles à facteurs dynamiques..... 7
1.2.2. Les équations de prévision avec ou sans prolongement des facteurs 8
1.3. L'étude en « pseudo-temps réel »..... 10
1.4. Les principales conclusions de l'étude précédente, leurs limites et les prolongements pouvant être apportés 11
1.4.1. Rappel des principales conclusions de l'étude précédente 11
1.4.2. Limites de l'étude et prolongements possibles 11

2. Spécifications plus générales des équations de prévision 13
2.1. Élimination systématique des facteurs non significatifs des équations de prévision 13
2.2. Prise en compte des retards des facteurs et de la variable cible pour la prévision 14
2.2.1. Introduction des retards des facteurs 14
2.2.2. Introduction des retards des taux de croissance du PIB 15
2.2.3. Introduction des retards des taux de croissance du PIB et des retards des facteurs 16
2.3. Modification de la prise en compte des valeurs trimestrielles des facteurs 17

3. Impact sur les performances en prévision du choix des blocs de variables utilisés 18
3.1. Présentation de la nouvelle base de données utilisée 18
3.2. Résultats 19
3.2.1. Résultats obtenus avec les 3 combinaisons de variables analogues de celles de l'étude Bessec-Doz 2011..... 20
3.2.2. Résultats obtenus en testant l'ensemble des 255 combinaisons de blocs de variables ... 21
3.2.3 Calcul des facteurs bloc par bloc 22

4. Impact sur les performances en prévision de la mise en œuvre d'un algorithme de sélection de variables fondé sur la corrélation de ces variables avec la variable d'intérêt (LARS)... 23
4.1. Principes théoriques de l'algorithme 23
4.2. Mise en œuvre pratique 25
4.3. Résultats 27
4.3.1. Impact de l'intégration des retards des exogènes, de la variable d'intérêt et du choix de la méthode de trimestrialisation sur les performances en prévision du modèle 27
4.3.2. Impact du paramétrage du LARS 28
4.3.3. Meilleures performances en prévision obtenues en fonction des stratégies de sélection de variables testées 30
4.3.4. Variables sélectionnées par le LARS en fonction de l'horizon de prévision 31
4.4. L'apport des variables financières dans la prévision 34
4.4.1. Les variables financières retenues dans la base de données 34
4.4.2. Les blocs de variables financières semblent pertinents à tous les horizons de temps 34
4.4.3. Ce constat est confirmé par l'application du LARS qui aboutit bien à la sélection de variables financières 35

Conclusion 36
Annexes 37

Résumé

Les modèles à facteurs sont de plus en plus utilisés pour la prévision de court terme du PIB par les banques centrales et les grands organismes internationaux. Ces modèles permettent en effet de résumer l'information apportée par un grand nombre de variables en un petit nombre de variables latentes appelées facteurs. Sous leur forme dynamique, ces modèles permettent rendre compte des évolutions conjointes des indicateurs qui les constituent. De plus, en recourant à des techniques d'estimation appropriées, il est possible de résoudre les problèmes posés par les différences de délais de publication des variables utilisées. De cette façon, il n'est pas nécessaire de développer différents modèles en fonction de la date de prévision et des variables disponibles à cette date-là.

Dans une précédente étude, avaient été examinées les performances en prévision de ces modèles pour la prévision du taux de croissance du PIB français sur des horizons courts, en utilisant une base constituée d'une centaine de variables parmi lesquelles des variables d'enquêtes, des indicateurs réels, des variables monétaires et financières et des indicateurs sur l'environnement international. Il a cependant été constaté que la présence d'un trop grand nombre de variables, incluant donc des variables sans pouvoir explicatif significatif pour le PIB, pouvait diminuer la qualité de la prévision. À partir d'une base de données élargie, nous étudions ici des stratégies de sélection de variables bâties pour maximiser la qualité des prévisions de la croissance. Ainsi, la constitution d'une base plus restreinte, correspondant à la combinaison optimale de blocs de variables construits par « dire d'expert » à partir de la base initiale, permettrait d'améliorer significativement les performances en prévision des modèles à tous les horizons. En revanche, le recours à un algorithme automatique de sélection des variables, liant variable d'intérêt et variables explicatives, ne semble pas améliorer significativement la qualité des prévisions de la croissance du PIB.

Abstract

Factor models have received increasing interest from central banks and international organizations due to their ability to forecast short-term activity. In these models, a large set of variables is summarized by a small set of unobservable variables, referred to as factors. This allows to take advantage of the information provided by the large set of initial variables. In their dynamic form, these models can also take into account the comovements of the underlying variables. Moreover, the estimation procedure can be adapted to handle the missing values at the end of the sample due to publication lags. Thus, it is no longer necessary to develop separate models for different forecasting dates, even if information sets differ from one date to another.

In a recent paper, the accuracy of these models in forecasting French GDP growth rate over short horizons was investigated, using a large data set of a hundred of variables including surveys, real, financial and international data. It appeared that using too large a data set, which could include irrelevant variables, could damage forecasting quality. In this paper, we use an extended data set, and we implement two distinct variable selection strategies designed to maximize forecasting accuracy. Restriction of the data set, achieved by keeping the best combination of expert-judgment-built variable subsets of the initial data set, happens to improve significantly the dynamic factor model forecasting performance at every horizon. On the contrary, restricting the number of variables using a selection procedure which aims at choosing variables according to their correlation to the forecasted series, doesn't seem to provide a real gain.

Introduction

Cette étude vise à prolonger les travaux menés par M. Bessec et C. Doz en 2011[1] pour mettre en œuvre un modèle de prévision à court terme de la croissance trimestrielle du PIB utilisant les modèles à facteurs dynamiques (MFD). Cette méthode tend en effet à se développer (banques centrales, OCDE, autres organismes internationaux…).

Les modèles à facteurs dynamiques ont pour but de fournir une synthèse de l'information conjoncturelle disponible à un moment donné, telle qu'elle ressort d'un large ensemble d'indicateurs (résultats d'enquêtes de conjoncture, indices réels d'activité, variables financières,…) disponibles à fréquence mensuelle ou trimestrielle. Cette synthèse prend la forme d'un nombre limité de facteurs communs dont la dynamique représente les évolutions conjointes de ces indicateurs.

Calculés en temps réel et éventuellement prolongés à un horizon plus ou moins lointain, ces facteurs sont ensuite utilisés comme variables explicatives dans des équations de prévision. Actuellement, la plupart des modèles à facteurs dynamiques mis en œuvre ont pour but la prévision du taux de croissance du PIB d'un pays ou de la zone euro (cf. travaux de la BCE par exemple[2]), mais leur utilisation dans d'autres domaines tend également à se développer (modèles de prévisions de l'évolution de la demande mondiale de l'OCDE, prévisions d'inflation, …).

Le succès que remportent les modèles à facteurs dynamiques ces dernières années s'explique par les avantages qu'ils procurent en comparaison d'autres techniques de prévision à court terme (étalonnages notamment). Parmi les principaux avantages, on peut citer :

- concilier deux objectifs *a priori* contradictoires, d'une part **mobiliser de manière quasi-exhaustive toute l'information disponible** et, d'autre part, éviter le « fléau de la dimension » puisque finalement un nombre limité de variables (les facteurs) est utilisé pour établir la prévision,
- **procéder à des prévisions en continu**, à mesure que sont incorporées des informations nouvelles et ce en utilisant une méthode dont la mise en œuvre est rapide et qui peut être employée même si certaines variables ne sont pas disponibles ou publiées,
- **disposer d'une méthode** pouvant être adaptée à la prévision de nombreuses variables économiques.

Cependant, diverses questions se posent lorsqu'on met en œuvre cette technique. En particulier, il est important de mesurer la sensibilité des prévisions au choix des variables utilisées pour construire les facteurs. Pour cette raison, une nouvelle étude a été lancée à l'automne 2011 avec pour objectif d'améliorer la précision et la qualité des prévisions de la croissance du PIB obtenues à partir des MFD déjà mis en œuvre au sein de la DG Trésor. Ce document de travail rend compte des tests et des progrès accomplis dans le cadre de cette étude. Il est organisé en 4 grandes parties :

La 1ère consiste en un rappel de la méthode adoptée par Bessec et Doz et des résultats obtenus. La 2ème complète cette étude en utilisant la même base de données et la même spécification du modèle mais en testant diverses spécifications alternatives de l'équation de prévision (suppressions des termes non significatifs dans l'équation de prévision, introduction des retards des facteurs et de l'endogène, modification de la prise en compte des valeurs trimestrielles des facteurs).

La 3ème partie propose une démarche d'évaluation systématique des ensembles de variables utilisés et de la forme de l'équation de prévision : en utilisant des blocs de données analogues à

[1] Marie Bessec, Catherine Doz (2011), "Prévision de court terme de la croissance du PIB français à l'aide de modèles à facteurs dynamiques", *Document de travail de la DG Trésor*, n°2011/01, Juillet.
[2] Marta Banbura, Domenico Giannone and Lucrezia Reichlin, December 2011, "Nowcasting", Oxford Handbook on Economic Forecasting, eds. Michael P. Clements and David F. Hendry: 2011.

ceux de l'étude précédente et en ajoutant quelques blocs de données complémentaires, une étude exhaustive est d'abord menée en utilisant toutes les combinaisons possibles de ces blocs, et toutes les spécifications possibles des équations de prévision. Cette étude est ensuite complétée par une approche dans laquelle les facteurs sont calculés séparément sur des blocs de variables de nature similaire (enquêtes, variables financières, indicateurs réels..) et utilisés ensuite simultanément dans les équations de prévision.

La 4ème partie étudie l'apport de l'utilisation d'un algorithme de sélection des variables (LARS) destiné à réduire la taille de la base sur laquelle sont calculés les facteurs ainsi que l'apport, en termes de précision de la prévision, de l'inclusion de variables financières dans la base pour les deux méthodes explorées dans ces deux dernières parties.

1. Utilisation des modèles à facteurs dynamiques pour la prévision de court terme : rappel de l'étude menée par M.Bessec et C.Doz en 2011

Dans cette partie sont rappelés les principaux aspects de l'étude menée par Bessec et Doz en 2011. On rappelle les horizons de prévision retenus, les principes d'utilisation des modèles à facteurs dynamiques, ainsi que le contexte d'une étude en « pseudo temps réel » et enfin les conclusions

1.1. Les horizons de prévision

Dans le cadre de cette étude, on s'intéresse à la prévision de l'évolution trimestrielle du PIB. Chaque horizon de prévision est alors défini par le nombre de mois à attendre avant l'obtention des premiers résultats du PIB publié par l'Insee environ 45 jours après la fin du trimestre concerné.

- Les prévisions pour le trimestre suivant (***Forecasting***) sont les prévisions du taux de croissance du trimestre *T* construites à partir des facteurs mensuels du trimestre *T-1*. On définit ainsi l'horizon H7 comme celui correspondant à la prévision faite à l'issue du mois 1 du trimestre *T-1*, l'horizon H6 à celle du mois 2 de *T-1* et H5 à celle du mois 3 de *T-1* ;
- Les prévisions du trimestre en cours (***Nowcasting***) consistent à prévoir la croissance à la fin de chaque mois de ce trimestre (H4 fin du mois 1, H3 fin du mois 2 et H2 fin du mois 3) ;
- Les prévisions pour le trimestre précédent (***Backcasting***) visent à estimer au cours du trimestre *T+1* la croissance du PIB du trimestre *T,* avant sa 1ère publication. Dans le cas qui nous occupe ici, cette prévision peut être réalisée à l'issue du 1er mois du trimestre *T+1*. Tant que les premiers résultats du PIB ne sont pas publiés, les « prévisions » en Backcasting ont un intérêt pour le conjoncturiste (cf. schéma 1).

En raison des différences de délais de publication des différentes variables qui sont utilisées pour extraire les facteurs mensuels, ceux-ci sont construits en fin de période alors même que certaines variables sont manquantes. Il faut en effet deux à trois mois supplémentaires pour que les variables mensuelles retenues dans l'échantillon soient toutes renseignées pour un mois donné. De ce fait, pour un trimestre cible donné, les facteurs incorporent, dans le cas du ***Backcasting***, davantage d'information conjoncturelle qu'il n'avait été possible d'en mobiliser précédemment lors des exercices de ***Nowcasting*** et de ***Forecasting***.

Schéma 1 : horizon des MFD dans le cas de la prévision du taux de croissance du PIB du trimestre T

FORECASTING / trimestre T			NOWCASTING / trimestre T			BACKASTING / trimestre T	
Données du Trimestre T-1			Données du Trimestre T			Données du Trimestre T+1	
Mois 1 / T-1	mois 2 / T-1 publications des Premiers résultats de T-2	Mois 3 / T-1	Mois 1 / T	mois 2 publications des Premiers résultats de T-1	Mois 3 / T	mois 1 / T+1	mois 2 / T+1
H7							
	H6						
		H5					
			H4				
				H3			
					H2		
						H1	

Source : DG Trésor.

Les 7 horizons mensuels cibles sont donc associés aux 2 ou 3 prévisions mensuelles (cf. schéma 2).

Schéma 2 : calendrier des prévisions pouvant être mises en œuvre chaque mois par des MFD

	A-1 / T4	A / T1	A / T2	A / T3	A / T4	A+1 / T1
Janvier	B (H1)	N (H4)	F (H7)			
Février		N (H3)	F (H6)			
Mars		N (H2)	F (H5)			
Avril		B (H1)	N (H4)	F (H7)		
Mai			N (H3)	F (H6)		
Juin			N (H2)	F (H5)		
Juillet			B (H1)	N (H4)	F (H7)	
Août				N (H3)	F (H6)	
Septembre				N (H2)	F (H5)	
Octobre				B (H1)	N (H4)	F (H7)
Novembre					N (H3)	F (H6)
Décembre					N (H2)	F (H5)

*Lecture : **B** pour **Backcasting**, **N** pour **Nowcasting** et **F** pour **Forecasting ;** entre parenthèses (horizon de prévision).*

Source : DG Trésor.

1.2. Utilisation des modèles à facteurs dynamiques pour la prévision

1.2.1 Les modèles à facteurs dynamiques

On rappelle ici très brièvement les principales notions sur les modèles à facteurs dynamiques. Une présentation plus détaillée est fournie dans la partie 2 du Document de Travail de Bessec et Doz.

De façon générale, les modèles à facteurs ont pour but de fournir une représentation parcimonieuse de l'information apportée par un grand nombre de variables lorsque ces variables sont corrélées. En particulier, les modèles à facteurs dynamiques rendent compte de la dynamique commune aux variables observées (ou des « co-mouvements » des variables observées). Plus précisément, on suppose que les variables observées peuvent être décrites en fonction d'un petit nombre de variables latentes inobservables, appelées facteurs, dont la dynamique rend compte de l'essentiel de la dynamique commune aux variables. Dans les modèles considérés ici, on suppose que le vecteur formé par ces facteurs latents admet une représentation VAR.

Formellement, si l'on note $x_t = (x_{1t}, \dots, x_{nt})'$, le vecteur des variables observées, et f_t le vecteur des facteurs, on suppose que :

$$x_t = \Lambda_0 f_t + \cdots + \Lambda_s f_{t-s} + e_t \qquad (1)$$

$$f_t = \sum_{i=1}^{p} A_i f_{t-i} + u_t \qquad (2)$$

avec (u_t) un bruit blanc, et e_t non corrélé à f_τ quelles que soient les dates *t* et τ.

Si f_t a q composantes, on dit que q est le nombre de facteurs dynamiques. On peut aussi réécrire le modèle sous la forme d'un modèle à facteurs statique :

$$\begin{cases} x_t = \Lambda F_t + e_t \\ F_t = A F_{t-1} + \underbrace{B\varepsilon_t}_{\varsigma_t} \quad \text{avec} \quad V(\varsigma_t) = \Sigma_\varsigma \end{cases}$$

en posant $F_t = (f_t', \dots, f_{t-s}')'$ et on retient un nombre r de facteurs statiques (c'est-à-dire le nombre de composantes de F_t) tel que $r = q(s+1)$. Dans l'étude, on n'utilise en pratique que cette représentation statique du modèle, et le nombre de facteurs dynamiques n'intervient que par les critères qui permettent de choisir la spécification du modèle.

1.2.2 Les équations de prévision avec ou sans prolongement des facteurs

Si on note y_t la variable à prévoir (ici la croissance du PIB), on s'intéresse à la prévision faite pour y_{T+h} à la date *T*, avec $h = -1$ (***Backasting***), $h = 0$ (***Nowcasting***) ou $h = 1$ (***Forecasting***). Cette prévision est ici obtenue par l'intermédiaire d'une équation reliant les valeurs de la variable à prévoir à celles des facteurs. Lorsque, comme c'est le cas ici, la variable à prévoir est une variable trimestrielle alors que les données utilisées, et donc aussi les facteurs estimés, sont à valeurs mensuelles, la prévision repose donc d'abord sur une *trimestrialisation* de ces facteurs puis sur une régression de la variable d'intérêt sur les facteurs *trimestrialisés* - régression dont la forme dépend de la méthode employée (cf. *infra*).

La valeur *trimestrialisée* f_t^Q du facteur à la date *t* est, dans le cadre de cette étude, calculée comme une moyenne arithmétique des valeurs estimées aux différents mois du trimestre. En pratique, cette moyenne sera calculée de façon différente suivant la position du mois dans le trimestre et suivant la méthode utilisée pour la prévision (cf. *infra*).

Deux types principaux d'équations de prévision sont proposés dans la littérature et ont été utilisées dans l'étude de 2011.

