



**HAL**  
open science

## Analyse et prévision du chômage

Clément Cariou

► **To cite this version:**

Clément Cariou. Analyse et prévision du chômage. Economies et finances. Nantes Université, 2022. Français. NNT: 2022NANU3011 . tel-03982723

**HAL Id: tel-03982723**

**<https://theses.hal.science/tel-03982723>**

Submitted on 10 Feb 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# THESE DE DOCTORAT DE

NANTES UNIVERSITE

ECOLE DOCTORALE N° 597  
*Sciences Economiques et sciences De Gestion*  
Spécialité : Sciences Économique

Par

**Clément CARIOU**

**Analyse et prévision du chômage**

**Thèse présentée et soutenue à l'Université de Nantes, le 16/12/2022**  
**Unité de recherche : LEMNA, IAE Nantes, Chemin de la Censive du Tertre – Bâtiments Erdre,**  
**44322 Nantes**

## **Rapporteurs avant soutenance :**

Yannick LE PEN      Maître de conférences – HDR, Université Paris-Dauphine  
Florent SARI        Professeur des Universités, Université Paris-Est Créteil

## **Composition du Jury :**

Président :      Isabelle CADORET      Professeur des Universités, Université de Rennes 1  
Examineurs : Isabelle CADORET      Professeur des Universités, Université de Rennes 1  
                 Yannick LE PEN      Maître de conférences – HDR, Université Paris-Dauphine  
                 Florent SARI      Professeur des Universités, Université Paris-Est Créteil

Dir. de thèse : Olivier DARNE      Professeur des Universités, Nantes Université

# Table des matières

Remerciements	4
Liste des figures	7
Liste des tables	9
Introduction générale	13
Première partie : Le cycle d'emploi	27
1 Analyse du cycle des demandeurs d'emploi français : une application nationale, régionale et départementale	27
1.1 Introduction . . . . .	27
1.2 Méthodologie . . . . .	32
1.2.1 Méthode de désaisonnalisation . . . . .	32
1.2.2 Analyse des cycles . . . . .	43
1.3 Données . . . . .	48
1.3.1 Analyse des points atypiques . . . . .	49
1.3.2 Analyse de la désaisonnalisation . . . . .	55
1.4 Application et résultats . . . . .	65
1.4.1 Le cycle des affaires . . . . .	65
1.4.2 Le cycle d'accélération . . . . .	72
1.4.3 Synchronisation cyclique . . . . .	82
1.5 Conclusion . . . . .	86
1.6 Annexes . . . . .	88
2 Analyse du cycle de l'emploi : une étude régionale	100

2.1	Introduction . . . . .	100
2.2	Résultats . . . . .	102
2.2.1	Datation des cycles d’emploi régionaux . . . . .	102
2.2.2	Caractéristiques . . . . .	105
2.2.3	Datation des points de retournement des cycles d’accélération régionaux de l’emploi . . . . .	109
2.2.4	Synchronisation cyclique régionale . . . . .	116
2.3	Conclusion . . . . .	120

**Deuxième partie : Forecasting french jobseeker 128**

**3 Forecasting French jobseekers with Google Trends 130**

3.1	Introduction . . . . .	131
3.2	Literature Review . . . . .	132
3.3	Methodology . . . . .	138
3.4	Data . . . . .	141
3.5	Empirical results . . . . .	143
3.5.1	Nowcasting results . . . . .	143
3.6	Conclusion . . . . .	145

**4 Nowcasting French jobseekers using data reduction methods<sup>1</sup> 158**

4.1	Introduction . . . . .	159
4.2	Data . . . . .	160
4.3	Methodology . . . . .	162
4.3.1	Penalized regressions . . . . .	162
4.3.2	Random forest . . . . .	165
4.3.3	Dynamic factor model . . . . .	167
4.4	Nowcasting scheme . . . . .	168
4.5	Empirical results . . . . .	169
4.5.1	Variable selection . . . . .	169
4.5.2	Nowcasting results . . . . .	170
4.6	Conclusion . . . . .	173

---

1. Ce chapitre est composé d’un travail co-écrit avec Olivier DARNE

<b>5</b>	<b>Are national or regional surveys useful for nowcasting jobseekers? The case of the French region of Pays-de-la-Loire<sup>2</sup></b>	<b>191</b>
5.1	Introduction . . . . .	191
5.2	Data . . . . .	194
5.3	Methodology . . . . .	195
5.3.1	Penalized regressions . . . . .	195
5.3.2	Random forest . . . . .	196
5.3.3	Dimensionality reduction approaches . . . . .	198
5.3.4	Dynamic factor model . . . . .	202
5.4	Nowcasting scheme . . . . .	203
5.5	Empirical results . . . . .	204
5.5.1	Variable selection . . . . .	204
5.5.2	Nowcasting results . . . . .	205
5.6	Conclusion . . . . .	207

---

2. Ce chapitre est composé d'un travail co-écrit avec Amélie CHARLES et Olivier DARNE

# Remerciements

Commencer ces remerciements par un paragraphe entièrement rédigé par une intelligence artificielle pourrait être un bel éloge à cette thèse qui parle entre autres de *Machine Learning*. Néanmoins, s'il y a bien une partie de cette thèse qui ne doit pas être codifiée par la norme du passé, c'est bien celle-ci.

En premier lieu, je voudrais remercier Olivier DARNE, mon directeur de mémoire de M1 puis de M2 et aussi mon directeur de thèse, car sans son engagement et sans son aide, cette thèse n'existerait pas. Je ne songeais même pas à faire une thèse après avoir obtenu mon Master, mais Olivier a su me convaincre en me donnant espoir dans mes capacités, l'amour pour l'enseignement a fait le reste.

Je souhaite remercier aussi mon jury de thèse composé d'Isabelle CADORET, Florent SARI et Yannick LE PEN. Les rapports de thèse de Yannick et Florent ont été un vrai soutien pour la préparation de la soutenance. La présidence d'Isabelle a grandement contribué à une soutenance de thèse très agréable.

Un grand merci à l'ensemble de mes collègues enseignants, enseignants-chercheurs et administratifs avec qui j'ai eu la chance de travailler. Un merci encore plus grand pour les collègues avec qui j'ai eu l'occasion de faire autres choses que travailler. Par conséquent, merci à Jean-Baptiste, Etienne, Eric, Nikola, Jules, Jeremy, Guillaume, Vincenzo, Eléazar, Racine, Pascaline, Houneyda, Anh, Minh, Lucie, Raphaël, Sophia, Théo, Aurélien et à tous ceux que j'oublie. C'est le moment où le lecteur se demande si la liste suit un ordre particulier. Ai-je classé par préférence ou par importance? Ou encore ai-je classé les prénoms en suivant un ordre chronologique? Je laisserai planer le doute. En revanche, je peux vous dire qu'il reste encore quelques noms qu'il me doit de citer.

Alors, merci à Pauline pour sa bonne humeur et son soutien indéfectible. Je me rappellerai toujours de cette capacité à maîtriser aussi bien les jeux de sociétés que la théorie des jeux. Mais je me souviendrai aussi de ses *escapes games*, des tisanes au miel (alors qu'elle n'aime pas le miel) ainsi que de son niveau d'escalade digne des plus grandes Kermit. Je voudrais remercier maintenant tous les membres de l'association des amateurs de (vrai) café de l'IAE Nantes.

Alors merci à Léo pour son rôle de second de cordée. Une expression n'aura jamais été aussi bien employée pour décrire son rôle sur le plan amical et professionnel. Je suis heureux de

partager actuellement son bureau même si son implication dans la vaisselle est limitée...

Un grand merci à Florian pour m'avoir secondé dans la lutte contre l'empire. Une référence à Star Wars s'imposait, car sur le plan professionnel Florian a été un véritable maître Jedi. Je pourrai citer son appui en économie du travail, microéconomie ainsi qu'en économétrie, mais il serait plus pertinent de mentionner nos conversations du matin (du midi, de l'après-midi et aussi du soir) car même si elles débordaient très légèrement du cadre du travail, elles étaient d'une aide précieuse et elles ont contribué indéniablement à la réussite de cette thèse. Je suis particulièrement fière d'avoir co-construit R5-B12 (un droïde bouilloire) avec lui.

Je voudrais aussi remercier ma famille qui m'a toujours soutenu dans l'accomplissement de cette thèse. L'éducation de ma mère et de mon père a sans aucun doute fait de moi ce que je suis aujourd'hui. Par ailleurs, l'aide précieuse de ma sœur résonnera toujours dans ma réussite puisque sans elle, l'obtention de ma licence aurait été plus compliquée. Je me dois aussi de remercier mes chats qui m'ont aidé à affronter ce monde en le rendant plus agréable, alors merci à Dobby, Dali, Neko, Ori et Mysti.

Pour terminer, il me reste un dernier remerciement à allouer à ma compagne. En effet, si l'on pense aux collègues et aux amis, c'est bien ma compagne qui était à mes côtés pendant l'intégralité de ce parcours de thèse. Merci de ne pas avoir fui quand tu as vu ce que c'était de faire une thèse. Merci de m'avoir aidé dans les moments de doutes. Merci de m'aimer comme tu le fais, car c'est un vrai carburant pour avancer. Enfin, merci pour le début d'une nouvelle histoire, une histoire qui s'écrira nous plus à deux, mais à quatre (enfin six avec les chats, c'est important les chats).



# Liste des figures

# Table des figures

1.1	<i>Effet des différents types de points atypiques</i> . . . . .	42
1.2	<i>Phases d'expansion et de contraction stylisées</i> . . . . .	45
1.3	<i>Série demandeurs d'emploi nationale catégorie A-Brute, spectre autorégressif et périodogramme</i> . . . . .	56
1.4	<i>Série demandeurs d'emploi nationale catégorie A-CVS-CJO par TRAMO-SEATS (a), Tendance-Cycle (b), composante saisonnière (c) et composante irrégulière (d)</i> . . . . .	57
1.5	<i>Comparaison DEFM Pôle Emploi CVS versus série ajustée par TRAMO-SEATS - National</i> . . . . .	59
1.6	<i>Comparaison Trend DEFM Pôle Emploi CVS versus Trend série ajustée par TRAMO-SEATS - National</i> . . . . .	60
1.7	<i>Série nationale des demandeurs d'emploi brute catégorie A national et Tendance-Cycle - (1996-2018)</i> . . . . .	61
1.8	<i>Série régionale des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	61
1.9	<i>Série départementale (Loire-Atlantique) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	62
1.10	<i>Série départementale (Maine-et-Loire) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	62
1.11	<i>Série départementale (Mayenne) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	63
1.12	<i>Série départementale (Sarthe) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	63
1.13	<i>Série départementale (Vendée) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)</i> . . . . .	64

1.14	<i>Comparaison du cycle des affaires français et du cycle des demandeurs d'emploi français - 1996-2018</i> . . . . .	71
1.15	<i>Taux de croissance mensuel de la série des demandeurs d'emploi national et taux de croissance lissé - (1996-2018)</i> . . . . .	73
1.16	<i>Taux de croissance mensuel de la série des demandeurs d'emploi des Pays de la Loire et taux de croissance lissé - (1996-2018)</i> . . . . .	74
1.17	<i>Taux de croissance mensuel de la série des demandeurs d'emploi de la Loire-Atlantique (44) et taux de croissance lissé - (1996-2018)</i> . . . . .	74
1.18	<i>Taux de croissance mensuel de la série des demandeurs d'emploi du Maine et Loire (49) et taux de croissance lissé - (1996-2018)</i> . . . . .	75
1.19	<i>Taux de croissance mensuel de la série des demandeurs d'emploi de la Mayenne (53) et taux de croissance lissé - (1996-2018)</i> . . . . .	75
1.20	<i>Taux de croissance mensuel de la série des demandeurs d'emploi de la Sarthe (72) et taux de croissance lissé - (1996-2018)</i> . . . . .	76
1.21	<i>Taux de croissance mensuel de la série des demandeurs d'emploi de la Vendée (85) et taux de croissance lissé - (1996-2018)</i> . . . . .	76
1.22	<i>Comparaison des cycles d'accélération de l'activité et des demandeurs d'emploi - France - 1996-2018</i> . . . . .	81
1.23	<i>Indice de Moran modifié – k=61</i> . . . . .	84
1.24	<i>Processus de l'algorithme de Bry-Boschan</i> . . . . .	88
1.25	<i>Série demandeurs d'emploi Pays de la Loire catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	89
1.26	<i>Série demandeurs d'emploi Loire-Atlantique (44) catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	89
1.27	<i>Série demandeurs d'emploi Maine et Loire (49) catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	90
1.28	<i>Série demandeurs d'emploi Mayenne (53) catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	90
1.29	<i>Série demandeurs d'emploi Sarthe (72) catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	91
1.30	<i>Série demandeurs d'emploi Vendée (85) catégorie A-Brute, spectre auto-régressif et périodogramme</i> . . . . .	91

1.31	<i>Datation des cycles des demandeurs d'emploi - série nationale</i> . . . . .	93
1.32	<i>Datation des cycles des demandeurs d'emploi - série Pays de la Loire</i> . . . . .	94
1.33	<i>Datation des cycles des demandeurs d'emploi - série Loire-Atlantique 44</i> . . . . .	95
1.34	<i>Datation des cycles des demandeurs d'emploi - série Maine et Loire 49</i> . . . . .	96
1.35	<i>Datation des cycles des demandeurs d'emploi - série Mayenne 53</i> . . . . .	97
1.36	<i>Datation des cycles des demandeurs d'emploi - série Sarthe 72</i> . . . . .	98
1.37	<i>Datation des cycles des demandeurs d'emploi - série Vendée 85</i> . . . . .	99
2.1	<i>Indice de Moran modifié des cycles régionaux de l'emploi, <math>k=61</math></i> . . . . .	118
3.1	DEFM and search intensity (seasonally adjusted) - Pôle Emploi and Google Trends . . . . .	142
3.2	Forecasting strategy - Authors . . . . .	143
3.3	CSSED GI 4m full period $h = 1$ - Authors . . . . .	155
3.4	CSSED GI 4m full period $h = 2$ - Authors . . . . .	155
3.5	CSSED GI 4m full period $h = 3$ - Authors . . . . .	156
3.6	CSSED GI 3m non-crisis period $h = 1$ - Authors . . . . .	156
3.7	CSSED GI 4m non-crisis period $h = 2$ - Authors . . . . .	157
3.8	CSSED GI 4m non-crisis period $h = 3$ - Authors . . . . .	157
4.1	French jobseeker growth rate - 2004M6-2021M12. . . . .	162
4.2	Example of a regression tree (cf. Hastie et al., 2001) . . . . .	166
4.3	Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the small database. . . . .	188
4.4	Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the large database. . . . .	188
4.5	Cumulative sum of squared forecast error (CUMSFE) from the small and large databases. . . . .	189
4.6	Cumulative squared prediction errors during the non-crisis period – 2016M1-2019M12 – with the small database. . . . .	189
4.7	Cumulative squared prediction errors during the crisis period – 2020M1-2021M12 – with the large database. . . . .	190
5.1	Pays-de-la-Loire and French jobseeker growth rate - 2004M6-2021M12. . . . .	193

5.2	PdL's jobseeker growth rate - 2004M6-2021M12. . . . .	195
5.3	Example of a regression tree (cf. Hastie et al., 2001) . . . . .	197
5.4	Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the regional database. . . . .	223
5.5	Cumulative sum of squared forecast error (CUMSFE) from the regional and large databases. . . . .	223

# Liste des tables

# Liste des tableaux

1.1	Modèles SARIMA prédéfinis . . . . .	37
1.2	Statistiques descriptives des données brutes, en milliers . . . . .	49
1.3	Sélection du meilleur modèle selon AIC . . . . .	50
1.4	Synthèse des <i>outliers</i> communs . . . . .	54
1.5	Synthèse des <i>outliers</i> spécifiques . . . . .	54
1.6	Statistiques descriptives des données CVS de Pôle Emploi, en millier . . . . .	58
1.7	Statistiques descriptives des données CVS selon l’auteur, en millier . . . . .	58
1.8	Datation du cycle d’emploi – 1996-2019 . . . . .	66
1.9	Caractéristiques du cycle de l’emploi – 1996-2019 . . . . .	69
1.10	Ratios moyens des durées et des amplitudes entre les phases cycliques . . . . .	70
1.11	Datation du cycle d’accélération de l’emploi – 1996-2019 . . . . .	77
1.12	Anticipation des points de retournement (creux) du cycle d’accélération de l’emploi par rapport au cycle de l’emploi . . . . .	78
1.13	Anticipation des points de retournement (pics) du cycle d’accélération de l’emploi par rapport au cycle de l’emploi . . . . .	79
1.14	Caractéristiques du cycle d’accélération de l’emploi – 1996-2019 . . . . .	80
1.15	Indices de concordances et coefficients de corrélation – Cycle de l’emploi . . . . .	82
1.16	Indices de concordances et coefficients de corrélation – Cycle d’accélération de l’emploi . . . . .	83
1.17	Datation des cycles d’emploi régionaux – 1996-2019 . . . . .	92
1.18	Datation des points de retournement du cycle des demandeurs d’emploi - série nationale . . . . .	93
1.19	Caractéristiques du cycle classique des demandeurs d’emploi - série nationale . . . . .	93
1.20	Datation des points de retournement du cycle des demandeurs d’emploi - série Pays de la Loire . . . . .	94

1.21	Caractéristiques du cycle classique des demandeurs d'emploi - série Pays de la Loire . . . . .	94
1.22	Datation des points de retournement du cycle des demandeurs d'emploi - série Loire-Atlantique 44 . . . . .	95
1.23	Caractéristiques du cycle classique des demandeurs d'emploi - série Loire-Atlantique . . . . .	95
1.24	Datation des points de retournement du cycle des demandeurs d'emploi - série Maine et Loire 49 . . . . .	96
1.25	Caractéristiques du cycle classique des demandeurs d'emploi - série Maine et Loire . . . . .	96
1.26	Datation des points de retournement du cycle des demandeurs d'emploi - série Mayenne 53 . . . . .	97
1.27	Caractéristiques du cycle classique des demandeurs d'emploi - série Mayenne	97
1.28	Datation des points de retournement du cycle des demandeurs d'emploi - série Mayenne 72 . . . . .	98
1.29	Caractéristiques du cycle classique des demandeurs d'emploi - série Sarthe 72	98
1.30	Datation des points de retournement du cycle des demandeurs d'emploi - série Vendée 85 . . . . .	99
1.31	Caractéristiques du cycle classique des demandeurs d'emploi - série Vendée 85	99
2.1	Statistiques descriptives des données brutes régionales de la DEFM catégorie A, en millier . . . . .	101
2.2	Datation des cycles d'emploi régionaux – 1996-2019 . . . . .	103
2.3	suite : Datation des cycles d'emploi régionaux – 1996-2019 . . . . .	104
2.4	Caractéristiques du cycle de l'emploi – 1996-2019 . . . . .	107
2.5	Suite : Caractéristiques du cycle de l'emploi – 1996-2019 . . . . .	107
2.6	Suite : Caractéristiques du cycle de l'emploi – 1996-2019 . . . . .	108
2.7	Statistiques descriptives des données brutes régionales de la DEFM catégorie A, en millier . . . . .	109
2.8	Anticipation moyenne des phases d'accélération sur les phases d'expansions	110
2.9	Anticipation moyenne des phases décélérations sur les phases d'expansions .	110
2.10	Datation des cycles d'emploi d'accélération régionaux – 1996-2019 . . . . .	112



2.11	Suite : Datation des cycles d'accélération de l'emploi régionaux – 1996-2019 .	113
2.12	Caractéristiques des cycles d'accélération de l'emploi – 1996-2019 . . . . .	114
2.13	Suite : Caractéristiques des cycles d'accélération de l'emploi – 1996-2019 . .	114
2.14	Suite : Caractéristiques des cycles d'accélération de l'emploi – 1996-2019 . .	115
2.15	Suite : Datation des cycles d'accélération de l'emploi régionaux – 1996-2019 .	117
3.1	Literature review . . . . .	150
3.2	Literature review . . . . .	151
3.3	Literature review . . . . .	152
3.4	Literature review . . . . .	153
3.5	Nowcast performance during the full out-of-sample period – 2017M5-2021M12 – stated value $R_{OOS}^2(AR)$ . . . . .	154
3.6	Nowcast performance during the non-crisis period – 2017M5-2019M12 – stated value $R_{OOS}^2(AR)$ . . . . .	154
3.7	Nowcast performance during the crisis period – 2020M1-2021M12 – stated value $R_{OOS}^2(AR)$ . . . . .	154
4.1	Variables names and abbreviations - 1/3 . . . . .	179
4.2	Variables names and abbreviations 2/3 . . . . .	180
4.3	Variables names and abbreviations -3/3 . . . . .	181
4.4	The most important variables during the full out-of-sample period – 2016M1- 2021M12 – with the small database. . . . .	181
4.5	The most important variables during the full out-of-sample period – 2016M1- 2021M12 – with the large database. . . . .	182
4.6	The most important variables during the non-crisis period – 2016M1-2019M12 – with the small database. . . . .	182
4.7	The most important variables during the crisis period – 2020M1-2021M12 – with the small database. . . . .	183
4.8	The most important variables during the non-crisis period – 2016M1-2019M12 – with the large database. . . . .	184
4.9	The most important variables during the crisis period – 2020M1-2021M12 – with the large database. . . . .	184

4.10	Nowcast performance during the full out-of-sample period – 2016M1-2021M12 – with the small database. . . . .	185
4.11	Nowcast performance during the full out-of-sample period - 2016M1-2021M12 – with the large database. . . . .	185
4.12	Nowcast performance during non-crisis period - 2016M1-2019M12 – with the small database. . . . .	186
4.13	Nowcast performance during crisis period - 2020M1-2021M12 – with the small database. . . . .	186
4.14	Nowcast performance during non-crisis period - 2016M1-2019M12 – with the large database. . . . .	187
4.15	Nowcast performance during crisis period - 2020M1-2021M12 – with the large database. . . . .	187
5.1	Variables names and abbreviations - 1/3 . . . . .	213
5.2	Variables names and abbreviations 2/3 . . . . .	214
5.3	Variables names and abbreviations -3/3 . . . . .	215
5.4	Variables names and abbreviations - regional level . . . . .	216
5.5	The most important variables during the full out-of-sample period – 2016M1- 2021M12 – with the small database. . . . .	216
5.6	The most important variables during the full out-of-sample period – 2016M1- 2021M12 – with the large database. . . . .	217
5.7	The most important variables during the non-crisis period – 2016M1-2019M12 – with the small database. . . . .	217
5.8	The most important variables during the crisis period – 2020M1-2021M12 – with the small database. . . . .	218
5.9	The most important variables during the non-crisis period – 2016M1-2019M12 – with the large database. . . . .	218
5.10	The most important variables during the crisis period – 2020M1-2021M12 – with the large database. . . . .	219
5.11	Nowcast performance during the full out-of-sample period – 2016M1-2021M12 – with the small database. . . . .	219

5.12	Nowcast performance during the full out-of-sample period - 2016M1-2021M12 – with the large database. . . . .	220
5.13	Nowcast performance during non-crisis period - 2016M1-2019M12 – with the small database. . . . .	220
5.14	Nowcast performance during crisis period - 2020M1-2021M12 – with the small database. . . . .	221
5.15	Nowcast performance during non-crisis period - 2016M1-2019M12 – with the large database. . . . .	221
5.16	Nowcast performance during crisis period - 2020M1-2021M12 – with the large database. . . . .	222



” Vis au présent, souviens-toi du passé, et ne crains pas l’avenir, car il n’existe pas et n’existera jamais. Seul compte l’instant présent. ” Christopher Paolini

À mes enfants bientôt de ce monde,

# Introduction générale

Le chômage sous sa forme conceptuelle existe certainement depuis que l'homme s'est constitué en société. Pourtant, il nous est impossible de remonter aux origines de l'Antiquité grecque afin d'en proposer un exemple. L'explication la plus évidente à ce manque d'érudition réside dans l'inexistence du terme "travail", dans la langue grecque ancienne. De fait, il paraît logique que l'idée de chômer et de rechercher un travail n'existât pas non plus, au grand dam de cette introduction de thèse de doctorat. Nous souhaitons faire un bond dans le futur, en passant par l'apparition du terme "chômer" en France au XIIe siècle et du terme "chômage" le siècle suivant. Nous continuons notre avance rapide du temps, retrouvant ainsi les chômeurs au sens moderne décrit par Zola dans sa nouvelle "Le chômage".

*"Mes enfants, il n'y a pas de travail aujourd'hui... Les commandes n'arrivent plus ; de tous les côtés, je reçois des contre-ordres, je vais rester avec de la marchandise sur les bras"*

Plus généralement, c'est l'apparition du chômage tel que nous le connaissons aujourd'hui. Le chômage lié à la naissance du salariat. Cette notion du chômage se rapproche du chômage conjoncturel décrit par John Maynard Keynes. Mais en dehors des causes conjoncturelles ou structurelles, nous retrouvons l'absence de travail pour une personne désireuse d'offrir son temps, le conduisant à un manque de rémunération et de ressources. C'est à partir du XIXe siècle que les économistes, sociologues et statisticiens s'emparent de la notion et l'étudient afin de la définir théoriquement et de la mesurer. Adam Smith, David Ricardo, Karl Marx n'utilisent pas le terme "chômage" et ne cherchent pas à l'établir comme un objet d'étude à part entière. C'est Alfred Marshall qui dessinera les premiers contours d'une analyse sur le chômage avec une conception classique du terme. Ce saut dans le temps et l'espace nous amène à la définition la plus aboutie de notre époque pour décrire le chômage, tel qu'il est perçu par notre société, celle du Bureau international du Travail (BIT). Le chômage comptabilise les personnes en âge de travailler (conventionnellement 15 ans ou plus) qui :

- N'ont pas travaillé, ne serait-ce qu'une heure, au cours de la semaine de référence ;
- Sont disponibles pour travailler dans les deux semaines ;
- Ont entrepris des démarches actives de recherche d'emploi dans le mois précédent, ou ont trouvé un emploi qui commence dans les 3 mois.

La mesure du chômage réalisée par l'Institut National des Statistiques et des Études Économiques (INSEE) suit ces critères. Cela permet l'instauration d'une mesure commune propice aux

comparaisons internationales et aux analyses macroéconomiques. Dans ce cas, le chômage découle d'une enquête (l'enquête Emploi) et est mis à disposition tous les trimestres. La France possède une seconde mesure du "chômage". La mesure de Pôle Emploi qui collecte et transmet les inscriptions des "demandeurs d'emploi" auprès de leur service. Nous nous concentrerons sur le taux de croissance des demandeurs d'emploi en fin de mois (DEFM). C'est une mesure administrative et déclarative dont la catégorie A se rapproche conceptuellement de la mesure de l'Insee. En effet, la catégorie A des DEFM rassemble toutes les personnes sans emploi, tenues d'accomplir des actes positifs de recherche d'emploi, et à la recherche d'un emploi quel que soit le type de contrat (CDI, CDD, à temps plein, à temps partiel, temporaire ou saisonnier). La statistique est disponible en fréquence mensuelle, ce qui amplifiera son intérêt à notre égard. En effet, la littérature présente le chômage comme une variable anticipant les fluctuations du cycle de l'activité économique. Dès lors, l'étude du cycle d'emploi devient un sujet d'importance majeur, notamment avec l'utilisation des données des demandeurs d'emploi de Pôle Emploi. Les données fournies par Pôle Emploi sont disponibles avec un délai plus court, d'environ 24 jours après la fin du mois de référence, alors que les données du chômage selon l'Insee sont disponibles avec un délai d'environ 48 jours, après la fin du trimestre de référence. La fréquence de la variable ainsi que ses caractéristiques intrinsèques correspondent davantage à ce que nous recherchons afin d'établir un diagnostic conjoncturel de l'emploi. Par conséquent, la variable des DEFM constituera le fil rouge de cette thèse, tant sur la partie analyse que prévision.

## **Analyse du cycle**

La notion de cycle est très ancienne, on la retrouve en Chine avec l'expression du Yin et du Yang, littéralement expansion et contraction, deux termes que nous retrouverons à maintes reprises. La notion de cycle telle que nous le concevons remonte aux travaux de Clément Juglar et de William Stanley Jevons, où l'intérêt est orienté sur l'analyse des cycles avec un recours aux séries statistiques afin de mesurer le cycle. L'analyse des cycles est exclusivement liée à l'analyse des fluctuations économiques et la recherche de leurs causes. Même si Moore (1914) et Persons (1919) ont la paternité de l'utilisation rigoureuse des statistiques sur des séries longues afin de caractériser les cycles économiques, ce sont les travaux de Mitchell (1913) et Burns et Mitchell (1946) qui ont inspiré nos travaux sur la mesure du cycle. Le cycle classique est défini comme l'alternance des phases d'expansion et des phases de



récession. Les phases d'expansion correspondent à une hausse de la série temporelle étudiée alors que les phases de récession correspondent à une baisse. Ces phases sont délimitées par des points de retournement qui correspondent à des extrema locaux de la série étudiée ; Les maxima locaux (pics) représentent les plus hauts niveaux de la série et les minima locaux (creux) correspondent au plus bas niveau de la série. Traditionnellement, l'analyse des cycles économiques est réalisée à partir des séries chronologiques trimestrielle.

Comme l'explique parfaitement Jacky Fayolle en 2003, l'étude des cycles connaît un clivage entre l'investigation théorique et l'analyse conjoncturelle. Fayolle identifie quatre époques différentes. La première époque, la cohabitation positiviste, fait référence à l'influence réciproque entre les deux champs d'étude. Nous retrouvons des auteurs d'importances comme Joseph Schumpeter qui donne la réplique à Clément Juglar dans leurs travaux d'inauguration de l'analyse cyclique. Cette époque concentre aussi les développements des fameux baromètres des fluctuations où nous retrouverons le célèbre baromètre de Harvard construit par Warren Persons à l'aune de la Première Guerre mondiale. L'idée est d'objectiver et de matérialiser les mouvements oscillatoires de l'économie dans son ensemble. Nous renverrons le lecteur intéressé vers une présentation synthétique de ces baromètres par Vidal (2000).

La seconde époque identifiée par Fayolle, désignée comme la séparation agnostique, s'inscrit dans le parti-pris empirique des travaux du National Bureau of Economic research (NBER) américain. La définition du cycle classique (business cycle) proposée par Burns et Mitchell (1946) reste suffisamment souple et malléable afin que tous les chercheurs puissent se l'approprier. Le cycle est ainsi un phénomène reproductible, s'inscrivant dans nos économies contemporaines caractérisées par la concurrence et le marché. Même si l'idée du cycle unique et non décomposable semblait suffisamment souple pour être acceptée par tous les conjoncturistes. Les faits stylisés observés lors de l'après-guerre ont à nouveau relancé le débat sur la décomposition entre tendance et cycle. Les deux dernières époques, identifiées comme la professionnalisation sous influence keynésienne et le cycle entre déterminisme et incertitude font respectivement écho à la seconde et à la première. L'histoire de la pensée économique dépeint une certaine dualité dans l'étude des cycles et l'analyse conjoncturelle. Cette dualité de la notion de cycle est très bien décrite dans l'ouvrage du Colloquium on the History of Business Cycle Analysis (2003).

Nous allons nous concentrer sur la définition et le cadre d'analyse qui prévaudra dans cette

thèse. Ainsi, à la suite des fondations posées par le NBER, d'autres types de cycle apparaissent en réponse à ce besoin de décrire les fluctuations conjoncturelles. Ilse Mintz inaugure en 1969 le cycle de croissance. Ce cycle peut se définir comme la différence entre une série de l'activité économique et sa tendance de long terme. L'objectif est donc d'estimer les périodes de croissance supérieure à la tendance de long terme et les périodes de croissance ralentie. Plus récemment, les conjoncturistes ont développé le cycle d'accélération afin d'analyser les phases de hausse (accélération) et les phases de baisse (ralentissement) du taux de croissance de l'activité.