Dans les premiers articles apparus dans la littérature, l'approche consiste à estimer par MCO le modèle :

$$y_{t+h} = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^Q + \varepsilon_{t+h} \qquad (3a)$$

et à calculer ensuite la prévision de y_{T+h} à la date *T* en utilisant la formule suivante :

$$\hat{y}_{T+h|T} = \hat{\delta}_0 + \sum_{i=1}^{r} \hat{\delta}_i f_{i,T}^Q \qquad (3b)$$

Lorsqu'on utilise cette approche dans le cadre de cette étude, on le fait en pratique en utilisant deux équations :

- pour la prévision du trimestre suivant (***Forecasting***), resp. courant (***Nowcasting***), à partir des équations (3a) et (3b) avec $h = 1$, *resp* $h = 0$;
- pour $h = -1$ (***Backcasting***), on estime l'équation (3a) pour $h = 0$, et on calcule la prévision sous la forme :

$$\hat{y}_{T-1|T} = \sum_{i=1}^{r} \hat{\delta}_i f_{i,T-1}^Q \qquad (3c)$$

Dans le cadre de cette première approche, aucune prévision des facteurs n'est effectuée. Les valeurs *trimestrialisées* des facteurs sont calculées comme la moyenne empirique des valeurs mensuelles qui ont été calculées en estimant le modèle à facteurs. En particulier, la valeur *trimestrialisée* retenue pour le trimestre *T* est la moyenne des valeurs mensuelles calculées pour ce trimestre : il s'agit donc d'une moyenne sur une, deux ou trois valeurs, suivant la date où est faite la prévision (mois 1, 2 ou 3 de *T*).

Une autre approche est spécifiquement liée au cadre dynamique et utilise l'estimation de la dynamique des facteurs. On estime d'abord par les moindres carrés ordinaires l'équation reliant le taux de croissance du PIB aux facteurs *trimestrialisés* qui lui sont contemporains :

$$y_t = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^Q + \varepsilon_t \qquad (4a)$$

(cette équation coïncide avec l'équation (3a) lorsque celle-ci est estimée avec $h = 0$).

On calcule ensuite une prévision de y_{T+h} à la date *T* à partir d'une prévision du facteur *trimestrialisé*. Si les facteurs vérifient un modèle de la forme : $f_t = \sum_{i=1}^{p} A_i f_{t-i} + \varepsilon_t$, on obtient une prévision mensuelle $f_{T+m|T}$ de f_{T+m} à la date *T* pour tous les mois *m* couvrant la fin du trimestre en cours (si la prévision est faite aux mois 1 ou 2 du trimestre *T*) et le trimestre suivant. Cette prévision s'obtient de façon récursive en utilisant les valeurs estimées des matrices A_i et des facteurs ; on calcule ensuite la moyenne des prévisions mensuelles obtenues, pour déterminer une prévision $f^{Q}_{i,T+h|T}$ du facteur *trimestrialisé*. On obtient alors une prévision de y_{T+h} à la date *T* en utilisant la formule suivante :

$$\hat{y}_{T+h|T} = \hat{\delta}_0 + \sum_{i=1}^{r} \hat{\delta}_i f^{Q}_{i,T+h|T} \qquad (4b)$$

Dans cette approche, la même formule est employée quelle que soit la valeur de h $(h = 1, 0\ ou - 1)$.

Dans leur étude, Bessec et Doz ont aussi utilisé une troisième approche, qui s'inspire des deux précédentes. Dans cette approche, on utilise les équations (3a) et (3b), mais en modifiant la démarche lorsque la date *T* à laquelle la prévision est faite correspond au 1er ou au 2e mois d'un trimestre : dans ce cas, l'équation (3a) est utilisée comme précédemment, mais à partir d'une prévision des facteurs sur les mois suivants du trimestre concerné, estimée par la représentation VAR des facteurs. La prévision est ensuite obtenue en appliquant l'équation (3b) au facteur *trimestrialisé* associé.

On peut donc résumer les trois méthodes employées par Bessec et Doz de la façon suivante :

- **Méthode 1 : utilisation d'une seule équation de prévision après prolongation des facteurs mensuels** (et leur *trimestrialisation*). On estime une seule équation (4a), qui relie les valeurs contemporaines du taux de croissance du PIB et des facteurs *trimestrialisés*. Les prévisions ***Forecasting***, ***Nowcasting*** et ***Backcasting*** sont calculées de façon identique (équation 4b) en utilisant les valeurs prévues ou estimées des facteurs.
- **Méthode 2 : une équation de prévision spécifique pour chaque horizon de prévision sans prolongement des facteurs mensuels**. La prévision ***Forecasting*** est alors établie à partir de l'équation (3a) associée à $h = 1$, tandis que les prévisions ***Nowcasting*** et ***Backcasting,*** sont obtenues à partir de l'équation (3a) associée à $h = 0$.
- **Méthode 3 : une équation de prévision spécifique pour chaque horizon de prévision mais en prolongeant les facteurs mensuels à l'horizon du trimestre en cours**. Il s'agit d'une version mixte des deux méthodes précédentes donnant des résultats différents seulement pour les 2 premiers mois du trimestre. Elle est identique à la méthode 2 dans le cas du 3e mois du trimestre.

Pour le ***Nowcasting***, la méthode 2 est identique à la méthode 1 dans le cas du mois 3 du trimestre, date à laquelle il n'y a pas besoin de prolonger les facteurs mensuels avant de les *trimestrialiser*; la méthode 3 n'a pas lieu d'être. Pour le ***Backasting***, toutes les méthodes sont équivalentes puisque cette prévision est calculée pendant le mois qui suit le trimestre à prévoir.

Le schéma 3 (ci-dessous) résume les développements précédents et présente entre les différentes méthodes d'estimation disponibles en fonction de l'horizon de prévision visé.

Schéma 3 : méthodes d'estimation des modèles de prévision de la croissance du PIB du trimestre T

Données disponibles à la fin		Horizons de prévision		Méthode d'estimation des facteurs mensuels		
T-1	*mois 1*	*H7*	"Forecasting"	Méthode 1	Méthode 2	Méthode 3
T-1	*mois 2*	*H6*		Méthode 1	Méthode 2	Méthode 3
T-1	*mois 1*	*H5*		Méthode 1	Méthode 2 = Méthode 3	
T	*mois 1*	*H4*	"Nowcasting"	Méthode 1	Méthode 2	
T	*mois 2*	*H3*		Méthode 1	Méthode 2	
T	*mois 3*	*H2*		Méthode 1 = Méthode 2		
T+1	*mois 1*	*H1*	"Backcasting"	Méthode 1 = Méthode 2		

Lecture :

- *Méthode 1 : une seule équation et les facteurs mensuels sont prolongés avant trimestrialisation ;*
- *Méthode 2 : deux équations, et les facteurs ne sont pas prolongés ;*
- *Méthode 3 : deux équations et les facteurs sont prolongés seulement sur le trimestre en cours.*

Source : DG Trésor.

1.3. L'étude en « pseudo-temps réel »

Pour évaluer les performances d'une technique de prévision, il est classique de calculer les prévisions qui auraient été fournies par le modèle avec les données qui étaient effectivement disponibles à une date donnée. Ce calcul est fait sur une plage de dates couvrant dix années et les prévisions obtenues sont ensuite comparées avec les vraies valeurs du taux de croissance du PIB aux dates concernées.

De façon plus précise, l'analyse est réalisée comme suit. L'échantillon utilisé pour estimer le modèle à facteurs contient 93 variables conjoncturelles publiées à fréquence mensuelle ou trimestrielle sur la période 1990T1-2009T3 (cf. annexe 2). Ces variables ont été mensualisées puis différenciées pour les *stationnariser* si nécessaire. On retire d'abord de l'échantillon les données correspondant aux 39 derniers trimestres (on ne conserve que les données de la période 1990T1-1999T4). Ce retrait est fait en reproduisant les conditions qui auraient été observées en temps réel, c'est-à-dire en tenant compte des délais de publication des variables. Le modèle à facteurs est estimé sur l'ensemble de données ainsi obtenu, les prévisions ***Forecasting***, ***Nowcasting***, et ***Backcasting*** sont calculées avec les trois méthodes présentées ci-dessus, et on calcule les erreurs de prévision associées. Ensuite l'échantillon est augmenté pas à pas, d'une observation à chaque étape, et pour chacun des sous-échantillons obtenus, tous les calculs sont répliqués : estimation du modèle à facteurs, calcul des prévisions et des erreurs de prévision. L'exercice est poursuivi jusqu'à ce que l'ensemble de l'échantillon disponible ait été utilisé.

Il s'agit cependant d'un exercice en « pseudo temps réel » : si l'on tient effectivement compte des délais de publications des séries, il n'est cependant pas possible de tenir compte des révisions apportées aux variables conjoncturelles au gré de leurs publications. Ces révisions peuvent être importantes dans le temps, notamment pour les variables ayant le plus d'influence sur la variable cible[3].

Cet exercice a été réalisé pour des modèles à facteurs estimés sur trois ensembles de données : le premier modèle est estimé en utilisant les 93 variables, le deuxième en n'utilisant que les indicateurs habituellement considérés par les conjoncturistes, *i.e.* les soldes d'enquête et les variables réelles, le 3e en utilisant toutes les variables sauf les variables réelles (dont les délais de publication sont les plus longs).

[3] C'est le cas des indices IPI ou de la consommation des ménages qui sont des inputs directs du calcul du PIB trimestriel.

1.4. Les principales conclusions de l'étude précédente, leurs limites et les prolongements pouvant être apportés

1.4.1 Rappel des principales conclusions de l'étude précédente

À l'issue de cette expérience, pour chaque ensemble de données et chaque méthode de prévision qui ont été considérés, on dispose d'une série d'erreurs de prévision pour chacun des horizons de prévision. Les précisions relatives des différentes méthodes d'estimation sont évaluées en analysant les critères d'évaluation classiques : le *Root Mean Square Forecast Error* (RMSFE) et *le Mean of the Absolute Forecast Errors* (MAFE) qui mesurent la dispersion des erreurs de prévision et leur amplitude moyenne. On procède en outre à un calcul permettant d'évaluer l'aptitude des modèles à prévoir les mouvements de la série d'intérêt (probabilité de bien capter en prévision le sens de variation du taux de croissance trimestriel du PIB). Les principales conclusions de cette étude étaient les suivantes :

- **La qualité des prévisions tirées des modèles MFD s'améliore de façon quasi-continue** à mesure que l'on s'approche de la date de publication des premiers résultats du PIB.
- **Concernant la méthode de prévision**, on constate que, pour le ***Forecasting***, les modèles les plus performants sont ceux qui prolongent les facteurs mensuels. Le prolongement des facteurs mensuels permet également d'améliorer la précision des prévisions établies au 1er mois du trimestre en cours et éventuellement au 2e mois de ce même trimestre.
- **Concernant le choix des variables** entrant dans la construction des facteurs mensuels, la prise en compte des variables réelles n'apporte un gain en termes de prévision qu'en toute fin de période (*i.e.* lorsqu'il s'agit de prévoir la croissance du PIB en ***Backcasting***), et lorsqu'elles sont associées aux résultats des enquêtes de conjoncture de l'Insee. Elles apparaissent éventuellement utiles au mois 3 de *T-1* (***Forecasting***) associées à toutes les autres variables de l'échantillon. Cependant, il faut préciser que les performances en prévision ainsi obtenues ne sont que très légèrement supérieures à celles qu'on obtient en ne retenant pas ces variables – modèles à 65 variables. Surtout, l'étude de Bessec-Doz montre tout l'intérêt d'introduire et de tester la pertinence pour la prévision de la croissance à court terme du PIB (***Forecasting*** et ***Nowcasting***) de variables rarement utilisées dans ce cadre (notamment des variables monétaires et financières ou bien représentatives de la conjoncture internationale).

1.4.2 Limites de l'étude et prolongements possibles

L'étude précédente comportait un certain nombre de limites auxquelles nous avons remédié dans cette étude :

1.4.2.1 Spécification du modèle

Dans l'étude menée par Bessec et Doz, la spécification du modèle (c'est-à-dire le nombre de facteurs dynamiques et statiques retenus, ainsi que l'ordre du VAR sur ces facteurs) a été définie une fois pour toutes, en utilisant l'ensemble de l'échantillon disponible et sans tenir compte des délais de publication des variables conjoncturelles[4]. En ce sens, l'analyse n'était pas entièrement faite en « pseudo-temps réel » : pour cela, il aurait fallu que la spécification du modèle soit, elle aussi, refaite à chaque date (*i.e.* que les critères permettant d'aboutir à la spécification soient recalculés sur chaque pseudo-échantillon, en tenant compte des délais de disponibilité des données).

[4] Ainsi, en considérant l'échantillon complet (période 1990T1-2009T3), les critères de Bai-Ng ont conduit à retenir un total de 10 facteurs communs statiques (cf. le document de travail de Bessec-Doz (2011) pour la description des critères utilisés).

1.4.2.2 Facteurs contribuant à la prévision

De même, les facteurs jugés non significatifs dans une équation de prévision ont été retirés de cette équation. Mais l'analyse de la significativité des facteurs a été faite en échantillon complet. Pour que cette analyse soit entièrement menée en « pseudo-temps réel », il aurait fallu que la significativité des facteurs dans les équations de prévision soit testée pour chacun des pseudo-échantillons.

1.4.2.3 Spécificité des prévisions au mois 1 du trimestre

L'étude menée par Bessec et Doz ne prenait pas en compte une spécificité des modèles estimés au mois 1 du trimestre liée au délai de publication du PIB (pour rappel : 45 jours après la fin du trimestre cible). En effet, le PIB du trimestre précédent n'est pas connu lors de l'estimation des prévisions faites au mois 1 d'un trimestre (que ce soit pour le ***Forecasting*** –, le ***Nowcasting*** ou le ***Backcasting***). Les modèles ne peuvent alors être estimés qu'en prenant en compte les résultats du PIB jusqu'au trimestre *T-2* (cf. le schéma 4 ci-dessous).

Schéma 4 : différents millésimes des comptes nationaux trimestriels disponibles chaque mois

	versions des comptes nationaux trimestriels disponibles	
Informations disponibles à la date du :	T-2	T-1
mois 1 de T	Résultats détaillés	
mois 2 de T		Premiers Résultats
mois 3 de T		Résultats détaillés

Lecture : au mois 1 de T, seuls les résultats détaillés des CNT de T-2 sont disponibles.
Source : DG Trésor.

Par ailleurs, d'autres prolongements peuvent être envisagés :

1.4.2.4 Spécification des équations de prévision (cf. section 2)

Les valeurs retardées du taux de croissance du PIB peuvent être introduites dans les équations de prévision puisqu'elles sont, de façon générale, souvent utilisées pour la prévision. De même, il peut être pertinent d'introduire les valeurs retardées des facteurs eux-mêmes dans les équations de prévision, puisque ces valeurs retardées des facteurs rendent compte des valeurs passées des variables présentes dans la base de données.

1.4.2.5 Sélection des variables utilisées pour construire les facteurs

L'étude précédente a montré qu'une base trop importante pouvait nuire à la précision des prévisions, mais la meilleure façon de la réduire n'est pas évidente. Il semble donc important d'étudier plus en détail l'impact de la composition de la base de données sur les performances en prévision des modèles à facteurs.

Sur une base de données étendue, obtenue en introduisant d'autres variables (variables financières), ou des transformations des variables initiales (évolutions des soldes d'enquête par exemple), nous étudions donc deux façons de procéder à la sélection d'une base restreinte sur laquelle sera appliqué le modèle à facteurs :

L'approche dite « par blocs » qui consiste à séparer la base en différentes catégories de variables, les blocs, et à tester différentes combinaisons de blocs afin de ne garder que ceux dont la présence dans la base contribue à améliorer les performances en prévision du modèle à un horizon donné (cf. section 3).

L'algorithme LARS, *Least Angle Regression Shrinkage*, qui sélectionne automatiquement les variables de la base en fonction de leur corrélation avec la variable d'intérêt (cf. section 4).

2. Spécifications plus générales des équations de prévision

Diverses spécifications ont été étudiées en mobilisant l'échantillon de l'étude précédente (Bessec et Doz 2011) et en suivant la même démarche d'estimation en « pseudo temps réel ». Ainsi, ces tests sont menés dans les conditions les plus proches possibles de celles ayant cours lors d'un exercice de prévision ***Forecasting***, ***Nowcasting*** et éventuellement ***Backcasting*** pouvant être mis en œuvre chaque mois à partir de modèles à facteurs dynamiques. Ce sont également les trois mêmes sous-ensembles de variables choisies précédemment qui ont été retenus dans la partie de l'étude présentée ici.

Pour l'évaluation des performances prédictives, trois échantillons différents avaient été testés :

- **MF1** : 93 variables (blocs 1 à 4 : toutes variables) ;
- **MF2** : 55 variables (blocs 1 et 2 : variables d'enquêtes Insee et variables réelles) ;
- **MF3** : 65 variables (blocs 1, 3 et 4 : toutes variables sauf réelles).

Nous présentons ci-dessous les résultats obtenus après prise en compte du problème soulevé au 1.4.2.3.

Tableau 1 : *Benchmarks* selon le protocole de l'étude Bessec-Doz 2011

Horizon	Trimestre cible	Mois du trimestre T	RMFSE	MAFE	Proba de capter le sens de variation	échantillon	Méthode d'estimation des facteurs trimestriels
H7	T+1	1	0,46	0,33	0,61	MF3	Méthode 3
H6	T+1	2	0,45	0,34	0,63	MF3	Méthode 1
H5	T+1	3	0,41	0,32	0,61	MF1	Méthode 2
H4	T	1	0,38	0,29	0,63	MF3	Méthode 1
H3	T	2	0,35	0,28	0,53	MF3	Méthode 1
H2	T	3	0,37	0,28	0,58	MF3	Méthode 1
H1	T-1	1	0,28	0,22	0,63	MF2	Méthode 1

Lecture : Sont présentés dans ce tableau, les « meilleurs » résultats des évaluations des prévisions obtenues « en pseudo temps réel » à partir des trois échantillons de variables mensuelles MF1, MF2 et MF3, au sens du Root Mean Forecasts Standard Errors (RMFSE). Sont également présentés ici les Mean Average Forecasts Errors (MAFE) et la probabilité estimée de capter le sens véritable de la variation du taux de croissance du PIB (hausse ou baisse). Les performances en prévision relatives de ces différents modèles ont été étudiées en testant chacune des trois méthodes de prévision présentées au point 1.2.2 de cette présente note.

- *Méthode 1 : une seule équation et les facteurs mensuels sont prolongés avant trimestrialisation ;*
- *Méthode 2 : deux équations, et les facteurs ne sont pas prolongés ;*
- *Méthode 3 : deux équations et les facteurs sont prolongés seulement sur le trimestre en cours.*

Source : DG Trésor.

Rappelons cependant que plusieurs combinaisons de variables ou différentes méthodes employées pour la prévision conduisent à des résultats très proches de ces « meilleurs » résultats présentés dans le tableau 1.[5]

2.1 Élimination systématique des facteurs non significatifs des équations de prévision

On présente ici les résultats obtenus en éliminant systématiquement dans les équations de prévision (***Backcasting***, ***Nowcasting*** et ***Forecasting***), les facteurs non significatifs au vu du test de Student. La procédure est itérée, le nombre de fois nécessaire, pour ne conserver

[5] Par exemple, pour l'horizon de prévision H6, la performance relative des prévisions obtenues en « pseudo temps réel » pour l'échantillon MF3 est pratiquement identique que l'on utilise la méthode 1 pour estimer les facteurs mensuels ou la méthode 3 (RMFSE = 0,46 ; MAFE = 0,33 et probabilité de capter le bon sens de la variation du PIB = 0,63)[5].

finalement que les variables pour lesquelles, le T de Student[6] est supérieur à 2. En revanche, la « significativité » de la constante n'est pas testée dans cette procédure[7]. Ceci est fait en « pseudo temps réel », c'est-à-dire pour chaque sous-échantillon concerné.

Les erreurs de prévision moyennes ainsi obtenues conduisent à des résultats finalement très comparables aux résultats présentés précédemment (Tableau 1).

Tableau 2 : précisions relatives des prévisions obtenues en ne conservant que les facteurs significatifs

Horizon	Trimestre cible	Mois du trimestre T	RMFSE	MAFE	Probabilité de capter le bon sens de variation	échantillon	Méthode d'estimation des facteurs trimestriels
H7	T+1	1	0,46	0,33	0,68	MF3	Méthode 2
H6	T+1	2	0,46	0,33	0,71	MF3	Méthode 3
H5	T+1	3	0,40	0,31	0,66	MF3	Méthode 2
H4	T	1	0,35	0,27	0,63	MF3	Méthode 2
H3	T	2	0,33	0,26	0,61	MF3	Méthode 1
H2	T	3	0,33	0,26	0,55	MF3	Méthode 1
H1	T-1	1	0,27	0,21	0,63	MF2	Méthode 1

Lecture : cf. tableau 1.
Source : DG Trésor.

2.2 Prise en compte des retards des facteurs et de la variable cible pour la prévision

Il peut être intéressant de tester d'autres équations de prévision faisant intervenir notamment les retards des facteurs et/ou de la variable endogène[8]. Pour ce faire, il est nécessaire de procéder à l'élimination systématique des variables non significatives, en raison du nombre de variables explicatives potentielles au regard du nombre d'observations effectivement mobilisables en « pseudo temps réel ».

2.2.1 Introduction des retards des facteurs

On teste ici des spécifications des équations de prévision incorporant les retards d'ordre 1 des facteurs. L'introduction de retards supplémentaires conduirait à utiliser un nombre trop élevé de variables explicatives (en particulier en début de période, lorsque l'on ne dispose que d'une quarantaine d'observations).

Les estimations des modèles s'expriment alors de la manière suivante, avec y_t la série à prévoir (à l'horizon h) :

$$y_{t+h} = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^{Q} + \sum_{j=1}^{r} \beta_j f_{j,t-1}^{Q} + \varepsilon_{t+h} \qquad \text{(3a bis)}$$

$$y_t = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^{Q} + \sum_{j=1}^{r} \beta_j f_{j,t-1}^{Q} + \varepsilon_t \qquad \text{(4a bis)}$$

L'intégration des retards à l'ordre 1 des facteurs permet également de tester dans les modèles l'apport en prévision, non seulement des niveaux des facteurs, mais aussi des variations

[6] Valeur de T ayant 5 % de chance d'être dépassée en valeur absolue avec n compris entre 40 et 80.
[7] De par la forme des modèles, la constante mesure la croissance moyenne observée au cours de la période sous-revue.
[8] Si l'introduction de retards des facteurs modifie peu la procédure, celle des retards de l'endogène oblige à tenir compte à nouveau de spécificités pour le 1er mois du trimestre.

trimestrielles conjoncturelles de ces derniers[9] au trimestre *T* puisque $\delta_i f_{i,t}^Q + \beta_i f_{i,t-1}^Q$ peut aussi s'écrire : $\delta_i f_{i,t}^Q + \varphi_i.(f_{i,t-1}^Q - f_{i,t-1}^Q)$.

Tableau 3 : précisions relatives des modèles intégrant les retards à l'ordre 1 des facteurs

Horizon	Trimestre cible	Mois du trimestre T	RMFSE	MAFE	Probabilité de capter le bon sens de variation	échantillon	Méthode d'estimation des facteurs trimestriels
H7	T+1	1	0,45	0,33	0,58	MF3	Méthode 3
H6	T+1	2	0,43	0,33	0,68	MF3	Méthode 3
H5	T+1	3	0,39	0,31	0,71	MF3	Méthode 2
H4	T	1	0,41	0,30	0,74	MF3	Méthode 2
H3	T	2	0,38	0,30	0,61	MF3	Méthode 1
H2	T	3	0,36	0,27	0,66	MF3	Méthode 1
H1	T-1	1	0,27	0,21	0,63	MF2	Méthode 1

Lecture : cf. tableau 1.
Source : DG Trésor.

Les résultats présentés dans le tableau 3 montrent que cette modification de la spécification des équations de prévision ne permet pas d'obtenir un gain en termes de précision par rapport à la procédure Bessec-Doz.

2.2.2 Introduction des retards des taux de croissance du PIB

Les équations de prévision testées jusqu'à maintenant n'intégraient pas de termes autorégressifs. Or les équations utilisées pour la prévision à court terme font parfois intervenir les valeurs passées de la variable à prévoir. Dans certains cas, il peut s'avérer en effet utile d'utiliser également l'information fournie par les taux de croissance passés pour mieux calibrer la prévision de croissance.

Comme pour l'estimation des équations au mois 1 du trimestre (cf. partie 1 de ce document), la prise en compte des évolutions passées de la croissance pour la prévision doit tenir compte des délais de publication du PIB. La significativité des endogènes retardées est testée, et elles peuvent être éventuellement écartées.

Les équations de prévisions ont été estimées sous la forme suivante :

$$y_{t+h} = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^Q + \sum_{k=1}^{2} \gamma_k y_{t+h-k} + \varepsilon_{t+h} \qquad \text{(3a ter)}$$

$$y_t = \delta_0 + \sum_{i=1}^{r} \delta_i f_{i,t}^Q + \sum_{k=1}^{2} \gamma_k y_{t-k} + \varepsilon_t \qquad \text{(4a ter)}$$

Les résultats obtenus pour cette classe d'équations (cf. les tableaux ci-dessous) sont pratiquement identiques à ceux des équations qui n'incorporent pas de retards des taux de croissance du PIB. Tout au plus le sens de variation de la croissance du PIB apparaît mieux prévu avec ce type de modèles.

[9] Cette forme est également testée avec succès dans les modèles d'étalonnages des résultats d'enquêtes de conjoncture.

Tableau 4 : précisions relatives des modèles intégrant les retards du taux de croissance du PIB

Horizon	Trimestre cible	Mois du trimestre T	RMFSE	MAFE	Probabilité de capter le bon sens de variation	échantillon	Méthode d'estimation des facteurs trimestriels
H7	T+1	1	0,47	0,34	0,71	MF3	Méthode 2
H6	T+1	2	0,44	0,33	0,74	MF3	Méthode 3
H5	T+1	3	0,39	0,30	0,68	MF3	Méthode 2
H4	T	1	0,35	0,28	0,68	MF3	Méthode 2
H3	T	2	0,32	0,25	0,74	MF3	Méthode 1
H2	T	3	0,34	0,26	0,66	MF2	Méthode 1
H1	T-1	1	0,27	0,21	0,66	MF2	Méthode 1

Lecture : cf. tableau 1.
Source : DG Trésor.

2.2.3 Introduction des retards des taux de croissance du PIB et des retards des facteurs

Au vu des résultats précédents, une forme encore plus générale des équations de prévision a également été testée, intégrant à la fois les retards des facteurs à l'ordre 1 et des retards de l'endogène (éventuellement complétés par les prévisions établies pour les trimestres précédents). Les équations (3a) et (4a) intègrent les retards des facteurs et de l'endogène tels qu'ils ont été introduits dans les équations (3a bis) et (3a ter), et dans les équations (4a bis) et (4a ter).

Les résultats ainsi obtenus apparaissent comparables à ceux qui l'ont été en introduisant seulement des retards des facteurs. Notamment, le sens de variation du taux de croissance du PIB est particulièrement bien rendu par ces modèles à tous les horizons de prévision. Quant à la précision des prévisions, elle n'a finalement pas été améliorée par rapport aux spécifications retenues dans l'étude précédente.

Tableau 5 : précisions relatives des modèles intégrant les retards du taux de croissance du PIB et des facteurs

Horizon	Trimestre cible	Mois du trimestre T	RMFSE	MAFE	Probabilité de capter le bon sens de variation	échantillon	Méthode d'estimation des facteurs trimestriels
H7	T+1	1	0,44	0,33	0,74	MF3	Méthode 3
H6	T+1	2	0,42	0,32	0,74	MF3	Méthode 3
H5	T+1	3	0,41	0,32	0,79	MF3	Méthode 2
H4	T	1	0,34	0,26	0,68	MF3	Méthode 1
H3	T	2	0,37	0,28	0,71	MF3	Méthode 2
H2	T	3	0,33	0,26	0,63	MF3	Méthode 1
H1	T-1	1	0,30	0,23	0,66	MF2	Méthode 1

Lecture : cf. tableau 1.
Source : DG Trésor.

Les résultats obtenus pour les différentes spécifications étudiées sont regroupés dans le tableau ci-dessous. Il apparaît qu'aucune spécification n'améliore significativement les performances pour tous les horizons de façon frappante.

Tableau 6 : précisions relatives des modèles présentés précédemment, en bleu foncé, les modèles les plus performants aux différents horizons

Horizon	Trimestre cible	Mois du trimestre T	*Benchmark* (cf. Tableau 1)	Modèle avec facteurs significatifs (cf. Tableau 2)	Modèle avec retards des facteurs (cf. Tableau 3)	Modèle avec retard de PIB (cf. Tableau 4)	Modèle intégrant les retards du PIB et des facteurs (cf. Tableau 5)
H7	T+1	1	**0,46**	0,46	0,45	0,47	0,44
H6	T+1	2	**0,45**	0,46	0,43	0,44	0,42
H5	T+1	3	**0,41**	0,40	0,39	0,39	0,41
H4	T	1	**0,38**	0,35	0,41	0,35	0,34
H3	T	2	**0,35**	0,33	0,38	0,32	0,37
H2	T	3	**0,37**	0,33	0,36	0,34	0,33
H1	T-1	1	**0,28**	0,27	0,27	0,27	0,30

Lecture : en bleu turquoise, les horizons et spécifications pour lesquels les RMSFE sont inférieurs au Benchmark, en bleu foncé, ceux pour lesquels les RMSE sont inférieurs de plus de 0,03 point.
Source : DG Trésor.

2.3 Modification de la prise en compte des valeurs trimestrielles des facteurs

La modélisation utilisée pour la prévision nécessite d'avoir recours à la *trimestrialisation* des facteurs calculés sur une base mensuelle. Cette *trimestrialisation*, consistant simplement à ce stade à prendre la moyenne trimestrielle des valeurs mensuelles des facteurs, peut être jugée arbitraire et susceptible d'entraîner la perte d'une partie de l'information. En effet, la moyenne trimestrielle des variables, en particulier lorsqu'elles représentent un flux et non un stock, ne semble pas pertinente pour résumer l'information disponible sur l'ensemble du trimestre. Mais la caractéristique même des facteurs, qui est d'agréger un grand nombre de variables de nature différente, ne permet pas de procéder beaucoup plus finement à l'étape de la *trimestrialisation*.

Dans ce contexte, il pourrait sembler plus rigoureux théoriquement de considérer non pas la moyenne trimestrielle des facteurs mais les trois variables à fréquence trimestrielle obtenues à partir d'un facteur mensuel correspondant aux valeurs prises aux premier, deuxième et troisième mois de chaque trimestre respectivement. Les coefficients de ces variables, estimés dans les équations de prévision du taux de croissance du PIB, ne seraient donc plus contraints et la contribution réelle de chaque mois d'un trimestre pourrait ainsi être correctement estimée. Dans ce but, pour chacune des méthodes 1,2[10] et 3 présentées et utilisées dans l'étude « Bessec-Doz », nous avons implémenté des variantes reposant non plus sur les facteurs *trimestrialisés* mais sur les trois variables trimestrielles obtenues à partir des valeurs mensuelles de chaque facteur.

L'implémentation de cette modélisation a abouti à des résultats moins satisfaisants qu'en utilisant une *trimestrialisation* plus simple. Deux arguments peuvent justifier cela : la démultiplication des variables explicatives dans l'équation de prévision qui la rend moins précise, et la forte probabilité de voir apparaître de la multicolinéarité entre les variables ainsi construites, ne présentant notamment plus des propriétés de quasi orthogonalité qui sont un bénéfice usuel de l'approche factorielle.

De plus, cette méthode, contrairement à son usage dans les étalonnages du taux de croissance du PIB construits sur les seuls soldes d'enquêtes, nécessite tout de même de prolonger certaines variables (celles qui ne sont pas encore disponibles à la date où est faite la prévision).

En conclusion, les résultats de cette section montrent que des spécifications plus générales des équations de prévision n'apportent que des gains minimes en termes de précision des prévisions. Il apparaît donc nécessaire d'explorer d'autres pistes d'amélioration, notamment la sélection des variables utilisées pour construire les facteurs.

[10] Dans le cas particulier de la méthode 2 qui n'utilise pas de prolongation des facteurs par VAR, on a en fait trois modèles puisqu'en fonction du mois du trimestre auquel on se trouve, on ne disposera que d'un, deux ou des trois mois des facteurs utilisés pour la prévision.

3. Impact sur les performances en prévision du choix des blocs de variables utilisés

L'étude menée par Bessec et Doz en 2011 avait montré que l'utilisation de toutes les variables d'une base de données pour construire un MFD ne constituait pas toujours la meilleure solution. Certains articles académiques avaient conduit aux mêmes types de conclusions[11].

Dans cette partie, on se propose d'évaluer les gains potentiels en termes de précision des prévisions obtenues en adaptant la base de données et en répliquant de manière plus systématique les principes de la méthode établie par Bessec-Doz 2011. De la sorte, il apparaît possible d'améliorer la précision des modèles MFD pour les exercices de prévision ***Nowcasting*** et ***Backcasting***.

3.1 Présentation de la nouvelle base de données utilisée

La base de données de l'étude Bessec-Doz (2011) était constituée de 93 variables, qui avaient été regroupées en quatre blocs selon la nature des données. Par rapport à cet ensemble, un premier tri a conduit à exclure de l'échantillon un certain nombre de variables et à en ajouter d'autres, par exemple, des transformations des soldes d'enquêtes, traditionnellement utilisées par les prévisionnistes.

Dans le détail, les quatre blocs de variables mensuelles considérés par Bessec et Doz sont désormais modifiés de la manière suivante :

- **Bloc 1** (**23 variables mensuelles, disponibles depuis janvier 1991** – cf. annexe 1) : *constitué des principaux soldes d'opinion issus des enquêtes mensuelles de conjoncture de l'Insee*. La plupart d'entre eux entrent dans la définition des indices de climats des affaires publiés par l'Insee. Y étaient ajoutées précédemment des variables d'enquêtes trimestrielles dont les résultats ne sont disponibles qu'avec retard. Ainsi, elles n'ont pas été reprises dans la nouvelle base de données.
- **Bloc 2** (**25 variables mensuelles disponibles depuis décembre 1990** - cf. annexe 1) : *constitué des variables réelles, dont la plupart sont utilisées en tant qu'inputs par les comptables nationaux pour construire les comptes trimestriels et déterminer le taux de croissance trimestriel du PIB.* Comme pour le bloc 1, certaines de ces variables n'ont pas été reprises dans la base de données en raison de leur disponibilité tardive tandis que d'autres ont été ajoutées en raison de leurs liens avec les résultats des comptes nationaux (données douanières du commerce extérieur).
- **Bloc 3** (**22 variables mensuelles disponibles depuis janvier 1991** – cf. annexe 1) : *constitué de variables nominales* monétaires et financières (taux d'intérêts, pente des taux, indices boursiers, etc.). De la même manière que pour les deux blocs précédents, certaines des variables prises en compte dans l'étude précédente ont été supprimées ou remplacées par d'autres dans la base de données.
- **Bloc 4** (**16 variables sont prises en compte dans ce bloc depuis janvier 1992** – cf. annexe 1) : *indicateurs de l'environnement international* (taux de change de l'euro et indicateurs conjoncturels des principaux partenaires économiques). Au contraire des autres blocs, aucune variable n'a été supprimée ou ajoutée par rapport à l'étude précédente.