De nos jours, la recherche académique, c'est très largement développée en conservant le principe de séparation. L'essor des mesures réalisées par les institutions nationales ou internationales a motivé la recherche académique sur ce sujet. Comme décrit auparavant, le premier organisme d'importance notoire est le NBER, mais d'autres organismes tels que le Center for Economic Policy Research (CPER), l'Economic Cycle Research Institute (ECRI), l'Organisation de Coopération et de Développement Économique (OCDE) ou encore les Banques Centrales construisent une analyse conjoncturelle des cycles européens et internationaux. Parmi cette multitude d'analyses conjoncturelles, certains pays se sont davantage emparés de la notion que d'autres. Nous pensons aux États-Unis avec la création du Dating Committee en 1978 sous l'égide du NBER qui propose une datation de référence des points de retournement du cycle classique de l'économie américaine. Cependant, il aura fallu attendre le 25 octobre 2021 pour que la France décide de se doter de son propre comité de datation des cycles de l'économie française (CDCEF) sous la supervision de l'Association Française de Science Économique (AFSE). Ces écarts d'intérêt sont moins visibles dans le monde académique même si la crise économique de 2008 a suscité un regain d'intérêt pour cette discipline. De nombreux auteurs se sont attelés à décrire les cycles (Anas et alii, 2008 ; McKay et Reis, 2008 ; Ferrara, 2009 ; Darné et Ferrara, 2011 ; Belaire-France et Peiró, 2012) ou à améliorer les aspects méthodologiques de l'analyse conjoncturelle (Mazzi et Savio, 2006 ; Artis et al., 1997 ; Harding, 2002 ; Artis et al., 2004 ; Harding et Pagan, 2002 et 2006 ; Camacho et Quiros, 2006).

Les aspects méthodologiques s'intéressent essentiellement aux caractéristiques des cycles considérés lors d'une étude. Les cycles sont ainsi disséqués en trois caractérisés, la durée moyenne des phases d'expansion et de récession, l'amplitude de ces mêmes phases et le pro-

fil de l'intensité de l'inflexion conjoncturelle aussi notifiée par l'indice de sévérité mis en avant par Harding et Pagan (2002). Depuis les deux dernières décennies, l'analyse cyclique s'accompagne d'une étude de la synchronisation et de la convergence des cycles.

## **La prévision**

Selon Alfred Sauvy dans "Conjoncture et prévision, 1969", la prévision permet d'annoncer, sous une forme ou l'autre, ce qui doit ou peut se passer dans l'avenir. La prévision est l'antichambre de la planification n'en déplaise à Friedrich Hayek, puisqu'elle risque d'inciter les pouvoirs publics à intervenir davantage. L'intérêt de la prévision économique réside dans la connaissance anticipée du futur, le plus souvent à court terme. Pourtant, cela reste encore aujourd'hui une tâche difficile. Comme dirait un grand maître selon Georges Lucas :

*"Difficile à voir. Toujours en mouvement est l'avenir"*

Pourquoi prévoir le niveau de chômage ? Quelles informations le prévisionniste délivre aux pouvoirs publics ? Les réponses sont évidentes, mais très vastes. Dans nos économies de marché, le chômage et l'emploi sont des indicateurs de la santé de l'économie. La démographie impact la population active qui impact elle-même le nombre de potentiels chômeurs. Au-delà du simple indicateur, le taux de chômage et surtout le nombre de demandeurs d'emploi conditionnent le montant des allocations chômage d'aide au retour à l'emploi. La prévision des DEFM indemnisés et dans une moindre mesure des DEFM intéressent les gestionnaires de l'assurance chômage, l'Union Nationale interprofessionnelle pour l'Emploi dans l'Industrie et le Commerce et Pôle Emploi. L'UNEDIC réalise ses propres prévisions financières afin d'éclairer les pouvoirs publics sur les potentielles modifications de budget alloué à l'assurance chômage dans les années à venir. En dehors de l'aspect purement comptable, les prévisions du chômage sont aussi un marqueur de la confiance des entreprises envers l'économie nationale. En effet, les tendances du chômage sont fortement corrélées à la confiance des entreprises. Ce constat sera un argument très utile pour justifier le choix de prévoir avec les enquêtes de conjonctures à disposition des chercheurs et de tous les acteurs économiques d'un pays. Comment les prévisions du chômage sont-elles réalisées ?

Il existe une multitude de techniques de prévision, le plus souvent issues de l'économétrie. Les grands principes restent les mêmes, quelle que soit la méthode utilisée. Dans un premier temps, on retrouve l'observation et la compréhension des faits passés. Dans un second

temps, le prévisionniste cherche à relier les variables économiques entre elles à l'aide d'une représentation sous forme d'un modèle plus ou moins complexe. Puis, vient la prévision après l'estimation de ces modèles en suivant les règles de l'économétrie. Il existe trois grandes familles de méthodes quantitatives assurant la fonction d'effectuer des prévisions économiques. La première approche est celle qui utilise les indicateurs synthétiques permettant de déterminer de manière anticipée les phases d'expansion ou de récession de l'activité économique d'un pays. Ceci s'inscrit dans la logique de notre première partie de thèse. Nous retrouvons parmi ces indicateurs de prévision le baromètre de Harvard, qui à lui seul présente les limites de ce type d'approche. L'institut Harvard rappelle que ce baromètre a été impeccable de prévoir la crise de 1929. Le principal problème de ces méthodes réside dans la juxtaposition d'indices synthétiques qui pouvaient accuser des corrélations positives et négatives en fonction de la période temporelle considérée.

Une seconde approche faisant écho à la première est la prévision des séries temporelles. Les prévisions sont fondées sur les propriétés statistiques des chroniques étudiées. Dans ce cas, le prévisionniste cherche à prévoir en captant l'information contenue dans la série, on parlera de processus stochastique. Quant à la troisième approche, elle s'inscrit dans les prévisions dites structurelles. Notre thèse se positionne dans la prévision des séries temporelles. Selon Arthur Charpentier, "l'étude des séries temporelles correspond à l'analyse statistique d'observations régulièrement espacées dans le temps". Depuis les découvertes de Yules et du processus autorégressif (AR) et de Slutsky avec le processus de moyenne mobile (MA), la discipline a été son apogée en 1970 avec l'ouvrage de Box et Jenkins "Time Serie Analysis, Forecasting and Control ». La démonstration des auteurs fait état que les processus de type ARMA pouvaient s'appliquer à de nombreux domaines. Dès lors, une littérature en prévision des séries temporelles s'amorçait. Le domaine est tellement vaste qu'il est difficile le présenter dans sa globalité. Cependant, nous pouvons nous concentrer sur la thématique de la prévision du chômage ainsi que sur les spécificités qui constituent cette thèse.

La prévision du chômage suit les mêmes règles que la prévision de l'activité économique. Le prévisionniste fait face à un double objectif. Le premier objectif consiste à récolter l'information la plus pertinente possible à des fins de prévision tout en conservant les principales hypothèses de stabilité des méthodes économétriques employées. Le second objectif est de définir un modèle qui améliore la prévision de la variable considérée. Ce devoir, de fournir des

prévisions fiables permet de combler, du moins en partie, l'incertitude de l'évolution future de l'économie. Cette incertitude est générée par le délai de publication de certaines variables d'intérêt permettant de mesurer l'état de l'économie, et plus précisément dans notre cas, du chômage. La prévision *via* les enquêtes de conjonctures apparaît alors comme un début très pertinent, notamment avec leurs délais de publication très précoces. Le premier article qui motiva la partie prévision de ce travail est celui de Boriss Siliverstovs (2012) sur l'apport des enquêtes de conjonctures à la prévision à court terme de l'emploi en Suisse. L'auteur cherche à répondre à l'ambivalence des résultats concernant le réel apport des enquêtes de conjoncture à la prévision. Hartle (1958) réalise une des premières études du sujet sur données canadiennes et aboutit à une conclusion négative. Abberger (2007b) conclut à un résultat positif sur données allemandes. La conclusion de Siliverstovs est positive. Les enquêtes de conjonctures possèdent un pouvoir informationnel pour la prévision de l'emploi. L'utilisation des enquêtes de conjoncture à des fins de prévision est donc un réel atout. D'une part, elles sont disponibles rapidement par rapport aux données administratives. D'autre part, l'information sur les sentiments et les opinions des entreprises délivrent une information non négligeable pour les objectifs de prévision.

Nos recherches de variables pertinentes à la prévision de l'emploi se sont rapidement tourné vers les webdata, ou en français, les données issues d'internet. Quoi de mieux, pour un chercheur, que de réussir à capter l'intensité de recherche d'emploi des individus à l'échelle d'un pays ou d'une région ? Et surtout en temps réel. Cela peut paraître angoissant de savoir que nos recherches internet nous définissent. Google, Amazon, Facebook et tous les grands de ce secteur nous connaissent probablement mieux que nos voisins, voir que de nous-mêmes. Cependant, cette connaissance peut être mise au service de la recherche. C'est ce que Hyunyoung Choi et Hal Varian ont tenté de démontrer avec une série d'articles mettant en avant l'utilisation de Google Trends (Choi et Varian, 2009, 2012). Google Trends, anciennement Google Insights, est une plateforme mise à disposition par Google. L'outil permet de connaître la fréquence à laquelle un terme de recherche a été tapé dans le moteur de recherche Google. La plateforme est d'autant plus puissante que le moteur de recherche Google est utilisé par 2,2 milliards d'individus chaque jour selon l'entreprise. Nous verrons au cours de cette thèse les fonctionnalités qu'offre Google Trends en matière de création de chroniques. Par ailleurs, la littérature sur le sujet à exploser ces dernières années, avec des applications sur la plupart des pays développés, c'est-à-dire, là où le moteur de recherche Google est le plus utilisé. (Askitas

et Zimmermann, 2009 ; Suhoy, 2009 ; Fondeur et Karamé, 2013 ; D’Amuri et Marcucci, 2017 ; Nagao et al., 2019).

L’ère du Big data a profondément changer le rapport aux données que pouvait entretenir un chercheur. L’abondance et la relative disponibilité permettent aux acteurs du monde académique d’accéder à beaucoup d’information. Il survient alors, le problème de traitement de cette nouvelle masse d’information. La statistique et l’économétrie cohabitent à merveille pour repousser les frontières du possible en matière de traitement de données. Nous nous intéresserons ici uniquement à ce que nous utilisons dans cette thèse. L’un des premiers moyens pour traiter cette masse de données est le raisonnement économique et l’avis d’un expert. Non, nous ne prendrons pas le prix au kilo des pommes de terre au Mexique afin de prévoir le chômage en France. Cependant, quand le nombre de variables pertinentes sélectionnées est grand, l’utilisation de certaines méthodes de sélection aboutit généralement à de bons résultats.

L’étude fondatrice de Stock et Watson (2002) posent les bases de la sélection de variables dans une optique de prévision. La sélection de variables revêt plusieurs formes allant des méthodes d’analyse en composante principale, des modèles à facteur et des techniques de régression pénalisées. Les régressions pénalisées prennent racine dans l’apprentissage automatique, le Machine Learning. Ce dernier se fonde sur des approches statistiques pour donner aux ordinateurs la capacité d’ ”apprendre” à partir de données. Le meilleur exemple reste l’apprentissage par une machine du jeu d’échecs. L’homme a appris à la machine les règles du jeu d’échecs (le modèle), puis la nourrit d’un historique important d’anciennes parties (les données), enfin la machine à estimer une fonction d’évaluation en maximisant son gain (l’objectif). Dans notre objectif de prévision, nous utiliserons des techniques issues du Machine Learning supervisée, c’est-à-dire que l’ordinateur doit réaliser une tâche d’apprentissage automatique avec un objectif de prévision à partir d’un modèle préexistant.

La sélection de variables *via* des méthodes économétriques ou de Machine Learning pour la prévision de l’activité reste majoritaire dans les premiers développements de la discipline. (Bai et Ng, 2008b ; Stock et Watson, 2012 ; Kim et Swanson, 2018 ; Babii et al., 2022). En outre, la littérature reste très réduite autour de la prévision du chômage *via* la sélection de variables ou l’utilisation de l’apprentissage automatique. (Aiken, 1995)

## Démarche et plan de thèse

Les enjeux liés à l'analyse conjoncturelle du chômage se sont considérablement renforcés. Les mesures des cycles économiques se concentrent sur les variables dites de références telles que le Produit Intérieur Brut ou l'activité en général. Le chômage n'est pas considéré comme une variable à part entière. Le Dating Committee mentionné en début d'introduction étudie la relation du cycle économique avec le cycle du taux de chômage et depuis peu, avec les demandes d'indemnisation des chômeurs. Nous avons vu que les préoccupations françaises en matière de cycle étaient plus récentes, bien qu'étudiées depuis plusieurs années dans le monde académique. Pourtant, la qualité du diagnostic conjoncturel engage la qualité des prévisions réalisées. Cette thèse s'intéresse à répondre à la problématique suivante : comment analyser de manière efficace la conjoncture au niveau du chômage ?

L'analyse conjoncturelle de l'emploi constitue un enjeu important sur les politiques publiques et économiques. Nous tenterons de répondre à cette question en deux parties. La première portera sur une analyse du cycle économique du chômage (chapitre 1 et 2). Les principaux objectifs de l'analyse cyclique sont de détecter en temps réel et d'anticiper autant que possible les fluctuations économiques du chômage (phases d'expansion et de contraction) en accordant une attention particulière aux points de retournement. Atteindre ces objectifs dépend fortement de la disponibilité d'une analyse fiable des principaux événements cycliques. En effet, lorsqu'on cherche, par exemple, à développer un indicateur de retournement cyclique, il est important de disposer d'une chronologie de référence des points de retournement des cycles à partir de laquelle on pourra juger de la qualité de l'indicateur. Ainsi, une datation précise des points de retournements passés et une description détaillée des mouvements cycliques (longueur, amplitude, symétrie) sont des éléments essentiels de cette analyse et développés dans cette thèse. De plus, une comparaison détaillée de la convergence et de la synchronisation des fluctuations cycliques parmi les départements de la région des Pays de la Loire, mais aussi entre les différentes régions françaises constitue un complément utile à l'analyse précédente afin d'identifier les comportements communs et idiosyncrasiques. La dimension régionale est une préoccupation importante pour les régions elles-mêmes, à l'échelle de la France afin de piloter les décideurs publics, mais aussi à l'échelle européenne. Depuis la création de la Communauté économique européenne, la politique régionale dans l'Union européenne (UE), connue sous l'appellation de politique de cohésion s'adresse à toutes les

régions et villes de l'UE afin de réduire les disparités régionales. Parmi ses objectifs, on retrouve la création d'emploi et la compétitivité des entreprises. À cet égard, on comprend tout l'intérêt que peut avoir un diagnostic conjoncturel jusqu'au niveau régional. Ce premier point sera traité à l'aide des données des demandeurs d'emploi en catégorie A, collectées par Pôle emploi.

La deuxième partie sera consacrée à l'élaboration d'un modèle de prévision mensuelle, du mois en cours, des demandeurs d'emploi en utilisant uniquement les données construites avec Google Trends (chapitre 3). Puis, nous réaliserons un second modèle de prévision mensuelle des demandeurs d'emploi au niveau national (Chapitre 4), mais également au niveau régional (Chapitre 5). Pour cela, on utilisera les enquêtes de conjoncture, notamment celles portant sur l'emploi et le chômage, réalisés par l'INSEE et la Banque de France. Ces enquêtes rendent compte des perceptions qu'ont les entreprises et les consommateurs de l'état actuel de la conjoncture ainsi que de ses évolutions récente et future. Leur publication précédant celle des données économiques, les décideurs peuvent s'en servir pour dresser assez tôt un premier état des lieux de l'état de la situation économique. Les nouvelles techniques d'apprentissage automatique et supervisé permettent l'amélioration des modèles de prévision. La sélection de variables sera donc un outil puissant afin d'améliorer la précision des prévisions des DEFM catégorie A.

Cette thèse possède un double objectif. Une contribution à la littérature existante sur l'analyse cyclique avec une approche originale en mettant l'accent sur la mesure du chômage de Pôle Emploi. Puis, après avoir analysé le contexte des cycles, notamment régionaux, faisant ressortir les enjeux locaux, nous cherchons à enrichir la connaissance académique en proposant un modèle de prévision des DEFM avec une identification des variables pertinentes pour une prévision de court terme aux niveaux national et régional. Plusieurs limites peuvent d'ores et déjà apparaître, la mesure du chômage choisie nous contraint à une analyse franco-française, peu généralisable à d'autres pays. Une deuxième et dernière limite seraient de ne considérer que le cas de la prévision en écartant l'intérêt de l'explication.



# Bibliographie

- [1] Abberger K. (2007b). "Qualitative Business Surveys and the Assessment of Employment : A Case Study for Germany", *International Journal of Forecasting*, vol. 23, n°2, pp. 249-258.
- [2] Anas J, Billio M, Ferrara L, Mazzi G-L (2008) A system for dating and detecting turning points in the euro area. *The Manchester School*, 76 (5), 549–577.
- [3] Artis M, Krolzig H-M, Toro J (2004) The European business cycle. *Oxford Economic Papers*, 56 (1), 1–44.
- [4] Artis M-J, Kontolemis Z-G, Osborn D-R (1997) Business cycles for G7 and European countries. *The Journal of Business*, 70 (2), 249–279.
- [5] Askitas, N., & Zimmermann, K. (2009). Google Econometrics and Unemployment Forecasting (No. 4201). Institute of Labour Economics (IZA).
- [6] Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304-317.
- [7] Babii, A., Ball, R.T., Ghysels, E., & Striaukas, J. (2022). Machine learning panel data regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3), 1094-1106.
- [8] Boriss S. (2012). L'apport des enquêtes de conjoncture auprès des entreprises à la prévision à court terme de l'emploi : une application en temps réel sur données suisses, *Economie & prévision* 1 (199), p. 93-109.
- [9] Box G-E, Jenkins G-M, Reinsel G (1970) *Time Series Analysis : Forecasting and Control*. Holden Day San Francisco.
- [10] Burns A-F, Mitchell W-C (1946) *Measuring Business Cycles*. NBER.

- [11] Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Google Inc, 1, 1-5.
- [12] Choi, H., & Varian, H. (2012). Predicting the present with Google Trends, *Economic Record*, 88, 2-9.
- [13] Camacho M, Perez-Quiros, G, Saiz L (2006) Are European business cycles close enough to be just one. *Journal of Economic Dynamics and Control*, 30 (9-10), 1687–1706.
- [14] D’Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801-816.
- [15] Darné O, Ferrara L (2011) Identification of slowdowns and accelerations for the Euro area economy. *Oxford Bulletin of Economics and Statistics*, 73 (3), 335–364.
- [16] Fayolle J. (2003), L’étude des cycles et l’analyse conjoncturelle dans l’histoire de la pensée économique dans le Colloquium on the History of Business Cycle Analysis, Luxembourg.
- [17] Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117-125.
- [18] Ferrara L (2009) Caractérisation et datation des cycles économiques en zone euro. *Revue Économique*, 60 (3), 703–712.
- [19] Harding D (2004) Non-parametric turning point detection, dating rules and the construction of the euro-zone chronology. Paper presented at the Eurostat. In Colloquium on modern tools for business cycle analysis.
- [20] Harding D, Pagan A (2002) Dissecting the cycle : A methodological investigation. *Journal of Monetary Economics*, 49 (2), 365–381.
- [21] Harding D, Pagan A (2006) Synchronization of cycles. *Journal of Econometrics*, 132 (1), 59–79.
- [22] Hartle D. (1958). “Predictions Derived from the Employment Forecast Survey”, *The Canadian Journal of Economics and Political Science / Revue canadienne d’Économie et de Science Politique*, vol. 24, n°3, pp. 373-390.
- [23] Jevons W.S. (1884), *Investigations in Currency and Finance*, Macmillan and co, London.
- [24] Juglar C, (1862) *Des Crises Commerciales et de leur retour périodique en France, en Angleterre et aux États-Unis*, Paris : Guillaumin.

- [25] Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2), 339-354.
- [26] Mazzi G, Savio G (2007) *Growth and Cycle in the Eurozone*. Palgrave-MacMillan.
- [27] McKay A, Reis R (2008) The brevity and violence of contractions and expansions. *Journal of Monetary Economics*, 55 (4), 738–751.
- [28] Mintz I, (1969) *Dating postwar business cycles : Methods and their application to Western Germany, 1950–67*. NBER.
- [29] Mitchell, W-C (1913) *Business Cycles*. Berkeley : University of California Press (pt. III reprinted as “Business cycles and their causes”, 1941).
- [30] Moore H-L, (1914) *Economic cycles : Their law and cause*. New York : The Macmillan Company.
- [31] Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, 30, 103-109.
- [32] Peiró A, Belaide-Franch J, Gonzalo M-T (2012) Unemployment, cycle and gender. *Journal of Macroeconomics*, 34 (4), 1167–1175.
- [33] Persons W-M, (1919) An index of general business conditions. *Review of Economic Statistics*, 1(1), 111-205.
- [34] Sauvy A., 1954, *La prévision économique*, Paris, PUF, coll. Que sais-je, 3ème édition ; 1ère ed. 1943 ; 5ème édition 1962.
- [35] Suhoy, T. (2009). Query indices and a 2008 downturn : Israeli data (No. 2009.06). Bank of Israel.
- [36] Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460), 1167-1179.
- [37] Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics*, 30(4), 481–493.
- [38] Vidal, J.F., 2000, *Dépression et retour de la prospérité, Les économies européennes à la fin du XIXe siècle*, L’Harmattan.

# Première partie : Le cycle d'emploi

# Chapitre 1

## Analyse du cycle des demandeurs d'emploi français : une application nationale, régionale et départementale

### 1.1 Introduction

De nos jours, l'analyse et la prévision des séries macroéconomiques sont au cœur des études économiques. En effet, l'analyse conjoncturelle s'est considérablement renforcée. Les décideurs politiques accordent beaucoup d'attention à ces enjeux de dimension nationale, régionale et même départementale. Jusqu'à très récemment, cet intérêt s'est principalement focalisé sur la croissance du PIB. Néanmoins, la connaissance de l'évolution des cycles du chômage et de ses points de retournement apparaît comme une contribution importante pour tous les décideurs publics, afin d'anticiper au maximum les variations de la conjoncture. La littérature au sujet de l'analyse cyclique de l'activité est très riche au niveau international et national. Ainsi, il est important de comprendre d'où vient l'analyse descriptive des fluctuations économiques avant d'entreprendre notre étude sur les cycles des demandeurs d'emploi français. C'est en 1920, avec la création du *National Bureau of Economic Research* (NBER) aux Etats-Unis, que Burns et Mitchell (1946) se sont intéressés à l'étude des cycles d'affaires américains en développant une définition de ces cycles. Le cycle des affaires est défini par ces auteurs comme une alternance de phases d'expansion et de phases de récession. Ces phases sont délimitées

par des points de retournement qui correspondent à des extrema locaux de la série étudiée. Les maxima locaux (pics) représentent les plus hauts niveaux de la série et les minima locaux (creux) correspondent au plus bas niveau de la série. Cette définition est encore largement reprise aujourd'hui dans la littérature. Dès lors, dater les points de retournements de l'activité est devenu un enjeu majeur du comité de datation (*Dating Committee*) afin de renseigner les dates d'entrée et de sortie d'une récession de l'économie américaine. Le NBER propose une datation officielle de la conjoncture américaine remontant jusqu'à l'année 1854. Cependant, d'autres organismes proposent des datations des points de retournements cycliques en utilisant les travaux du NBER comme référence. On peut par exemple penser au *Center for Economic Policy Research* (CEPR) qui officie sur des données européennes afin d'établir une chronologie des points de retournement au niveau européen. De nos jours, des organismes tels que l'Organisation de Coopération et de Développement Économiques (OCDE) ou la Banque Centrale Européenne (BCE) construisent leurs propres chronologies des cycles classiques ou des cycles de croissance (Ferrara, 2009). D'autres instituts s'investissent dans l'analyse des fluctuations économiques, comme par exemple l'*Economic Cycle Research Institute* (ECRI) qui propose une chronologie mensuelle des points de retournement pour 21 pays, dont la France. Par ailleurs, l'aspect méthodologique de la datation a fortement intéressé le monde académique. On peut citer par exemple Mazzi et Savio (2006) ou Artis *et alii* (1997) qui proposent une chronologie mensuelle des points de retournement de l'activité à partir de l'indice de production industrielle, ou encore Anas *et al.* (2008) pour une revue de la littérature sur le sujet.

Les premières études descriptives du cycle classique de l'activité économique ont mis en avant deux faits stylisés. Le premier est un co-mouvement des séries macroéconomiques, c'est-à-dire une évolution concomitante des séries des grands agrégats économiques. Le deuxième fait stylisé est la non-linéarité des séries macroéconomiques. En d'autres termes, le comportement des séries diffère grandement entre les phases de récession économique et celles d'expansion économique. Le développement de la discipline autour de la description des cycles économiques a donné naissance au cycle de croissance<sup>1</sup> et au cycle d'accélération. Le cycle de croissance peut se définir comme la différence entre une série de l'activité économique et sa tendance de long-terme<sup>2</sup>. L'objectif est donc d'estimer les périodes de croissance supérieure

---

1. Voir Mintz (1969)

2. Il existe plusieurs méthodes pour estimer cet écart, aussi appelé *output gap*. La littérature présente

à la tendance de long-terme et les périodes de croissance ralentie. L'étude des cycles de croissance est largement documentée dans la littérature (Anas et Ferrara, 2004; Zarnowitz et Ozyildirim, 2006). Le cycle d'accélération permet d'analyser les phases de hausse (accélération) et des phases de baisse (ralentissement) du taux de croissance de l'activité. La fréquence plus élevée des points de retournement par rapport aux points de retournement du cycle classique est perçue par les conjoncturistes comme un atout afin d'obtenir une information plus rapide sur l'évolution de la tendance de la série. Certains articles récents traitent cette question, comme par exemple Darné et Ferrara (2011) qui identifient les phases d'accélération et de ralentissement de l'activité en zone euro. Harding (2004) et Artis et alii (2004) présentent une analyse des cycles d'accélération européens et une méthodologie de datation des cycles d'affaires, respectivement.

La définition des cycles est un premier point important mais la caractérisation de ces derniers l'est tout autant. Le NBER proposait déjà à l'époque d'inspecter les caractéristiques du cycle des affaires avec la règle des 3D (*Duration, Depth et Diffusion*). L'idée principale, reprise par Harding et Pagan (2002), est d'évaluer la durée moyenne des phases et l'amplitude moyenne des phases et des cycles. Le comportement asymétrique entre les phases de récession et d'expansion nécessite l'étude de l'approximation triangulaire des mouvements cumulés d'une phase, plus communément appelé l'indice de sévérité, mesurant le gain (la perte) lié à une phase d'expansion (de récession) de l'économie. McKay et Reis (2008) montrent que les phases de contractions dans l'emploi sont plus courtes et plus violentes que les phases d'expansions sur données américaines. Peiró et Belaire-Franche (2012) étudient l'asymétrie dans la relation entre le cycle du chômage et le cycle de l'activité. La plupart de ces articles utilisent la définition du NBER pour qualifier le cycle des affaires. En outre, nous avons choisi d'utiliser la méthode non paramétrique (algorithme de Bry et Boschan, 1971) pour des raisons de transparence et de simplicité dans la réplique de notre travail. Cette méthode apparaît comme une bonne entrée en matière dans la réalisation d'une chronologie des points de retournement des cycles d'emploi.

---

de nombreuses méthodes d'estimation de la tendance de long terme, aussi appelé production potentielle dans le cadre d'études sur l'activité. Nous retrouvons ainsi les techniques des filtres spectraux (filtrage Hodrick et Prescott, 1997), des modèles à composantes inobservables ainsi que des fonctions de production et des méthodes économétriques du type vecteurs autorégressifs structurels (SVAR). L'objectif premier de ces méthodes est d'estimer une composante tendancielle de long terme afin d'isoler la composante de court terme qui correspondrait aux cycles de croissance.

L'intérêt d'estimer plusieurs datations cycliques à travers différents niveaux géographiques est d'introduire la notion de synchronisation des cycles. La synchronisation des cycles est devenue un enjeu majeur depuis le développement de la théorie de la zone monétaire optimale en 1960 par Robert Mundell. La synchronisation cyclique peut se définir comme la similarité des comportements (durée, ampleur) de la variable macroéconomique étudiée, entre les différents territoires observés. Depuis la création de l'Union Economique et Monétaire Européenne, cette question de la synchronisation a été traitée empiriquement dans de nombreux articles. Harding et Pagan (2001), Artis et alii (2004) et Camacho et Quiros (2006) étudient la synchronisation des cycles classiques européens sur l'activité économique. En parallèle, Harding et Pagan (2002, 2006) se sont concentrés sur l'étude de la synchronisation des cycles en développant l'indice de concordance, afin de mesurer le degré de synchronisation.

Nous avons pu voir que l'analyse du cycle des affaires était en grande partie réservée à l'activité économique d'un pays ou d'une zone géographique (Labondance, 2013). L'analyse cyclique du chômage reste un sujet peu traité. À ce jour et à notre connaissance, aucune n'étude existe sur la datation des points de retournement des cycles classique et d'accélération des demandeurs d'emploi<sup>3</sup>. En outre, aucun article traite d'une comparaison sur données régionales et départementales.

Dans un premier temps, nous tenterons de justifier à travers un raisonnement économique les *outliers* détectés par TRAMO et RegARIMA. Nous utilisons le logiciel recommandé par Eurostats et la Banque Centrale Européenne (BCE), JDemetra+, afin de corriger les séries de leurs variations saisonnières. L'objectif est ainsi d'obtenir la meilleure estimation possible de la composante Tendence-Cycle, afin de déterminer une datation précise des points de retournement conjoncturels ainsi qu'une description détaillée des mouvements cycliques (longueur, amplitude, symétrie)<sup>4</sup>. Par conséquent, nous proposons une datation des cycles de l'emploi à différents niveaux géographiques, nommément la région Pays de la Loire et ses cinq départements, à savoir la Loire-Atlantique, le Maine-et-Loire, la Mayenne, la Sarthe et la Vendée.

---

3. Nous noterons dorénavant le cycle classique des demandeurs d'emploi "cycle de l'emploi" afin d'alléger la lecture.

4. Le fait de désaisonnaliser nous-même les séries brutes nous permet d'obtenir directement la composante de tendance-cycle, alors que les séries corrigées des variations saisonnières (CVS) fournies par Pôle Emploi comportent, en plus de la tendance-cycle, une composante irrégulière. Cette dernière peut perturber la datation cyclique.



La section 2 définira la méthodologie et les *outliers* détectés. La section 3 exposera les données utilisées, alors que la section 4 présentera les résultats de l'étude empirique. La section 5 conclura notre propos.

## 1.2 Méthodologie

Nos travaux peuvent se rattacher en premier lieu à l'idéologie théorique développée par William Stanley Jevons et Clément Juglar, où l'intérêt est orienté sur l'analyse des cycles avec un recours aux séries statistiques afin de les mesurer. L'analyse des cycles est exclusivement liée à l'analyse des fluctuations économiques et la recherche de leurs causes. Même si Moore (1914) et Persons (1919) ont la paternité de l'utilisation rigoureuse des statistiques sur des séries longues afin de caractériser les cycles économiques, ce sont les travaux de Mitchell (1913) et Burns et Mitchell (1946) qui ont inspiré nos travaux sur la mesure du cycle de l'emploi. Dans cette approche, la mesure du cycle de l'emploi correspondrait à une étude des fluctuations d'un cycle spécifique que l'on comparerait au cycle de référence de l'économie, les différentes théories du cycle seraient alors surtout utilisées pour sélectionner les faits pertinents à retenir sous l'angle de la fluctuation économique. Dans ce chapitre, nous souhaitons mesurer le cycle de l'emploi comme un cycle de référence au même titre que le cycle des fluctuations économiques. En effet, nous souhaitons étudier le cycle de l'emploi en proposant un cadre de référence sur lequel on pourra s'appuyer pour de futures recherches. Nous pensons que la variable des demandeurs d'emploi, et plus généralement le chômage méritent une analyse cyclique à part entière. L'analyse de cette variable n'a pas vocation à expliquer directement les fluctuations du cycle économique, mais tenterait plutôt d'approfondir les connaissances du comportement intrinsèque du chômage. Par ailleurs, la comparaison avec le cycle de l'activité est aussi très instructive, tout comme la mise en relation de nos résultats avec certaines théories du cycle.