[11] Voir Boivin-Ng, (2006), "Are more data always better for Factor analysis ?", *Journal of Econometrics* 132, 169-194.

En outre, on a fait le choix de retenir sous la forme de blocs de données complémentaires certaines transformations des variables déjà utilisées pour construire les facteurs, en l'occurrence les soldes d'opinion issus des enquêtes de conjoncture de l'Insee. Ces soldes étaient, pour la plupart, déjà intégrés en niveau dans l'échantillon. Leurs transformations suivantes ont été ajoutées :

- **Bloc 5** : *variations mensuelles des soldes d'opinion des enquêtes de conjoncture de l'Insee*, qui sont utilisées sous cette forme dans les outils de prévision développés par le bureau PREV3 de la DG Trésor (étalonnages des enquêtes) avec un certain succès. **Au total, ce bloc est constitué de 20 variables, disponibles depuis février 1991** (cf. annexe 1).
- **Bloc 6** : *soldes d'opinion des enquêtes de conjoncture de l'Insee au carré signé (à savoir les soldes pris au carré mais dont le signe est conservé*) : ceci permet d'introduire dans les modèles des éléments de non-linéarité[12]. Ces non-linéarités peuvent contribuer à rendre compte du caractère particulier de la crise de 2008-2009 au cours de laquelle les soldes d'opinion ont atteint des niveaux historiquement bas (au cours du premier trimestre 2009), assez proches de ceux observés lors du précédent épisode récessif (du printemps 1992 au printemps 1993) tandis que le recul du PIB avait alors été moins marqué. Ce bloc est également constitué de 20 variables, disponibles depuis janvier 1991 (cf. annexe 1).

Enfin, d'autres blocs de variables conjoncturelles ont été ajoutés à l'échantillon. Il s'agit des résultats des enquêtes mensuelles de la Banque de France auprès des chefs d'entreprises de l'industrie manufacturière, auxquels a été ajouté l'indice de climat des affaires dans les services marchands. Les questions posées par la Banque de France, différentes de celles des enquêtes Insee, apportent un éclairage complémentaire. Ces données présentent l'avantage d'être disponibles rapidement (le 10 du mois *M+1*) sur une période aussi longue que celle couverte par les autres blocs de variables retenus)[13].

Ainsi deux blocs contenant ces variables ont également été ajoutés à l'échantillon :

- **Bloc 7** (**12 variables disponibles depuis septembre 1991** - cf. annexe 3) *Soldes d'opinion et indices de climat des affaires issus des enquêtes de conjoncture réalisées par la Banque de France auprès des chefs d'entreprises*.
- **Bloc 8** (**11 variables disponibles depuis octobre 1991** - cf. annexe 3). *évolution mensuelle des soldes d'opinion et des indices de climat des affaires issus des enquêtes de conjoncture réalisées par la Banque de France auprès des chefs d'entreprises*. Comme pour les résultats des enquêtes Insee, ces variables correspondent aux différences premières des soldes d'opinion mensuels.

3.2 Résultats

Pour prolonger l'étude, la démarche de sélection des données, préalable à l'utilisation des MFD, a été systématisée. Plutôt que de tester les performances en prévision des seules trois combinaisons de blocs de variables de l'étude de 2011, ce sont toutes les combinaisons possibles des huit blocs de variables définis dans la section précédente qui ont été étudiées systématiquement, à chaque horizon de prévision (H1 à H7), soit 255 combinaisons constituées de un à huit blocs de variables.

[12] Ce type d'approche avait été retenue par Bai and Ng, (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics*, 146.

[13] La disponibilité des données sur une période suffisamment longue est le principal critère de choix et d'intégration d'une variable dans la base de données mensuelle. De ce point de vue, les indices d'activité PMI, qui fournissent une information précoce aussi utile que les résultats d'autres enquêtes de conjoncture pour la prévision du taux de croissance du PIB, ne peuvent être utilisés car ils ne sont disponibles que sur une période courte (depuis le printemps 1998). Il en va de même pour des données financières et monétaires qui ne sont disponibles que depuis la création de la zone euro (janvier 1999).

Toutefois, pour que les performances en prévision des modèles réalisés à partir de différentes sélections de données soient bien comparables, nous n'avons pas utilisé tels quels les résultats publiés dans l'étude précédente. Nous avons construit un nouveau *benchmark* en tenant compte des points soulevés en 1.4.2. et des modifications de la base décrites en 3.1 : la spécification et l'élimination des variables non significatives ont donc été refaites en « pseudo temps réel » pour chaque sous-échantillon, la spécificité propre au premier mois de chaque trimestre a été respectée, et ce *benchmark* a été construit en utilisant la nouvelle composition des blocs 1 à 4[14].

3.2.1 Résultats obtenus avec les 3 combinaisons de variables analogues à celles de l'étude Bessec-Doz 2011

Ces résultats (cf. tableaux 6) constituent un nouveau *benchmark* permettant de relativiser l'amélioration relative de la précision des prévisions qui pourra être obtenue en mobilisant d'autres variables ou combinaisons de variables que celles définies précédemment.

Tableau 6 : modèles les plus précis au sens du RMFSE et MAFE pour le nouveau *benchmark*

	Horizon	RMFSE	Combinaison	Méthode	Proba. sens de variation
Forecasting (m1 / T-1)	**H7**	0,38	71	M3	52 %
Forecasting (m2 / T-1)	**H6**	0,38	71	M3	52 %
Forecasting (m3 / T-1)	**H5**	0,34	71	M2=M3	64 %
Nowcasting (m1 / T)	**H4**	0,34	71	M1	57 %
Nowcasting (m2 / T)	**H3**	0,33	11	M1	67 %
Nowcasting (m3 / T)	**H2**	0,29	11	M1=M2	69 %
Backcasting (m1 / T+1)	**H1**	0,24	11	M1=M2	71 %

Lecture des tableaux :
M1 : une seule équation et les facteurs mensuels sont prolongés avant trimestrialisation ;
M2 : deux équations pour les trimestres courant et suivant, facteurs non prolongés ;
M3 : deux équations pour les trimestres courant et suivant avec prolongement des facteurs (trimestre en cours).
Source : DG Trésor.

Tableau 7 : nature des variables composant les blocs retenus dans les modèles jugés les plus précis[15]

Horizon de prévision	Combinaison	Enquêtes Insee, en niveau	Enquêtes Insee, en variation	Enquêtes Insee, au carré (signé)	Indices réels	Indices nominaux	Indices étrangers	Enquêtes BdF, en niveau	Enquêtes BdF, en variation
H7	71	x				x	x		
H6	71	x				x	x		
H5	71	x				x	x		
H4	71	x				x	x		
H3	11	x			x				
H2	11	x			x				
H1	11	x			x				

Source : DG Trésor.

[14] La construction de ce benchmark a donc impliqué de réduire la période d'estimation, celle-ci commence désormais au mois de février 1992, date à laquelle débute la série mensuelle la plus récente dans la nouvelle base de données.

[15] Parmi les 255 combinaisons de blocs de variables, la combinaison n°11 correspond aux blocs de variables de l'échantillon MF2 de l'étude précédente (*i.e.* composé des enquêtes de conjoncture et des variables réelles), la combinaison n°71 aux blocs de variables de l'échantillon MF3 et la combinaison n°93 à l'échantillon MF1, non représentée ici puisque les modèles fondés sur cet échantillon ont toujours été dominés par d'autres modèles.

3.2.2 Résultats obtenus en testant l'ensemble des 255 combinaisons de blocs de variables

Comme indiqué précédemment, ce sont 255 combinaisons de blocs de variables qui peuvent être testées de manière à isoler, à chaque horizon, celle qui permet d'obtenir les prévisions les plus précises en « pseudo temps réel » à chaque horizon de prévision (H7 à H1). Les résultats des « meilleures » combinaisons de variables sont présentés dans le tableau 7 ci-dessous.

Tableau 8 : modèles les plus précis au sens du RMFSE parmi les 255 combinaisons

	Horizon	RMSFE	Blocs
Forecasting (m1 / T-1)	H7	0,35	Variables nominales, données dures et enquêtes Insee
Forecasting (m2 / T-1)	H6	0,35	Variables nominales, indicateurs internationaux et enquêtes Insee
Forecasting (m3 / T-1)	H5	0,32	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m1 / T)	H4	0,28	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m2 / T)	H3	0,21	Données dures, enquêtes Banque de France
Nowcasting (m3 / T)	H2	0,23	Données dures, indicateurs internationaux, enquêtes Banque de France
Backcasting (m1 / T+1)	H1	0,21	Données dures, indicateurs internationaux, enquêtes Banque de France

Source : DG Trésor.

Il convient tout d'abord de noter que les précisions des prévisions obtenues avec d'autres combinaisons de blocs de variables sont parfois proches de celles présentées ci-dessus. Ainsi, il est probable que si l'on modifie encore la structure des blocs de données testés, d'autres combinaisons de variables que celles présentées ici puissent être sélectionnées aux différents horizons de prévision.

Il semble que cette méthode très systématique permette de significativement améliorer les RMSFE à tous les horizons. Les performances en prévision des modèles se révèleraient donc relativement sensibles à la composition de la base. Cependant, il ne paraît pas souhaitable de changer de variables à chaque horizon, voire à chaque ré-estimation du modèle.

Comme d'autres combinaisons de blocs de variables que celles présentées dans les tableaux ci-dessus permettent d'obtenir des prévisions de qualité similaire à un horizon donné, il peut sembler opportun d'utiliser les mêmes combinaisons de variables pour la prévision à des horizons comparables afin de garder une certaine cohérence et homogénéité dans les blocs de variables utilisés au mois le mois. Ceci conduit à la sélection de blocs de variables proposée dans le tableau 9.

Tableau 9 : modèles finalement retenus

	Horizon	RMSFE	Blocs
Forecasting (m1 / T-1)	H7	0,38	Variables nominales, indicateurs internationaux et enquêtes Insee
Forecasting (m2 / T-1)	H6	0,35	Variables nominales, indicateurs internationaux et enquêtes Insee
Forecasting (m3 / T-1)	H5	0,33	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m1 / T)	H4	0,28	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m2 / T)	H3	0,22	Données dures, indicateurs internationaux, enquêtes Banque de France
Nowcasting (m3 / T)	H2	0,23	Données dures, indicateurs internationaux, enquêtes Banque de France
Backcasting (m1 / T+1)	H1	0,21	Données dures, indicateurs internationaux, enquêtes Banque de France

Source : DG Trésor.

On retrouve des résultats analogues à ceux qui avaient été obtenus dans l'étude de Bessec-Doz (2011) : aux horizons de prévisions éloignés, les variables pertinentes sont les enquêtes, les variables nominales et les indicateurs internationaux. L'inclusion des données dures n'améliore les prévisions qu'aux horizons rapprochés (à partir du mois 2 du trimestre en cours). En revanche, l'étude systématique qui a été menée permet de mettre en évidence un apport

spécifique des enquêtes de la Banque de France, ainsi que l'apport des indicateurs internationaux pour presque tous les horizons de prévision.

3.2.3 Calcul des facteurs bloc par bloc

Une alternative à la méthode présentée précédemment consiste à estimer un modèle à facteurs sur chacun des blocs de variables, et à utiliser l'ensemble ou une partie des facteurs ainsi estimés pour calculer les prévisions (en pratique on ne retient alors qu'un petit nombre de facteurs pour chaque bloc de variables). L'interprétation des variables explicatives intervenant dans les équations de prévision est plus intuitive, puisque les facteurs sont obtenus par regroupements de grandes catégories de variables et certains articles de la littérature ont montré que, dans certains contextes, ce type d'approche pouvait améliorer les prévisions[16].

Les principes de modélisation sont les mêmes que précédemment, que ce soit pour le calcul des facteurs, des tests de performance en prévision ou encore l'estimation des équations de prévision. En revanche, les facteurs sont ici calculés de façon indépendante pour chacun des différents blocs de variables dont la description a été donnée précédemment. En raison du nombre élevé de facteurs qui peuvent être potentiellement utilisables à chaque horizon de prévision, il a été décidé d'éliminer systématiquement les facteurs jugés non significatifs[17], et de ne pas utiliser tous les blocs disponibles.

Les résultats de cette étude peuvent être résumés comme suit :

En moyenne sur la période d'estimation, les équations qui fournissent les prévisions les plus précises donnent des résultats moins bons que ceux obtenus à partir de modèles dont les facteurs synthétisent l'information contenue dans la base de données globale et ce à tous les horizons de prévision.

Les modèles construits à partir des premiers facteurs de différents blocs de données peinent davantage à prévoir le sens de variation de la croissance au trimestre le trimestre, notamment pour les horizons de prévision proches (***Nowcasting*** et ***Backcasting***).

La méthode de prévision jugée la plus performante consiste à prolonger les facteurs obtenus pour chaque bloc de données à l'horizon de prévision et à utiliser la même équation de prévision quel que soit l'horizon visé (méthode 1).

Les données utilisées changent selon l'horizon de prévision (***forecasting***, ***nowcasting*** et ***backcasting***), conclusion qui rejoint celles de l'étude Bessec-Doz (2011).

La prise en compte des retards des facteurs et de l'endogène ne semble pas en mesure d'améliorer significativement la précision relative des prévisions. En revanche, cela semble permettre d'un peu mieux prévoir le sens de variation du PIB ce qui rejoint les conclusions de la partie 2 de cette étude.

Remarquons que cette modélisation, consistant à utiliser les premiers facteurs de différents blocs de variables, permet d'une certaine façon d'éliminer les blocs de variables qui ne contiennent finalement pas ou peu d'information pertinente. Leurs premiers facteurs ne sont alors pas retenus dans l'équation de prévision lorsque l'on procède à l'élimination des facteurs non significatifs. Cette façon de procéder permet donc d'effectuer - indirectement – une sélection des variables à prendre en compte dans des modèles à facteurs. Celle-ci dépend toutefois du choix préalable de la structure des blocs de données.

[16] Voir par exemple Massimiliano Marcellino, James H. Stock, et Mark W. Watson (2003), "Macroeconomic forecasting in the Euro area : Country specific versus area-wide information", *European Economic Review*, Elsevier, vol. 47, pages 1-18, February.

[17] Une méthode alternative aurait pu consister à mettre en œuvre les méthodes de sélections des variables utilisées dans les modèles d'étalonnages des d'enquêtes de conjoncture (méthode GETS ou procédures STEPWISE notamment).

4. Impact sur les performances en prévision de la mise en œuvre d'un algorithme de sélection de variables fondé sur la corrélation de ces variables avec la variable d'intérêt (LARS)

Nous considérons dans ce chapitre une alternative à la « *méthode par blocs* » présentée précédemment, et qui consiste en une automatisation de la sélection des variables pertinentes pour la prévision. Cette sélection aboutit à la constitution d'une base restreinte que l'on utilise ensuite pour faire des prévisions par MFD.

En effet, si le choix des blocs peut se justifier au regard de l'expertise du conjoncturiste et se mettre en œuvre simplement, ainsi que faciliter une cohérence d'ensemble entre les différents modèles à partir de jeux de blocs en partie communs, on pourrait tout de même considérer comme arbitraire, dès lors qu'on adopte l'idée selon laquelle trop de variables nuit à la précision[18] de la prévision, de conserver l'ensemble des variables d'un même bloc ou, inversement, de rejeter l'ensemble des variables des blocs non retenus. De plus, si la prévision de la croissance de l'activité à court terme a fait l'objet de suffisamment d'attention pour que l'expertise permette un choix rationnel et pertinent des variables contributrices à un horizon et un mois de prévision donné, cette approche pourrait constituer un véritable apport pour la prévision de variables pour lesquelles on ne dispose pas à l'heure actuelle de modèles satisfaisants, et permettrait *ex post* de trouver un sens économique aux variables retenues par l'algorithme de sélection statistique.

Il existe différents algorithmes de sélection de variables qui consistent, en général, à garder à chaque itération la variable explicative la plus corrélée à la variable endogène (au premier tour de boucle) ou au résidu obtenu après estimation de l'endogène en utilisant les premières variables retenues (non nécessairement par régression linéaire). Dans la lignée de l'approche proposée par Bai et Ng[19], notre attention s'est portée sur l'algorithme LARS[20] qui présente l'avantage de pouvoir retenir différentes variables, corrélées à l'endogène, pouvant présenter une forte corrélation entre elles, ce que d'autres algorithmes n'autorisent pas toujours.

L'utilisation d'un algorithme automatique de sélection de variables permet de raffiner la démarche adoptée jusqu'ici, en ce sens où l'on passe d'une approche où l'on teste toutes les configurations possibles à une approche faisant appel à une mesure objective de la pertinence de la présence des variables dans la base par le truchement de tests statistiques. Comme dit plus haut, cela présente l'avantage d'être systématique, et donc d'être applicable à la prévision de toute variable d'intérêt. Avant de détailler plus avant les résultats obtenus, nous allons tout d'abord revenir sur ce qui fait la spécificité méthodologique du LARS par rapport à d'autres algorithmes.

4.1 Principes théoriques de l'algorithme

Pour illustrer l'intérêt méthodologique du LARS, commençons par présenter deux autres algorithmes. Le *forward stepwise regression* procède par itération, en sélectionnant comme $k+1^{ème}$ variable, la variable la plus corrélée avec le résidu de la régression de la variable d'intérêt sur les k premières variables retenues. Ainsi, par construction, une variable assez corrélée avec un ou plusieurs des régresseurs déjà retenus ne pourra pas *a priori* être retenue, ce qui peut être jugé regrettable dans la mesure où certaines variables corrélées (telles que le solde d'opinion relatif à la production passée dans une enquête de conjoncture et l'indice de production industrielle par exemple) peuvent toutes deux s'avérer tout à fait pertinentes dans le cadre de la prévision de la croissance de l'activité.