### 1.2.1 Méthode de désaisonnalisation

Dans cette section, nous présenterons de manière simple toutes les méthodes que nous utilisons. Pour commencer, une série temporelle peut se décomposer en plusieurs composantes inobservables. Les travaux en astronomie de Buys & Ballot (1847) ont introduit deux schémas de décomposition d'une chronique temporelle. Le schéma additif est présenté avec l'équation suivante :

$$X_t = TC_t + S_t + I_t \tag{1.1}$$

L'équation suivante décrit quant à elle le schéma multiplicatif :

$$X_t = TC_t \times S_t \times I_t \quad (1.2)$$

où  $TC$  correspond à la tendance-cycle,  $S$  à la composante saisonnière et  $I$  à la composante irrégulière. De nos jours, le schéma de décomposition est plus complexe puisqu'il comprend des effets déterministes tels que les effets de calendriers et les *outliers*. Notre schéma est de la forme suivante :

$$X_t = TC_t + S_t + O_t + EC_t + I_t \quad (1.3)$$

où  $O$  correspond aux *outliers* et  $EC$  aux effets de calendriers (effets de Pâques, jours ouvrables, fêtes mobiles). Nous avons présélectionné le schéma additif, qui correspond davantage à nos données, afin de décomposer nos séries de demandeurs d'emploi. Les méthodes TRAMO-SEATS et X-13ARIMA sont nos outils d'ajustement saisonnier et de décomposition. Ils sont détaillés ci-après.

## TRAMO-SEATS

Le programme TRAMO-SEATS fait partie des méthodes dites paramétriques. Les méthodes paramétriques d'ajustement saisonnier sont basées sur l'utilisation des modèles explicites qui permettent la désaisonnalisation des séries temporelles. Le programme TRAMO (*Time Series Regression with ARIMA Noise, Missing Observations, and Outliers*) permet, dans un premier temps, de corriger la série des éventuelles valeurs manquantes, des *outliers* et des effets de jours ouvrables et calendaires, ainsi que de traiter la stationnarité de la série.

Le programme TRAMO estime un premier modèle avec des erreurs suivant un processus ARIMA (*Autoregressive Integrated Moving Average*). Les paramètres du modèle de régression sont estimés grâce au maximum de vraisemblance (ou *via* les moindres carrés inconditionnels/conditionnels). Pour résumer, étant donné le vecteur d'observations :

$$z = (z_{t_1}, \dots, z_{t_M}) \quad (1.4)$$

où  $0 < t_1 < \dots < t_M$ , le programme ajuste le modèle de régression suivant :

$$z_t = y_t' \beta + v_t \quad (1.5)$$

où  $\beta = (\beta_1, \dots, \beta_n)'$  est un vecteur de coefficients de régression ;  $y'_t = (y_{1t}, \dots, y_{nt})$  représente  $n$  variables de régression et  $v_t$  suit un processus SARIMA( $p, d, q$ )( $P, D, Q$ ) $_s$  (*Seasonal-ARIMA*) général :

$$\phi(B)\delta(B)v_t = \theta(B)a_t, \quad (1.6)$$

où  $B$  est l'opérateur de retard ;  $\phi(B), \delta(B)$  et  $\theta(B)$  sont des polynômes finis dans  $B$  et  $a_t \sim N(0; \sigma_a^2)$  un bruit blanc.  $\phi(B), \delta(B)$  et  $\theta(B)$  ont la forme multiplicative suivante :

$$\delta(B) = (1 - B)^d(1 - B^s)^D \quad (1.7)$$

$$\phi(B) = (1 + \phi_1 B + \dots + \phi_p B^p)(1 + \Phi_1 B^s + \dots + \Phi_P B^{s \times P}) \quad (1.8)$$

$$\theta(B) = (1 + \theta_1 B + \dots + \theta_q B^q)(1 + \Theta_1 B^s + \dots + \Theta_Q B^{s \times Q}) \quad (1.9)$$

où le polynôme  $\delta(B)$  est le polynôme AR non stationnaire en  $B$ , avec l'ordre de différenciation  $d \leq 3$  et  $D \leq 2$ .  $\phi(B)$  est le polynôme AR stationnaire en  $B$ , qui est lui même le produit d'un polynôme AR stationnaire dans  $B$  et d'un polynôme AR saisonnier stationnaire en  $B^s$ .  $\theta(B)$  est le polynôme MA inversible dans  $B$ , qui est lui même le produit d'un polynôme MA inversible dans et d'un polynôme MA saisonnier inversible en  $B^s$ . L'ordre des polynômes est contraint par JDemetra+, tel que  $p, q \leq 3$  et  $P, Q \leq 1$ ,  $s$  représentant le nombre d'observations par année. L'ordre final du polynôme non saisonnier est sélectionné au regard du critère d'information BIC (*Bayesian Information Criterion*), au travers de la méthode de Hannan et Rissanen (1982).

La première étape de la procédure automatique de TRAMO consiste à tester si une spécification en log-niveau est nécessaire. Ce test est basé sur une estimation du maximum de vraisemblance du paramètre  $\lambda$  de la transformation de Box-Cox, c'est-à-dire transformer la distribution afin de tendre vers la normalités. La seconde étape détecte et corrige les *outliers* un par un *via* l'approche de Tsay (1986). En outre, plusieurs régressions multiples sont utilisées, comme dans Chen et Liu (1993), pour détecter les *outliers* fallacieux. La procédure de sélection ou de rejet des *outliers* est similaire à une procédure de régression par étapes (*stepwise regression*) afin de sélectionner la meilleure équation de régression. L'algorithme teste la présence d'effets calendaires *via* une régression d'un modèle Airline SARIMA( $p, d, q$ )( $P, D, Q$ ) $_s$ . Pour finir, la méthode de Hannan et Rissanen (1982) est utilisée pour sélectionner le modèle SARIMA le plus adapté. Cette méthode consiste à sélectionner le modèle qui minimise le critère d'infor-

mation BIC avec les contraintes sur les polynômes vues précédemment ayant pour objectif la sélection d'un modèle parcimonieux et équilibré. C'est-à-dire, un modèle capable d'avoir un bon pouvoir explicatif tout en utilisant un petit nombre de paramètres, pour le côté parcimonieux. Un modèle équilibré est un modèle pour lequel l'ordre des polynômes AR et de différenciation est égal à celui du polynôme MA.

Le programme SEATS (*Signal Extraction in ARIMA Time Series*) appartient à la classe des méthodes AMB (*ARIMA-Model-Based*). Hillmer et Tiao (1983) ont contribué au développement de cette méthode, qui consiste à décomposer une série temporelle en plusieurs composantes inobservables afin d'extraire le signal saisonnier dans le cadre de la correction des variations saisonnières. Le programme SEATS commence par ajuster un modèle SARIMA à la série chronologique. Par ailleurs, TRAMO suit la même procédure et se substitue à SEATS si les deux programmes sont utilisés conjointement. Notons  $x_t$  la série originale (ou sa transformation en log) et,

$$z_t = \delta(\beta)x_t, \quad (1.10)$$

où  $z_t$  représente la série différenciée,  $\beta$  l'opérateur de retard et  $\delta(\beta)$  l'ordre de différenciation de  $x_t$  dans le but d'atteindre la stationnarité.

$$\delta(\beta) = \nabla^d \nabla_s^D, \quad (1.11)$$

où  $\nabla^d = 1 - B^d$  et  $\nabla_s^D = (1 - B^s)^D$  représente la différence saisonnière de la période  $s$ . Le modèle de série différenciée peut être écrit comme suit :

$$\phi(B)(z_t - \bar{z}) = \theta(\beta)a_t, \quad (1.12)$$

où  $\bar{z}$  est la moyenne des  $z_t$ ,  $a_t$  est un bruit blanc normalement distribué avec une moyenne nulle et de variance  $\sigma_a^2$ ,  $\phi(B)$  et  $\delta(B)$  sont respectivement des polynômes autorégressif (AR) et de moyenne mobile (MA) dans  $B$ , qui peuvent être exprimés dans la forme multiplicative comme le produit d'un polynôme non saisonnier dans  $B$  et d'un polynôme saisonnier dans  $B^s$ ; ainsi on peut écrire :

$$\phi(B) = \phi_r(B)\phi_s(B^s), \quad (1.13)$$

$$\theta(B) = \theta_r(B)\theta_s(B^s). \quad (1.14)$$

On peut réécrire le modèle sous une forme complète et détaillée :

$$\phi_r(B)\phi_s(B^s)\nabla^d\nabla_s^D x_t = \theta_r(B)\theta_s(B^s)a_t + c, \quad (1.15)$$

et, dans sa forme simplifiée,

$$\Phi(B)x_t = \theta(B)a_t + c, \quad (1.16)$$

où  $\Phi(b) = \phi(B)\delta(B)$  représente le polynôme autorégressif complet, intégrant toutes les racines unitaires. Notons que, si  $p$  correspond à l'ordre de  $\phi(B)$  et  $q$  l'ordre de  $\theta(B)$ , alors l'ordre de  $\Phi(B)$  est  $P = p + d + (D \times s)$ .

Le programme décompose la série, qui suit le modèle (1.9), c'est-à-dire la série linéarisée par TRAMO, en plusieurs composantes. Le domaine des fréquences permet l'utilisation d'un spectre partitionné en plusieurs spectres additifs qui sont associés aux différentes composantes. La condition canonique sur les composantes tendance, saisonnière et cyclique identifie une décomposition unique, Hillmer et Tiao (1982); C'est-à-dire que tout bruit blanc supplémentaire est ajouté à la composante irrégulière. La décomposition canonique produit des composantes tendanciennes et saisonnières aussi stables que possible puisqu'elle maximise la variance de l'irrégulier et minimise la variance des autres composantes. L'estimation des composantes par les estimateurs des erreurs moyennes quadratiques minimums est obtenue *via* le filtre de Weiner-Kolmogorov. Les programmes TRAMO et SEATS fournissent une méthode complète dite de *model-based* pour prévoir ou pour extraire le signal dans une série temporelle univariée. Pour une description complète des programmes TRAMO-SEATS, on pourra se référer à Gómez and Maravall (1996, 2001a) et Maravall et alii (2015).

## X-13ARIMA

La méthode X-13ARIMA est une méthode hybride ou semi-paramétrique qui prend ses origines dans les méthodes nommées "style X-11" de Shiskin, Young, et Musgrave (1967) puis avec Box et Jenkins (1970) et leur ouvrage sur les modèles ARIMA (AutoRegressive Integrated Moving Average). La méthode la plus récente de X-13ARIMA est développée par la

division de recherche du U.S. Bureau of the Census<sup>5</sup>. Le programme possède aussi un pré-traitement (Regression ARIMA, RegARIMA) des données qui permet de corriger les *outliers*, les variations de jours ouvrables et les effets de calendrier. Le modèle RegARIMA peut s'écrire sous la forme suivante :

$$(1 - B)^d(1 - B^s)^D y_t = \sum_i \beta_i (1 - B)^d(1 - B^s)^D x_{it} + w_t. \quad (1.17)$$

où  $w_t$  suit un modèle ARMA stationnaire,  $x_{it}$  sont des variables de régression du modèle RegARIMA,  $y_t$  correspond à la série et  $(1 - B)^d(1 - B^s)^D$  est l'opérateur de différence du modèle SARIMA. A l'origine, X-11 estime cinq modèles SARIMA prédéfinis afin d'extrapoler la série dans le but de la prolonger. Les cinq modèles disponibles sont présenté dans la Table 1.1 ci dessous :

TABLE 1.1 – Modèles SARIMA prédéfinis

$(0,1,1)(0,1,1)_{12}$
$(0,1,2)(0,1,1)_{12}$
$(2,1,0)(0,1,1)_{12}$
$(0,2,2)(0,1,1)_{12}$
$(0,1,2)(0,1,1)_{12}$

La procédure de sélection des cinq meilleurs modèles a changé avec les versions plus récentes de X-12 et X-13ARIMA. Maintenant, les cinq meilleurs modèles suivent le même processus que le programme TRAMO, c'est-à-dire que le module RegARIMA procède comme TRAMO en identifiant l'ordre de différenciation et le modèle SARIMA. Pour ce faire, la procédure d'Hannan-Rissanen (1982) est utilisée. En outre, l'ordre des polynômes est contraint de la même façon que dans TRAMO et le modèle minimisant le critère BIC est sélectionné. Même si les deux méthodes sont sensiblement identiques, certaines différences persistent. La méthode RegARIMA possède des extensions qui lui sont propres. Nous pouvons évoquer, par exemple, le traitement de l'effet de l'année bissextile ou la détection de la durée de l'effet de Pâques. Les différences notables par rapport à TRAMO résident dans le fait que X-13ARIMA utilise l'estimation exacte du maximum de vraisemblance alors que TRAMO utilise l'estimation approximative du maximum de vraisemblance. X-13ARIMA utilise des va-

---

5. Le Bureau du recensement des États-Unis est une administration publique américaine dépendant du département du Commerce.

leurs critiques plus grandes pour l'inclusion des paramètres d'*outliers*<sup>6</sup>. De plus, X-13ARIMA utilise le critère d'information de Akaike corrigé (AICC) au lieu du coefficient individuel de la statistique de Student pour l'inclusion des paramètres des jours ouvrables (*trading day, holiday effect*) (Findley, 2005). Ainsi, il est normal d'observer des identifications de modèles SARIMA différentes, notamment dans un contexte où la série étudiée est sujette à certains effets déterministes.

Bien que le module de pré-traitement soit similaire entre TRAMO et RegARIMA, le programme d'ajustement saisonnier est quant à lui totalement distinct. X-13ARIMA utilise une succession de moyennes mobiles symétriques et asymétriques (filtrage en cascade), ainsi que plusieurs filtres provenant de modèles ARIMA ajustés aux données. C'est ainsi que nous pouvons synthétiser l'algorithme X-11 en 8 étapes.

Étape 1 : Estimation de la tendance avec une moyenne mobile  $2 \times 12$  :

$$TC_t = M_{2 \times 12}(X_t) \quad (1.18)$$

Étape 2 : Estimation de la composante saisonnière-irrégulière :

$$(S_t + I_t) = X_t - TC_t \quad (1.19)$$

Étape 3 : Estimation de la composante saisonnière par une moyenne mobile  $3 \times 3$  :

$$S_t - M_{3 \times 3}[(S_t + I_t)] \quad (1.20)$$

La composante saisonnière est centrée grâce à une moyenne mobile  $2 \times 12$

$$\tilde{S}_t = S_t - M_{2 \times 12}(S_t) \quad (1.21)$$

Étape 4 : Estimation de la série corrigée des variations saisonnières (CVS) :

$$SA_t = (TC_t - I_t) = X_t - \tilde{S}_t \quad (1.22)$$

---

6. Les programmes TRAMO et RegARIMA calcul un t-statistique pour chaque *outlier* potentiel et le compare à une valeur critique prédéfinie. Pour X-13ARIMA la valeur critique est de 4 alors que la valeur critique pour TRAMO est de 3.5.



Etape 5 : Estimation de la tendance *via* une moyenne mobile d'Henderson à 9, 13 ou 23 termes :

$$TC_t^{(2)} = H_{9, 13, 23}(SA_t) \quad (1.23)$$

Etape 6 : Estimation de la composante saisonnière-irrégulière :

$$(S_t + I_t)^{(2)} = X_t - TC_t^{(2)} \quad (1.24)$$

Etape 7 : Estimation de la composante saisonnière par une  $3 \times 5$  :

$$S_t^{(2)} - M_{3 \times 5}[(S_t + I_t)^{(2)}] \quad (1.25)$$

La composante saisonnière est centrée grâce à une moyenne mobile  $2 \times 12$

Etape 8 : Estimation de la série CVS :

$$SA_t^{(2)} = (TC_t - I_t)^{(2)} = X_t - \tilde{S}_t^{(2)} \quad (1.26)$$

Pour une description complète de la méthode X-13ARIMA-SEATS, on pourra se référer à U.S. Census Bureau (2016).

## Détection des *outliers*

L'analyse cyclique d'une série chronologique repose sur des méthodes de décomposition. L'objectif de ces méthodes est de pouvoir extraire une composante Tendence-Cycle qui se veut moins "bruitée" qu'une série corrigée des variations saisonnières (CVS), étant donné que la composante irrégulière est éliminée. Cependant, avant d'utiliser ces méthodes, il convient de correctement nettoyer la série des éventuels *outliers* qui viendraient perturber le bon fonctionnement des méthodes de désaisonnalisation.

La détection des valeurs aberrantes est un sujet largement traité dans la littérature de l'ajustement saisonnier des séries temporelles. Mehrhoff (2018) nous fait remarquer que les valeurs extrêmes ne doivent pas gêner l'estimation de la composante saisonnière. Dès 1963, Shisken, Young, et Musgrave se sont intéressés à l'identification et la correction automatique des *outliers* en développant le programme Census X-11. Les *outliers* étaient identifiés à partir de la série saisonnière en utilisant un seuil qui correspondait à plusieurs moyennes quadratiques d'un intervalle ajusté à la série. Box et Tiao (1975) utilisaient des modèles d'interventions pour identifier les différents types d'*outliers*, ainsi que les localisations de leurs impacts. Par la suite, Hillmer, Bell, et Tiao (1983) ont développé une méthode de traitement des *outliers* basée sur un modèle à ajustement saisonnier. Ils ont montré comment les erreurs du modèle pouvaient être utilisées dans le but d'identifier les *outliers*<sup>7</sup>.

Depuis le développement des méthodes automatiques de désaisonnalisation, les méthodes d'identification et de correction des *outliers* se sont améliorées. La méthode de détection des valeurs aberrantes de X-13ARIMA-SEATS est largement inspirée de celle de Chang et Tiao (1983), Chang et alii. (1988) et Otto et Bell (1990). Cette approche est similaire à la régression pas-à-pas, puisque les *outliers* sont considérés comme des variables candidates à la régression. En d'autres termes, la signification de ces variables est testée sur l'ensemble des points de la série *via* une comparaison de la statistique de Student afin de sélectionner les *outliers* qui doivent être incorporés dans la régression.

Les *outliers* peuvent prendre différentes formes. La littérature retient plusieurs types d'*outliers*, notamment définis par Fox (1972). Pour commencer, il y a les *Additive Outliers* (AO). Ces points atypiques affectent une seule observation de la série. Notons par exemple, que sur une

---

7. Voir Burman et Otto (1988), pour une présentation de ces méthodes.

série temporelle de fréquence mensuelle, un AO agira sur un seul mois, car l'effet aura disparu le mois suivant. Le deuxième *outlier* se nomme *Level Shift* (LS). Le LS a un effet permanent sur la série, c'est-à-dire qu'il modifie le niveau de la série de manière permanente. Le *Temporary Change* (TC) est un *outlier* qui agit sur la série comme un LS, cependant l'effet de cet *outlier* se résorbe et la série retrouve son "état" normal plus ou moins rapidement, l'effet est donc temporaire sur la série. Enfin, le dernier *outlier* pris en compte est plus récemment étudié dans la littérature que les précédents. C'est Kaiser et Maravall (1999) qui ont exposé l'intérêt de traiter le *Seasonal Outlier* (SO), plutôt que d'intégrer l'*Innovative Outliers* (IO) qui correspond à un choc qui impacte les observations suivantes d'une série en se propageant *via* les pondérations de la représentation en moyenne mobile du modèle ARIMA. Le SO correspond à une rupture dans le profil saisonnier de la série. Cela correspond à un changement de niveau, mais qui intervient sur le profil saisonnier plutôt que sur l'ensemble de la série. Les auteurs ont démontré que l'ajout de cet *outlier* dans la procédure de détection automatique offrait de meilleurs résultats d'estimation, à condition que l'on travaille à partir de la série brute saisonnière.

TRAMO et RegARIMA utilisent une méthode de détection séquentielle, dont l'approche est la suivante.

Soit  $y_t$  une série temporelle décrite par un modèle ARIMA(p,d,q)

$$\phi(B)\delta(B)y_t = \theta(B)a_t, \quad (1.27)$$

où  $\phi(B)$ ,  $\delta(B)$  et  $\theta(B)$  sont des polynômes d'ordre respectifs (p, d, q).  $B$  est un opérateur de retard, soit  $B^j y_t = y_{t-j}$ . Quant à  $a_t$ , c'est un processus de bruit blanc,  $a_t \sim N(0, \sigma_a^2)$ . On peut modéliser les valeurs extrêmes *via* des polynômes de régression. Partant du principe que la série contient  $k$  *outliers*, voici comment transcrire leurs effets :

$$y_t^* = \sum_{j=1}^k \xi_j(B)\omega_j I_t^{\tau_j} + y_t, \quad (1.28)$$

où  $y_t^*$  représente la série observée contenant des *outliers* ;  $y_t$  suit un processus ARIMA comme dans l'équation (24) ;  $\omega_j$  correspond à l'impact de l'*outlier* au temps  $t = \tau_j$  ;  $I_t^{\tau_j}$  est une variable indicatrice qui prend la valeur de 1 pour  $t = \tau_j$  et 0 sinon ; et  $\xi_j(B)$  est un polynôme

déterminant la dynamique du point atypique se produisant au temps  $t = \tau_j$ , selon le schéma suivant :

$$AO : \xi_j(B) = 1, \quad (1.29a)$$

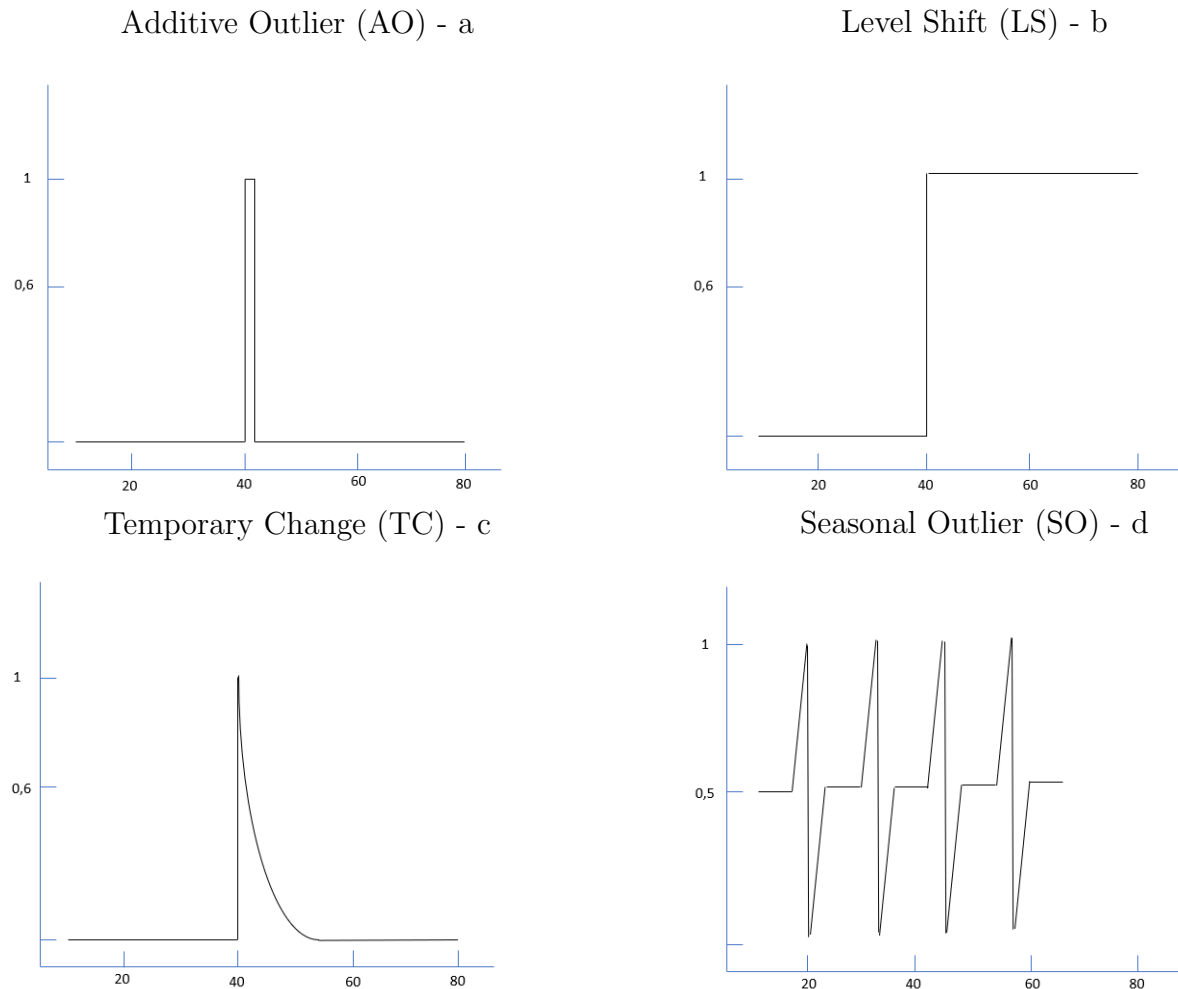
$$LS : \xi_j(B) = \frac{1}{(1 - B)}, \quad (1.29b)$$

$$TC : \xi_j(B) = \frac{1}{(1 - \delta B)}, 0 < \delta < 1 \quad (1.29c)$$

$$SO : \xi(B) = \frac{1}{S}; S = 1 + B + \dots + B^{s-1}. \quad (1.29d)$$

où S représente un opérateur d'agrégation annuel<sup>8</sup>, du point de vue de la représentation (16).

FIGURE 1.1 – *Effet des différents types de points atypiques*



Comme le montre la figure 1.1, les effets des AO et TC sont donc transitoires, alors que l'effet

8. Voir Kaiser et Maravall (1999) pour une présentation détaillée des *outliers* saisonniers.

du LS est permanent. Par ailleurs, le modèle SARIMA identifié par la procédure automatique n'a pas d'incidence sur l'impact de ces *outliers*. En revanche, l'effet du SO dépend du modèle SARIMA identifié. L'exemple de SO présenté correspond selon Kaiser et Maravall (1999) à un *outlier* purement saisonnier.

## 1.2.2 Analyse des cycles

### Datation des points de retournement

L'algorithme de Bry et Boschan (1971) est une méthode de datation du cycle des affaires sur une série corrigée de ses variations saisonnières (CVS) qui permet de dater les points de retournement du cycle classique ainsi que du cycle d'accélération. La datation du cycle classique s'effectue sur la série CVS-CJO, alors que la chronologie du cycle d'accélération est réalisée sur le taux de croissance de la série corrigée de ses variations saisonnière et corrigée des jours ouvrables (CVS-CJO) originale. Pour le cycle d'accélération, nous appliquons au préalable un filtre d'Hodrick-Prescott avec un  $\lambda = 13.9$  représentant un arbitrage entre la composante cyclique et le lissage de la tendance. Gómez (2001) a démontré que ce type de filtre se comportait comme un filtre passe-bas (*low-pass filter*) et qu'il éliminait, de ce fait, les fluctuations supérieures à un an. Le fonctionnement de l'algorithme est simple. Il détecte un pic à la date  $t$  si la condition suivante est vérifiée :

$$(\Delta_k Y_t, \dots, \Delta Y_t) > 0, (\Delta Y_{t+1}, \dots, \Delta_k Y_{t+k}) < 0$$

et détecte un creux à la date  $t$  si la condition suivante est vérifiée :

$$(\Delta_k Y_t, \dots, \Delta Y_t) < 0, (\Delta Y_{t+1}, \dots, \Delta_k Y_{t+k}) > 0$$

avec  $\Delta_k$  l'opérateur de différenciation tel que  $\Delta_k Y_t = Y_t - Y_{t-k}$ , où  $k = 5$  pour des données mensuelles, représentant les six mois consécutifs afin que l'algorithme identifie une expansion ou une contraction. Le paramétrage<sup>9</sup> s'effectue sur le logiciel R *via* le package *BCDating* ainsi que sur Microsoft Excel.

---

9. L'algorithme de Bry-Boschan est présenté en annexe 1.24.

## Caractéristiques des cycles

L'étude des cycles classiques par le National Bureau of Economic Research (NBER) présente une étude des caractéristiques des cycles en plus de la datation de ces derniers. Ainsi, la règle des 3D est utilisée comme méthodologie pour dater les phases et les cycles mais aussi pour analyser leurs caractéristiques. On distingue donc la durée (*Duration*), l'amplitude (*Depth*) et la diffusion (*Diffusion*)<sup>10</sup>. La durée d'une phase est mesurée, en mois pour une série mensuelle, entre deux points de retournement (pic à creux pour une phase de contraction et creux à pic pour une phase d'expansion), la durée sera notée  $D_i$  pour la durée de la  $i$ -ème phase. L'amplitude calcule la distance entre deux points de retournement, l'amplitude est en valeur absolue et sera notée  $A_i$  pour l'amplitude de la  $i$ -ème phase. En suivant Harding et Pagan (2002), nous nous intéresserons à la durée des phases et des cycles, à leur amplitude, aux comportements asymétriques de ces phases et pour terminer nous calculerons un indice de sévérité, noté  $S$ , qui s'apparente à la mesure du "triangle approximation to the cumulative movements".

$$S_i^{E;C} = 0.5 \times D_i^{E;C} \times A_i^{E;C} \quad (1.30)$$

avec  $A_i^E = |Y_{t,pic} - Y_{t,creux}|$  et  $A_i^C = |Y_{t,creux} - Y_{t,pic}|$ ,  $E$  signifie expansion et  $C$  contraction.  $Y_t$  correspond à la valeur de la série au moment où intervient le pic ou le creux. Les deux auteurs nous expliquent que le chemin actuel d'une phase ne correspond pas exactement à un triangle rectangle et proposent une mesure qui permet d'estimer la forme d'une phase d'expansion ou de contraction, c'est l'*Excess cumulated movements* :

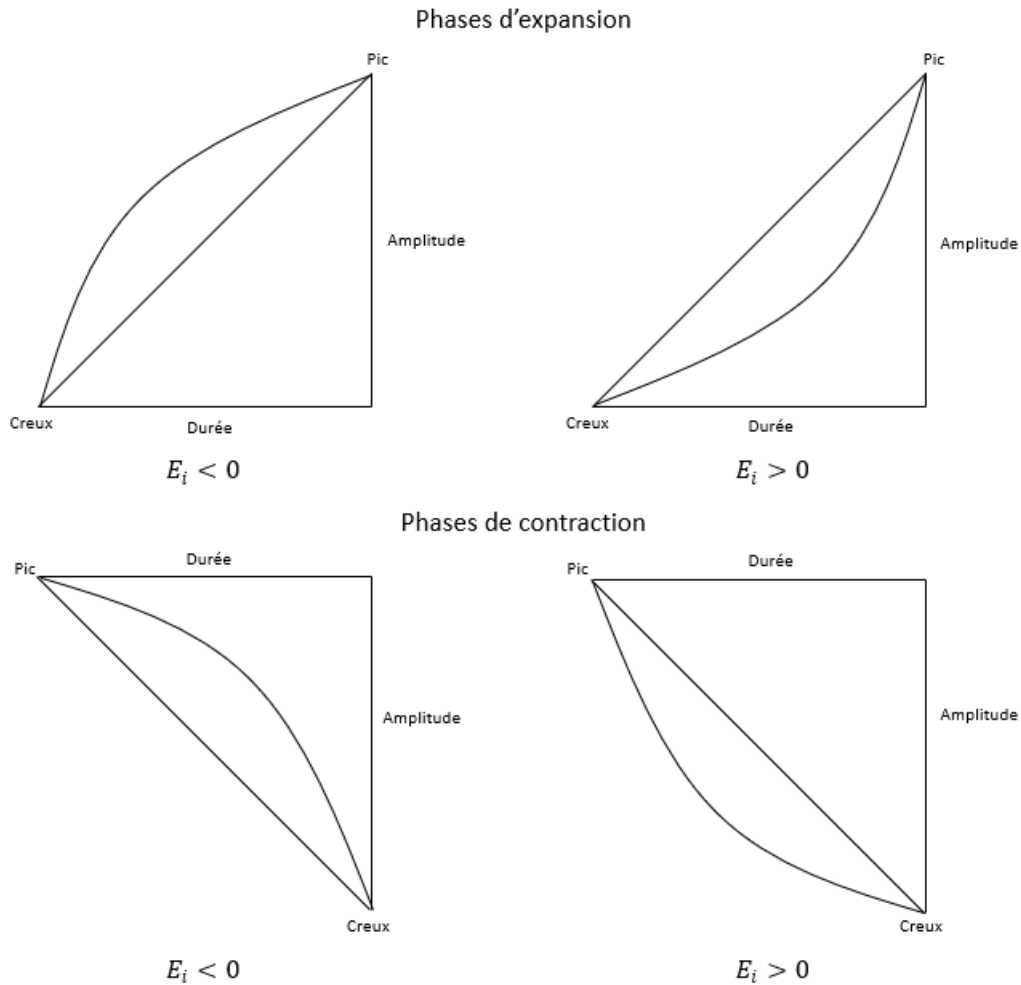
$$E_i = (S_i - C_i)/D_i \quad (1.31)$$

avec  $C_i = \left| \sum_i^{D_i} (y_i - y_0) \right| - 0.5 \times A_i$ , et  $y$  les valeurs de la série.

---

10. La notion de diffusion est relative au fait que le cycle doit être diffusé à travers les différents secteurs d'un même pays ou d'une même zone économique. Cependant, comme nous étudions le cycle d'une seule série, nous ne nous intéresserons pas à la diffusion.

FIGURE 1.2 – Phases d'expansion et de contraction stylisées



Source : Graphique inspiré de Harding et Pagan (2002) et de Majetti (2012)

Note : Les hypoténuses des triangles rectangles correspondent à l'indice de sévérité, les courbes représentent le chemin réel de la phase.

L'indice de l'*Excess* mesure l'intensité de l'inflexion conjoncturelle des deux types d'une phase. On observe sur la Figure 1.2 que durant une phase d'expansion, si l'*Excess* est négatif, alors l'augmentation est rapide et brutale (forme concave) et il y a un retour progressif à l'approche du point de retournement. Inversement, un *Excess* positif est synonyme d'une expansion progressive et moins violente (forme convexe). Au contraire, lors d'une phase de contraction, un *Excess* négatif évoque une phase baissière progressive et douce (forme concave) et un *Excess* positif renseigne d'une baisse rapide et marquée avec un retour progressif à l'approche du point de retournement (forme convexe).

## Synchronisation des cycles

La mesure de la synchronisation peut être atteinte par plusieurs moyens. On entend par synchronisation des cycles, le fait que les cycles de plusieurs séries temporelles suivent un schéma identique. On parle de co-mouvements quand les mêmes cycles de plusieurs séries évoluent dans le même sens. Nous utilisons deux méthodes afin de calculer le degré de synchronisation entre nos séries de demandeurs d'emploi, ainsi qu'une méthode appliquée uniquement aux données départementales. La première méthode est le coefficient de corrélation de Pearson. Le coefficient de corrélation s'écrit de la façon suivante :

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} \in [0, 1] \quad (1.32)$$

où  $x$  et  $y$  sont des séries temporelles. Par ailleurs, Harding et Pagan (2002, 2006) proposent une équation très simple pour calculer le degré de synchronisation entre deux séries temporelles. Cette équation est nommée, indice de concordance :

$$IC = \frac{1}{T} \left\{ \sum_{t=1}^T S_{x_t} S_{y_t} + \sum_{t=1}^T (1 - S_{x_t})(1 - S_{y_t}) \right\} \quad (1.33)$$

où  $S_{x_t}$  et  $S_{y_t}$  sont des variables binaires représentant les cycles des séries  $x$  et  $y$ . Cette variable est égale à 1 quand le cycle est en phase d'expansion et 0 quand il est en phase de contraction. Si  $IC = 1$  alors les séries sont procycliques, c'est-à-dire que les deux cycles évoluent dans le même sens. En revanche, si  $IC = 0$  alors les séries sont contracycliques, c'est-à-dire que les cycles évoluent dans des sens opposés.

Dans le but d'apporter une mesure nouvelle du co-mouvement et de la synchronisation des séries temporelles, Stock et Watson (2008) ont développé une mesure basée sur l'indice spatial de Moran (Moran, 1950), appelé l'indice de Moran modifié. L'intérêt d'utiliser ce type de mesure n'est pas tant la dimension spatiale, mais l'étude du co-mouvement des séries dans le temps à travers les cinq départements considérés ici. En effet, l'indice fournit bien une mesure de la corrélation spatiale entre plusieurs départements mais le poids de chacun est normalisé à 1, c'est-à-dire que l'accent est donné à la corrélation à travers le temps, plutôt que sur la situation géographique des départements. Les auteurs utilisent cette méthode sur des données mensuelles des permis de construire aux U.S.A entre 1969 et 2007. De plus, Bandrés et alii (2017) utilisent ce même indice sur des données annuelles du PIB de plusieurs



pays européens. Nous appliquerons cet indice sur les départements de la région Pays de la Loire. L'indice spatial de Moran modifié se calcule à partir de l'expression suivante :

$$\widehat{I}_i = \frac{\sum_{i=1}^N \sum_{j=1}^{i-1} \widehat{cov}(y_{it}, y_{jt}) / N(N-1)/2}{\sum_{i=1}^N \widehat{var}(y_{it}) / N} \quad (1.34a)$$

$$\widehat{cov}(y_{it}, y_{jt}) = \frac{1}{k} \sum_{s=t-10}^{t+10} (y_{is} - \overline{y_{it}})(y_{js} - \overline{y_{jt}}) \quad (1.34b)$$

$$\widehat{var}(y_{it}) = \frac{1}{k} \sum_{s=t-10}^{t+10} (y_{is} - \overline{y_{it}})^2 \quad (1.34c)$$

$$\overline{y_{it}} = \frac{1}{k} \sum_{s=t-10}^{t+10} y_{is} \quad (1.34d)$$

où  $y_{it}$  est le taux de croissance du nombre de demandeurs d'emploi du département  $i$  à la période  $t$ . La fenêtre glissante  $k = 21$  est sélectionnée en suivant Stock et Watson (2008), qui travaillent aussi sur des données mensuelles<sup>11</sup>. Nous avons cinq séries départementales donc  $N = 5$ .

La littérature comprend plusieurs indices de synchronisation qui ne peuvent malheureusement pas tous être utilisés dans notre étude. Mink et alii (2012) proposent un indice de synchronisation entre les *output gaps* de plusieurs pays européens comparant uniquement si les cycles possèdent le même signe sur les mêmes périodes. Les auteurs proposent aussi un indice de similarité étudiant le co-mouvement en prenant en compte l'amplitude des différentes phases cycliques. Samarina et alii (2015) et Comunale (2017) reprennent ces mesures spécifiques à l'étude du cycle de croissance. Par conséquent, ces mesures ne sont pas appropriées dans le cadre de notre étude sur les cycles classiques et d'accélération des demandeurs d'emploi.

---

11. Bandrés et alii (2017) suggèrent une fenêtre glissante de  $k = 5$  pour des données annuelles et précisent qu'augmenter la taille de la fenêtre au-delà conduit à des résultats similaires avec une forme plus lissée de l'indice. Nous sommes arrivés à la même conclusion après plusieurs études de l'indice avec des fenêtres glissantes plus grandes et plus réduites. Dans le cas d'une fenêtre plus petite, l'indice se comporte mal avec les données mensuelles tandis qu'un indice plus grand produit des résultats similaires et conduit à écarter un trop grand nombre d'observations, limitant le pouvoir informationnel de cet indice.

## 1.3 Données

La France possède deux mesures statistiques du chômage, qui sont notamment suivies par les décideurs publics pour rendre compte de l'état du marché de l'emploi et plus généralement de l'économie. D'une part, l'Insee mesure le taux de chômage de la population active *via* une enquête trimestrielle auprès d'un panel d'environ 100000 logements distincts constituant un échantillon représentatif des ménages français, appelé l'*enquête Emploi*. Le questionnaire de l'enquête est harmonisé au niveau international puisque la définition utilisée est celle du Bureau International du Travail (BIT). Elle fait partie des *Labor Force Survey* définies par l'Union européenne. D'autre part, la mesure statistique du chômage de Pôle Emploi, consiste à agréger le nombre de demandeurs d'emploi en fin de mois<sup>12</sup> (DEFM) inscrits auprès de leur service. Par conséquent, un chômeur au sens du BIT n'est pas forcément inscrit à Pôle Emploi et inversement. La mesure du BIT est largement utilisée au niveau international, néanmoins la mesure de Pôle Emploi vient compléter le panorama des statistiques du chômage français. La mesure de Pôle Emploi donne un autre regard sur le chômage et correspond à une statistique mensuelle avec un fort pouvoir informatif sur la structure du chômage en France. Les individus inscrits administrativement en catégorie A sont représentatifs d'une population sans emploi et en recherchant un activement. Cette statistique mensuelle est critiquée en raison de sa volatilité liée à plusieurs facteurs, tels que la modification du suivi et de l'accompagnement du demandeur d'emploi, du comportement d'inscription ou encore de la mise en place du Revenu de Solidarité Active (RSA). Cependant, la fréquence des données reste un point fort de la qualité de l'information fournie par la mesure, le confinement lié au COVID-19 l'a récemment démontré. Ainsi étudier le cycle de l'emploi n'est pas un exercice anodin, dans le sens où l'information recherchée apportera toujours une connaissance précieuse sur les caractéristiques de cette mesure.

Les données des demandeurs d'emploi en catégorie A sont collectées par Pôle Emploi. Elles

---

12. Nous nous concentrons sur la catégorie A de Pôle Emploi, c'est-à-dire les demandeurs d'emploi n'ayant pas travaillé dans le mois précédent leur actualisation et tenus d'accomplir des actes positifs de recherche d'emploi. Selon Pôle Emploi, la classification des demandeurs d'emploi peut se décomposer en 5 catégories (nous avons déjà évoqué la catégorie A). "La catégorie B rassemble les personnes ayant exercé une activité réduite de 78 heures maximum par mois, tenues d'accomplir des actes positifs de recherche d'emploi » ; la catégorie C rassemble "les personnes ayant exercé une activité réduite de plus de 78 heures par mois, tenues d'accomplir des actes positifs de recherche d'emploi » ; la catégorie D rassemble "les personnes sans emploi, qui ne sont pas immédiatement disponibles, non tenues d'accomplir des actes positifs de recherche d'emploi (demandeurs d'emploi en formation, en maladie)" ; et la catégorie E rassemble "les personnes pourvues d'un emploi, non tenues d'accomplir des actes positifs de recherche d'emploi ».

couvrent une période courant de janvier 1996 à décembre 2019, les données sont régionales (Pays de la Loire) et départementales (44-Loire Atlantique, 49-Maine-et-Loire, 53-Mayenne, 72-Sarthe, 85-Vendée). Le tableau suivant (Table 1.2) présente leurs statistiques descriptives.

TABLE 1.2 – Statistiques descriptives des données brutes, en milliers

	Nationale	PDLL	44	49	53	72	85
Min.	1858	826,4	31,6	18,9	5,2	13,7	12,4
Max.	3742	189	70,53	44,2	12	31,1	33,1
Moyenne	2874	138	53,6	31,5	8,5	22,7	21,5
Médiane	2818	135,6	52,7	30,5	8,4	22,4	20,6
Ecart-type	453,31	24,97	9,17	5,7	1,6	4,39	4,84
Skewness	-0,046	0,03	-0,13	0,14	0,05	0,01	0,37
v1	-0,31	0,26	-0,92	1,02	0,36	0,10	2,57
Kurtosis	2,099	2,1	2,16	2,19	2,15	1,8	2,35
v2	-3,03	-3,01	-2,81	-2,69	-2,83	-4,04	-2,16
Jarque-Bera	0,009	0,009	0,01	0,01	0,01	0,0002	0,003

Source : Auteur, données Pôle Emploi.

Note : PDLL correspond à la région Pays de la Loire et les départements sont : Loire-Atlantique (44), Maine-et-Loire (49), Mayenne (53), Sarthe (72), Vendée (85). Les valeurs du test de Jarque-Bera correspondent à la p-value. Si  $p < 0,05$  alors on rejette l'hypothèse nulle de normalité. Les statistiques de test du skewness et du kurtosis sont mesurés avec les  $|v1| = \frac{(skewness-0)}{\sqrt{6/T}}$  et  $|v2| = \frac{(Kurtosis-3)}{\sqrt{24/T}}$ , avec  $T$  le nombre d'observations.

Le test de Jarque-Bera indique qu'aucune variable ne suit une loi normale. La moyenne et la médiane sont relativement éloignées ce qui signifie une légère dispersion des données. De plus, les valeurs des écarts-types suggèrent aussi une dispersion des données autour de la moyenne. Le coefficient d'asymétrie (skewness) indique qu'aucune série n'est statistiquement symétrique et que l'une d'entre elles, la série de la Loire-Atlantique (44), est oblique à droite de la médiane, tandis que les autres séries sont très légèrement obliques à gauche de la médiane. Le kurtosis de chaque série est statistiquement inférieur à 0 ce qui signifie que les distributions sont platikurtiques, c'est-à-dire que la distribution est aplatie.

### 1.3.1 Analyse des points atypiques

Les *outliers* sont des valeurs atypiques qui biaisent la modélisation SARIMA et par conséquent l'exercice d'ajustement saisonnier. Nous avons vu que les pré-traitements TRAMO et RegARIMA s'occupaient de les identifier et de les écarter des séries brutes, afin d'obtenir une chronique corrigée des variations saisonnières de bonne qualité. Après avoir paramétré les deux méthodes de pré-traitement et de désaisonnalisation *via* le logiciel JDemetra+, nous sélectionnons la méthode qui minimise le critère d'Akaike. Le tableau 1.3 reprend les modèles sélectionnés pour chacune des séries étudiées.

TABLE 1.3 – Sélection du meilleur modèle selon AIC

Séries	Méthode	Modèle	AIC
DE Nationale Brute	TRAMO-SEATS	$(0, 2, 1)(0, 1, 1)_{12}$	6009,98
DE Nationale Brute	X-13ARIMA	$(0, 2, 1)(0, 1, 1)_{12}$	6010,90
DE PDLL Brute	TRAMO-SEATS8	$(2, 1, 0)(0, 1, 1)_{12}$	4732,50
DE PDLL Brute	X-13ARIMA	$(0, 2, 1)(0, 1, 1)_{12}$	4664,63
DE 44 Brute	TRAMO-SEATS	$(1, 1, 0)(0, 1, 1)_{12}$	4236,51
DE 44 Brute	X-13ARIMA	$(1, 1, 0)(0, 1, 1)_{12}$	4135,09
DE 49 Brute	TRAMO-SEATS	$(0, 1, 1)(0, 1, 1)_{12}$	4065,03
DE 49 Brute	X-13ARIMA	$(0, 2, 1)(0, 1, 1)_{12}$	4037,60
DE 53 Brute	TRAMO-SEATS	$(1, 1, 0)(0, 1, 1)_{12}$	3500,57
DE 53 Brute	X-13ARIMA	$(1, 1, 0)(0, 1, 1)_{12}$	3501,23
DE 72 Brute	TRAMO-SEATS	$(3, 1, 0)(0, 1, 1)_{12}$	3806,32
DE 72 Brute	X-13ARIMA	$(0, 2, 1)(0, 1, 1)_{12}$	3823,75
DE 85 Brute	TRAMO-SEATS	$(1, 1, 0)(0, 1, 1)_{12}$	3854,08
DE 85 Brute	X-13ARIMA	$(0, 1, 0)(0, 1, 1)_{12}$	3826,30

Source : Auteur.

Note : Les modèles SARIMA sélectionnés présentent tous un diagnostic d'ajustement "Good" selon JDemetra+. Les diagnostics sont en annexe.

D'après le tableau, les modèles les plus ajustés à nos chroniques nationale et départementales (53 et 72) sont obtenus *via* TRAMO-SEATS, alors que les meilleurs modèles pour les autres séries départementales sont obtenus *via* X-13ARIMA. Par ailleurs, nous pouvons noter que les modèles SARIMA ajustés pour les séries nationale, régionale ainsi que celle du département du Maine et Loire (49) sont identiques quelque soit la méthode de désaisonnalisation sélectionnée. De plus, les départements de la Loire-Atlantique (44) et de la Mayenne (53) possèdent un modèle SARIMA identique. Nous avons pu voir dans la section 1.2.1 que les deux programmes étaient quasi-similaires dans leurs pré-traitements, mais qu'ils présentaient certaines différences, d'où une différence observée entre les modèles identifiés, qui engendre de ce fait des écarts dans l'estimation de ces modèles. Nous remarquons finalement que les différences sont au niveau des polynômes non saisonniers car tous les modèles SARIMA ont la même structure saisonnière, à savoir un polynôme *Seasonal integrated moving average SIMA* $(0, 1, 1)_{12}$ .

Les *outliers* déstabilisent le bon fonctionnement de l'ajustement saisonnier, les corriger est donc d'une importance majeure. Expliquer un *outlier* consiste à comprendre la réaction extrême de la série au travers de l'actualité contemporaine du point atypique, le plus souvent d'ordre économique, politique ou statistique. Les tableaux de synthèse 1.4 et 1.5 permettent d'observer que certains *outliers* sont présents dans plusieurs séries à la fois, tandis que certains sont spécifiques à une seule série.

Dans un premier temps, nous allons présenter les *outliers* saisonniers dans leur ensemble,

sans suivre la chronologie de leurs apparitions. En effet, les points atypiques saisonniers sont plus particuliers que les autres *outliers*. Nous savons que les séries des demandeurs d'emploi connaissent une saisonnalité caractérisée par une baisse du nombre d'inscriptions vers la fin du deuxième trimestre et une hausse significative des inscriptions à Pôle Emploi à partir du quatrième trimestre. Les SO détectés en août 1998 dans la région des Pays de la Loire et en août 1999 dans le département de la Loire-Atlantique (44) correspondent à une rupture dans le profil saisonnier. Une explication probable de cette rupture proviendrait d'une forte utilisation des contrats courts pendant la période estivale. De plus, l'arrivée des jeunes diplômés sur le marché de l'emploi est aussi une raison qui est souvent avancée pour expliquer ce phénomène saisonnier.

En outre, nous remarquons qu'un SO est détecté en août 2015 pour l'ensemble des séries étudiées, excepté pour les départements du Maine et Loire (49) et de la Mayenne (53) qui observent ce SO en 2014 et 2016, respectivement. Les arguments développés précédemment peuvent s'appliquer ici, néanmoins un phénomène plus récent est souvent mis en avant par Pôle Emploi. En effet, Pôle Emploi constate depuis quelques années une forte variabilité dans les désinscriptions ayant pour motif le défaut d'actualisation, potentiellement dû à la période de vacances estivale. Ainsi, cela a pour conséquence d'impacter directement les chiffres des demandeurs d'emploi. Cette nouvelle baisse récurrente en septembre s'apparenterait à un mois transitoire dans lequel les demandeurs d'emploi terminent un contrat de travail ou ne prennent pas le temps de se réinscrire en bonne et due forme sur les listes de Pôle Emploi.

Revenons aux années 2000 et reprenons l'explication des points atypiques en suivant la chronologie de leurs apparitions. Le dynamisme des années 2000 a impacté positivement la croissance économique et par conséquent le niveau du chômage. Un *outlier* a été identifié en septembre/octobre 2000 pour l'ensemble de nos séries, excepté pour la série des demandeurs d'emploi nationale et la série du département de la Mayenne (53). Les points atypiques sont de type TC et LS en fonction des zones géographiques considérées. En ce sens, l'effet de ce regain d'activité est temporaire dans certaines zones alors qu'il est permanents pour les autres.<sup>13</sup> Nous pouvons noter, qu'un LS est détecté en septembre et en octobre 2000 pour le département de la Sarthe, ce qui pourrait correspondre à un grand changement dans la tendance de la série, suite à une forte baisse du nombre de demandeurs d'emploi dans ce

---

13. Effet temporaire pour le département du Maine et Loire (49), effet permanent pour la région Pays de la Loire et les départements de la Loire-Atlantique (44), Sarthe (72)

département et à cette période.

La crise économique de 2008 a eu un effet négatif, relativement constant dans les séries des demandeurs d'emploi. Par conséquent, il n'est pas étonnant d'observer un *outlier* de type LS en novembre 2008 pour la région des Pays de la Loire ainsi que les départements de la Loire-Atlantique (44) et de la Vendée (85).

Le dernier point atypique en commun est un AO localisé en février 2016 dans les séries régionale et départementale Sarthoise (72). Ce point atypique est explicité, selon le gouvernement de l'époque et Pôle Emploi, comme un nombre important de réinscriptions à la suite des défauts d'actualisation lors du mois précédent. Pôle Emploi évoque très souvent ce phénomène et prévient que l'effet est variable, pouvant être tantôt à la hausse, tantôt à la baisse.

Nous venons de présenter les *outliers* communs à la plupart des séries chronologiques étudiées. Maintenant, nous allons détailler les points atypiques spécifiques à une série de demandeurs d'emploi. La plupart des *outliers* spécifiques à une série sont identifiés dans la série du département de la Loire-Atlantique (44).

Les deux premiers *outliers* sont identifiés en mai 1997 (LS) et juillet 1997 (TC). Ces deux points atypiques évoquent un changement permanent et un changement temporaire dans la tendance de la série. La raison est probablement liée à la crise du système monétaire Européen, qui a créé une pression économique avant le passage à l'euro.

En août 1998, le programme RegARIMA identifie un AO, toujours dans la série des demandeurs d'emploi de la Loire-Atlantique (44). À cette époque, le climat économique du département est marqué par une forte hausse du chômage et une grève des chômeurs qui consistait à occuper certaines agences de L'Association pour l'emploi dans l'industrie et le commerce (Assedic)<sup>14</sup>. Néanmoins, une raison plus évidente, en raison de la date à laquelle est localisée le point atypique, correspond à un effet "coupe du monde". La recrudescence des emplois liée à la coupe du monde de la FIFA<sup>15</sup> de 1998 (fin juin, début juillet) est retombée dès la fin de la compétition.

---

14. Il s'agit de l'ancien nom de Pôle Emploi avant la fusion de l'ANPE (l'Agence Nationale Pour l'Emploi) et des Assedic fin 2008.

15. Fédération Internationale du Football Association.

L'année 1999 est une mauvaise année en terme de chômage. Cependant, une première baisse du nombre des demandeurs d'emploi est amorcée afin que la tendance passée retrouve un niveau moins élevé. C'est ainsi qu'un LS est détecté en avril 1999 dans la série de la Loire-Atlantique (44). Les raisons de ce LS peuvent être dues au retour de la croissance économique, ou encore aux politiques d'emploi de l'époque (emploi-jeune par exemple). Dans une logique inverse, un AO est identifié en août 1999 dans le département de la Sarthe. Durant cette période, le taux de chômage fait un bond de plus de 11% ayant pour cause la suppression des emplois dans l'usine Moulinex (Roupenel-Fuentes, 2015).

Le programme décèle un *outlier* de type TC en septembre 2003, toujours dans le département de la Loire-Atlantique (44). La consommation des ménages est faible durant cette période et la croissance est chahutée, en ce sens, cela pourrait être une explication d'un changement temporaire dans la tendance de la série.

Toujours dans le même département, le mois de mars 2011 est sujet à une valeur aberrante (AO). Le rebond de la création d'emplois est mis en cause, favorisant la baisse du nombre de demandeurs d'emploi.

L'*outlier* de septembre 2013 correspond au "Bug SFR". Ce bug avait empêché la relance des e-mails de réinscription des demandeurs d'emploi, entraînant une chute considérable des inscriptions fin août 2013 avant une remontée très forte dès le mois suivant.

Pour terminer, le dernier *outlier* détecté est un LS situé en mars 2016 dans le département de la Vendée (85). À cette époque, la Vendée connaît de fortes embauches grâce à son réseau de petites et moyennes entreprises (PME), ce qui influence la tendance à la baisse de la série du nombre des demandeurs d'emploi.

TABLE 1.4 – Synthèse des *outliers* communs

Date	Type	Séries	Explication probable
Août-1998, Août-1999	SO	PDLL, 44	Apparition des pics au mois d'août, plus de CD
Septembre-2000, Octobre-2000	LS/TC	PDLL, 44, 49, 72	Dynamisme des années 2000,
Novembre-2008	LS/TS	PDLL, 44, 85	période de croissance économique Effet de la crise économique de 2008
Août-2014, Août-2015, Août-2016	SO	PDLL, 44, 49, 53, 72, 85	Modification du profil saisonnier suite à l'utilisation accrue des contrats courts (CDD, contrats saisonniers) et suite à la sortie des jeunes du système scolaire se présentant sur le marché de l'emploi.
Février-2016	AO	PDLL, 72	Défaut d'actualisation Phénomène de réinscriptions importantes suite aux défauts d'actualisation le mois précédent

Source : Auteur, JDemetra+

TABLE 1.5 – Synthèse des *outliers* spécifiques

Date	Type	Séries	Explication probable
Mai-1997, Juin-1997	LS/TS	44	Hausse du nombre de demandeurs d'emploi, crise du système monétaire eu- ropéen
Août-1998	AO	PDLL, 44	Forte hausse du chômage et grève des chômeurs, effet coupe du monde
Avril-1999	LS	44	Effet d'une légère reprise suite à une période de récession économique
Août-1999	AO	72	Fort augmentation mensuelle (creux saisonnier)
Septembre-2003	TC	44	Période de faible consommation et croissance économique chahutée

Source : Auteur, JDemetra+

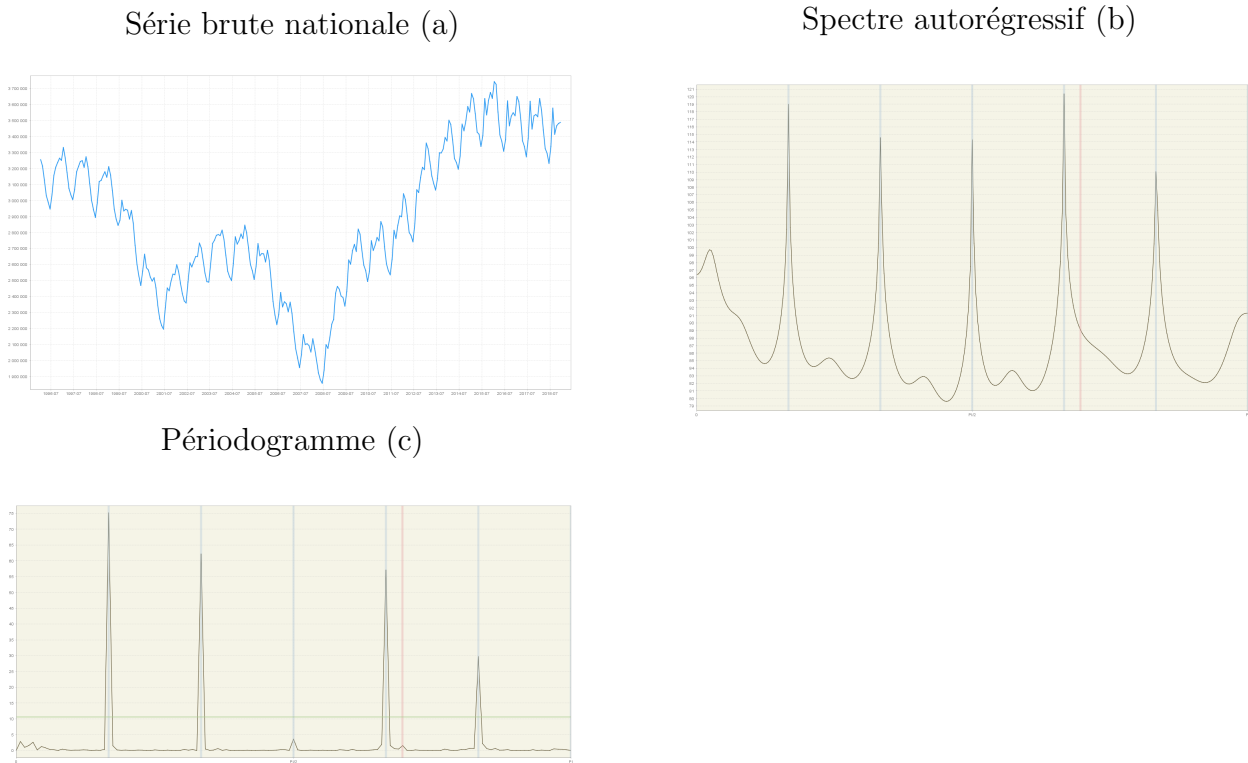


### 1.3.2 Analyse de la désaisonnalisation

Nous avons fait le choix de corriger les variations saisonnières et les effets des jours ouvrables (CVS-CJO) nous-même. En effet, les séries de demandeurs d'emploi en catégorie A aux niveaux régional et départemental sont disponibles corrigées des variations saisonnières et des jours ouvrables depuis le début de l'année 2018. Néanmoins, les séries CVS-CJO comportent une composante irrégulière qui peut rendre difficile la détection des points de retournements du fait de la volatilité, comme ce fut le cas durant l'année 1996. Afin d'obtenir une tendance-cycle lisse, permettant ainsi de mieux détecter les points de retournements, nous décomposons la série brute des demandeurs d'emploi en utilisant le logiciel JDemetra+ recommandé par Eurostat. Cet outil permet de recourir à la méthode TRAMO-SEATS proposée par Gómez et Maravall (1996) et à la méthode du U.S. Census Bureau, X-13ARIMA que nous avons présentée dans la section méthodologique.

Avant toute manipulation, nous devons vérifier que nos chroniques possèdent une saisonnalité, pour ce faire, nous observons visuellement la forme de la série. Puis, nous analysons les spectres autorégressifs et les périodogrammes. Il faut s'assurer que toutes nos chroniques sont sujettes à un comportement saisonnier, en observant graphiquement les séries brutes, puis en étudiant les spectres autorégressifs et les périodogrammes de chaque série.

FIGURE 1.3 – Série demandeurs d'emploi nationale catégorie A-Brute, spectre autorégressif et périodogramme

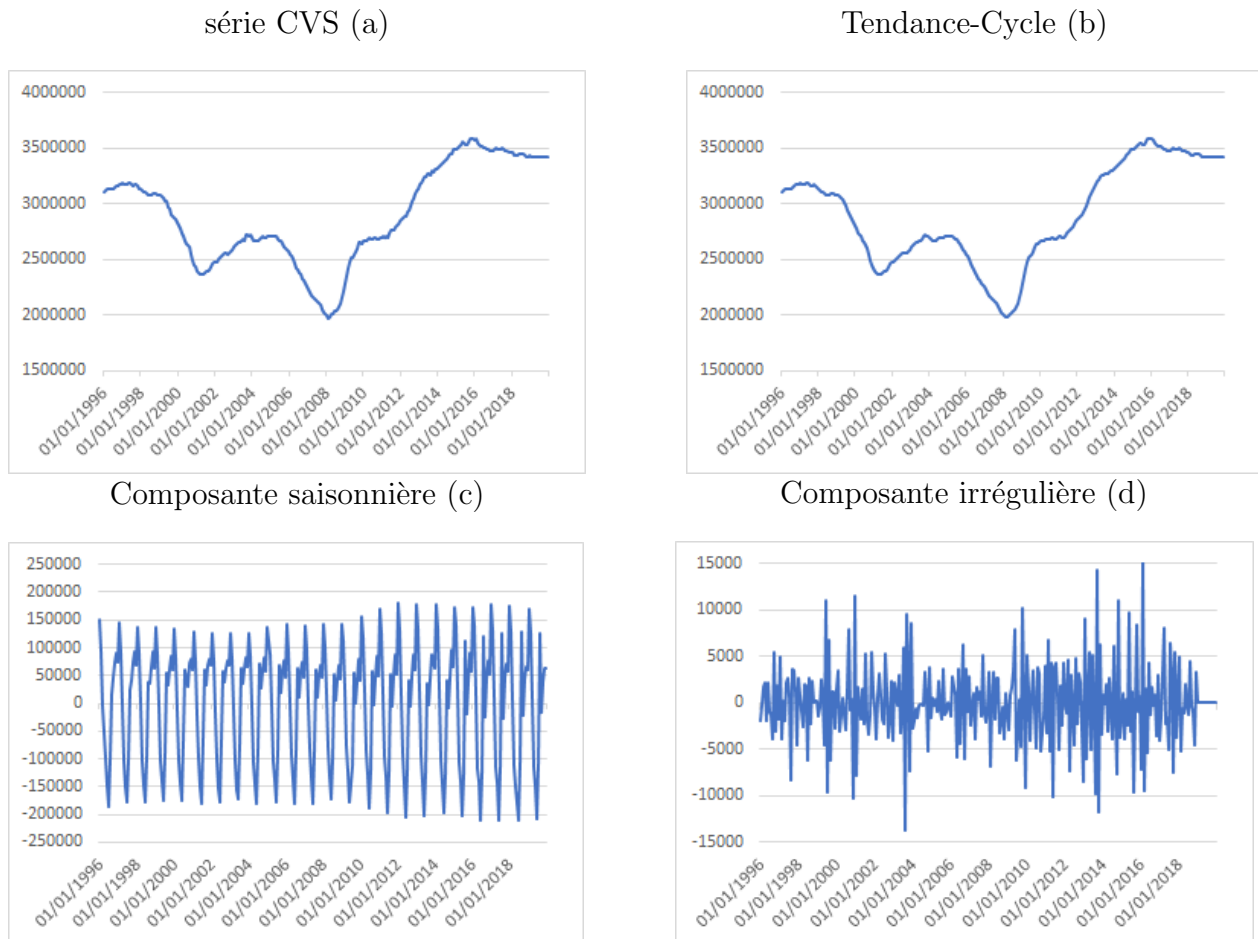


Le spectre autorégressif et le périodogramme présentent bien des pics aux fréquences saisonnières mensuelles (lignes grises Fig. 1.3 (b) et (c)). Les séries brutes, les spectres autorégressifs et les périodogrammes de la série des Pays de la Loire et des départements (44, 49, 53, 72, 85) sont présentées en annexe.