[18] Jean Boivin, Serena Ng, (2006), "Are more data always better for factor analysis ?" *Journal of Econometrics* 132.
[19] Jushan Bai, Serena Ng, (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics*, 146.
[20] *Least Angle Regression Shrinkage.*

Pour remédier à ce problème, l'algorithme *forward stagewise* permet de sélectionner les variables plus finement au prix d'un grand nombre d'itérations. En effet, à chaque itération, l'ajustement de l'endogène est recalculé en ajoutant à l'estimateur obtenu à l'étape précédente une portion infinitésimale de la dernière variable sélectionnée. À chaque étape, on dispose ainsi d'un nouveau résidu et on cherche, parmi l'ensemble des variables, la variable la plus corrélée au résidu ainsi construit. La valeur ajustée de l'endogène est alors actualisée, comme décrit précédemment, en associant à cette variable un coefficient de valeur absolue infinitésimale (choisie à l'avance) et dont le signe est celui de la corrélation entre cette dernière et le résidu. Si la variable avait déjà été sélectionnée précédemment, cela revient à augmenter, en valeur absolue, le coefficient associé à ce régresseur.

$$\hat{y}_{i+1} = \hat{y}_i + \varepsilon \, sign(\hat{c}_k) \, \hat{x}_k$$

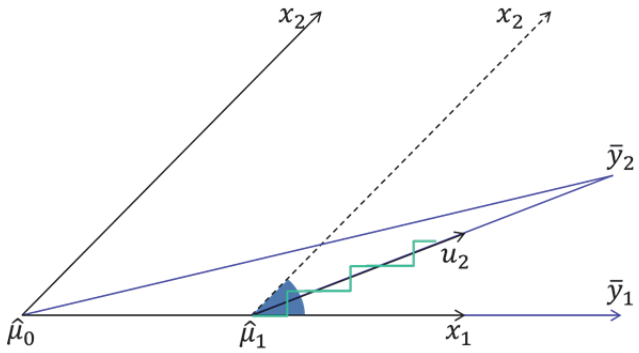
$$\text{où } \hat{c}_k = \hat{x}_k{}'(y - \hat{y}_i)$$

Chaque itération n'aboutit donc pas toujours à l'intégration d'une nouvelle variable à la sélection, et il n'est pas possible de prévoir le nombre d'itérations nécessaires à la sélection d'un nombre de variables préalablement fixé.

Le LARS présente l'avantage d'éviter ces tâtonnements puisqu'il permet de calculer directement la distance à parcourir le long de la $k^{ième}$ direction avant de rencontrer une nouvelle variable équi-corrélée avec le résidu. Il actualise ensuite l'ajustement en se déplaçant dans la direction formant le même angle avec chacune des variables sélectionnées (une direction généralisant la notion de bissectrice dans le plan).

Comme on peut le voir sur le graphique ci-dessous, les différentes étapes du *forward stagewise* se concrétisent par une oscillation autour de la solution du LARS puisqu'à chaque itération, on se déplace dans la direction d'une des variables sélectionnées et non dans la direction de la combinaison optimale de ces variables.

Figure 1 : sélection de deux variables par les algorithmes LARS et Stagewise



Lecture : Soit $\bar{y}_2$ la projection de y, sur le plan défini par les variables x_1 et x_2. Le LARS s'initialise à $\hat{\mu}_0 = 0$, le résidu $\bar{y}_2 - \hat{\mu}_0$ est plus corrélé avec x_1.que x_2. L'ajustement est donc actualisé comme suit : $\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 x_1$ où $\hat{\gamma}_1$ est choisi de telle sorte que le résidu $\bar{y}_2 - \hat{\mu}_1$ forme le même angle avec x_1 et x_2. À l'étape suivant, l'ajustement devient $\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$, obtenu en se déplaçant le long de la bissectrice de x_1 et x_2, de vecteur unité u_2. La trajectoire en escalier représente les ajustements successifs calculés par l'algorithme Stagewise. On voit qu'elle converge vers la solution du LARS lorsque ε tend vers 0.
Source : illustration inspirée de « Least Angle Regression », de B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, (2004), «The Annals of Statistics», 32.

La démarche suivie dans le cadre de la mise en œuvre du LARS permet d'obtenir le même nombre d'itérations que dans l'algorithme *forward stepwise regression* tout en procédant à une sélection plus proche de celle du *forward stagewise*. Notons que l'utilisation d'une approche factorielle par la suite permet de gérer les éventuels problèmes de multicolinéarité.

Les algorithmes LASSO et LARS-EN que l'on peut rencontrer dans la littérature sont des cas particuliers du LARS. Ils ont la particularité de pouvoir s'écrire explicitement comme un problème de minimisation des carrés des résidus sous contrainte.

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 \quad \text{sous contrainte} \quad (1-\alpha)|\beta| + \alpha\|\beta\|^2 \leq t^{21}$$

Dans notre contexte, le vecteur solution est interprété comme suit : si β_i est nul, la variable *i* de la base n'a pas été retenue par l'algorithme, dans le cas contraire, elle a été retenue. L'idée étant de s'approcher le plus possible de la solution optimale des moindres carrés linéaires en faisant une concession sur le nombre de variables utilisées. On favorise donc la parcimonie au détriment de la précision de l'ajustement.

Dans le cas du LASSO, la pénalisation consiste à majorer la solution en utilisant la norme L1, ce qui revient à poser $\alpha = 0$. Cela peut entraîner l'annulation de façon assez agressive de la plupart des composantes de β. C'est bien là l'intérêt principal de cet algorithme par rapport à une majoration par la norme L2 (ou norme euclidienne) qui n'entraînera quasiment jamais l'annulation totale d'un coefficient, et s'avère donc peu efficace pour sélectionner les données (*Ridge regression*)[22].

Le LASSO peut donc s'avérer trop parcimonieux et se montre très sensible au choix de la borne *t*. En revanche la régression *Ridge* sera peu sensible au choix de *t* mais ne permettra pas de réellement réduire le nombre de variables retenues. À cet égard, le LARS-EN ou *Elastic Net* peut leur être préféré. En effet, il fait intervenir les deux normes, par le biais de pondérations, et permet ainsi de ne pas rejeter trop systématiquement les variables et de ne pas être trop sensible au choix du paramètre *t*. Il présente également l'avantage de pouvoir être réécrit sous la forme d'un LASSO, avec un changement de variable adéquat, ce qui rend son implémentation aisée. Cette variante du LARS est notamment préconisée quand le nombre de variables initial est supérieur au nombre d'observations. C'est donc cet algorithme que nous avons utilisé.

4.2 Mise en œuvre pratique

La mise en œuvre pratique de l'algorithme présente des avantages mais aussi des inconvénients. Son principal avantage réside dans le faible coût de calcul de l'algorithme, il n'est donc pas problématique d'agrandir la base de données à volonté. On ajoute donc, pour chaque variable, un certain nombre de retards afin que le LARS, dans son procédé de sélection, détermine automatiquement avec quel retard une variable pertinente peut maximiser la corrélation et, autrement dit, tirer le meilleur parti pour la prévision du caractère avancé ou coïncident des variables disponibles.

Cependant, l'utilisation d'un algorithme de sélection, quel qu'il soit, soulève une difficulté importante liée à la disponibilité, à des fréquences différentes, des données utilisées pour le calcul des facteurs, en général mensuelles, et de l'endogène, publiée à un rythme trimestriel. Ainsi, l'apport d'un algorithme de sélection des variables par rapport au modèle à facteurs traditionnel qui calcule les facteurs indépendamment de la variable à prévoir, peut s'en trouver significativement réduit ici, puisqu'on ne peut pas exploiter, lors de cette étape déterminante, toute l'information disponible dans la base de données. En effet, bien que les facteurs soient finalement toujours calculés sur une base mensuelle constituée des variables sélectionnées *via* le LARS, l'étape de la sélection ne peut se faire, quant à elle, que sur données de même fréquence, soit ici trimestrielle. Non seulement, il s'ensuit une perte d'information et potentiellement un biais dans la sélection des variables, mais en plus, cela suggère de réfléchir au concept de *trimestrialisation* des données qui n'est pas anodin. Dans la littérature, lorsque cela est précisé par les auteurs, ce qui est loin d'être systématique, il semblerait que la

[21] La norme L1 d'un vecteur correspond à la somme des valeurs absolues des composantes. Dans le cas d'un scalaire, il s'agit simplement de sa valeur absolue. *t* est ici une valeur fixée à l'avance.

[22] Une intéressante illustration graphique de ce constat peut être trouvée en figure 2 de « Least Angle Regression », de B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, (2004), *The Annals of Statistics*, 32.

trimestrialisation consiste simplement à prendre la valeur moyenne d'une variable sur les trois mois d'un même trimestre. Nous optons pour deux variantes : la *trimestrialisation* par une moyenne simple et la *trimestrialisation* par la formule de passage du taux de croissance mensuel en taux de croissance trimestriel[23].

Au-delà de cette limite, la mise en œuvre du LARS complexifie significativement l'implémentation du modèle. En effet, pour chacune des trois méthodes présentées précédemment, il est nécessaire d'appliquer le LARS à des arguments compatibles avec la logique de la méthode choisie. Par exemple, les méthodes 1 et 3 reposent sur la régression de l'endogène sur des facteurs mensuels prolongés par filtre de Kalman sur le trimestre en cours ou suivant avant qu'ils ne soient *trimestrialisés*. Il s'agit donc, à l'étape du LARS, de comparer l'endogène à des variables explicatives qui soient cohérentes avec la façon dont les facteurs seront ensuite construits. Le cas échéant, on prolonge donc la base de données mensuelle initiale par filtre de Kalman, avant de la *trimestrialiser,* et d'appliquer le LARS[24]. Ce qui nécessite, pour établir la représentation espace-état[25], de faire une première estimation de facteurs, sur la base tout entière. On pourrait, pour simplifier, prolonger les données par acquis ou encore par ARMA mais cela pourrait biaiser la sélection de données et ne semble pas pertinent au regard des équations de prévision utilisées par la suite. A noter que dans l'ensemble, ces procédures augmentent considérablement le temps de calcul de l'algorithme, alors que la rapidité de la mise en œuvre était initialement considérée comme un des avantages d'une telle approche.

Ces aspects purement techniques mis de côté, l'utilisation des résultats obtenus en sélection les variables de cette façon doit être réalisée avec circonspection. En effet, pour chaque mois et chaque horizon de prévision, il est vraisemblable que l'ensemble de variables sélectionné par le LARS diffère. L'approche par blocs permettait, en choisissant des « jeux de blocs » proches, au prix d'une légère dégradation de la précision de la prévision, de conserver une certaine cohérence dans le modèle et ainsi de réduire la volatilité de la prévision. La mise en œuvre d'un algorithme de sélection automatique suggère de ne pas interférer dans la sélection des variables et ne permet donc pas de garantir un minimum de stabilité et de continuité entre les différents modèles mensuels de prévision d'un même trimestre.

Enfin, l'utilisation d'un tel algorithme n'est pas compatible avec d'autres développements que l'on pourrait souhaiter réaliser à moyen terme sur le modèle. Par exemple, si l'on souhaite utiliser l'approche MIDAS[26] pour mieux tirer parti de l'information contenue dans les variables financières, cela suppose de calculer un ou plusieurs facteurs financiers à partir d'une base de variables à fréquence quotidienne, ce qui ne revêt pas de difficulté particulière si l'on procède par analyse en composantes principales. Ceci ne peut pas se concilier avec la mise en œuvre du LARS pour deux raisons. D'abord, il ne semble pas rigoureux d'appliquer deux fois le LARS à la même endogène pour obtenir simultanément deux sélections de variables (financières et non financières). Mais surtout, la *trimestrialisation* indispensable à l'implémentation du LARS rend caduque l'intérêt même de faire du MIDAS par la suite, puisque le MIDAS trouve sa

[23] À noter que nous pourrions à ce stade *trimestrialiser* les variables mensuelles, comme évoqué au 2.3 ; en construisant à partir d'une variable mensuelle, trois variables trimestrielles correspondant aux valeurs prises aux mois 1, 2 et 3 de chaque trimestre. Mais cela serait incompatible avec l'utilisation, par la suite, de la base mensuelle pour le calcul des facteurs, puisque la probabilité que le LARS sélectionne les trois variables trimestrielles construites à partir de la même variable mensuelle serait faible. Nous n'avons donc pas retenu cette approche.

[24] Pour être parfaitement rigoureux, il faudrait prendre en compte, pour les variables financières, le jour exact auquel on fait tourner le modèle et prolonger, le cas échéant, la variable pour obtenir sa valeur mensuelle estimée qui ne devrait pas *a priori* égaler sa valeur historique exacte. Nous considérons que nous faisons tourner le modèle en fin de mois, après obtention des enquêtes, donc nous faisons comme si toutes les valeurs de la variable financière étaient disponibles pour le mois considéré.

[25] Rappelons que la représentation espace-état contient les équations d'état reliant les séries mensuelles aux facteurs ou « composantes inobservables » (initialisées par analyse en composantes principales), et des équations de transition, modélisant l'évolution de ces composantes inobservables suivant un processus vectoriel autorégressif.

[26] Les modèles MIDAS (*mixed-data sampling*) visent à modéliser une variable diffusée à une fréquence faible (ici le PIB trimestriel) grâce à des indicateurs disponibles à plus haute fréquence, et sans avoir recours à un lissage, source de perte d'information. Techniquement, une forme fonctionnelle est spécifiée afin de pondérer les différents retards des variables à haute fréquence. Cette spécification, qui dépend seulement de deux paramètres, a l'avantage de permettre d'introduire un grand nombre de retards en limitant le nombre de paramètres à estimer, permettant ainsi de conserver un modèle parcimonieux.

légitimité dans l'exploitation de l'information disponible à une fréquence plus élevée que celle de la variable d'intérêt[27].

4.3 Résultats

Le modèle à facteurs tel que nous l'avons implémenté doit être paramétré. Pour déterminer le meilleur modèle de prévision à chaque horizon, nous avons comparé les performances en prévision du modèle en fonction de différents paramétrages. Parmi les options possibles, nous en distinguons de deux types :

- Les paramètres du modèle général
- Les paramètres du LARS

Nous avions déjà relevé précédemment que la solution du LARS correspondait à la solution du problème formel de minimisation suivant :

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|^2 \quad sous\ contrainte \quad (1-\alpha)|\beta|_1 + \alpha\|\beta\|^2 \leq t$$

Il est équivalent de calculer le vecteur β minimisant le lagrangien suivant :

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_1|\beta|_1 + \lambda_2\|\beta\|^2$$

Par la suite, le paramétrage du LARS se fera en fixant le nombre de variables sélectionnées et le λ_2.

4.3.1 Impact de l'intégration des retards des exogènes, de la variable d'intérêt et du choix de la méthode de *trimestrialisation* sur les performances en prévision du modèle

Quelle que soit l'approche retenue pour le choix des variables, la mise en œuvre de la méthode nécessite de fixer certains paramètres auxquels nous ne souhaitons pas donner des valeurs arbitraires : il faut donc tester différentes valeurs des paramètres et comparer les performances en prévision des modèles ainsi paramétrés. Nous ne fixons donc pas d'*a priori* sur la pertinence d'intégrer un retard de l'endogène dans l'équation de prévision et nous testons la performance en prévision du modèle avec et sans retard de l'endogène. De même, nous avions évoqué l'aspect arbitraire et pas toujours justifiable d'une *trimestrialisation* par une simple moyenne. Nous comparons donc les résultats de modèles reposant sur cette méthode de *trimestrialisation* (aussi bien dans la mise en œuvre du LARS que dans la *trimestrialisation* finale des facteurs), avec ceux obtenus avec des modèles pour lesquels la *trimestrialisation* est réalisée en appliquant la formule de passage du taux de croissance mensuel au taux de croissance trimestriel.

Tableau 10 : performance en prévision des modèles en intégrant ou non des retards des variables à la base de données avant application du LARS

		Sans retard des exogènes			Avec retards des exogènes		
	Horizon	RMSFE	Nombre de variables optimal	λ_2 optimal	RMSFE	Nombre de variables optimal	λ_2 optimal
Forecasting (m1 / T-1)	H7	0,43	150	0,3	0,47	150	0,1
Forecasting (m2 / T-1)	H6	0,40	30	0,2	0,44	30	0,2
Forecasting (m3 / T-1)	H5	0,35	150	0,1	0,43	80	0,3
Nowcasting (m1 / T)	H4	0,31	150	0,2	0,33	50	0,4
Nowcasting (m2 / T)	H3	0,26	80	0,3	0,31	80	0,2
Nowcasting (m3 / T)	H2	0,26	80	0,3	0,29	50	0,3
Backcasting (m1 / T+1)	H1	0,27	80	0,4	0,27	100	0,1

Source : DG Trésor.

[27] Sauf à modifier la logique même de l'algorithme LARS, ce qui pourrait être envisageable mais nécessiterait un important travail de formalisation.

Il semblerait que le fait de sensiblement réduire la base avant de lui appliquer le LARS améliore les performances en prévision à presque tous les horizons. On pensait que le LARS pourrait sélectionner la variable retardée la plus adéquate dans une base comportant plusieurs retards, mais il est possible au contraire que le LARS, qui sélectionne des variables corrélées à la variable expliquée pouvant être corrélées entre elles, capture une variable et plusieurs de ses retards sans que cela participe à améliorer la prévision.