Les procédures automatiques TRAMO et RegARIMA rendent stationnaires les séries chronologiques, puis traitent les éventuelles valeurs manquantes, en terminant par corriger les séries de leurs valeurs atypiques, afin de réaliser un ajustement saisonnier le moins perturbé possible par ces valeurs extrêmes, qui ne refléteraient pas la tendance longue de chaque série. Nous pouvons à présent observer le résultat de cette analyse. Rappelons que, SEATS utilise la série linéarisée par TRAMO et ajuste un modèle SARIMA, puis, SEATS fait appel à la méthode ARIMA Model-Based (AMB) afin d'obtenir les composantes saisonnière, tendance-cycle et irrégulière. D'une autre manière, X-13ARIMA emploie une succession de filtres sur la série nettoyée par RegARIMA. Dans les deux cas, nous obtenons une série CVS-CJO de bonne qualité. L'utilisation des séries brutes saisonnières permet d'obtenir directement la composante Tendance-Cycle grâce aux méthodes de désaisonnalisation. Cette composante

est naturellement plus lisse que la série CVS puisque la composante irrégulière est retirée. Les graphiques suivants reprennent les résultats de l'ajustement saisonnier.

FIGURE 1.4 – Série demandeurs d'emploi nationale catégorie A-CVS-CJO par TRAMO-SEATS (a), Tendance-Cycle (b), composante saisonnière (c) et composante irrégulière (d)



Nous pouvons observer sur la Figure 1.4, la série CVS et la Tendance-Cycle prolongée de douze mois. Ce sont des prévisions réalisées par le logiciel afin d'éviter les effets de bord. Un effet de bord est un biais du comportement aux extrémités de la série à la suite d'une manipulation de celle-ci. Par exemple, les séries peuvent être sujettes à de fortes révisions qui impacteraient les données en fin de série. En outre, nous remarquons que la Tendance-Cycle est plus lisse que la série CVS. On peut constater que la composante irrégulière présente un comportement très fluctuant.

Nous avons choisi de placer les autres graphiques relatifs aux séries régionales et départementales en annexe pour une meilleure lecture du document. Cependant, il est intéressant d'effectuer une analyse statistique des séries CVS que nous avons calculées en les comparant aux séries

CVS de Pôle Emploi, qui corrige ses séries des variations saisonnières *via* la méthode X-12ARIMA.

TABLE 1.6 – Statistiques descriptives des données CVS de Pôle Emploi, en millier

	Nationale	PDLL	44	49	53	72	85
Min.	1976	90,3	34,9	20,4	5,3	15	14,1
Max.	3585	178,1	66,7	41,3	11,2	29,6	29,5
Moyenne	2875	138,1	53,7	31,5	8,5	22,7	21,5
Médiane	2766	132,2	50,9	29,6	8,3	22	21
Ecart-type	441,3	23,8	8,6	5,5	1,5	4,2	4,4
Skewness	-0,07	-0,013	-0,213	0,1	0,03	-0,0001	0,3
v1	-0,50	-0,09	-1,46	0,72	0,24	0,00	2,10
Kurtosis	1,99	1,9	2	1,98	2	1,64	1,92
v2	-3,38	-3,72	-3,34	-3,41	-3,35	-4,60	-3,63
Jarque-Bera	0,0092	0,0009	0,001	0,002	0,003	0,00002	0,0001

Source : Auteur, données Pôle Emploi.

Note : PDLL correspond à la région Pays de la Loire et les départements sont : Loire-Atlantique (44), Maine et Loire (49), Mayenne (53), Sarthe (72), Vendée (85). Les valeurs du test de Jarque-Bera est la p-value. Si  $p < 0.5$  alors on rejette l'hypothèse nulle de normalité.

TABLE 1.7 – Statistiques descriptives des données CVS selon l'auteur, en millier

	Nationale	PDLL	44	49	53	72	85
Min.	1974	90,4	34,9	20,3	5,3	15	13,9
Max.	3587	179,9	67,1	41,5	11,2	29,8	29,9
Moyenne	2874	138	53,6	31,5	8,5	22,7	21,5
Médiane	2766	132,1	50,9	29,6	8,3	22	21
Ecart-type	440,9	23,7	8,6	5,5	1,5	4,2	4,4
Skewness	-0,07	-0,01	-0,2	0,1	0,03	-0,002	0,31
v1	-0,50	-0,07	-1,39	0,72	0,22	-0,01	2,14
Kurtosis	1,99	1,9	2	1,98	2	1,6	1,93
v2	-3,39	-3,72	-3,36	-3,41	-3,37	-4,60	-3,61
Jarque-Bera	0,002	0,0009	0,001	0,002	0,003	0,00002	0,0001

Source : Auteur, données Pôle Emploi.

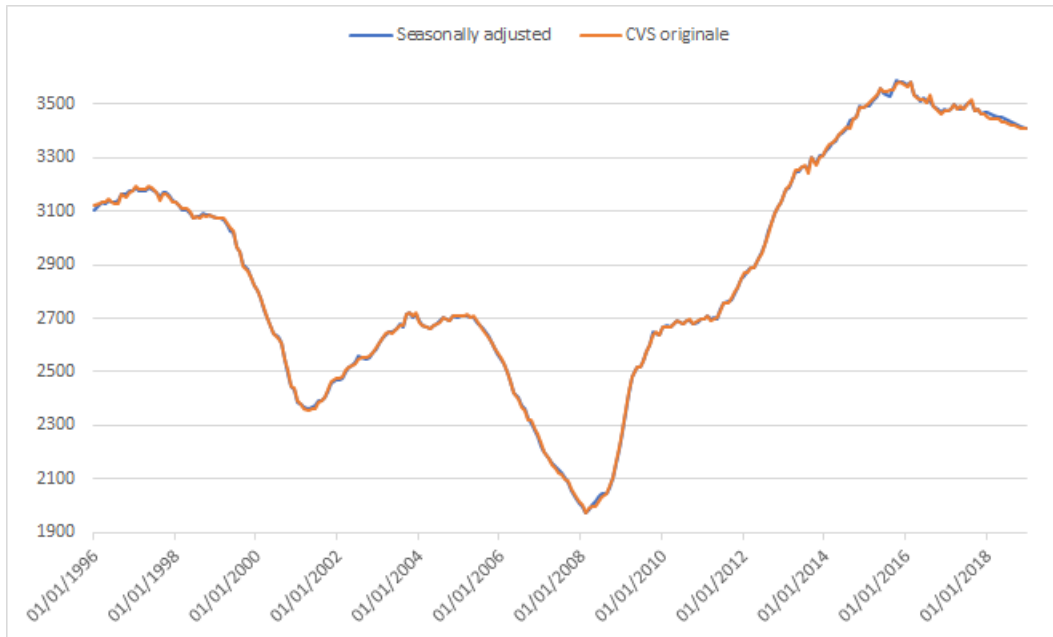
Note : Méthode TRAMO-SEATS pour Nationale et 72, méthode X-13ARIMA pour les autres. PDLL correspond à la région Pays de la Loire et les départements sont : Loire-Atlantique (44), Maine et Loire (49), Mayenne (53), Sarthe (72), Vendée (85). Les valeurs du test de Jarque-Bera est la p-value. Si  $p < 0.5$  alors on rejette l'hypothèse nul de normalité.

L'étude des statistiques descriptives (Table 1.6 et 1.7) met en lumière la performance de nos méthodes de désaisonnalisation, dans le sens où l'ensemble des mesures présentées sont sensiblement proches les unes des autres. On observe que le département de la Loire-Atlantique (44) possède une moyenne plus élevée du nombre de demandeurs d'emploi, suivi par les départements du Maine et Loire (49), de la Sarthe (72), de la Vendée (85) et enfin de la Mayenne (53). On peut classer les départements selon la médiane et nous retrouvons le même ordre que pour la moyenne, à l'exception de la Sarthe (72) et la Vendée (85) qui échangent leur place. Au niveau des coefficients d'asymétrie, la Loire-Atlantique (44) et la Sarthe (72) possèdent une distribution étalée à gauche, tandis que les autres départements ont une distribution étalées à droite, que ce soit pour les séries CVS de Pôle Emploi ou celles

que nous avons calculées. Dans tous les cas, le coefficient d'aplatissement renseigne que toutes les séries comparées sont platikurtiques. Par ailleurs, aucun changement dans les conclusions de ces mesures entre les séries Pôle Emploi et les nôtres n'est observé.

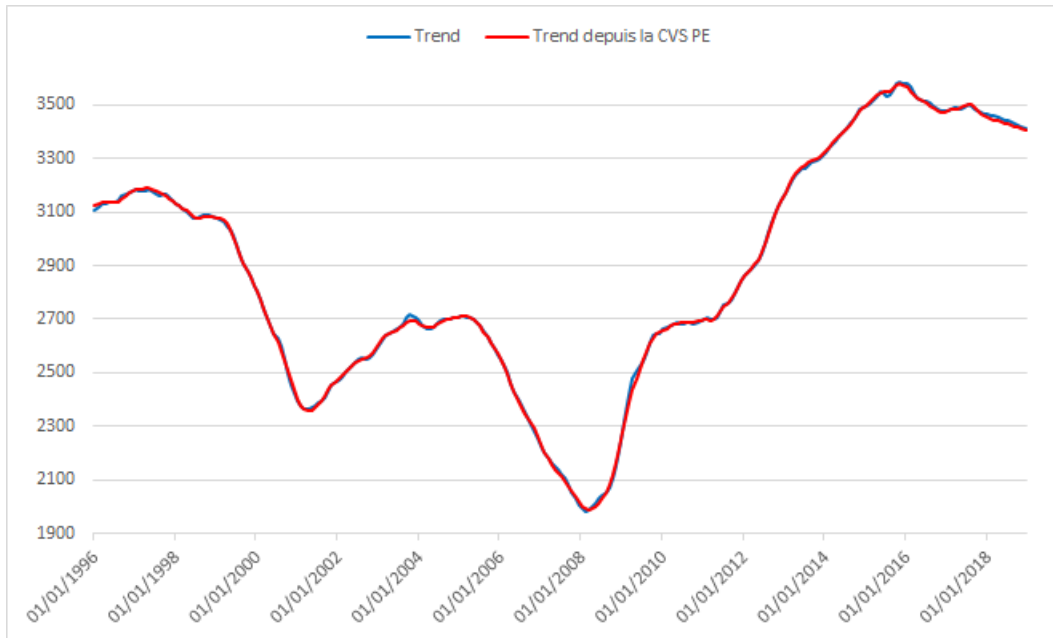
Le graphique suivant permet de comparer directement la série CVS de Pôle Emploi et la série CVS estimée *via* TRAMO-SEATS. Les deux séries sont similaires. Ce qui confirme la qualité et l'utilisation de la méthode de désaisonnalisation. On peut néanmoins observer que notre série ajustée est légèrement plus lisse que la série originale de Pôle Emploi. Ainsi, l'obtention d'une série CVS plus lisse que celle de Pôle Emploi permet d'obtenir une composante Tendence-Cycle plus lisse et par conséquent d'améliorer la datation des points de retournement.

FIGURE 1.5 – *Comparaison DEFM Pôle Emploi CVS versus série ajustée par TRAMO-SEATS - National*



A titre de comparaison, nous avons appliqué la méthode de désaisonnalisation sur la série CVS de Pôle Emploi afin d'obtenir la composante Tendence-Cycle. Nous remarquons que notre estimation de la Tendence-Cycle de notre série ajustée est quasiment identique à la Tendence-Cycle estimée à partir de la série de Pôle Emploi (Figure 1.5). Cependant, on peut remarquer que notre composante Tendence-Cycle reste plus lisse, justifiant son utilisation dans la datation cyclique.

FIGURE 1.6 – *Comparaison Trend DEFM Pôle Emploi CVS versus Trend série ajustée par TRAMO-SEATS - National*



Les graphiques 1.6 présentent les séries saisonnières superposées aux composantes Tendance-Cycle calculées par une des deux méthodes de désaisonnalisation. Un premier commentaire général consiste à dire que la forme des différentes séries temporelles est similaire entre les différents niveaux géographiques considérés. La série nationale est, tout de même, moins fluctuante que la série régionale ou les séries départementales<sup>16</sup>. De plus, on observe que la Tendance-Cycle est beaucoup plus lisse que la série brute. En effet, on comprend pourquoi les méthodes de datation des points de retournement s'appliquent mieux sur cette composante que sur la série brute. (Figure 1.7 à 1.13).

Enfin, il est intéressant d'étudier les différences de tendance entre les cinq départements de la région Pays de la Loire. Nous pouvons remarquer que certains creux ou pics sont davantage marqués d'une série à l'autre. La datation et l'analyse des mouvements cycliques va nous informer sur les différences et sur les similarités de ces séries de demandeurs d'emploi.

16. En annexe, nous avons présenté les coefficients de variation de chaque série.

FIGURE 1.7 – Série nationale des demandeurs d'emploi brute catégorie A national et Tendence-Cycle - (1996-2018)

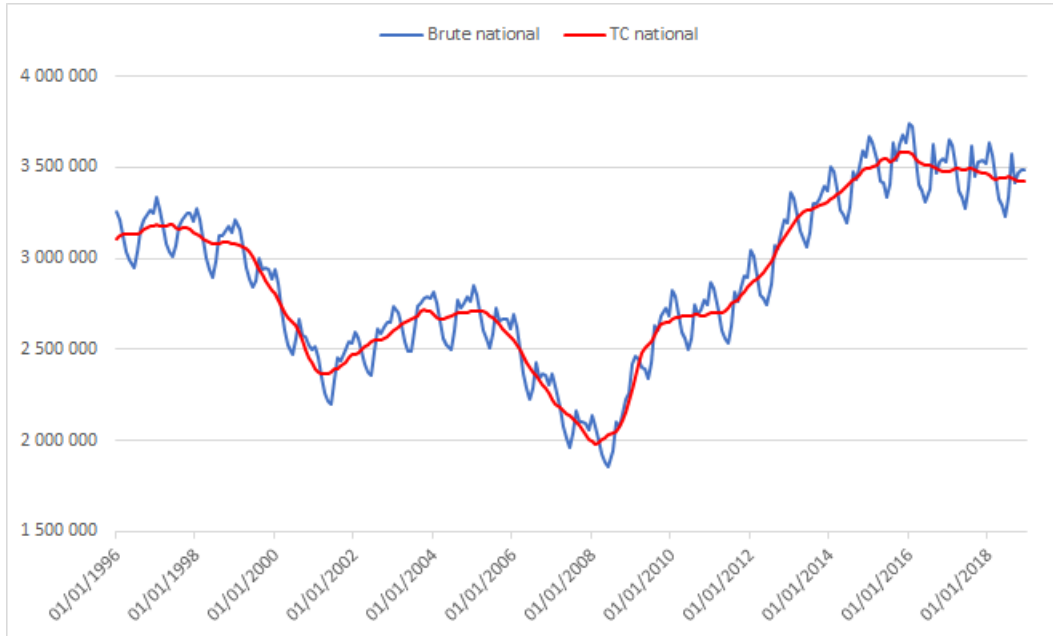


FIGURE 1.8 – Série régionale des demandeurs d'emploi brute catégorie A et Tendence-Cycle - (1996-2018)

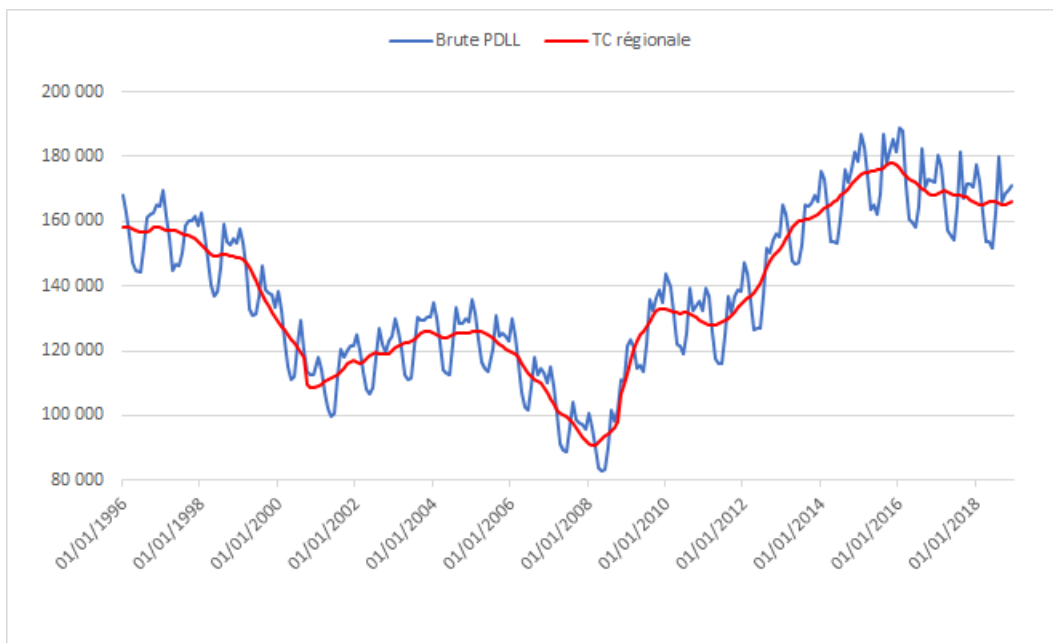


FIGURE 1.9 – Série départementale (Loire-Atlantique) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)

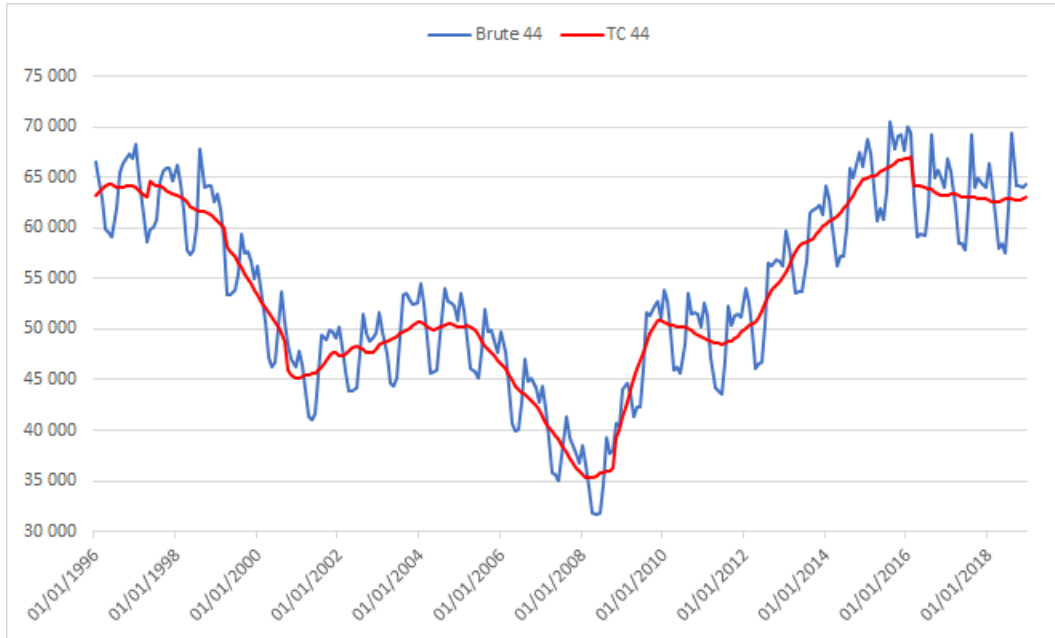


FIGURE 1.10 – Série départementale (Maine-et-Loire) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)

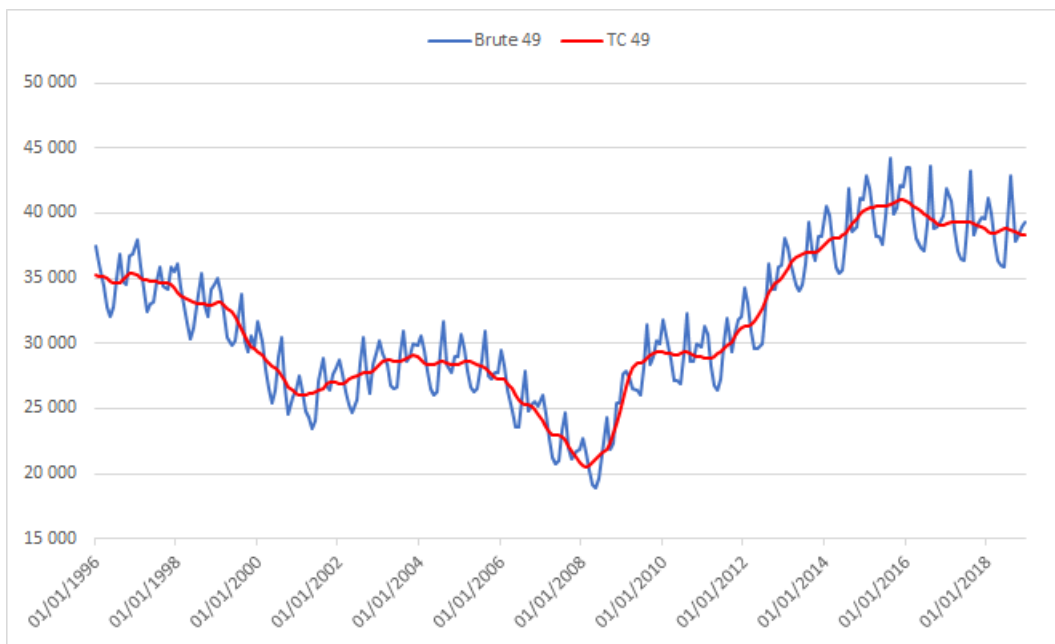




FIGURE 1.11 – Série départementale (Mayenne) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)

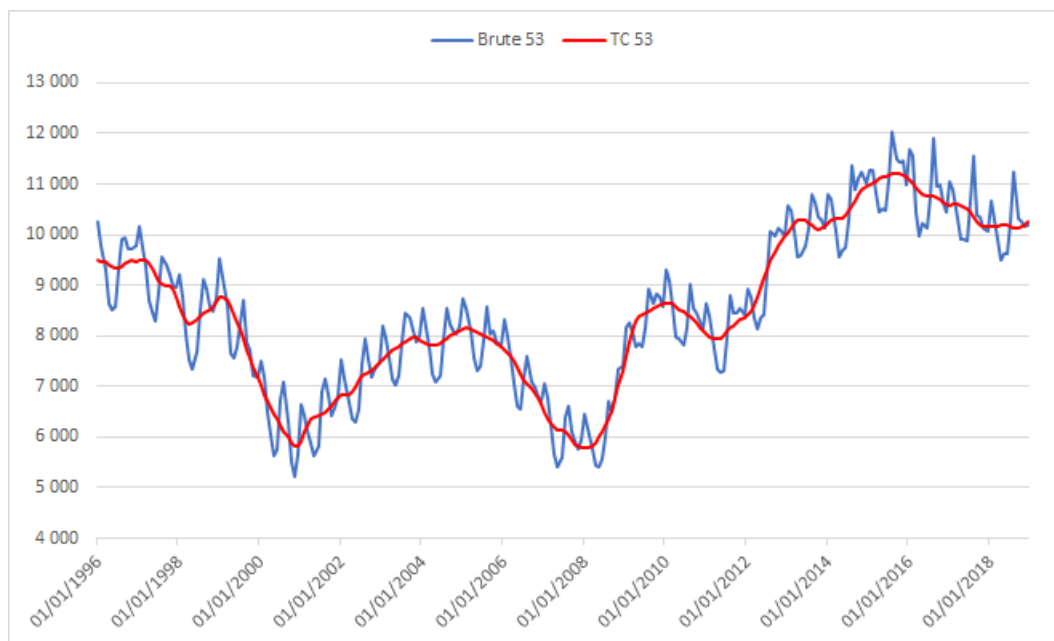


FIGURE 1.12 – Série départementale (Sarthe) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)

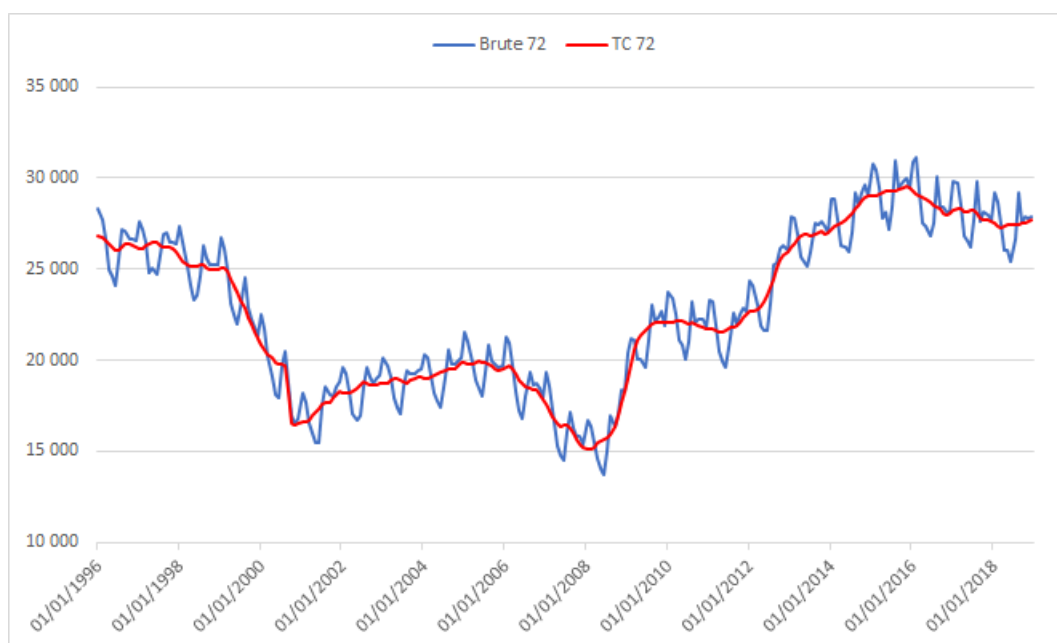
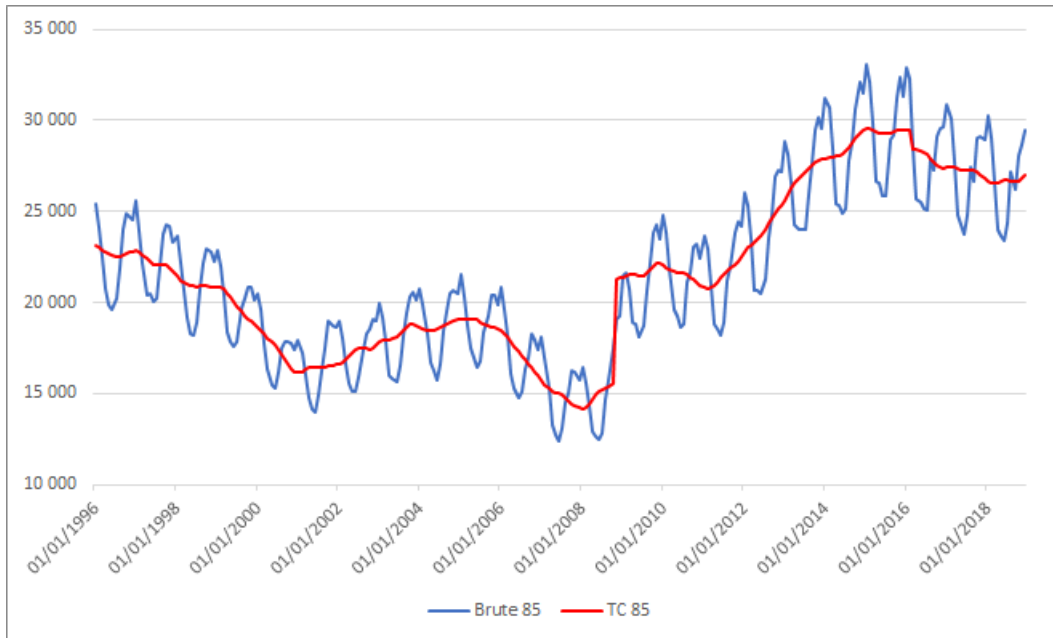


FIGURE 1.13 – Série départementale (Vendée) des demandeurs d'emploi brute catégorie A et Tendance-Cycle - (1996-2018)



## 1.4 Application et résultats

### 1.4.1 Le cycle des affaires

Les premiers résultats présentés dans cette section ont pour objectif de proposer une chronologie des points de retournement des cycles classique et d'accélération des demandeurs d'emploi au niveau national, ainsi qu'au niveau de la région des Pays de la Loire et de ses départements. De plus, une étude de la synchronisation des cycles est réalisée afin d'estimer empiriquement dans quelle proportion de temps les cycles sont dans une même phase.

#### Datation des points de retournement

La datation du cycle des affaires de la série des demandeurs d'emploi (catégorie A) nationale est obtenue à partir de l'algorithme de Bry-Boschan appliqué sur la Tendance-Cycle de la série que nous avons estimée précédemment. En outre, nous avons remarqué qu'il était pertinent de travailler sur une série corrigée plutôt que sur la série officielle publiée par Pôle Emploi. En effet, la Tendance-Cycle, correspondant à la série CVS-CJO allégée de sa composante irrégulière, est plus lisse que la série CVS-CJO de Pôle Emploi. Par conséquent, la datation des points de retournement est facilitée. Cependant, la datation des pics et des creux est basée sur la série CVS de Pôle Emploi, puisque l'objectif est de fournir une datation officielle des cycles des demandeurs d'emploi français. Nous avons fait le choix de "marquer" les phases d'expansion plutôt que les phases de contraction, car les séries de demandeurs d'emploi sont contracycliques, et nous voulons mettre en avant les phases du cycle où le nombre de demandeurs d'emploi augmente, afin d'identifier plus aisément dans quels contextes ces augmentations interviennent. Conformément à littérature, nous n'étudions pas les points de retournement présents dans les 6 premiers et les 6 derniers mois de notre chronologie, dans le but d'éviter les faux pics ou les faux creux qui ne possèdent pas de signification cyclique en raison du manque de disponibilité des données avant le début et après la fin d'une série chronologique.