Tableau 11 : performance en prévision des modèles en fonction de l'intégration ou non de retards de la variable expliquée

		Avec retard du PIB			Sans retard du PIB		
	Horizon	RMSFE	Nombre de variables optimal	λ_2 optimal	RMSFE	Nombre de variables optimal	λ_2 optimal
Forecasting (m1 / T-1)	H7	0,43	150	0,3	0,43	150	0,3
Forecasting (m2 / T-1)	H6	0,40	50	0,1	0,40	30	0,2
Forecasting (m3 / T-1)	H5	0,35	150	0,2	0,35	150	0,1
Nowcasting (m1 / T)	H4	0,31	150	0,3	0,31	150	0,2
Nowcasting (m2 / T)	H3	0,26	80	0,3	0,26	80	0,3
Nowcasting (m3 / T)	H2	0,26	80	0,3	0,26	80	0,3
Backcasting (m1 / T+1)	H1	0,27	80	0,4	0,27	80	0,4

Source : DG Trésor.

À la lecture de ce tableau, il apparaît que retenir ou non des retards de la variable expliquée ne change globalement pas beaucoup les résultats, quel que soit l'horizon. En revanche, on peut voir que cela influence le paramétrage optimal du LARS, ce qui nous donne un premier aperçu de la difficulté principale de l'implémentation du LARS, à savoir l'absence de robustesse d'un paramétrage optimal qui suppose d'avoir recours à des valeurs arbitraires.

Tableau 12 : performance en prévision des modèles en fonction du choix de la méthode de trimestrialisation

		Avec taux de croissance mensuel			Avec trimestrialisation par la moyenne		
	Horizon	RMSFE	Nombre de variables optimal	λ_2 optimal	RMSFE	Nombre de variables optimal	λ_2 optimal
Forecasting (m1 / T-1)	H7	0,43	150	0,3	0,43	150	0,3
Forecasting (m2 / T-1)	H6	0,40	50	0,1	0,40	30	0,2
Forecasting (m3 / T-1)	H5	0,35	150	0,1	0,35	150	0,2
Nowcasting (m1 / T)	H4	0,31	150	0,3	0,31	150	0,2
Nowcasting (m2 / T)	H3	0,26	80	0,3	0,26	80	0,3
Nowcasting (m3 / T)	H2	0,26	80	0,3	0,26	80	0,3
Backcasting (m1 / T+1)	H1	0,27	80	0,4	0,27	80	0,4

Source : DG Trésor.

La méthode de *trimestrialisation* la plus efficace semble être la méthode utilisant les taux de croissance mensuels. Mais cela se voit pour quelques horizons seulement et le gain en termes de performance en prévision n'est pas très prononcé. On constate ici encore, qu'en figeant un paramètre tel que la méthode de *trimestrialisation*, le paramétrage optimal du LARS se modifie.

4.3.2 Impact du paramétrage du LARS

Le LARS peut ici être paramétré de deux façons : on peut donner un poids plus ou moins important à la pénalisation de type LASSO (valeur absolue) ou à celle de type *Ridge Regression* (norme euclidienne), entraînant une sélection plus ou moins agressive des variables. En outre, on peut fixer arbitrairement le nombre final de variables retenues.

Tableau 13 : impact du nombre de variables retenues et du choix du paramètre λ_2

	λ_2	0,1		0,2	
	Horizon	RMSFE	Nombre de variables optimal	RMSFE	Nombre de variables optimal
Forecasting (m1 / T-1)	H7	0,43	50	0,43	50
Forecasting (m2 / T-1)	H6	0,40	50	0,40	30
Forecasting (m3 / T-1)	H5	0,35	150	0,35	150
Nowcasting (m1 / T)	H4	0,31	150	0,31	150
Nowcasting (m2 / T)	H3	0,27	80	0,28	50
Nowcasting (m3 / T)	H2	0,27	80	0,27	80
Backcasting (m1 / T+1)	H1	0,28	50	0,28	80

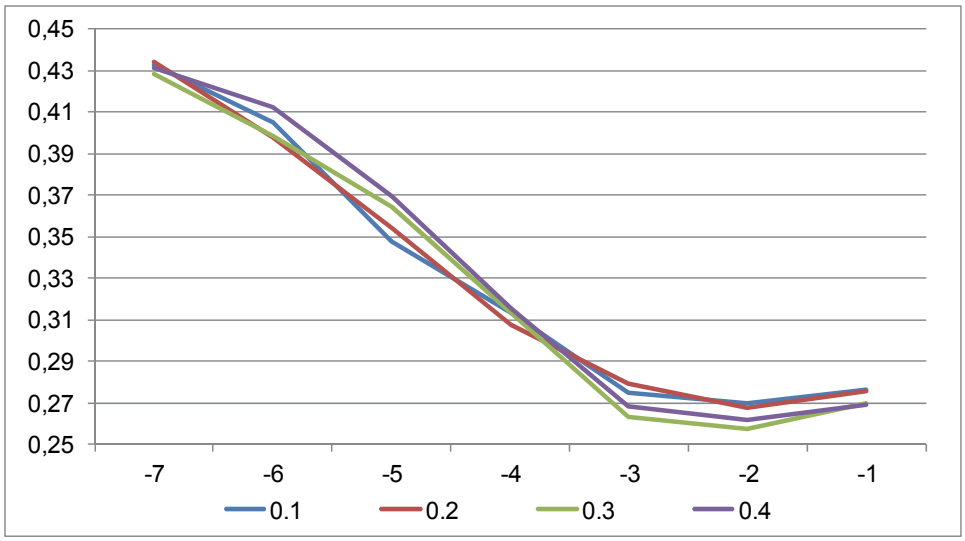
	λ_2	0,3		0,4	
	Horizon	RMSFE	Nombre de variables optimal	RMSFE	Nombre de variables optimal
Forecasting (m1 / T-1)	H7	0,43	150	0,43	80
Forecasting (m2 / T-1)	H6	0,40	30	0,41	50
Forecasting (m3 / T-1)	H5	0,36	30	0,37	150
Nowcasting (m1 / T)	H4	0,31	150	0,32	150
Nowcasting (m2 / T)	H3	0,26	80	0,27	80
Nowcasting (m3 / T)	H2	0,26	80	0,26	80
Backcasting (m1 / T+1)	H1	0,27	80	0,27	80

Source : DG Trésor.

Il semble que l'on obtienne, dans l'ensemble, des résultats similaires quelles que soient les valeurs de λ_2 dans l'intervalle considéré. Sans que les évolutions soient suffisamment importantes pour en tirer des conclusions robustes, il apparaîtrait tout de même qu'un λ_2 plus petit permette d'obtenir de meilleurs résultats à un horizon lointain et que le contraire soit observé à un horizon plus proche de la publication du PIB. Mais ces considérations sont à prendre avec prudence car les résultats sont relativement proches.

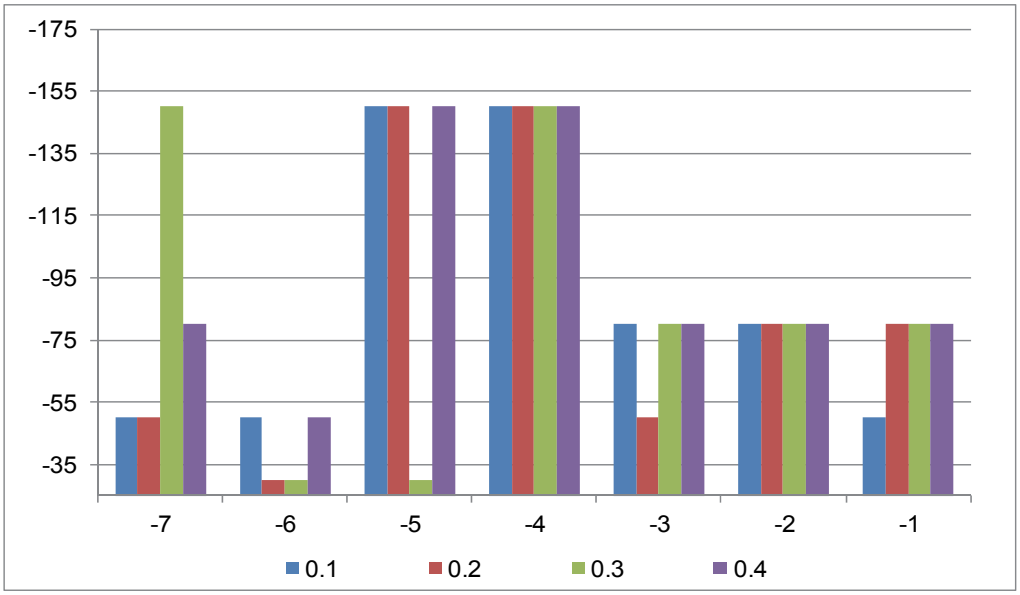
En outre, il semblerait que l'on ait besoin de moins de variables à mesure que l'on se rapproche de la publication, ce qui semble raisonnable car les variables entrant dans les méthodologies des comptes trimestriels (données dures) deviennent disponibles. Il n'est cependant pas essentiel d'avoir un grand nombre de variables à une grande distance du PIB, probablement en raison du nombre finalement relativement faible d'indicateurs très avancés.

Graphique 1 : sensibilité du RMSFE des meilleurs modèles de prévision aux différents horizons de prévision, en fonction de λ_2



Source : DG Trésor.

Graphique 2 : nombre de variables sélectionnées pour les meilleurs modèles de prévision aux différents horizons de prévision, en fonction de λ_2



Source : DG Trésor.

4.3.3 Meilleures performances en prévision obtenues en fonction des stratégies de sélection de variables testées

On observe que la mise en œuvre du LARS donne des résultats moins précis que l'approche retenue dans la section 3. Celle-ci teste en effet toutes les combinaisons possibles de blocs qui ont été choisis et ordonnés de façon éclairée, et permet donc, à la fois de tester un grand nombre de configurations dans un temps finalement plus court, mais également de ne pas rejeter des combinaisons à tort. Parallèlement, le LARS, bien que son intérêt repose sur le fait qu'il devrait choisir les variables avec plus de pertinence, peut parfois échouer à les sélectionner en raison des différents obstacles méthodologiques préalablement cités.

Tableau 14 : comparaison des performances en prévision des modèles en fonction de la méthode de sélection des variables

		LARS		Blocs	
	Horizon	RMSFE	Méthode	RMSFE	Méthode
Forecasting (m1 / T-1)	H7	0,43	2	0,35	2
Forecasting (m2 / T-1)	H6	0,40	2	0,35	2
Forecasting (m3 / T-1)	H5	0,35	1	0,32	1
Nowcasting (m1 / T)	H4	0,31	1	0,28	2
Nowcasting (m2 / T)	H3	0,26	1	0,21	2
Nowcasting (m3 / T)	H2	0,26	1	0,23	1
Backcasting (m1 / T+1)	H1	0,27	1	0,21	1

Source : DG Trésor.

4.3.4 Variables sélectionnées par le LARS en fonction de l'horizon de prévision

Tableau 15 : 20 variables sélectionnées le plus souvent par le LARS en fonction de l'horizon de prévision[28]

Horizon	-7	-6	-5	-4	-3	-2	-1
1	carnets de commande	euro_yuan	production passée	stocks	conso manufacturière	conso manufacturière	conso manufacturière
2	carnets de commande étrangers	perspectives personnelles - bâtiment	stocks	carnets de commande étrangers	conso biens durables	conso biens durables	conso biens durables
3	activité passée - service	niveau des effectifs	activité passée - service	perspectives personnelles	conso automobile	conso automobile	conso automobile
4	demande prévue - service	euro_yen	demande prévue - service	demande prévue - service	ventes de détail	ventes de détail	ventes de détail
5	perspectives personnelles - bâtiment	effectifs prévus - commerce de détail	effectifs passés - bâtiment	commandes - commerce de détail	BE IPI	BE IPI	BE IPI
6	effectifs passés - bâtiment	ventes passées - commerce de détail	jugement des carnets de commande - bâtiment	effectifs prévus - commerce de détail	C5 IPI	C5 IPI	C5 IPI
7	tuc - bâtiment	Nikkei	commandes - commerce de détail	niveau de vie passé	TIimmo	TIimmo	TIimmo
8	opportunité d'achats	perspectives personnelles - bâtiment	effectifs prévus - commerce de détail	niveau de vie prévu	TI1Y (gov. Bond)	TI1Y (gov. Bond)	TI1Y (gov. Bond)
9	immatriculations	vix	situation fin. Passée - ménages	immatriculations	IPC	IPC	IPC
10	conso autres produits manuf.	fin1	niveau de vie passé	conso biens durables	euro_yen	euro_yen	euro_yen
11	ventes de détail	évolution prévue de la production (M+12)	opportunité d'achats	conso automobile	euro_yuan	euro_yuan	euro_yuan
12	ventes de détail	prix3	immatriculations	conso équipement logement	chômage US	chômage US	chômage US
13	chômage	situation fin. Prévue - ménages	conso manufacturière	conso textile	effectifs passés - bâtiment	effectifs passés - bâtiment	effectifs passés - bâtiment
14	BE IPI	IPC	conso biens durables	ventes de détail	production passée	production passée	TI10Y (gov. Bond)
15	C1 IPI	fin4	conso automobile	ventes de détail	situation fin. Passée - ménages	situation fin. Passée - ménages	production passée
16	C3 IPI	niveau des carnets de commandes	conso équipement logement	chômage	exports	exports	situation fin. Passée - ménages
17	C5 IPI	perspectives personnelles - services	conso autres produits manuf.	emplois vacants	jugement des carnets de commande - bâtiment	jugement des carnets de commande - bâtiment	exports
18	DE IPI	Situation de trésorerie	ventes de détail	BE IPI	conso équipement logement	conso équipement logement	jugement des carnets de commande - bâtiment
19	vix	spreadUS	ventes de détail	CZ IPI	CZ IPI	CZ IPI	conso équipement logement
20	prets	effectifs passés - bâtiment	ventes de détail	C2 IPI	CL1 IPI	CL1 IPI	CZ IPI

Source : DG Trésor.

[28] Lorsque cela n'est pas précisé, le solde d'opinion d'une enquête correspond au point de vue des industriels.

Tableau 16 : les secteurs couverts par les 20 variables sélectionnées le plus souvent par le LARS en fonction de l'horizon de prévision

Horizon	-7	-6	-5	-4	-3	-2	-1
1	industrie	taux de change	industrie	industrie	conso manufacturière	conso manufacturière	conso manufacturière
2	industrie	bâtiment	industrie	industrie	conso manufacturière	conso manufacturière	conso manufacturière
3	services	Industrie	services	industrie	conso manufacturière	conso manufacturière	conso manufacturière
4	services	taux de change	services	services	ventes de détail	ventes de détail	ventes de détail
5	bâtiment	commerce de détail	bâtiment	commerce de détail	industrie	industrie	industrie
6	bâtiment	commerce de détail	bâtiment	commerce de détail	industrie	industrie	industrie
7	bâtiment	bourse	commerce de détail	ménages	taux d'intérêt	taux d'intérêt	taux d'intérêt
8	ménages	bâtiment	commerce de détail	ménages	taux d'intérêt	taux d'intérêt	taux d'intérêt
9	immatriculation	bourse	ménages	immatriculation	prix	prix	prix
10	conso manufacturière	financier	ménages	conso manufacturière	taux de change	taux de change	taux de change
11	ventes de détail	Industrie	ménages	conso manufacturière	taux de change	taux de change	taux de change
12	ventes de détail	prix	immatriculation	conso manufacturière	États-Unis	États-Unis	États-Unis
13	chômage	ménages	conso manufacturière	conso manufacturière	bâtiment	bâtiment	bâtiment
14	industrie	prix	conso manufacturière	ventes de détail	industrie	industrie	taux d'intérêt
15	industrie	financier	conso manufacturière	ventes de détail	ménages	ménages	industrie
16	industrie	industrie	conso manufacturière	chômage	commerce extérieur	commerce extérieur	ménages
17	industrie	services	conso manufacturière	emploi	bâtiment	bâtiment	commerce extérieur
18	industrie	industrie	ventes de détail	industrie	conso manufacturière	conso manufacturière	bâtiment
19	bourse	taux d'intérêt	ventes de détail	industrie	industrie	industrie	conso manufacturière
20	monnaie	bâtiment	ventes de détail	industrie	industrie	industrie	industrie

Source : DG Trésor.

Tableau 17 : catégories des 35 variables sélectionnées le plus souvent par le LARS en fonction de l'horizon de prévision

Horizon	-7	-6	-5	-4	-3	-2	-1
1	enquêtes Insee	international	enquêtes Insee	enquêtes Insee	réel	réel	réel
2	enquêtes Insee	enquête Insee	enquêtes Insee	enquêtes Insee	réel	réel	réel
3	enquêtes Insee	enquête BdF	enquêtes Insee	enquêtes Insee	réel	réel	réel
4	enquêtes Insee	international	enquêtes Insee	enquêtes Insee	réel	réel	réel
5	enquêtes Insee	enquête Insee	enquêtes Insee	enquêtes Insee	réel	réel	réel
6	enquêtes Insee	enquête Insee	enquêtes Insee	enquêtes Insee	réel	réel	réel
7	enquêtes Insee	nominal	enquêtes Insee	enquêtes Insee	nominal	nominal	nominal
8	enquêtes Insee	enquête Insee	enquêtes Insee	enquêtes Insee	nominal	nominal	nominal
9	réel	nominal	enquêtes Insee	réel	nominal	nominal	nominal
10	réel	nominal	enquêtes Insee	réel	international	international	international
11	réel	enquête BdF	enquêtes Insee	réel	international	international	international
12	réel	nominal	réel	réel	international	international	international
13	réel	enquête Insee	réel	réel	enquête Insee	enquête Insee	enquête Insee
14	réel	nominal	réel	réel	enquête Insee	enquête Insee	nominal
15	réel	nominal	réel	réel	enquête Insee	enquête Insee	enquête Insee
16	réel	enquête BdF	réel	réel	réel	réel	enquête Insee
17	réel	enquête Insee	réel	réel	enquêtes Insee	enquêtes Insee	réel
18	réel	enquête BdF	réel	réel	réel	réel	enquêtes Insee
19	nominal	nominal	réel	réel	réel	réel	réel
20	nominal	enquête Insee	réel	réel	réel	réel	réel
21	nominal	international	réel	réel	international	international	réel
22	nominal	nominal	réel	réel	enquête Insee	enquête Insee	international
23	nominal	enquête Insee	réel	réel	enquête Insee	enquête Insee	enquête Insee
24	international	enquête Insee	réel	réel	enquête Insee	enquête Insee	enquête Insee
25	international	enquête BdF	réel	réel	enquête BdF	enquête BdF	enquête BdF
26	international	nominal	réel	nominal	enquête BdF	enquête BdF	enquête BdF
27	international	international	réel	nominal	nominal	nominal	nominal
28	international	enquête BdF	réel	nominal	nominal	nominal	nominal
29	international	nominal	réel	nominal	enquêtes Insee	enquêtes Insee	enquêtes Insee
30	international	international	nominal	nominal	enquêtes Insee	enquêtes Insee	enquêtes Insee
31	international	enquête Insee	nominal	nominal	réel	réel	réel
32	enquête Insee	enquête Insee	nominal	nominal	nominal	nominal	enquête Insee
33	enquête Insee	réel	nominal	nominal	enquête Insee	enquête Insee	enquêtes Insee
34	enquête Insee	nominal	nominal	nominal	enquêtes Insee	enquêtes Insee	enquête Insee
35	enquête Insee	enquête Insee	nominal	nominal	enquête Insee	enquête Insee	enquête Insee

Source : DG Trésor.

4.4 L'apport des variables financières dans la prévision

En supplément des données dures lorsqu'elles sont disponibles et des données d'enquêtes qui en fournissent la teneur tout en étant publiées plus rapidement, il semble naturel d'utiliser des variables financières comme indicateurs avancés de la conjoncture macroéconomique, en particulier dans un contexte marqué par la crise financière de 2008 et la crise des dettes souveraines qui a suivi. L'étude Bessec-Doz (2011) ainsi que les résultats de la section précédente ont mis en évidence la pertinence de ces variables pour prévision.

Cela étant, ce type de données n'est pas directement utilisé en comptabilité trimestrielle et l'on peut donc se demander si leur absence dégraderait significativement les prévisions.

La comparaison des modèles incluant et n'incluant pas de variables financières, ainsi que l'application du LARS à la base de données élargie, permettent d'apporter une réponse à cette question, et de préciser les horizons auxquels ces variables ajoutent de la précision à la prévision.

4.4.1 Les variables financières retenues dans la base de données

Près d'une quarantaine de variables financières sont intégrées à la base de données. Parmi lesquelles des variables nominales : monétaires et financières (taux d'intérêts, pente des taux, indices boursiers, etc.) et des indicateurs de l'environnement international (taux de change de l'euro et indicateurs conjoncturel des principaux partenaires économiques).

À noter toutefois, qu'il n'est pas aisé d'agrandir la base de données financières dans la mesure où un certain nombre de variables ne sont disponibles qu'à partir des années 2000, avec la création de l'euro.

4.4.2 Les blocs de variables financières semblent pertinents à tous les horizons de temps

Il n'y a qu'au mois 2 en ***nowcasting***, qu'ils ne semblent pas apporter d'information supplémentaire. Pour le ***backcasting,*** et bien davantage encore pour le ***forecasting***, ils apportent une information qui permet d'améliorer sensiblement les performances en prévision du modèle.

Tableau 18 : combinaison de blocs fournissant le modèle le plus performant en termes de précision de la prévision, hors variables financières

	Horizon	RMSFE	Combinaison	Blocs
Forecasting (m1 / T-1)	H7	0,39	29	Enquêtes Insee et Banque de France
Forecasting (m2 / T-1)	H6	0,39	92	Données dures, enquêtes Insee et Banque de France
Forecasting (m3 / T-1)	H5	0,37	147	Données dures, enquêtes Insee et Banque de France
Nowcasting (m1 / T)	H4	0,31	68	Enquêtes Insee et Banque de France
Nowcasting (m2 / T)	H3	0,21	24	Données dures, enquêtes Banque de France
Nowcasting (m3 / T)	H2	0,24	24	Données dures, enquêtes Banque de France
Backcasting (m1 / T+1)	H1	0,22	26	Données dures, enquêtes Insee

Source : DG Trésor.

Tableau 19 : combinaison de blocs fournissant le modèle le plus performant en termes de précision de la prévision, variables financières incluses

	Horizon	RMSFE	Combinaison	Blocs
Forecasting (m1 / T-1)	H7	0,35	101	Variables nominales, données dures et enquêtes Insee
Forecasting (m2 / T-1)	H6	0,35	119	Variables nominales, indicateurs internationaux et enquêtes Insee
Forecasting (m3 / T-1)	H5	0,32	140	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m1 / T)	H4	0,28	87	Variables nominales, enquêtes Insee et Banque de France
Nowcasting (m2 / T)	H3	0,21	24	Données dures, enquêtes Banque de France
Nowcasting (m3 / T)	H2	0,23	84	Données dures, indicateurs internationaux, enquêtes Banque de France
Backcasting (m1 / T+1)	H1	0,21	84	Données dures, indicateurs internationaux, enquêtes Banque de France

Source : DG Trésor.

On constate que les blocs non financiers sont proches dans les deux cas de figure, pour le ***nowcasting*** et le ***backcasting*** notamment. Pour le ***forecasting***, en l'absence d'indicateurs financiers, il semble que les données dures ressortent plus facilement, sûrement en raison de l'information qu'elles fournissent sur le trimestre futur à partir des acquis des variables à l'issue des derniers mois disponibles. Cependant, la présence de variables nominales ou des indicateurs internationaux, associés aux seules enquêtes, semble permettre l'amélioration sensible des RMSFE à ces horizons.

Plus généralement, les variables financières permettent d'améliorer la performance en prévision des modèles à tous les horizons. Les indicateurs internationaux sont particulièrement pertinents pour le ***backcasting***. Publiés avec retard, ils doivent néanmoins apporter de l'information consistante sur le commerce extérieur. Les variables nominales interviennent davantage pour le ***forecasting*** ou le ***nowcasting*** en début de trimestre, en l'absence d'information disponible autre que celle fournie par les enquêtes.

4.4.3 Ce constat est confirmé par l'application du LARS qui aboutit bien à la sélection de variables financières

Il apparaît clairement dans le tableau 17 que les variables financières (en bleu foncé) font partie des variables les plus souvent sélectionnées, et ce à tous les horizons de prévision, qu'il s'agisse de taux d'intérêt, de taux de change ou de prix.

Elles sont sélectionnées au même titre que les enquêtes aux horizons les plus lointains, et parfois privilégiées aux données dures. Elles apportent donc une réelle plus-value en l'absence d'information quantitative, mais elles sont tout de même encore sélectionnées lorsque l'on se rapproche de la publication du PIB. Ce constat milite donc en faveur d'un intérêt particulier à porter aux variables financières, et à la meilleure façon de les exploiter dans les outils de prévision.

Conclusion

Selon Bessec et Doz (2011), il convenait d'utiliser des MFD sur des ensembles de données différents selon les horizons de prévision visés, de manière à améliorer la précision des estimations obtenues. Celle-ci s'améliore ainsi fortement à mesure que « l'on se rapproche de la date de la publication du taux de croissance du PIB ; quelques mois avant cette publication … le meilleur modèle est alors celui estimé à partir des variables réelles et des variables d'enquêtes de l'Insee ».

Les prolongements apportés à cette étude confirment cette conclusion puisque, quelle que soit l'approche de sélection de variables retenue, « par blocs » ou par l'application d'un algorithme de sélection automatique de variables tel que le LARS, on constate que l'ensemble d'information jugé optimal pour la prévision à un horizon donné évolue.

De la composition des blocs optimaux, on constate que le recours aux données dures n'est réellement efficace qu'à partir du premier mois du trimestre en cours ; avant cela, les enquêtes, les variables financières et les indicateurs internationaux suffisent. On note cependant deux résultats intéressants : les enquêtes interviennent dans les combinaisons de blocs optimales à tous les horizons, manifestant donc autant de pertinence en tant qu'indicateurs avancés qu'en tant qu'indicateurs retardés. Par ailleurs, les variables financières semblent réellement améliorer les performances en prévision des modèles et ce, à tous les horizons de prévision, ce qui encourage à leur porter un intérêt particulier dans d'éventuels développements futurs de ces modèles (MIDAS, facteurs financiers..).

Concernant le recours à un algorithme de sélection automatique de variables tel que le LARS, il ressort de cette étude qu'il ne permet apparemment pas, pour la prévision du taux de croissance du PIB, d'améliorer les performances en prévision des modèles par rapport à la méthode dite « par blocs ». Bien que l'algorithme repose sur une rationalité statistique, puisqu'il sélectionne les variables de la base en tenant compte de leur corrélation avec la variable expliquée, sa mise en œuvre dégrade les performances en prévision de ces modèles par rapport aux modèles plus simples obtenus par l'approche par « blocs » qui fournissent déjà d'excellentes performances.

Malgré tout, cet algorithme peut sembler tout à fait adapté pour prévoir l'évolution d'une variable à partir d'une large base de données pour laquelle la catégorisation pertinente des variables apparaîtrait moins évidente et le caractère automatique de l'algorithme pourrait permettre de suppléer l'absence d'une expertise comparable à celle que les prévisionnistes ont développé avec le temps pour la prévision de court terme du taux de croissance du PIB.

Annexe 1 : Détail des blocs de différents blocs de variables

Bloc 1: enquêtes mensuelles de l'Insee

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
enquête	industrie	production passée (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	industrie	stocks (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	industrie	carnets com. globaux (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	industrie	carnets de com. étrangers (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	industrie	persp. personnelles de prod. (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	industrie	persp. générales de prod. (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	services	activité passée (niv. solde opinion)*	T/M	88T1 à 00T2 puis juin	M+0	Insee
enquête	services	activité prévue (niv. solde opinion)*	T/M	88T1 à 00T2 puis juin	M+0	Insee
enquête	services	demande prévue (niv. solde opinion)*	T/M	88T1 à 04T2 puis juil.04	M+0	Insee
enquête	bâtiment	activité passée (niv. solde opinion)*	T/M	80T1 à 93T3 puis sept	M+0	Insee
enquête	bâtiment	activité prévue (niv. solde opinion)*	T/M	80T1 à 93T3 puis sept	M+0	Insee
enquête	bâtiment	effectifs passés (niv. solde opinion)*	T/M	80T1 à 93T3 puis sept	M+0	Insee
enquête	bâtiment	jugement sur carnets de com. (dif. solde opinion)*	T/M	80T1 à 93T3 puis oct.	M+0	Insee
enquête	bâtiment	taux d'Utilisation des Capacités (dif. en %)*	T/M	80T1 à 93T3 puis oct.	M+0	Insee
enquête	commerce de détail	persp. générales d'activité (niv. solde opinion)	M	janvier 1991	M+0	Insee
enquête	commerce de détail	ventes passées (niv. solde opinion)	M	janvier 1991	M+0	Insee
enquête	commerce de détail	commandes (niv. solde opinion)	M	janvier 1991	M+0	Insee
enquête	commerce de détail	emploi prévu (niv. solde opinion)	M	janvier 1991	M+0	Insee
enquête	ménages	évol. passée situation fin. perso (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	ménages	évol. prévue situation fin. perso. (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	ménages	évol. passée niveau de vie (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	ménages	évol. prévue niveau de vie (niv. solde opinion)	M	janvier 1980	M+0	Insee
enquête	ménages	opportunité d'achats importants (dif. solde opinion)	M	janvier 1980	M+0	Insee
enquête	*industrie*	*demande - évolution passée (niv. solde opinion)*	*T*	*1980T1*	*T+0*	*Insee*
enquête	*industrie*	*demande - évolution prévue (niv. solde opinion)*	*T*	*1980T1*	*T+0*	*Insee*
enquête	*services*	*résultats d'exploitation passés (Niv. solde opinion)*	*T*	*1988T1*	*T+0*	*Insee*
enquête	*services*	*résultats d'exploitation prévus (niv. solde opinion)*	*T*	*1988T1*	*T+0*	*Insee*

Lecture : en italique figurent pour mémoire les variables trimestrielles utilisées dans l'étude précédente (Bessec-Doz, 2011), non reprises dans le prolongement de l'étude présentée ici ; le symbole T/M indiquent les résultats d'enquêtes tout d'abord réalisées à un rythme trimestriel puis mensuel à partir de la date indiquée. Les soldes trimestriels ont été mensualisés par simple interpolation linéaire.

Source : DG Trésor.

Les quatre soldes d'opinion trimestriels n'ont pas été reconduits dans l'étude présentée ici et raison de leur périodicité atypique (disponibilité immédiate au mois 1 du trimestre, l'information étant ensuite indisponible aux mois 2 et 3 du trimestre) et du problème du choix de la méthode d'interpolation des données manquantes par la suite. De plus, ces variables apportent finalement peu d'information supplémentaire. D'une part, elles sont fortement corrélées entre elles et, d'autre part, elles le sont avec d'autres variables mensuelles également issues des enquêtes de conjoncture de l'Insee. Au total, ce sont 23 variables mensuelles qui sont prises en compte dans ce bloc de données, disponibles sur longue période (au moins depuis janvier 1991).

Bloc 2: variables réelles d'activité

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
réel	ménages	immatriculation véhicules neufs (croissance)	M	janvier 1980	M+1	Insee
réel	ménages	conso. de prod. Manuf. (croissance)	M	février 1980	M+1	Insee
réel	ménages	conso. de biens durables (croissance)	M	février 1980	M+1	Insee
réel	ménages	conso. d'automobile (croissance)	M	février 1980	M+1	Insee
réel	ménages	conso. de biens d'équip. Logements (croissance)	M	février 1980	M+1	Insee
réel	ménages	conso. de textile (croissance)	M	février 1980	M+1	Insee
réel	ménages	conso. autres prod. manuf. (croissance)	M	février 1980	M+1	Insee
réel	ménages	ventes au détail hors automobile (croissance)	M	février 1990	M+2	BdF
réel	**ménages**	**vente au détail de produits alimentaires**	**M**	**décembre 1990**	**M+2**	**BdF**
réel	**ménages**	**vente au détail prod. ind. hors alim. (croissance)**	**M**	**décembre 1990**	**M+2**	**BdF**
réel	ménages	taux de chômage (variation sur un mois)	M	janvier 1980	M+2	Insee-
réel	ménages	nombre d'emplois vacants (variation sur un mois)	M	février 1989	M+2	Eurostat
réel	industrie	IPI – industrie totale (BE) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI - industrie manuf (CZ) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI - ind. agricoles et alimentaires (C1) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – cokéfaction et raffinage (C2) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – équipements électriques (C3) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – automobile (CL1) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – transports hors auto (CL2) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – autres produits industriels (C5) (croissance)	M	février 1990	M+2	Insee
réel	industrie	IPI – ind. extractives, énergie et eau (DE)	M	février 1990	M+2	Insee
réel	industrie	IPI – construction (F) (croissance)	M	février 1990	M+2	Insee
réel	**comex**	**exportations en valeur (croissance)**	**M**	**décembre 1990**	**M+1**	**Insee**
réel	**comex**	**importations en valeur (croissance)**	**M**	**décembre 1990**	**M+1**	**Insee**
réel	**comex**	**solde commercial – en valeur (variation**	**M**	**décembre 1990**	**M+1**	**Insee**
réel	*ménages*	*taux de chômage (-25 ans) - (variation mensuelle*	*M*	*janvier 1983*	*M+3*	*Insee*
réel	*bâtiment*	*mises en chantier total (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*mises en chantier collectif (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*mises en chantier individuel (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*mises en chantier résidentiel (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*permis construire total (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*permis construire collectif (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*
réel	*bâtiment*	*permis construire individuel (croissance, CVS)*	*M*	*janvier **1994***	*M+1*	*SOeS*

Lecture : en italique figurent pour mémoire les variables trimestrielles utilisées dans l'étude précédente (Bessec-Doz, 2011), non reprises dans le prolongement de l'étude présentée ici ; en gras, figurent les variables mensuelles ajoutées au bloc de variable des variables « réelles ».

Source : DG Trésor.

Le taux de chômage des jeunes (selon la définition du BIT) est publié avec beaucoup de retard et n'est a *priori* pas pris en compte dans les estimations des facteurs mensuels en fin de période. En outre, il connaît des évolutions très erratiques d'un trimestre sur l'autre. De la même manière, les données de la construction neuve (permis de construire et mises en chantier) ne sont plus retenues dans l'échantillon même si ces données sont utilisées dans les comptes nationaux pour déterminer les montants de la FBCF des différents agents en construction et entrent ainsi dans le calcul du PIB. Même en ayant pris soin de les désaisonnaliser, ces données mensuelles connaissent des évolutions mensuelles heurtées. En outre, leur prise en considération dans les comptes nationaux intervient avec retard (principe des « grilles délais » et lissage des résultats sur plus d'un an), un délai *a priori* incompatible avec les horizons de prévision des MFD.

À l'inverse, l'échantillon a été complété en introduisant des variables supplémentaires : il s'agit des indices de volume des ventes au détail dans le commerce alimentaire et dans le commerce non alimentaire, et des flux mensuels du commerce extérieur (importations et exportations en valeur et solde) publiés par les Douanes. Au total, ce bloc comprend 25 variables mensuelles disponibles sur longue période (au moins depuis décembre 1990).