Le tableau 1.8 synthétise les chronologies des points de retournement du cycle classique des demandeurs d'emploi (catégorie A). De manière générale, l'étude des cycles des affaires des demandeurs d'emploi fait apparaître deux grandes phases d'expansion, toutes zones géographiques confondues. La première phase d'expansion du cycle se localise vers

TABLE 1.8 – Datation du cycle d’emploi – 1996-2019

	HUR	Nationale	PDLL	44	49	53	72	85
Pic	1996M11	1997M5						
Creux						1998M3		
Pic						1999M1		
Creux	2001M1	2001M4	2000M11	2000M11	2000M11	2001M3	2000M11	2000M11
Pic	2001M8							
Creux	2003M1							
Pic	2003M12	2003M10	2003M11	2004M1	2003M10	2003M9		
Creux	2005M1	2004M4	2004M6			2004M4		
Pic	2006M1	2005M3	2005M2			2005M3	2005M5	2005M4
Creux	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2
Pic	2009M12		2010M1	2010M5		2010M1	2010M5	2010M1
Creux	2011M3		2011M3	2011M5		2011M6	2011M5	2011M1
Pic	2015M5	2015M11	2015M11	2016M2	2016M2	2015M11	2015M11	2015M1
Creux			2016M11					
Pic			2017M8					
Creux			2018M2	2018M2		2018M9	2018M2	2018M3

Source : Auteur, données Pôle Emploi.

la fin de l’année 2000 et début 2001, et s’expliquerait par un ralentissement de la croissance économique française et européenne à la suite de l’éclatement de la bulle internet. L’Insee évoque une création d’emploi en forte baisse, avec seulement 183 000 emplois créés en mars 2002 contre 498 000 en mars 2001, soit une baisse de 315 000 créations d’emplois. La deuxième phase d’expansion du cycle est parfaitement identifiée en février 2008 dans toutes les séries observées. La crise financière et économique marque ses effets dès le début 2008 dans les séries des demandeurs d’emploi. A titre de comparaison, l’organisme *Economic Cycle Research Institute* estime l’entrée en récession de l’activité française en février 2008 et rejoint donc notre propre estimation sur le cycle des demandeurs d’emploi. Nous avons aussi appliqué l’algorithme de datation de Bry-Boschan sur la série mensuelle du taux de chômage

harmonisé<sup>17</sup> (HUR) calculé par l'OCDE. Le taux de chômage harmonisé utilise la définition du BIT calculée en France par l'Insee. Elle définit les chômeurs comme des personnes en âge de travailler qui sont sans emploi, disponibles pour travailler et recherchant activement du travail. L'indicateur est mesuré à partir de cette définition commune du chômage. L'HUR correspond donc au pourcentage de chômeurs dans la population active, le tout corrigé des variations saisonnières. La comparaison de cette datation avec celle du niveau national de notre étude permet d'apprécier une similarité au niveau de la localisation de la plupart des pics et des creux de la chronologie, notamment le creux de février 2008.

Une dernière phase d'expansion marquante, commune à la plupart des séries étudiées, se situe en 2011 à la suite de la crise européenne. Ce creux n'est pas visible dans la série nationale des demandeurs d'emploi, en raison d'une trop forte inertie de la phase d'expansion précédente qui correspondait aux effets de la crise de 2008 sur le niveau du chômage.

Par ailleurs, les différences qu'on peut observer entre la région Pays de la Loire et la France, ou les différences entre départements, prennent deux formes. La première différence intéressante correspond au décalage de datation entre les débuts et les fins des phases d'expansion ou de contraction des cycles des séries étudiées. Par exemple, l'expansion de février 2008 est commune à tous les cycles calculés, alors que l'expansion du début des années 2000 varie entre novembre 2000 et avril 2001.

La deuxième forme de différence réside dans l'existence de cycles supplémentaires dans certaines séries, qui ne se retrouvent pas dans l'ensemble de l'échantillon. Par conséquent, une interprétation économique de ces différences peut prendre la forme d'une intuition sur les caractéristiques favorables d'une zone géographique en matière de réduction du chômage. En effet, calculer et étudier le cycle classique des demandeurs d'emploi jusqu'au niveau départemental permet de capter une information propre à chaque territoire. Par exemple, la région Pays de la Loire possède une phase de contraction du cycle entre janvier 2010 et mars 2011, qui n'est pas présente sur le cycle national. De ce fait, on peut penser que la région possède une caractéristique, qui favorise la baisse du nombre de demandeurs d'emploi. Par ailleurs, si nous poursuivons notre raisonnement et que nous étudions le cycle des départements de cette région, on observe que cette phase de contraction est présente dans tous

---

17. OECD (2019), Harmonised unemployment rate (HUR) (indicator). doi : 10.1787/52570002-en (consulté le 20 juin 2019).

les départements de la région, à l'exception du Maine-et-Loire. Par conséquent, la conclusion de notre intuition permettra de dire que, outre le Maine-et-Loire, l'ensemble des départements ont contribué ou bénéficié de caractéristiques favorables à la baisse localisée du chômage.

Au contraire, nous pouvons aussi penser que d'autres régions impactent plus négativement le nombre de demandeurs d'emploi national, ce qui engendrerait une augmentation constante d'année en année. Par conséquent, le point de retournement relatif à la crise européenne de 2011 est masqué par la longue phase d'expansion du cycle des demandeurs d'emploi français. Plus récemment, un creux est détecté dans la série nationale en novembre 2016 et le pic de cette phase d'expansion est localisé en août 2017. La baisse prolongée du nombre de demandeurs d'emploi durant l'année 2016 s'est interrompue au profit d'une remontée durable des inscriptions. Cette phase d'expansion du cycle des demandeurs d'emploi nationaux intervient dans un contexte où les chiffres de Pôle Emploi sont très volatils d'un mois à l'autre. Une telle augmentation peut être le résultat de transferts entre catégories des demandeurs d'emploi. Par exemple, on sait que l'année 2016 a été marquée par une baisse du nombre des demandeurs d'emploi en catégorie A au profit de la catégorie D<sup>18</sup>, catégorie utilisée pour regrouper les demandeurs d'emploi en formation. De ce fait, les demandeurs d'emploi en formation n'ont probablement pas retrouvé un emploi durable, ce qui expliquerait en partie, l'augmentation des inscriptions en catégorie A.

La situation actuelle montre que la série nationale ainsi que les séries des départements du Maine-et-Loire (49) et de la Mayenne<sup>19</sup> (53) des demandeurs d'emploi de catégorie A sont en phase de contraction, à savoir une phase de diminution de demandeurs d'emploi. Nous pouvons noter que la série du taux de chômage harmonisé de l'OCDE est aussi en période de contraction suite au dernier pic identifié en mai 2015. En revanche, les séries de la région Pays de la Loire, du département de la Loire-Atlantique (44), de la Sarthe (72) et de la Vendée (85) présentent depuis début 2018 une phase d'expansion du cycle des demandeurs d'emploi.

---

18. Sur la période novembre 2016 à août 2017, la catégorie D de Pôle Emploi voit ses effectifs diminuer de 14%.

19. L'algorithme de Bry-Boschan détecte un creux en septembre 2018 dont on ne tiendra pas compte étant donné la proximité de ce point de retournement avec la fin de l'échantillon, décembre 2018, soit un point de retournement présent dans les 6 derniers mois de la série.

## Caractéristiques du cycle des demandeurs d’emploi français

Le premier constat évident d’après le tableau 1.9, est une asymétrie entre les phases d’expansion et les phases de contraction du cycle des demandeurs d’emploi à tous les niveaux géographiques étudiés. Nos résultats mettent en évidence des durées moyennes et des amplitudes moyennes plus grandes en phase d’expansion qu’en phases de contraction. En d’autres termes, une expansion du cycle des demandeurs d’emploi est plus longue et plus marquée qu’une période de contraction. L’inertie présente dans les séries du chômage pourrait être responsable d’une partie de cette asymétrie.

TABLE 1.9 – Caractéristiques du cycle de l’emploi – 1996-2019

		HUR	Nationale	PDLL	44	49	53	72	85
<b>Expansion</b>	Durée	21,4	36,75	32,50	41,67	64,50	27,20	46,00	41,33
	Amplitude	0,98	518,43	28,07	13,50	12,55	2,00	6,29	6,47
	<i>Excess</i>	-0,02	26,58	0,77	0,36	0,68	0,10	0,57	0,15
<b>Contraction</b>	Durée	25	26	21	29,33	53	24	25	29
	Amplitude	1,36	437,70	14,52	7,28	9,23	1,64	2,75	3,11
	<i>Excess</i>	-0,21	-41,93	-0,53	-0,23	-1,25	-0,002	-0,20	-0,20

Note : La durée moyenne est en mois, l’amplitude moyenne est en milliers de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

A titre de comparaison, nous avons calculé les caractéristiques du cycle classique de la série du taux de chômage harmonisé de l’OCDE. Nous remarquons que la durée moyenne est plus longue en période de contraction, et l’amplitude moyenne la plus importante relève aussi des phases de contraction. Les résultats sont donc opposés aux mesures de notre série nationale. Une raison de cette différence réside dans l’utilisation par l’OCDE de la définition du chômage du BIT plutôt que dans la mesure de Pôle Emploi. Les écarts de mesure additionnés au fameux ”halo autour du chômage”<sup>20</sup>, peuvent fournir des mesures très éloignées. Selon Pôle Emploi et l’Insee, l’écart de la mesure est passé de 164 000 à 800 000 entre 2012 et 2017.

En se concentrant sur les départements de la région Pays de la Loire, que peuvent nous apprendre les caractéristiques des cycles des demandeurs d’emploi de ces territoires ? L’asymétrie en faveur des phases d’expansion est présente dans chaque département étudié. Par ailleurs, les ratios entre la durée des phases d’expansion et de contraction sont relativement proches,

20. Le halo autour du chômage rassemble tous les individus se rapprochant de la définition du chômage sans être comptabilisés comme chômeurs en raison d’une violation d’une condition essentielle de la définition du BIT.

même si l'on peut observer que le département de la Sarthe (72) possède le plus haut ratio. En d'autres termes, les phases d'expansion sont beaucoup plus longues que les phases de contraction. Dans une même démarche, nous pouvons observer les ratios entre les amplitudes des phases d'expansion et les phases de contraction. Les résultats sont moins homogènes puisque l'on remarque que la Sarthe (72) et la Vendée (85) possède un ratio élevé par rapport aux autres départements de la région, c'est-à-dire que les phases d'expansion sont plus amples et profondes que les phases de contraction. (Tableau 1.10)

TABLE 1.10 – Ratios moyens des durées et des amplitudes entre les phases cycliques

	PDLL	44	49	53	72	85
Ratio durée	1,55	1,42	1,22	1,13	1,84	1,43
Ratio amplitude	1,93	1,85	1,36	1,22	2,29	2,08

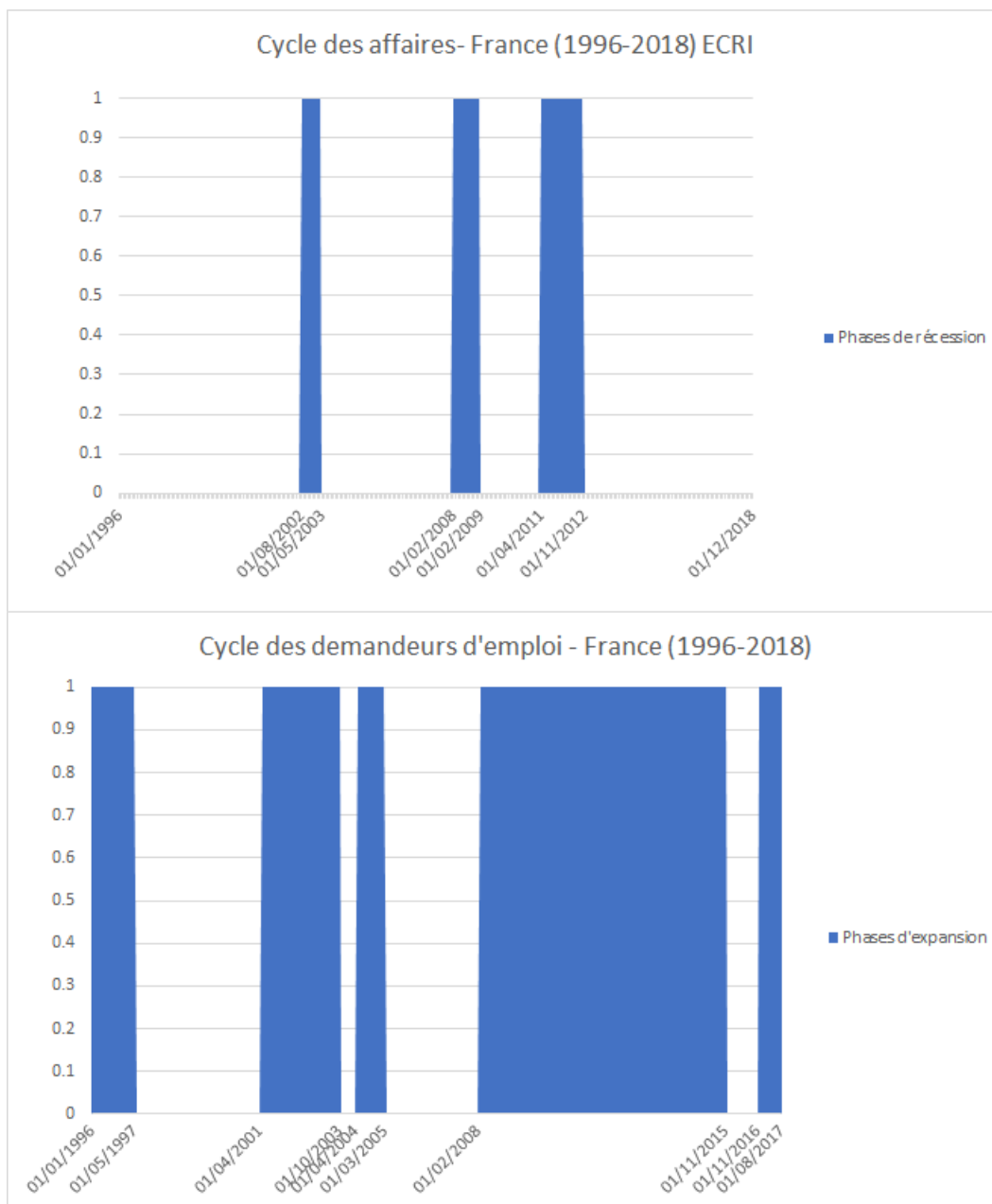
Note : Les deux ratios sont calculés avec la mesure de l'expansion au numérateur et la mesure de la contraction au dénominateur.

Concernant l'indice d'*Excess*, nous pouvons constater que l'ensemble des départements étudiés possèdent des phases de contraction et d'expansion progressives, au regard d'un indice d'*Excess* positif pour les phases d'expansion et un indice d'*Excess* négatif pour les phases de contraction. Cependant, certaines phases du cycle sont estimées comme brutales si nous étudions l'indice d'*Excess* pour chacune des phases identifiées. Nous avons reporté en annexe cette analyse.

En outre, nous pouvons comparer graphiquement le cycle des affaires français calculé par l'*ECRI* à notre cycle national des demandeurs d'emploi.



FIGURE 1.14 – *Comparaison du cycle des affaires français et du cycle des demandeurs d'emploi français - 1996-2018*



Note : Entre 1996 et 2018, le cycle classique de l'activité française dévoile 3 récessions (2002-2003, 2008-2009, 2011-2012). On aperçoit que le cycle des demandeurs d'emploi présente aussi des périodes d'expansion sur ces dates.

Un commentaire intéressant de cette comparaison (Tableau 1.14 consiste à mettre en avant la durée des phases d'expansion du cycle des demandeurs d'emploi comparativement aux phases de récessions du cycle de l'activité. On observe même qu'à partir de février 2008 les deux cycles rentrent respectivement dans une phase de récession et dans une phase d'expansion. Cependant, la récession, pour le cycle des affaires de l'activité économique, se termine en

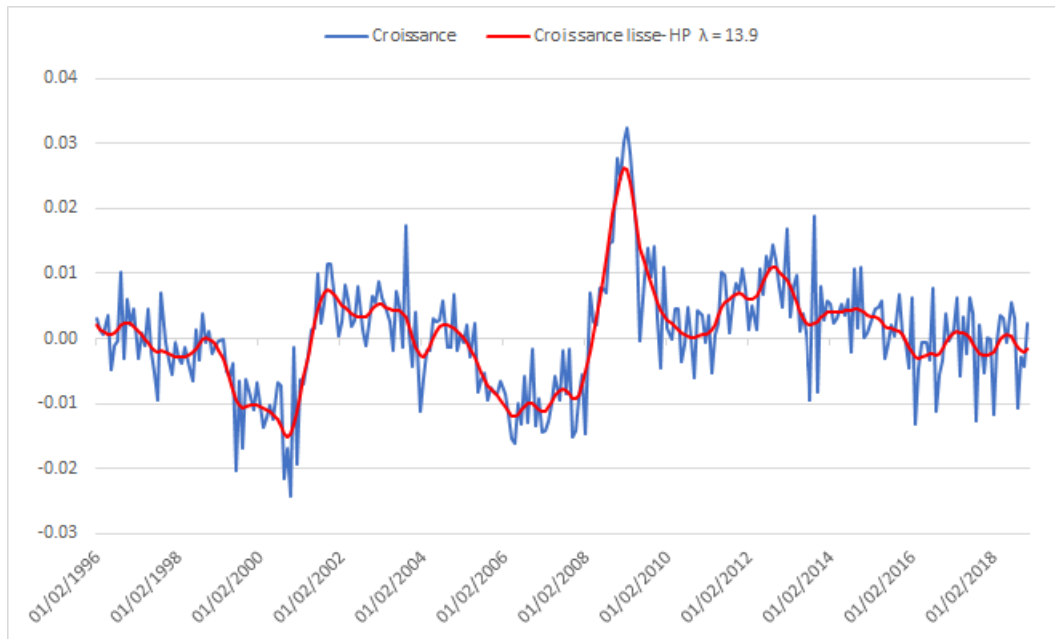
février 2009 alors qu’aucune phase de contraction n’est visible dans le cycle des demandeurs d’emploi. Il faudra attendre fin 2015 pour apercevoir une période de contraction. Durant cette période, le cycle classique de l’activité économique française connaît une autre récession entre 2011 et 2012 selon ECRI. Par conséquent et prenant en compte un fait bien connu dans la littérature de l’économie du travail, le chômage est persistant et connaît donc une certaine forme d’inertie. Il est donc probable que le nombre de demandeurs ne baisse pas suffisamment pour être enregistré comme une phase de contraction du cycle lorsque l’activité connaît une période d’expansion économique. L’arrivée de la nouvelle récession de 2011 vient à nouveau enrayer la baisse du chômage qui ne rencontrera son pic qu’en novembre 2015, soit 3 ans après la date de sortie de récession de la France, toujours selon ECRI. Au contraire, nous pouvons évoquer le décalage entre le début des phases de récession du cycle des affaires de l’économie et le début des phases d’expansion du cycle des demandeurs d’emploi. En effet, la récession des années 2002-2003 est visible dans le cycle classique des demandeurs d’emploi dès avril 2001. En revanche, la date de la crise de 2008 est identique dans les deux chronologies et la récession de 2011 n’est pas directement visible dans la datation du cycle des demandeurs d’emploi.

### 1.4.2 Le cycle d’accélération

Nous allons calculer le cycle d’accélération à partir du taux de croissance des séries CVS-CJO du nombre de demandeurs d’emploi en catégorie A fourni par Pôle Emploi. Un filtre Hodrick-Prescott avec un paramètre de lissage à faible calibrage, soit  $\lambda = 13.9$  est appliqué, ce qui correspond à supprimer les fluctuations ayant une fréquence supérieure à 12 mois. Comme nous l’avons déjà expliqué dans la méthodologie, ce calibrage permet d’imiter un filtre passe-bas et de supprimer les fluctuations de court de terme qui correspondent à du bruit dans la série et, par conséquent, gênent l’exercice de datation du cycle d’accélération. Par ailleurs, nous avons appliqué ce même filtre aux séries CVS en taux de croissance, obtenues par TRAMO-SEATS. L’idée est de dater le cycle d’accélération à partir des séries que nous avons estimées pour identifier de potentiels écarts de datation entre la série CVS de Pôle emploi et celle de TRAMO-SEATS. Nous avons noté que seule la chronologie du cycle d’accélération du département de la Loire-Atlantique (44) possède de légères différences, en fonction de la série CVS utilisée. Ces écarts peuvent s’expliquer au travers du grand nombre d’*outliers* qui sont détectés et corrigés par TRAMO. Nous présentons la comparaison des deux datations

en annexe.

FIGURE 1.15 – *Taux de croissance mensuel de la série des demandeurs d'emploi national et taux de croissance lissé - (1996-2018)*



La figure 1.15 représente la série nationale en taux de croissance superposée au lissage de la série par le filtre *low-pass* permettant d'améliorer la lecture des points de retournement au travers de l'algorithme Bry-Broschan et de ce fait, d'améliorer la précision de la chronologie. Les figures suivantes représentent l'ensemble des taux de croissances et des taux de croissances lissées de notre échantillon. Les lissages sont réalisés avec le logiciel R et le package *mFilter*. (Figure 1.16 à 1.21)

FIGURE 1.16 – *Taux de croissance mensuel de la série des demandeurs d'emploi des Pays de la Loire et taux de croissance lissé - (1996-2018)*

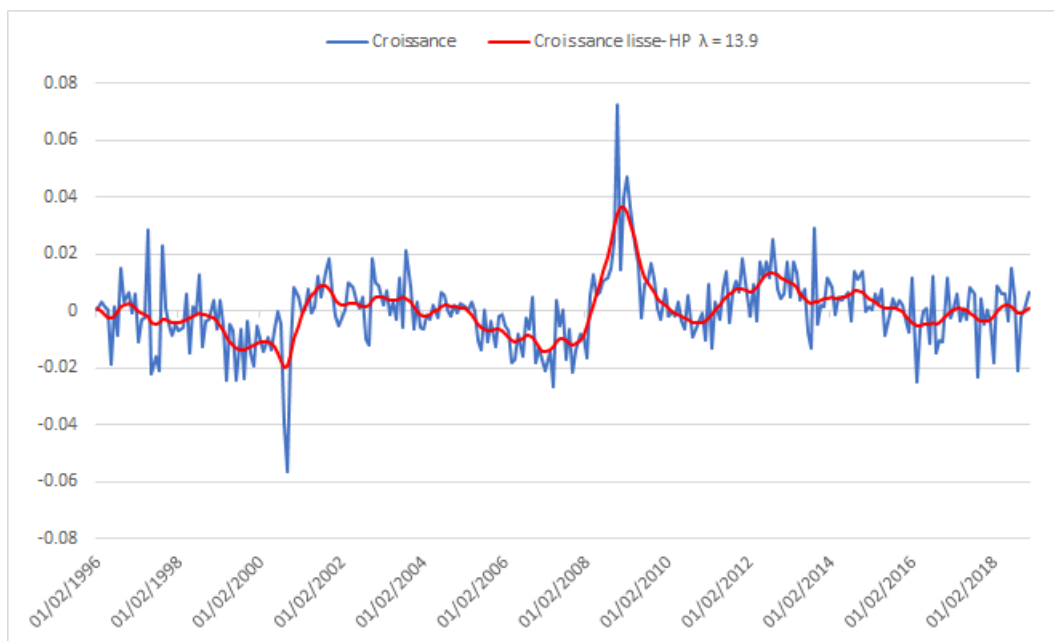


FIGURE 1.17 – *Taux de croissance mensuel de la série des demandeurs d'emploi de la Loire-Atlantique (44) et taux de croissance lissé - (1996-2018)*

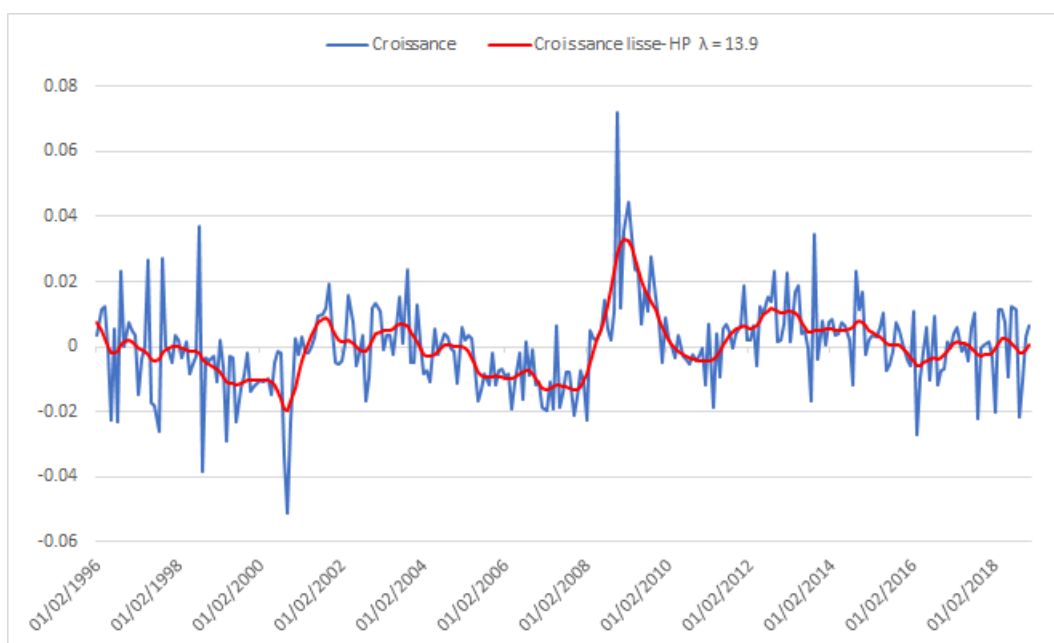


FIGURE 1.18 – *Taux de croissance mensuel de la série des demandeurs d'emploi du Maine et Loire (49) et taux de croissance lissé - (1996-2018)*

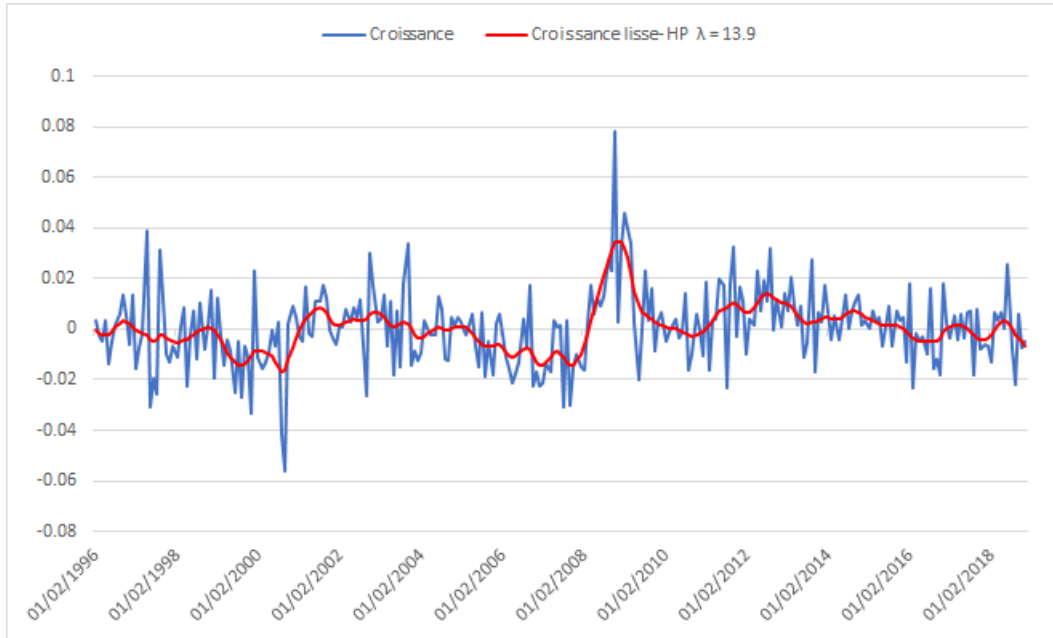


FIGURE 1.19 – *Taux de croissance mensuel de la série des demandeurs d'emploi de la Mayenne (53) et taux de croissance lissé - (1996-2018)*

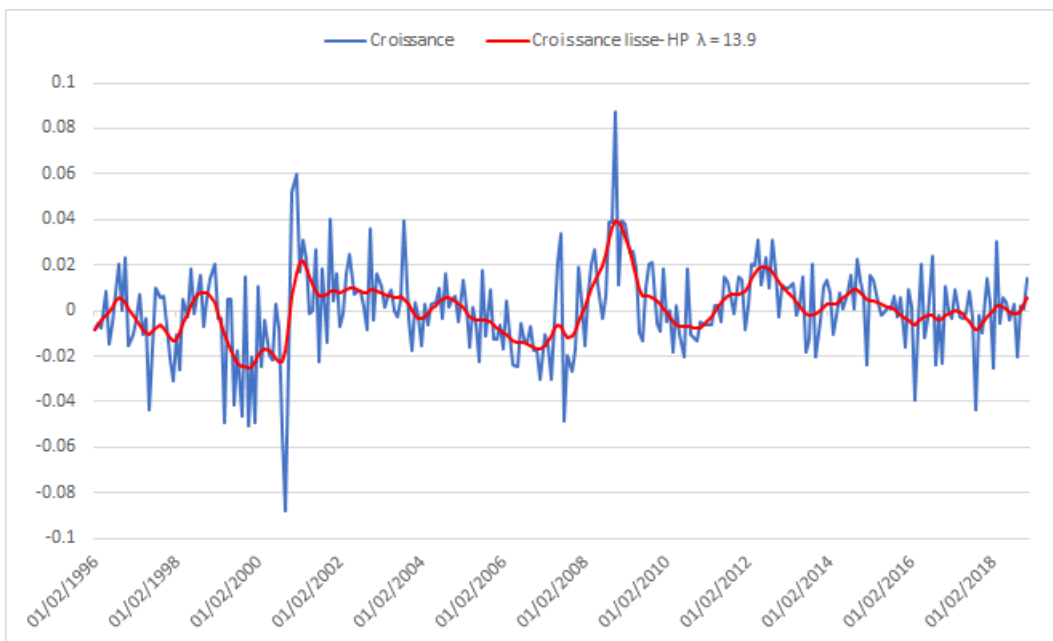


FIGURE 1.20 – Taux de croissance mensuel de la série des demandeurs d'emploi de la Sarthe (72) et taux de croissance lissé - (1996-2018)

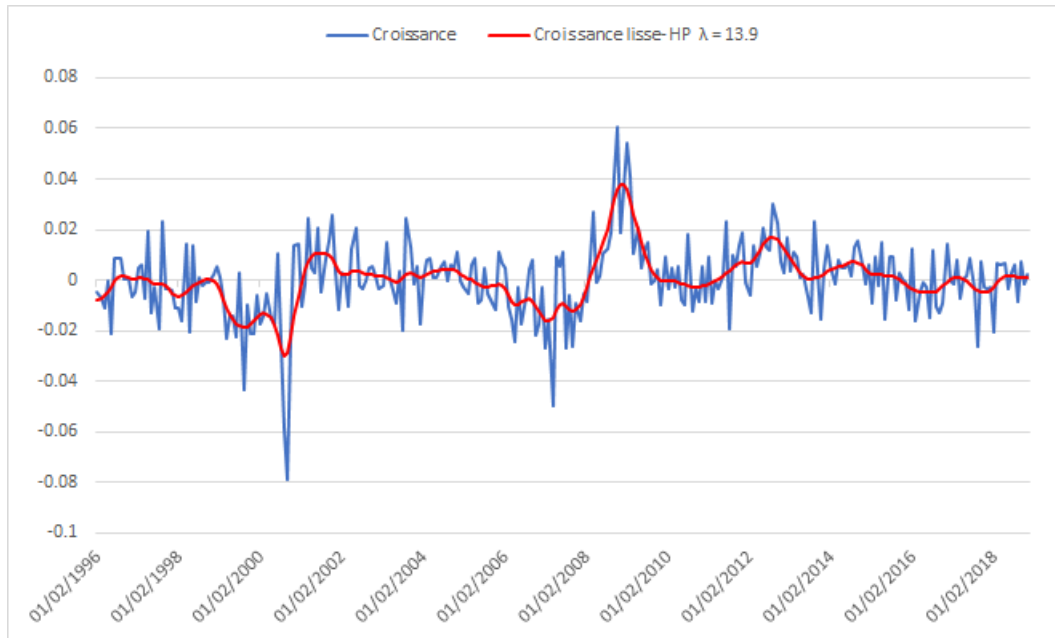
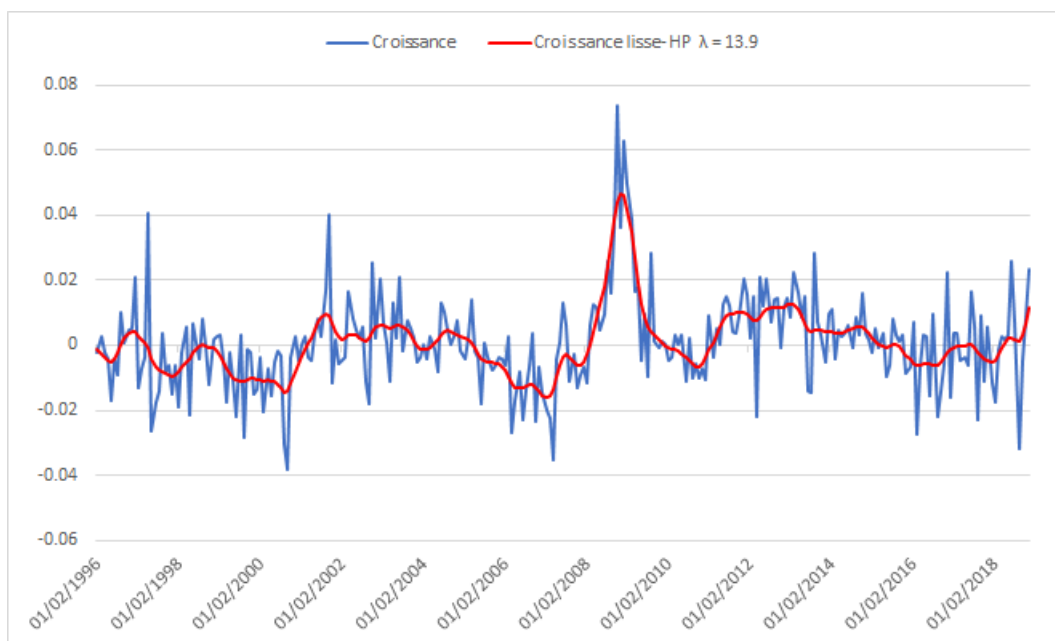


FIGURE 1.21 – Taux de croissance mensuel de la série des demandeurs d'emploi de la Vendée (85) et taux de croissance lissé - (1996-2018)



## Datation des points de retournement

La chronologie des points de retournement du cycle d'accélération est beaucoup plus importante en nombre de pics et de creux, car le taux de croissance lissé utilisé pour dater les points de retournement est plus fluctuant. (Tableau 1.11).