Bloc 3: variables nominales

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
nominal	bourse	CAC 40 - (croissance mensuelle)	J / M	août 1987	M+0	Nyse Euronext Paris
nominal	bourse	SP500 - (croissance mensuelle)	J / M	février 1980	M+0	Standard and Poors
nominal	bourse	FTSE - (croissance mensuelle)	J / M	février 1984	M+0	FTSE
nominal	bourse	DAX - (croissance mensuelle)	J / M	janvier 1980	M+0	Frankfurt SE
nominal	bourse	€urostoxx50 - (croissance mensuelle)	J / M	janvier 1987	M+0	Financial Time
nominal	bourse	nikkei - (croissance mensuelle)	J / M	janvier 1981	M+0	OSE
nominal	bourse	indice de volatilité Vix (log)	J / M	janvier 1990	M+0	CBOE
nominal	monnaie	prêts bancaires (valeur, croissance cvs)	M	février 1980	M+2	BdF
nominal	taux intérêt	taux d'intérêt immob. (dif. mensuelle)	M	janvier 1980	M+1	BdF
nominal	taux intérêt	taux d'intérêt à 3 mois (dif. mensuelle)	M	janvier 1980	M+1	FMI
nominal	taux intérêt	taux d'intérêt à 1 an (dif. mensuelle)	M	janvier 1980	M+1	BdF
nominal	taux intérêt	taux d'intérêt à 10 ans (dif. mensuelle)	M	février 1989	M+0	Daily press
nominal	taux intérêt	pente des taux France (12 ans / 2 ans)	M	janvier 1990	M+0	Global insight
nominal	taux intérêt	pente des taux US (12 ans / 2 ans)	M	janvier 1990	M+0	Global insight
nominal	prix	or (croissance mensuelle)	J / M	janvier 1980	M+0	LBMA
nominal	prix	pétrole (croissance)	M	février 1980	M+1	FMI
nominal	prix	matières premières (croissance)	M	janvier 1980	M+1	FMI
nominal	prix	IPC (variation mensuelle)	M	février 1990	M+1	Insee
nominal	**bourse**	**indice SB120 (croissance mensuelle)**	**J / M**	**janvier 1991**	**M+0**	**Data insight**
nominal	**taux**	**taux Euromarket 1 mois (dif. mensuelle)**	**J / M**	**janvier 1991**	**M+0**	**Data insight**
nominal	**taux**	**taux Euromarket 6 mois (dif. mensuelle)**	**J / M**	**janvier 1991**	**M+0**	**Data insight**
nominal	**taux**	**taux bons d'État à 5 ans (dif. mensuelle)**	**J / M**	**janvier 1991**	**M+0**	**Data insight**
nominal	*bourse*	*PER (actions US) – (croissance)*	*M*	*janvier 1980*	*M+1*	*Standard and Poors*
nominal	*monnaie*	*M1 (croissance, cvs)*	*M*	*janvier 1980*	*M+2*	*BdF*
nominal	*monnaie*	*M2 (croissance, cvs)*	*M*	*janvier 1980*	*M+2*	*BdF*
nominal	*monnaie*	*M3 (croissance, cvs)*	*M*	*janvier 1980*	*M+2*	*BdF*

Lecture : en italique figurent pour mémoire les variables trimestrielles utilisées dans l'étude précédente (Bessec-Doz, 2011), non reprises dans le prolongement de l'étude présentée ici ; en gras, figurent les variables mensuelles ajoutées au bloc de variable des variables « nominales ».

Source : DG Trésor.

L'indice d'évolution du Price Earning Ratio des actions aux États-Unis n'est plus disponible sur la période récente et n'est donc plus retenu dans l'échantillon. En outre, les évolutions mensuelles des grandes catégories de la masse monétaire (M1, M2 et M3) ont été également exclues du bloc de variables nominales en raison de leur forte volatilité et de leur faible apport à la constitution des 10 facteurs retenus dans l'étude précédente. On peut en effet mettre en doute la pertinence de ces variables pour la prévision de l'évolution de l'activité réelle, comme l'ont déjà montré d'autres études[29]. Tout d'abord, ces quantités ne sont pas homogènes sur l'ensemble de la période (masses monétaires en franc puis libellées en euros dans le cadre d'un marché unique de capitaux). En outre, la mise en œuvre après la récession de 2008-2009 de facilités monétaires pour le secteur financier européen est probablement à l'origine d'une déconnexion forte des évolutions réelles et de celles des agrégats monétaires.

Dans le bloc des variables nominales, ont été ajoutés un indice boursier supplémentaire (évolution mensuelle de l'Indice SBF120) et des variables représentatives de l'évolution des taux d'intérêt à court terme (variations mensuelles des taux d'intérêt interbancaires à 1 mois et à 6 mois) et à moyen terme (variation mensuelle du taux d'intérêt des emprunts d'État à 5 ans). Au total, ce bloc comprend 22 variables mensuelles disponibles sur longue période (au moins depuis janvier 1991).

[29] Dont notamment : "Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP" *Oxford bulletin of economics and statistics,* 72, 4 (2011) 0305-9049 par M. Marcellino (*European University Institute*) et C. Schumacher (*BundesBank*).

Bloc 4: variables représentatives de l'environnement international

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
international	change	euro/dollar (croissance)	M	janvier 1980	M+0	BCE
international	change	euro / livre sterling (croissance)	M	janvier 1980	M+0	BCE
international	change	euro/yen (croissance)	M	janvier 1980	M+0	BCE
international	change	euro/yuan (croissance)	M	janvier 1980	M+0	BCE
international	change	taux effectif de l'euro (croissance)	M	janvier 1980	M+2	FMI
international	change	taux effectif réel de l'euro (croissance)	M	janvier 1980	M+2	FMI
international	Allemagne	IFO : situation actuelle (niv. solde opinion)	M	janvier 1980	M+0	IFO
international	Allemagne	IFO : perspectives d'activité (niv. solde opinion)	M	janvier 1980	M+0	IFO
international	Allemagne	ZEW : situation actuelle (niv. solde opinion)	M	décembre 1991	M+0	ZEW
international	Allemagne	ZEW : persp. économiques (niv. solde opinion)	M	décembre 1991	M+0	ZEW
international	Allemagne	IPI manufacturé (croissance)	M	février 1991	M+2	Destatis
international	US	ventes de détail en valeur (croissance)	M	février 1992	M+1	Census Bureau
international	US	IPI manufacturier (croissance)	M	janvier 1980	M+1	Fed
international	US	emploi (variation mensuelle)	M	janvier 1980	M+1	BLS
international	US	taux de chômage (variation mensuelle)	M	janvier 1980	M+1	BLS
international	US	ISM manufacturier (niv. solde opinion)	M	janvier 1980	M+1	ISM

Source : DG Trésor.

Bloc 5 : enquêtes mensuelles de l'Insee en différence

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
Enq. Dif	industrie	production passée (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	industrie	stocks (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	industrie	carnets commandes globaux (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	iIndustrie	carnets de commandes étrangers (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	industrie	persp. personnelles de production (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	industrie	persp. générales de production (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	services	activité passée (dif. solde opinion)*	M	88T1 à 00T2 puis juin 00	M+0	Insee
Enq. Dif	services	activité prévue (dif. solde opinion)*	M	88T1 à 00T2 puis juin 01	M+0	Insee
Enq. Dif	services	demande prévue (dif. solde opinion)*	M	88T1 à 00T2 puis juin 02	M+0	Insee
Enq. Dif	bâtiment	activité passée (dif. solde opinion)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
Enq. Dif	bâtiment	activité prévue (dif. solde opinion)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
Enq. Dif	bâtiment	effectifs passés (dif. solde opinion)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
Enq. Dif	comm détail	persp. générales d'activité (dif. solde opinion)*	M	février 1991	M+0	Insee
Enq. Dif	comm détail	ventes passées (dif. solde opinion)*	M	février 1991	M+0	Insee
Enq. Dif	comm détail	commandes (dif. solde opinion)*	M	février 1991	M+0	Insee
Enq. Dif	comm détail	emploi prévu (dif. solde opinion)*	M	février 1991	M+0	Insee
Enq. Dif	ménages	évol. passée situation fin. perso (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	ménages	évol. prévue situation fin. perso (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	ménages	évol. passée niveau de vie (dif. solde opinion)*	M	février 1980	M+0	Insee
Enq. Dif	ménages	évol. prévue niveau de vie (dif. solde opinion)*	M	février 1980	M+0	Insee

Source : DG Trésor.

Certains soldes d'opinion étaient déjà pris en compte sous la forme de différences premières dans l'échantillon (« Jugement sur l'état des carnets de commandes » et « taux d'utilisation des capacités productives » pour l'enquête de conjoncture auprès des chefs d'entreprise dans l'industrie du bâtiment ; « opportunité de réaliser des achats importants » pour l'enquête mensuelle de conjoncture auprès des ménages) dans le bloc 1, et n'ont donc pas été repris dans ce « nouveau » bloc de données.

Bloc 6 : enquêtes mensuelles de l'Insee au carré (signé)

Bloc	Secteur	Série	Fréq	Début de la série	Délais	Source
enq. carrée	industrie	production passée (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	industrie	stocks (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	industrie	carnets commandes globaux (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	industrie	carnets de commandes étrangers (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	industrie	persp. personnelles de production (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	industrie	persp. générales de production (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	services	activité passée (carré signé)*	M	88T1 à 00T2 puis juin 00	M+0	Insee
enq. carrée	services	activité prévue (carré signé)*	M	88T1 à 00T2 puis juin 00	M+0	Insee
enq. carrée	services	demande prévue (carré signé)*	M	88T1 à 04T2 puis juil.04	M+0	Insee
enq. carrée	bâtiment	activité passée (carré signé)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
enq. carrée	bâtiment	activité prévue (carré signé)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
enq. carrée	bâtiment	effectifs passés (carré signé)*	M	80T1 à 93T3 puis sept 93	M+0	Insee
enq. carrée	comm. détail	persp. générales d'activité (carré signé)*	M	janvier 1991	M+0	Insee
enq. carrée	comm. détail	ventes passées (carré signé)*	M	janvier 1991	M+0	Insee
enq. carrée	comm. détail	commandes (carré signé)*	M	janvier 1991	M+0	Insee
enq. carrée	comm. détail	emploi prévu (carré signé)*	M	janvier 1991	M+0	Insee
enq. carrée	ménages	évol. passée situation fin. perso (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	ménages	évol. prévue situation fin. perso (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	ménages	évol. passée niveau de vie (carré signé)*	M	janvier 1980	M+0	Insee
enq. carrée	ménages	évol. prévue niveau de vie (carré signé)*	M	janvier 1980	M+0	Insee

Source : DG Trésor.

Là encore, n'ont été retenus que les carrés signés (*i.e.* le solde d'opinion multiplié par sa valeur absolue) des soldes d'opinion présents en niveau dans le bloc 1 en niveau. Ce « nouveau » bloc est également constitué de 20 variables disponibles sur longue période (au moins depuis janvier 1991).

Bloc 7 : enquêtes de conjoncture de la Banque de France

Type	Secteur	Série	Fréq	Début de la série	Délais	Source
enq_BDF	industrie	TUC productives – (dif. mensuelle)	M	janvier 1991	M+1	BdF
enq_BDF	industrie	évol. passée de la prod. (M-1) – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	évol. globale des commandes – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	évol. des commandes étrangères – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	évol. des stocks de produits finis – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	niveau des effectifs – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	niveau stocks de produits finis – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	niv. carnets de commandes – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	évol. prévue de la prod. (M+12) – (solde opinion)	M	décembre 1990	M+1	BdF
enq_BDF	industrie	climat des affaires (moyenne 100)	M	janvier 1987	M+1	BdF
enq_BDF	service	climat des affaires (moyenne 100)	M	avril 1989	M+1	BdF
enq_BDF	industrie	situation de trésorerie – (solde opinion)	M	octobre 1991	M+1	BdF

Source : DG Trésor.

Le « taux d'utilisation mensuel des capacités productives » dans l'industrie est pris en compte en évolution mensuelle. Au total, ce bloc est constitué de 12 variables mensuelles, disponibles au moins depuis octobre 1991.

Bloc 8 : enquêtes de conjoncture de la Banque de France en différence

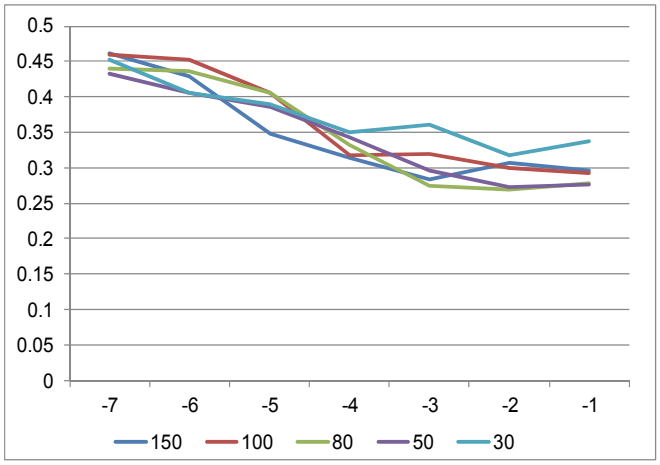
Type	Secteur	Série	Fréq	Début de la série	Délais	Source
enq_BDF_DIF	industrie	évol. passée de la prod. (M-1) – (dif. solde opinion)	M	février 1991	M+1	BdF
enq_BDF_DIF	industrie	évol. globale des commandes – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	évol. des commandes étrangères – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	évol. des stocks de produits finis – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	niveau des effectifs – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	niveau stocks de produits finis – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	niv. carnets de commandes – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	évol. prévue de la prod. (M+12) – (dif. solde opinion)	M	janvier 1991	M+1	BdF
enq_BDF_DIF	industrie	climat des affaires (dif. indice en moyenne = 100)	M	février 1987	M+1	BdF
enq_BDF_DIF	service	climat des affaires (moyenne 100)	M	mai 1989	M+1	BdF
enq_BDF_DIF	industrie	situation de trésorerie – (dif. solde opinion)	M	novembre 1991	M+1	BdF

Source : DG Trésor.

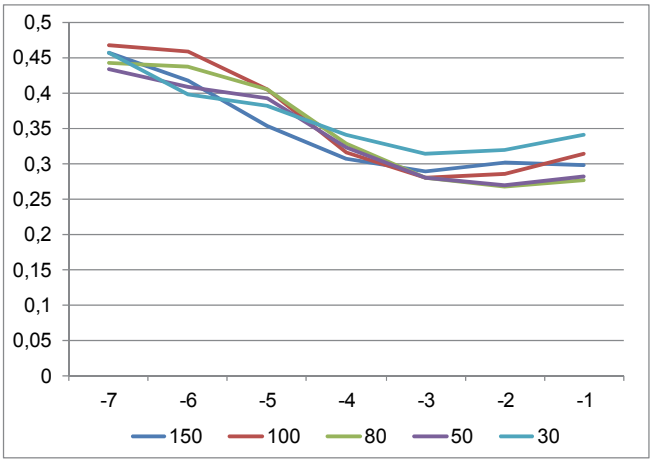
Le « taux d'utilisation mensuel des capacités productives » déjà pris en compte en évolution dans le bloc de variables 7 n'est pas repris ici. Au total, ce sont 11 variables mensuelles qui sont retenues ici, toutes disponibles depuis octobre 1991.

Annexe 2 : Sensibilité du RMSFE au paramétrage du modèle dans la mise en œuvre du LARS

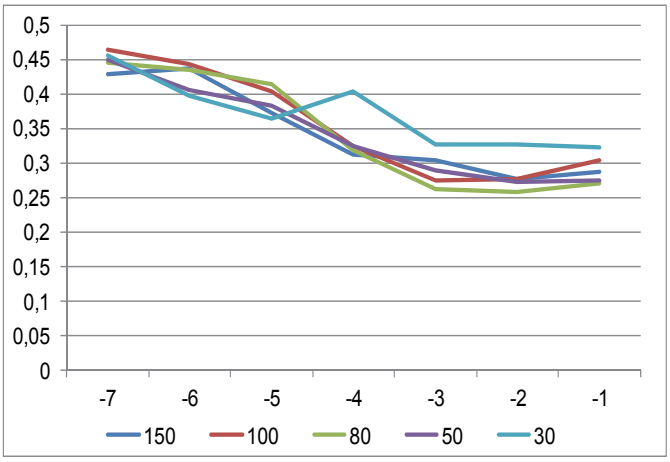
Graphique 3 : sensibilité du RMSFE en fonction du nombre de variables retenues et de l'horizon de prévision, pour $\lambda_2 = 0{,}1$



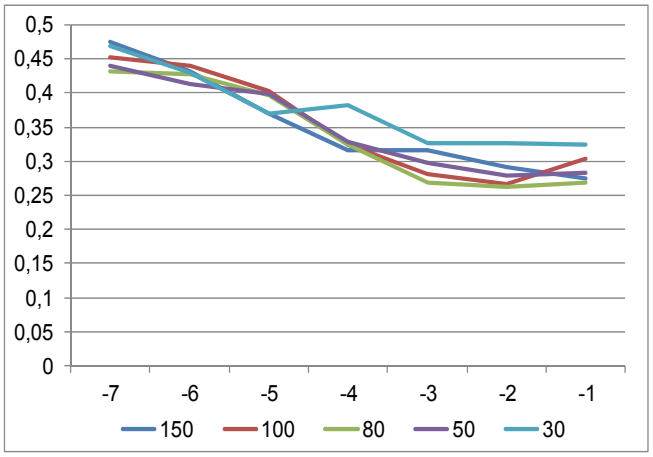
Graphique 4 : sensibilité du RMSFE en fonction du nombre de variables retenues et de l'horizon de prévision, pour $\lambda_2 = 0{,}2$



Graphique 5 : sensibilité du RMSFE en fonction du nombre de variables retenues et de l'horizon de prévision, pour $\lambda_2 = 0{,}3$



Graphique 6 : sensibilité du RMSFE en fonction du nombre de variables retenues et de l'horizon de prévision, pour $\lambda_2 = 0{,}4$



Source : DG Trésor.