TABLE 1.11 – Datation du cycle d'accélération de l'emploi – 1996-2019

	HUR	Nationale	PDLL	44	49	53	72	85
Pic		1996M11	1996M9		1996M10	1996M1	1996M9	1997M1
Creux	1997M10	1998M3	1997M6	1999M4	1998M5	1998M1	1998M5	
Pic	1998M5	1998M10	1998M8	2000M4	1998M12	1998M9	1999M1	
Creux	2000M9	2000M11	2000M10	2000M10	2000M10	1999M11	2000M10	2000M10
Pic	2001M6	2001M10	2001M10	2001M10	2002M11	2001M1	2001M11	2001M10
Creux	2003M1	2004M1	2003M11	2002M9		2003M11	2001M8	
Pic	2003M9	2004M8	2004M7	2003M9		2002M11		
Creux	2005M1	2006M5				2003M11	2003M8	2003M12
Pic	2005M9				2005M5	2004M9	2004M12	2004M7
Creux	2007M10		2007M4	2008M2	2006M11	2007M4	2007M4	2007M4
Pic	2009M2	2009M2	2008M11	2008M11	2008M11	2008M11	2008M11	2008M11
Creux	2010M7	2010M10	2011M1	2011M3	2010M9	2010M11	2010M9	2010M7
Pic	2012M11	2012M9	2012M9	2012M9	2012M9	2012M5	2012M9	2013M12
Creux	2013M10	2013M8	2013M8	2013M8	2013M10	2013M10	2013M11	2013M12
Pic	2014M8	2014M11	2014M9	2014M9	2014M7	2014M10	2014M10	2014M11
Creux	2015M10	2016M3	2016M3	2016M3	2016M11	2016M3	2016M3	2016M3
Pic	2016M8	2017M3	2016M12	2017M3	2017M5	2016M12	2016M12	2017M7
Creux		2017M9	2017M9	2017M9	2017M11	2017M9	2017M9	2018M2

Source : Auteur, données Pôle Emploi.

La lecture et la compréhension des cycles d'accélération sont plus compliquées que celle des cycles classiques. Dans un premier temps, on remarque que les datations estimées sont cohérentes d'une zone géographique à une autre. Par exemple, la chronologie des points de retournement du cycle d'accélération des demandeurs d'emploi nationaux est proche de la

chronologie du taux de chômage harmonisé, ce qui nous laisse penser que notre estimation à partir de la mesure du nombre de demandeurs d'emploi de Pôle Emploi fonctionne bien.

De plus, on remarque que les points de retournement du cycle d'accélération apparaissent toujours de façon anticipée par rapport aux points de retournement du cycle classique. Précédemment, nous avons observé dans le cycle des demandeurs d'emploi trois grandes phases d'expansions qui correspondaient aux épisodes de crises économiques. Nous avons déjà noté le caractère précoce du cycle des demandeurs d'emploi sur le cycle des affaires de l'économie française. Nous constatons avec l'étude du cycle d'accélération des demandeurs d'emploi que les points de retournement correspondant à ces épisodes expansionnistes interviennent aussi de manière anticipée. Le contenu informatif du cycle d'accélération est développé ici, en étudiant le décalage moyen entre les pics du cycle classique et les pics du cycle d'accélération, de même pour les creux de ces cycles.

Selon ECRI, la France est en période de récession entre août 2002 et mai 2003, puis entre février 2008 et février 2009 ainsi qu'entre mars 2011 et novembre 2012. Il est donc très aisé de rapprocher ces dates de récession économique de l'activité avec nos périodes d'expansion du cycle des demandeurs d'emploi. Nous retiendrons quatre grandes périodes d'expansions du cycle classique des demandeurs d'emploi présenté dans les graphiques suivants.

TABLE 1.12 – Anticipation des points de retournement (creux) du cycle d'accélération de l'emploi par rapport au cycle de l'emploi

	HUR	Nationale	PDLL	44	49	53	72	85
2000-2001	3 mois	5 mois	1 mois	1 mois	5 mois	9 mois	1 mois	4 mois
2004	0 mois	3 mois	7 mois	non présent	non présent	8 mois	non présent	non présent
2008	4 mois	21 mois	10 mois	0 mois	15 mois	10 mois	10 mois	10 mois
2010-2011	15 mois	non présent	2 mois	2 mois	non présent	7 mois	8 mois	6 mois
Moyenne	5,5 mois	9,7 mois	5 mois	1 mois	10 mois	8,5 mois	6,3 mois	6,7 mois

Note : La mention "non présent" signifie que le cycle de l'emploi ne possède pas de point de retournement autour de cette date. Si l'on indique "0 mois", cela signifie que les points de retournement se produisent en même temps.  
Source : Auteur, données Pôle Emploi.

Les points de retournement du cycle d'accélération de la France pour la période d'expansion (creux) apparaissent, en moyenne, 9.7 mois avant les points de retournement du cycle classique des demandeurs d'emploi. On constate que certains départements possèdent des moyennes d'anticipation des creux plus élevées que d'autres. Par exemple, le département



de la Loire-Atlantique (44) enregistre des creux du cycle d'accélération 1 mois en moyenne avant le creux du cycle classique. En revanche, le département de la Mayenne (53) enregistre une moyenne de 8,5 mois sur l'anticipation des creux entre le cycle d'accélération et le cycle classique. Il est important de noter qu'anticiper les points de retournement du cycle classique permet d'obtenir une information sur le comportement à venir de la série. Ainsi, observer un point de retournement (creux) dans le cycle d'accélération d'une série est un signal sur le point de retournement dans le cycle classique des demandeurs d'emploi. (Tableau 1.12). De plus, nous pouvons proposer une période latente durant laquelle se produira le prochain creux, en se reportant aux moyennes calculées pour chaque zone géographique étudiée. Maintenant, nous pouvons réaliser le même exercice sur les sorties des phases d'expansion (pics). (Tableau 1.13).

TABLE 1.13 – Anticipation des points de retournement (pics) du cycle d'accélération de l'emploi par rapport au cycle de l'emploi

	HUR	Nationale	PDLL	44	49	53	72	85
2000-2001	2 mois	24 mois	25 mois	27 mois	11 mois	32 mois	non présent	non présent
2004	4 mois	7 mois	7 mois	non présent	non présent	6 mois	5 mois	9 mois
2008	10 mois	non présent	14 mois	18 mois	non présent	14 mois	18 mois	14 mois
2010-2011	9 mois	12 mois	14 mois	17 mois	19 mois	13 mois	12,3 mois	8,3 mois
Moyenne	6,3 mois	14,3 mois	15 mois	20,7 mois	15 mois	16,3 mois	6,3 mois	6,7 mois

Note : La mention "non présent" signifie que le cycle de l'emploi ne possède pas de point de retournement autour de cette date.  
Source : Auteur, données Pôle Emploi.

Nous remarquons que les points de retournement du cycle d'accélération se produisent, eux aussi, plus tôt que ceux du cycle classique. Néanmoins, les anticipations moyennes des pics sont beaucoup plus élevées que l'anticipation moyenne des creux. De ce fait, l'interprétation est beaucoup plus difficile, notamment en raison des longues période d'expansion du cycle classique qui effacent certains points de retournement (pics). Par conséquent, on observe, notamment dans les départements, des moyennes très élevées, comme en Loire-Atlantique (44) qui enregistre une moyenne d'anticipation des pics de 20,7 mois. Bien que la durée moyenne des anticipations des points de retournement soit grande, le caractère informatif sur la tendance à venir du cycle classique reste un élément à ne pas écarter.

## Caractéristiques des cycles d'accélération

Nous utilisons les mêmes mesures que précédemment pour évaluer les caractéristiques du cycle d'accélération. Ainsi, la durée, l'amplitude et l'indice d'*Excess* seront calculées ici. (Tableau 1.14).

TABLE 1.14 – Caractéristiques du cycle d'accélération de l'emploi – 1996-2019

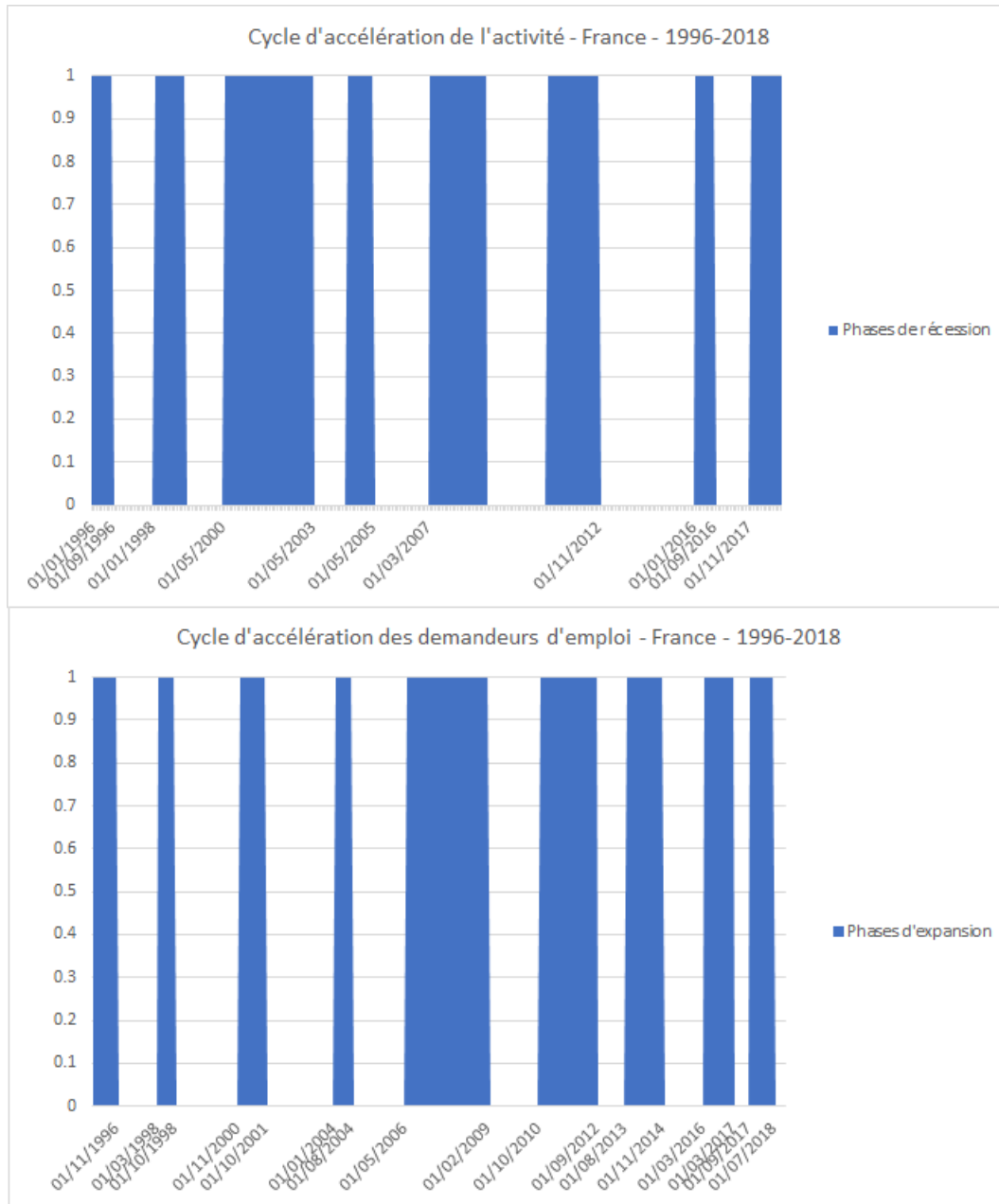
		HUR	Nationale	PDLL	44	49	53	72	85
<b>Expansion</b>	Durée	13	16,43	14,57	13,57	17,14	14,25	15,28	17,17
	Amplitude	0,033	0,023	0,046	0,048	0,049	0,063	0,053	0,051
	<i>Excess</i>	0,000	0,000	0,00	0,000	-0,001	-0,005	0,003	-0,004
<b>Contraction</b>	Durée	19,57	19	20,63	20	18,5	16,11	20	23,28
	Amplitude	0,035	0,023	0,042	0,047	0,046	0,063	0,051	0,049
	<i>Excess</i>	-0,001	-0,002	0,003	-0,002	0,000	0,003	-0,004	0,000

Note : La durée moyenne est en mois, l'amplitude moyenne est en millier de demandeurs d'emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L'*Excess* n'a pas d'unité, seul le signe est commenté ici. L'amplitude de l'HUR est en point de pourcentage.

On remarque que le cycle d'accélération est plus symétrique et linéaire que le cycle classique, notamment en matière d'amplitude. En effet, on observe moins d'écart entre l'amplitude des phases d'accélération (*i.e.* expansion) et celles de décélération (*i.e.* contraction). Cependant, les phases de décélération sont en moyenne plus longues que celles d'accélération. Ainsi, le taux de croissance du nombre de demandeurs d'emploi connaît davantage de baisses consécutives que de hausses. La différence d'asymétrie présente entre le cycle d'accélération et le cycle classique permettent surtout de comprendre que l'amplitude des phases d'accélération à un impact important sur les phases d'expansion du cycle classique et par conséquent sur la tendance de la série. En d'autres termes, bien que la durée moyenne des phases d'accélération soit plus courte que celle des phases de décélération, on constate que la durée moyenne des phases d'expansion du cycle classique, qui les suivent est plus longue et persistante que les phases de contraction.

Par ailleurs, l'indice de l'*Excess* est variable d'une zone géographique à une autre, mais aucune différence majeure n'est à noter. En outre, les départements du Maine-et-Loire (49), de la Mayenne (53) et de la Vendée (85) connaissent des trajectoires assez brutales dans les phases d'accélération. À l'image de la section sur le cycle classique, nous allons comparer notre chronologie des points de retournement du cycle d'accélération des demandeurs d'emploi avec le cycle d'accélération de l'activité française calculé par l'ECRI.(Figure 1.22).

FIGURE 1.22 – Comparaison des cycles d'accélération de l'activité et des demandeurs d'emploi - France - 1996-2018



La littérature de l'analyse cyclique de l'activité a mis en avant l'aspect informatif du cycle d'accélération sur le cycle classique puisque ses points de retournement se produisent avant ceux du cycle classique. Nous remarquons graphiquement que les points de retournement du cycle d'accélération des demandeurs d'emploi se produisent avant, pendant ou après ceux du cycle d'accélération de l'activité économique française. Par conséquent, il est difficile d'identifier un comportement précoce du cycle d'accélération des demandeurs par rapport au cycle d'accélération de l'activité économique. Cependant, on peut observer qu'après la

période de la crise économique de 2008, le cycle d'accélération des demandeurs d'emploi est un signal sur le cycle classique de l'activité, excepté pour la phase de décélération de mars 2016.

### 1.4.3 Synchronisation cyclique

#### Indice de concordance et coefficient de corrélation

Il est important d'étudier le degré de synchronisation des cycles entre deux séries afin d'apporter une complète connaissance de la relation entre les cycles des demandeurs d'emploi. (Tableau 1.15). Notre objectif est de vérifier si les cycles des demandeurs d'emploi sont intégrés entre eux en fonction des différents niveaux géographiques étudiés. La mesure du coefficient de corrélation de Pearson viendra compléter cette analyse.

TABLE 1.15 – Indices de concordances et coefficients de corrélation – Cycle de l'emploi

		Indice de concordance						
		Nationale	PDLL	44	49	53	72	85
<b>Coefficient de corrélation</b>	Nationale	1	0,78	0,75	0,85	0,78	0,78	0,76
	PDLL	0,99	1	0,92	0,84	0,91	0,95	0,91
	44	0,97	0,97	1	0,89	0,86	0,93	0,86
	49	0,99	0,99	0,95	1	0,82	0,83	0,81
	53	0,94	0,96	0,87	0,96	1	0,89	0,85
	72	0,97	0,99	0,94	0,98	0,96	1	0,92
	85	0,92	0,94	0,83	0,95	0,97	0,94	1

Note : Le coefficient de corrélation correspond à la partie inférieure gauche et l'indice de concordance à la partie supérieure droite.

Les indices de concordance entre le nombre de demandeurs d'emploi au niveau national, régional et départemental sont élevés. Le cycle régional semble synchronisé à 78% avec le cycle national, ce qui met en lumière, d'un autre côté, une légère désynchronisation. Les départements possèdent des indices de concordances similaires. Par ailleurs, les départements de la région Pays de la Loire sont davantage synchronisés avec la région, la valeur de l'indice le plus faible est de 0.84 entre le département du Maine et Loire (49) et la région. Ces constats évoquent une bonne intégration des différents départements de la région entre eux mais aussi avec la région elle-même. De plus, la région Pays de la Loire semble posséder un cycle du nombre des demandeurs d'emploi qui évolue dans le même sens que le cycle national, sans être pour autant parfaitement synchrone. La région Pays de la Loire possède ses propres caractéristiques lui permettant d'agir sur son nombre de demandeurs d'emploi. L'attractivité de la région en matière d'activité et d'emploi peut ainsi expliquer en partie les phases de contraction qui viendraient perturber la "parfaite" synchronisation des cycles. La mesure

du coefficient de corrélation, bien que moins précise, permet de confirmer le sentiment de synchronisation des cycles classiques.

TABLE 1.16 – Indices de concordances et coefficients de corrélation – Cycle d'accélération de l'emploi

		Indice de concordance						
		Nationale	PDLL	44	49	53	72	85
Coefficient de corrélation	Nationale	1	0,81	0,74	0,79	0,71	0,78	0,85
	PDLL	0,96	1	0,70	0,76	0,82	0,86	0,78
	44	0,95	0,97	1	0,64	0,67	0,65	0,69
	49	0,95	0,98	0,93	1	0,77	0,78	0,73
	53	0,83	0,89	0,81	0,89	1	0,83	0,70
	72	0,92	0,97	0,91	0,95	0,90	1	0,78
	85	0,92	0,96	0,90	0,95	0,87	0,92	1

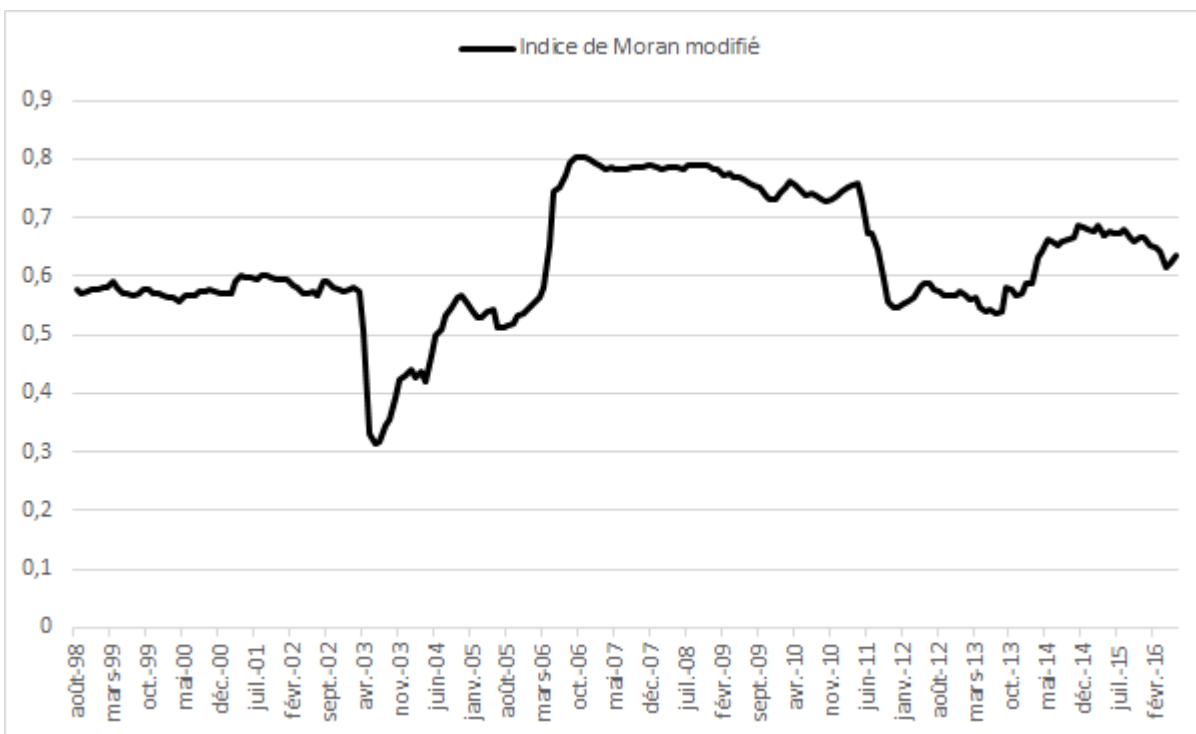
Note : Le coefficient de corrélation correspond à la partie inférieur gauche et l'indice de concordance à la partie supérieur droite.

Le cycle d'accélération des demandeurs d'emploi de la France et de la région Pays de la Loire est synchronisé à 81%. De manière générale, les départements de la région Pays de la Loire sont aussi synchronisés avec le cycle d'accélération national. Cependant, l'indice de concordance entre les cycles d'accélération montrent des départements moins synchrones avec la région, 0.70 pour la plus petite valeur entre la Loire-Atlantique et la région. Par ailleurs, la Loire-Atlantique apparaît comme le département le moins synchronisé avec les autres départements. Quant aux coefficients de corrélation calculés sur les taux de croissance lissés, nous pouvons affirmer qu'ils sont tous très élevés. (Tableau 1.16).

### Indice de Moran modifié

À présent, concentrons-nous sur les cinq départements de la région Pays de la Loire afin de mesurer le degré de synchronisation des cycles de l'emploi départementaux. Pour ce faire, nous utilisons une mesure de l'indice de Moran modifié. Cet indice de Moran modifié capture les co-mouvements des séries à travers le temps. L'indice de Moran modifié (Figure 1.23) est calculé à partir de la corrélation des taux de croissance en utilisant une fenêtre glissante de 61 mois. En effet, Stock et Watson (2008) utilisent une fenêtre de 21 trimestres pour des données trimestrielles et Bandrés, Gadea Rivas et Gómez (2017) utilisent une fenêtre de 5 années pour des données annuelles.

FIGURE 1.23 – Indice de Moran modifié – k=61



Source : Auteur, données Pôle Emploi.

Note : "k" correspond à la taille de la fenêtre roulante utilisée pour calculer l'indice.

La synchronisation des co-mouvements entre départements est en moyenne de 0,62, et le niveau de l'indice reste fluctuant sur l'ensemble de la période. On observe de nombreuses pertes de synchronisation autour des points de retournement du cycle de l'emploi, reflétant notamment les écarts de datation. C'est ainsi que nous pouvons émettre l'hypothèse qu'autour des points de retournement (creux), évocateurs d'une récession économique et donc d'une augmentation du nombre des demandeurs d'emploi, une perte de synchronisation s'en suit quasiment systématiquement. En effet, chaque département va, dans un premier temps, réagir différemment aux effets d'une récession économique en s'appuyant sur ses propres forces ou en subissant ses faiblesses, avant que les mesures locales d'emploi viennent harmoniser cette synchronisation. Par ailleurs, nous pouvons remarquer qu'après la crise de 2008, l'indice moyen est plus élevé que la période précédant la crise économique. On mesure une moyenne de 0,58 pour la période allant de 1997 à 2008, puis une moyenne de 0,67 pour la période entre 2009 et 2017<sup>21</sup>. Une raison probable de cette meilleure synchronisation départementale serait la fusion de l'ANPE et des Assedic le 20 décembre 2008 pour devenir Pôle Emploi. L'intégration des conseillers indemnisation et des conseillers placement des demandeurs d'emploi aurait

21. La fenêtre glissante coupe 10 observations en début et en fin de série.

permis aux régions de mieux s'organiser autour des politiques d'emploi. Il est important de noter que l'augmentation moyenne du niveau de synchronisation avant et après 2008 peut être aussi expliquée par des effets spatiaux intrinsèques aux départements étudiés. On peut conclure que la tendance de l'indice de Moran modifié semble à la hausse sur les dernières années de l'échantillon, même si la fin des données est tronquée en raison de la taille de la fenêtre roulante.

## 1.5 Conclusion

Cet article tente de proposer une chronologie de référence des cycles de l'emploi entre 1996 et 2019 de la région Pays de la Loire et de ses départements. Le cycle étudié est le cycle classique au sens de Mitchell et Burns (1946), calculé à partir de nos estimations de la composante Tendence-Cycle de la série brute du nombre de demandeurs d'emploi en catégorie A collectée par Pôle Emploi. L'algorithme appliqué pour la datation des points de retournement est celui de Bry et Boschan (1971). Une brève revue de la littérature a permis de démontrer que l'algorithme fonctionnait très bien pour établir une chronologie à partir d'un modèle univarié. Notre analyse empirique confirme les conclusions théoriques de Juglar (1862) et de Schumpeter (1939). En effet, ces deux auteurs s'accordent sur le fait que les périodes de récession économique sont relativement de courtes durées, environ 1 à 2 ans pour les "embarras commerciaux". Cependant, la théorie développée par Juglar ne considère pas les effets collatéraux d'une récession voire d'une crise sur le chômage. Or, notre comparaison entre le cycle du PIB réel et le cycle de l'emploi régional et départemental permet de mettre en lumière un cycle de l'emploi caractérisé par une forte inertie, notamment en période d'expansion. Cela renforce donc notre conviction d'étudier le chômage comme une variable à part entière. L'analyse de la chronologie des points de retournement du cycle de l'emploi permet d'identifier trois périodes de contraction majeures qui correspondent aux récessions économiques connues. Une fois la chronologie des points de retournement établie, nous avons étudié les caractéristiques des cycles afin d'en déduire plusieurs faits stylisés. Le premier constat fait écho aux résultats avancés dans la littérature, le cycle de l'emploi est asymétrique et non linéaire. L'étude des caractéristiques de durée, d'amplitude et d'*Excess* met en lumière des phases d'expansion plus longues et plus amples que les phases de contraction. L'indice d'*Excess* permet d'estimer les trajectoires des phases d'expansion et de contraction. Ainsi, nous avons remarqué que les cycles de l'emploi, à tous les niveaux géographiques étudiés, connaissent des périodes d'expansion et de contraction progressives tous au long de la phase. Pour terminer, nous nous sommes intéressés à la synchronisation des cycles de l'emploi. L'estimation de l'indice de concordance permet de mesurer l'intégration des cycles entre eux. Il en ressort donc que les départements de la région Pays de la Loire sont synchrones entre eux, pour un minimum de 81 %, et on observe aussi, que les cycles de l'emploi des départements s'inscrivent, pour un minimum de 84 %, dans les mêmes phases d'expansion ou



de contraction que le cycle régional. Par ailleurs, nous avons utilisé l'indice de Moran modifié des cinq départements de la région. Nos analyses ont décelé une augmentation, en tendance, de l'indice moyen, notamment après la crise économique de 2008. La raison probable de cette augmentation est la fusion de l'ANPE et des Assedic, suivie d'un renforcement du pouvoir des régions en matière de formation et d'orientation professionnelle. Nous avons aussi remarqué que les pertes de synchronisation étaient localisées autour des points de retournements du cycle de l'emploi. L'écart de datation des points de retournement entre les cycles de l'emploi départementaux est, selon nous, une cause majeure de la perte de synchronisation des mouvements des séries des demandeurs d'emploi. Ceci étant, la cause des écarts de datation pourrait prendre la forme de composantes intrinsèques à chaque département qui viendraient différencier les territoires, notamment en termes de réponses aux problématiques d'emploi. Lors de futures recherches, du point de vue méthodologique, nous pourrions être amenés à confirmer nos résultats en utilisant l'approche paramétrique de datation des points de retournement, avec les modèles à changement de régime markovien. Par ailleurs, un indicateur de retournement probabiliste peut aussi faire suite au travail déjà réalisé dans cet article. De plus, la détermination de la (dé)synchronisation des départements de la région Pays de la Loire au travers d'une analyse plus fine du territoire peut aussi être très informative. Du point de vue empirique, une extension pourrait se porter également sur une comparaison entre les différentes régions françaises, et également, étudier si la recomposition récente des régions a eu une incidence en termes de demandeurs d'emploi. Pour conclure, nous envisageons d'étudier en détail la prévision de la série mensuelle du nombre des demandeurs d'emploi de Pôle Emploi, ainsi que la série du trimestriel du chômage de l'Insee.

## 1.6 Annexes

FIGURE 1.24 – *Processus de l’algorithme de Bry-Boschan*

<b>Programme de sélection des points de retournement</b>	
I.	Détermination des valeurs extrêmes et de substitutions.
II.	Détermination des cycles avec une moyenne mobile de 12 mois (valeurs extrêmes évincées). A. Identification des points supérieurs (inférieurs) à 5 mois de chaque côté. B. Application de l’alternance des points de retournement face à de multiples points supérieurs ou inférieurs. Le plus élevé (bas) est choisi.
III.	Détermination des points de retournement dans la courbe de Spencer (valeurs extrêmes évincées). A. Identification de la plus grande (plus petite) valeur entre plus ou moins 5 mois du point de retournement sélectionné dans une moyenne mobile de 12 mois. B. Application de la durée minimum de 15 mois pour un cycle afin d’éliminer les cycles trop courts.
IV.	Détermination des points de retournement correspondant dans une moyenne mobile de court terme d’ordre 3 à 6 mois, dépendant de MCD ( <i>months of cyclical dominance</i> ). A. Identification de la valeur la plus élevée (faible) entre plus ou moins 5 mois du point de retournement sélectionné dans la courbe de Spencer.
V.	Détermination des points de retournement dans les séries non lissées. A. Identification de la valeur la plus élevée (faible) entre plus ou moins 4 mois, ou le terme MCD, n’importe lequel est plus grand, du point de retournement dans la moyenne mobile de court terme. B. Elimination des points de retournement dans les 6 mois au début et à la fin de la série. C. Elimination des pics (creux) à la fin de la série qui sont trop faible (élevé) que les valeurs proches de la fin. D. Elimination des cycles qui ont une durée inférieure à 15 mois. E. Elimination des phases qui ont une durée inférieure à 5 mois.
VI.	Déclaration des points de retournement finaux.

La figure 29 reprend la procédure de datation des points de retournement de l’algorithme de Bry et Boschan (1971).

FIGURE 1.25 – Série demandeurs d'emploi Pays de la Loire catégorie A-Brute, spectre auto-régressif et périodogramme

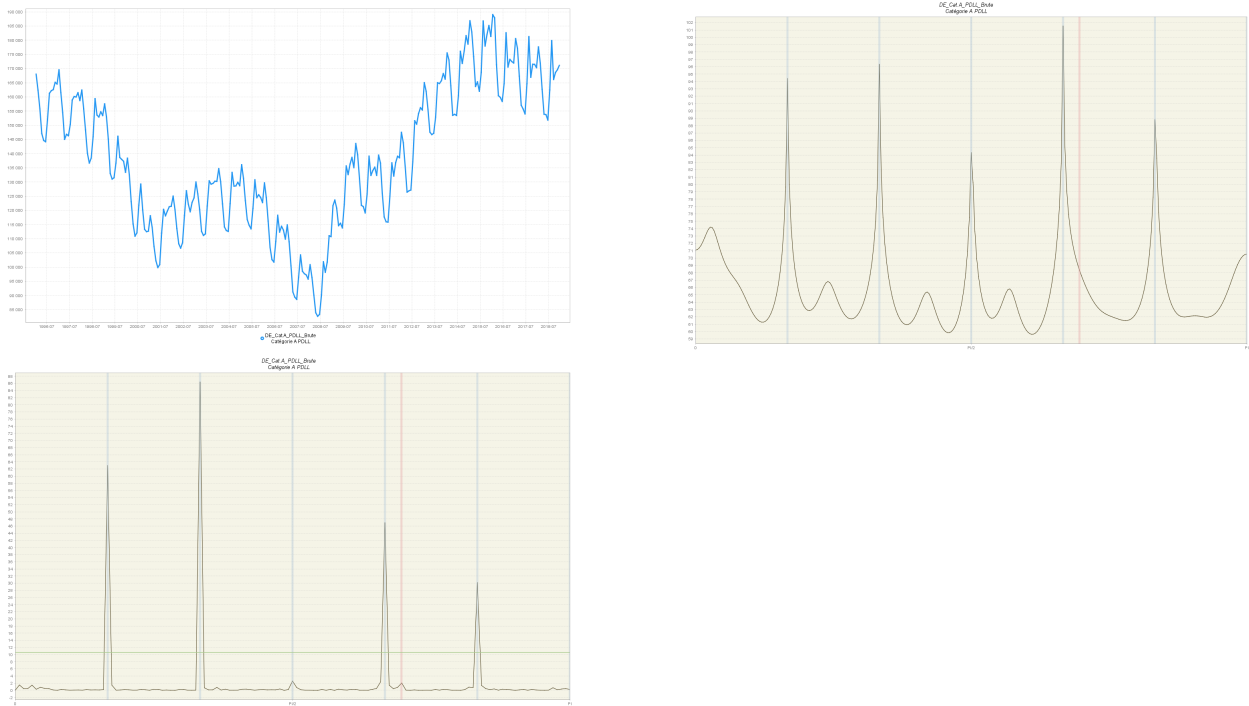


FIGURE 1.26 – Série demandeurs d'emploi Loire-Atlantique (44) catégorie A-Brute, spectre auto-régressif et périodogramme

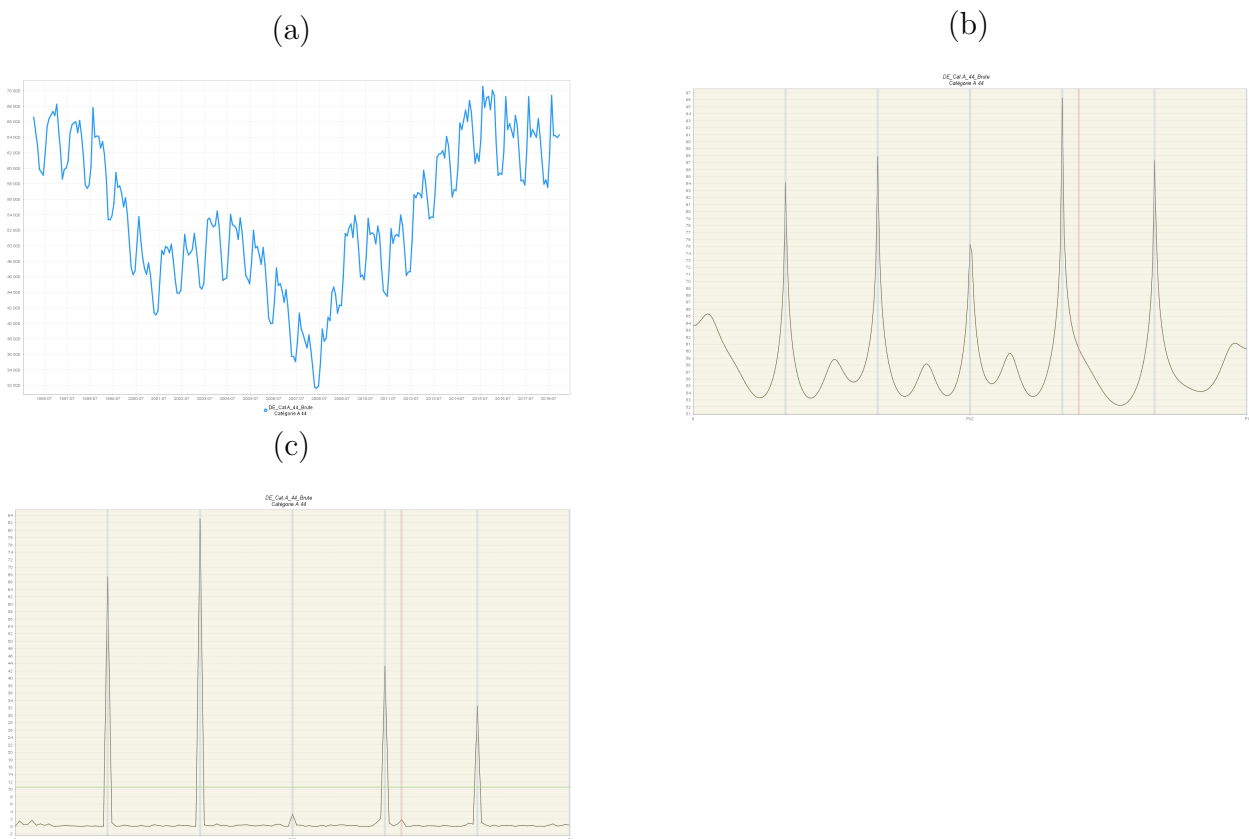


FIGURE 1.27 – Série demandeurs d'emploi Maine et Loire (49) catégorie A-Brute, spectre auto-régressif et périodogramme

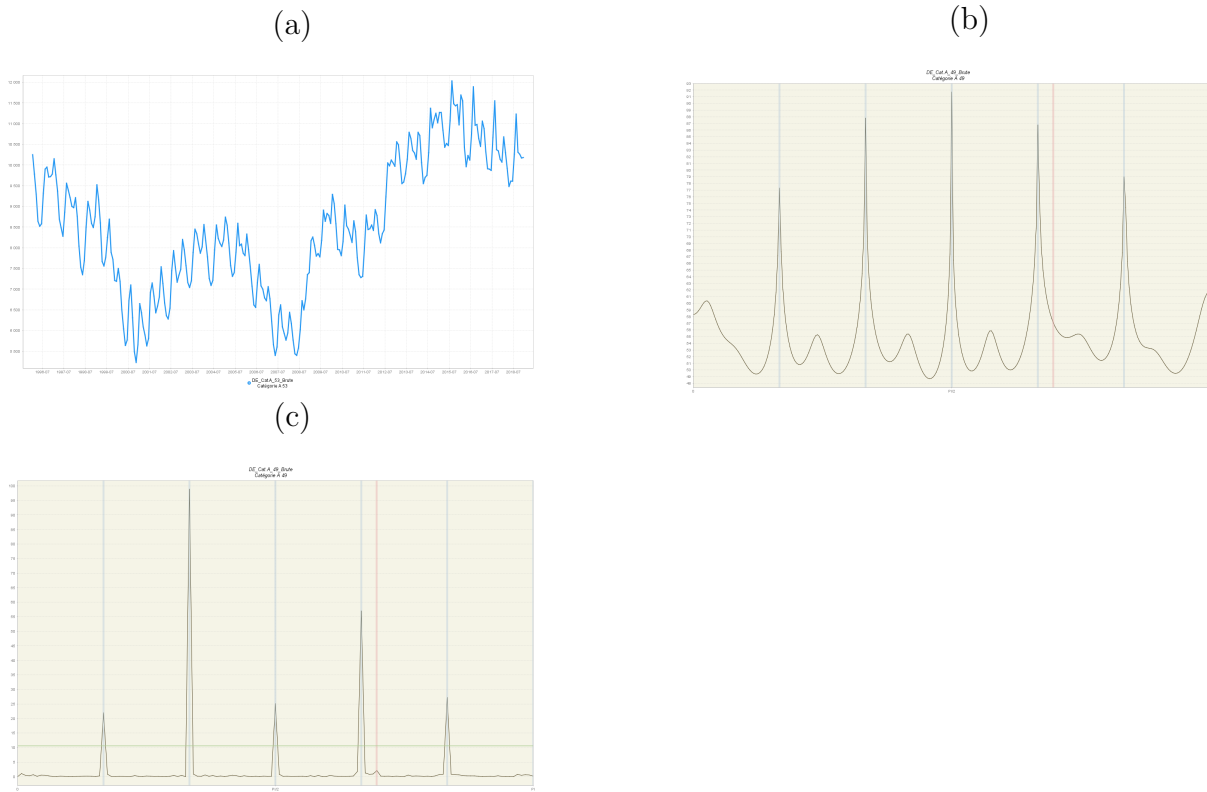


FIGURE 1.28 – Série demandeurs d'emploi Mayenne (53) catégorie A-Brute, spectre auto-régressif et périodogramme

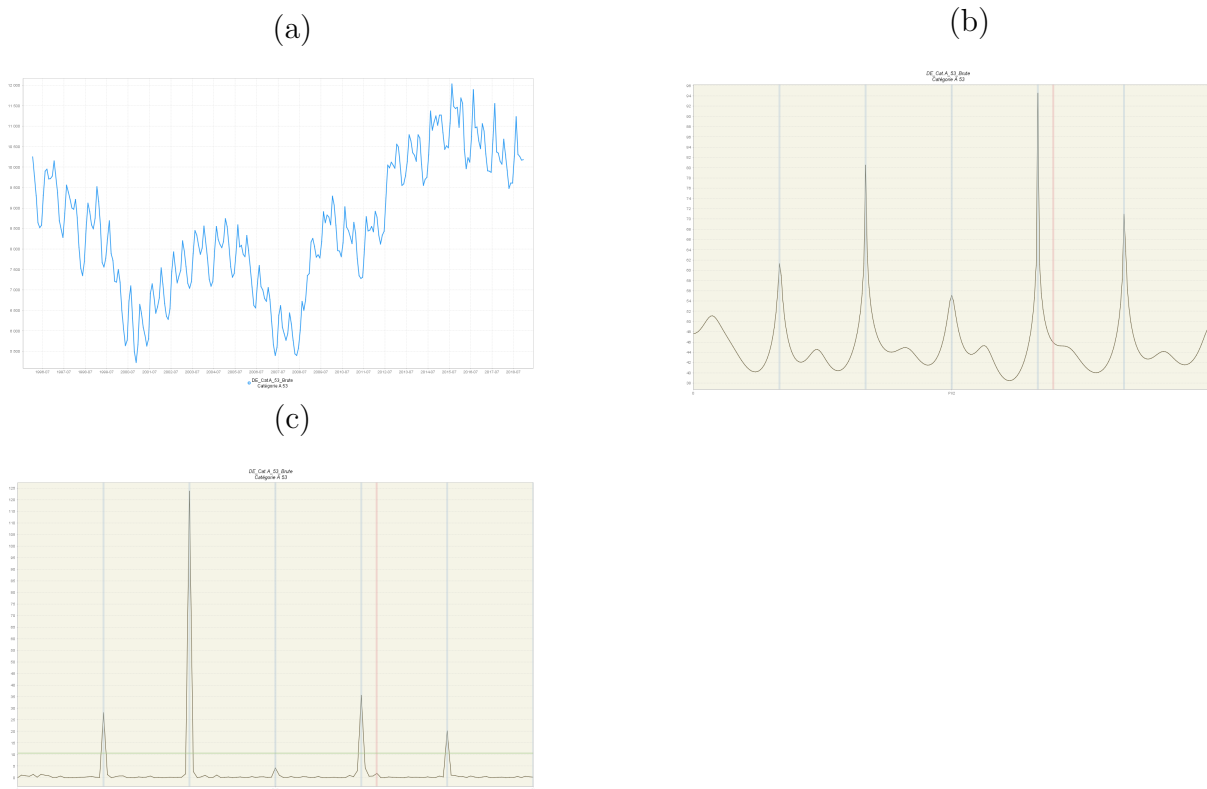


FIGURE 1.29 – Série demandeurs d'emploi Sarthe (72) catégorie A-Brute, spectre auto-régressif et périodogramme

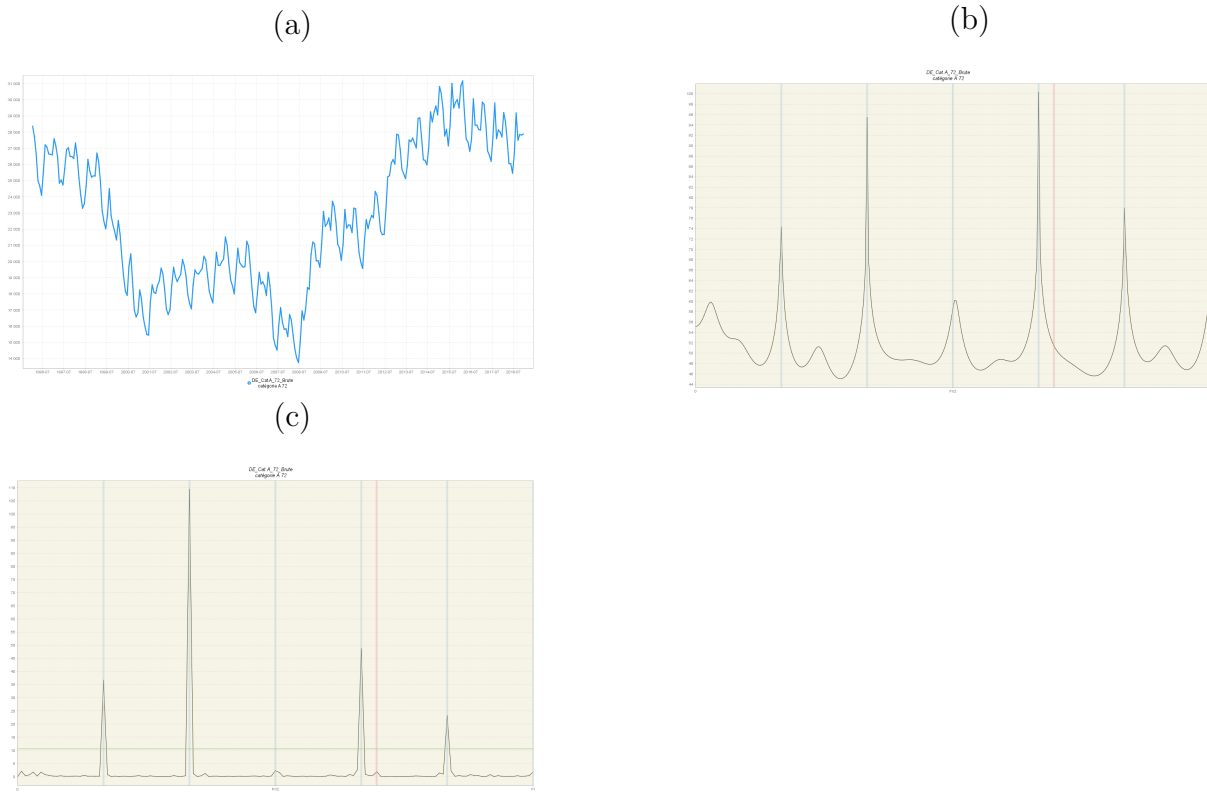


FIGURE 1.30 – Série demandeurs d'emploi Vendée (85) catégorie A-Brute, spectre auto-régressif et périodogramme

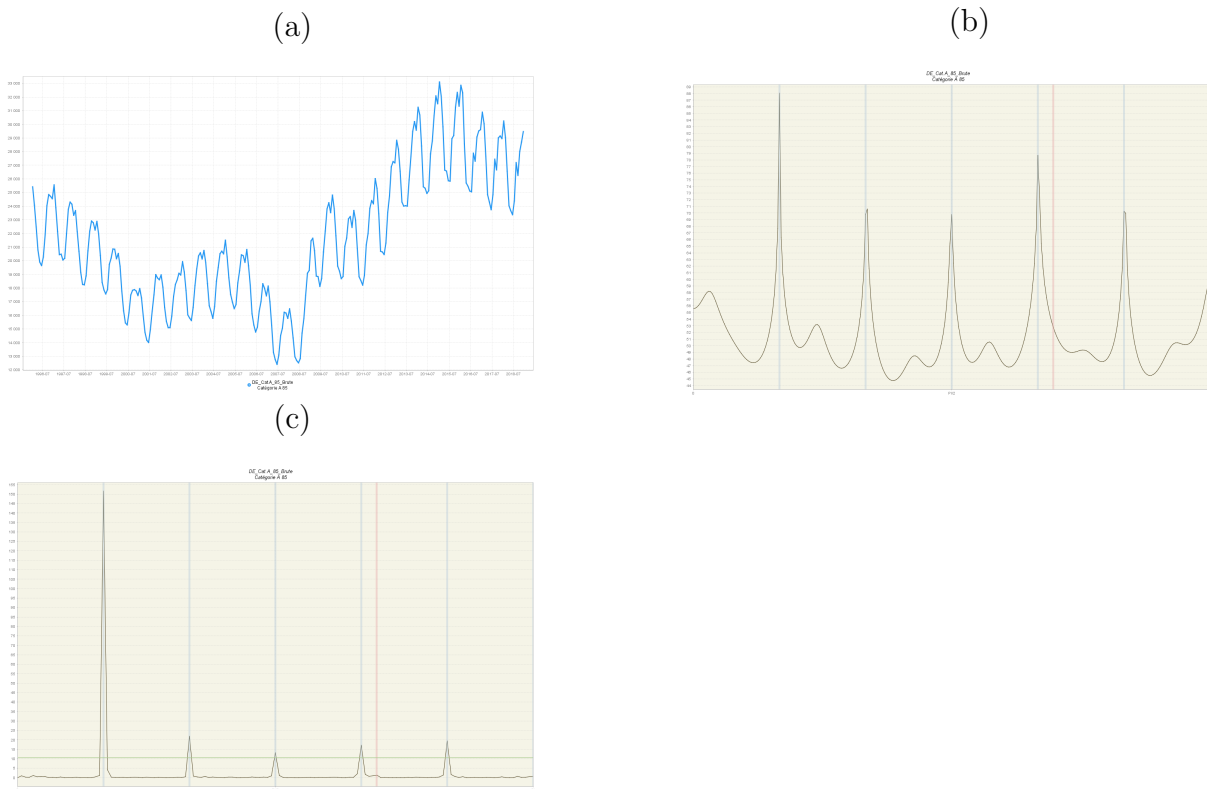


TABLE 1.17 – Datation des cycles d’emploi régionaux – 1996-2019

	Loire-Atlantique – PE	Loire-Atlantique – TS
Creux	1999M4	1997M8
Pic	2000M4	
Creux	2000M10	
Pic	2001M10	2001M10
Creux	2002M9	
Pic	2003M9	
Creux		2004M4
Pic		
Creux	2008M2	
Pic	2008M11	2008M7
Creux	2001M3	2011M3
Pic	2012M9	2011M12
Creux	2013M8	
Pic	2014M9	
Creux	2016M3	2016M3
Pic	2017M3	
Creux	2017M9	

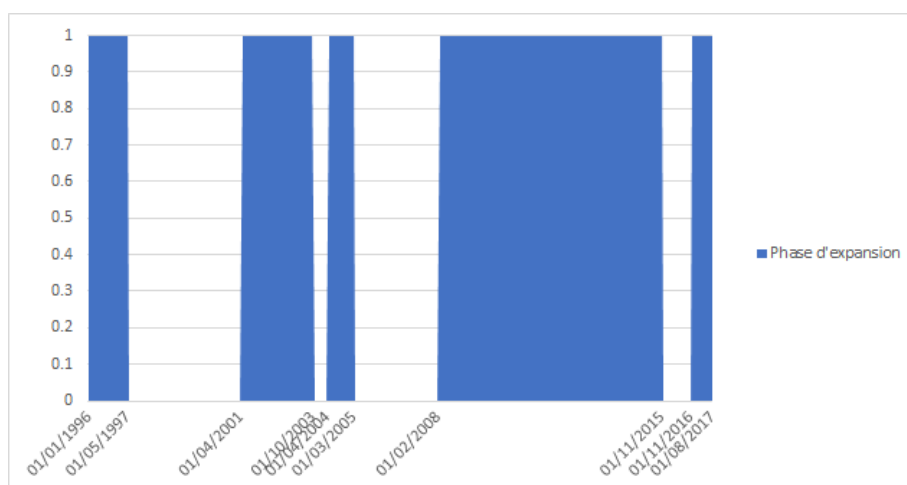
Source : Auteur, données Pôle Emploi.

Les figures de 1.25 à 1.30 représentent la série brute étudiée (a), le spectre autorégressif(b) et le périodogramme (c). Si le spectre autorégressif ou le périodogramme présente des pics au niveau des bandes grises alors la série est sujette à une forme de saisonnalité.

Le tableau 2.2 compare les datations des points de retournement du département de la Loire-Atlantique (44) en fonction de la Tendence-Cycle utilisée. La colonne "PE" fait référence à la TC de la série CVS de Pôle emploi qui est utilisé dans l’exercice de datation *via* l’algorithme de Bry-Boschan, alors que la colonne "TS" fait référence à la TC de la série CVS que nous avons calculé *via* TRAMO SEATS. Nous avons renseigné que le département de la Loire-Atlantique (44) car nous avons observé aucun changement de datation dans les autres séries avec les deux TC comparées.

Les figures et tableaux restant reprennent un à un la chronologie des points de retournement du cycle classique des demandeurs d’emploi à tous les niveaux géographiques étudiés. (Figures 1.18 à 1.37 et tableaux 1.19 à 1.31).

FIGURE 1.31 – *Datation des cycles des demandeurs d'emploi - série nationale*



Lecture : Les phases d'expansions correspondent à une augmentation du nombre de demandeurs d'emploi.

TABLE 1.18 – *Datation des points de retournement du cycle des demandeurs d'emploi - série nationale*

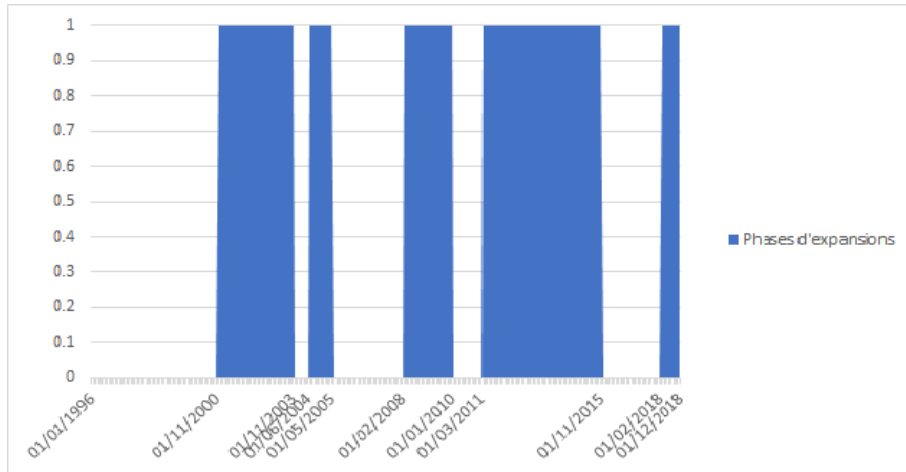
Phase cycle	Début	Fin	durée en mois	Amplitude
Expansion	NA	1997M5	NA	NA
Contraction	1997M5	2001M4	47	838.6
Expansion	2001M4	2003M10	30	362.5
Contraction	2003M10	2004M4	6	57.6
Expansion	2004M4	2005M3	11	50.2
Contraction	2005M3	2008M2	35	736.6
Expansion	2008M2	2015M11	93	1604.8
Contraction	2015M11	2016M11	12	118
Expansion	2016M11	2017M8	9	56.2
Contraction	2017M8	NA	NA	NA

TABLE 1.19 – *Caractéristiques du cycle classique des demandeurs d'emploi - série nationale*

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	35.75	518.43	9266.85
Contraction	25.00	437.70	5471.25

Lecture : la durée moyenne est en mois, l'amplitude moyenne et la sévérité sont en milliers.

FIGURE 1.32 – *Datation des cycles des demandeurs d’emploi - série Pays de la Loire*



Lecture : Les phases en bleu correspondent aux phases d’expansion du cycle des demandeurs d’emploi. En d’autres termes, cela correspond à une augmentation du nombre de demandeurs d’emploi.

TABLE 1.20 – *Datation des points de retournement du cycle des demandeurs d’emploi - série Pays de la Loire*

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	2000M11	NA	NA
Expansion	2000M11	2003M11	36	17030
Contraction	2003M11	2004M6	7	1720
Expansion	2004M6	2005M5	11	2450
Contraction	2005M5	2008M2	33	36100
Expansion	2008M2	2010M1	23	42580
Contraction	2010M1	2011M3	14	5150
Expansion	2011M3	2015M11	56	50220
Contraction	2015M11	2018M2	27	15160
Expansion	2018M2	NA	NA	NA

TABLE 1.21 – *Caractéristiques du cycle classique des demandeurs d’emploi - série Pays de la Loire*

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	31.50	28070.00	442102.50
Contraction	20.25	14532.50	147141.56



FIGURE 1.33 – Datation des cycles des demandeurs d’emploi - série Loire-Atlantique 44

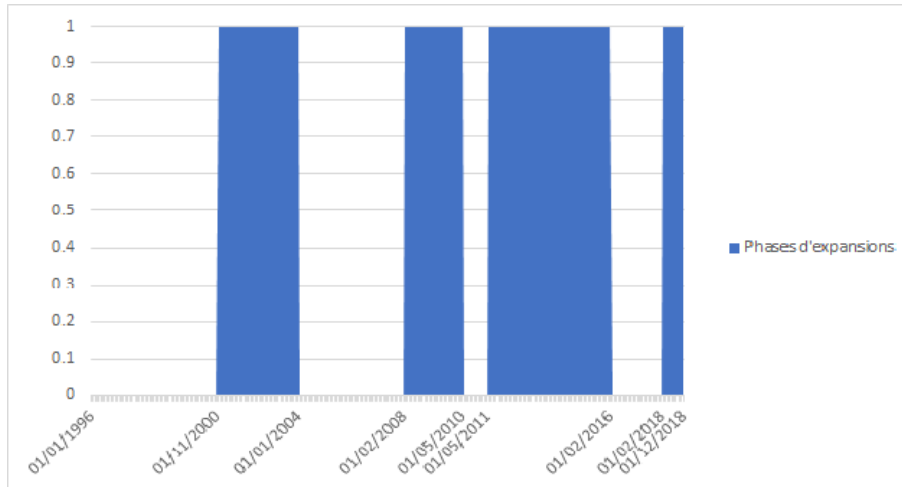


TABLE 1.22 – Datation des points de retournement du cycle des demandeurs d’emploi - série Loire-Atlantique 44

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	2000M11	NA	NA
Expansion	2000M11	2004M1	38	6120
Contraction	2004M1	2008M2	49	16010
Expansion	2008M2	2010M5	27	15810
Contraction	2010M5	2011M5	12	2440
Expansion	2011M5	2016M2	57	18120
Contraction	2016M2	2018M2	24	3400
Expansion	2018M2	NA	NA	NA

La datation des points de retournement de la série de la Loire-Atlantique est similaire mais pas identique à celle de la région Pays de la Loire.

TABLE 1.23 – Caractéristiques du cycle classique des demandeurs d’emploi - série Loire-Atlantique

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	40.67	13350.00	271450.00
Contraction	28.33	7283.33	103180.56

FIGURE 1.34 – *Datation des cycles des demandeurs d'emploi - série Maine et Loire 49*

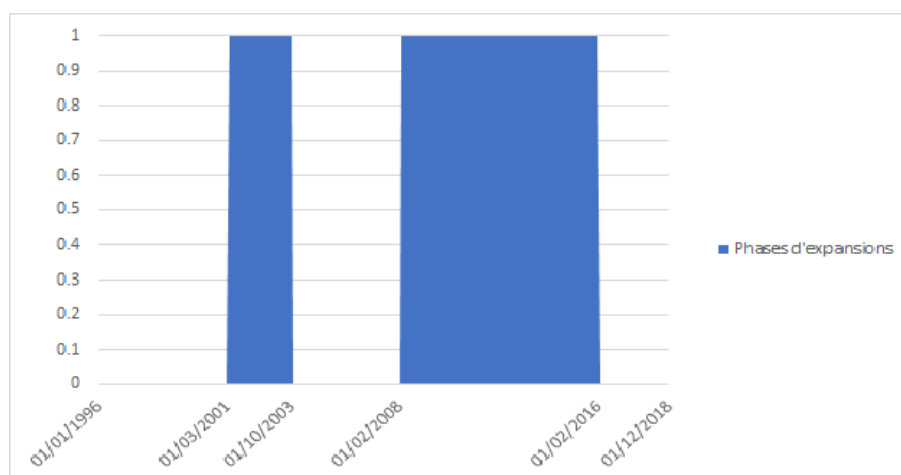


TABLE 1.24 – *Datation des points de retournement du cycle des demandeurs d'emploi - série Maine et Loire 49*

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	2001M3	NA	NA
Expansion	2001M3	2003M10	31	4180
Contraction	2003M10	2008M2	52	9230
Expansion	2008M2	2016M2	96	20930
Contraction	2016M2	NA	NA	NA

TABLE 1.25 – *Caractéristiques du cycle classique des demandeurs d'emploi - série Maine et Loire*

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	63.50	12555.00	398621.25
Contraction	52.00	9230.00	239980.00

FIGURE 1.35 – Datation des cycles des demandeurs d'emploi - série Moyenne 53

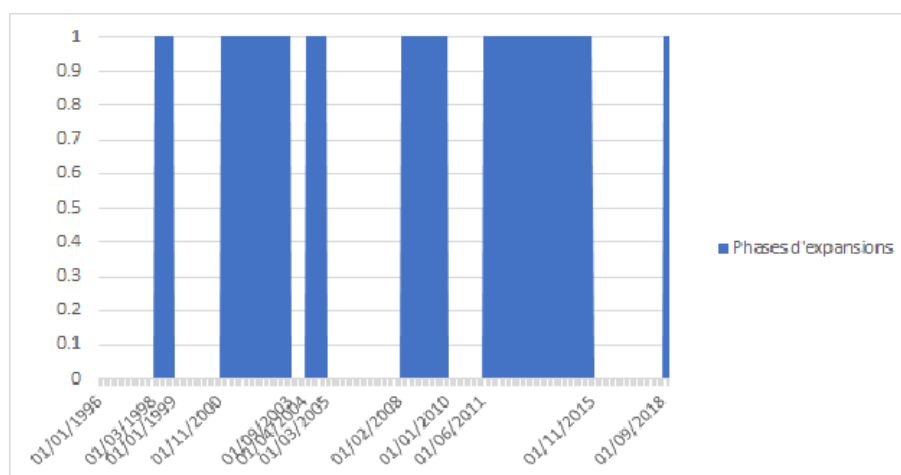


TABLE 1.26 – Datation des points de retournement du cycle des demandeurs d'emploi - série Moyenne 53

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	1998M3	NA	NA
Expansion	1998M3	1999M01	10	640
Contraction	1999M01	2000M11	22	3460
Expansion	2000M11	2003M9	34	2720
Contraction	2003M9	2004M4	7	310
Expansion	2004M4	2005M3	11	410
Contraction	2005M3	2008M2	35	2440
Expansion	2008M2	2010M1	23	2970
Contraction	2010M1	2011M6	17	780
Expansion	2011M6	2015M11	53	3300
Contraction	2015M11	2018M9	34	1200
Expansion	2018M9	NA	NA	NA

TABLE 1.27 – Caractéristiques du cycle classique des demandeurs d'emploi - série Moyenne

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	26.20	2008.00	26304.80
Contraction	23.00	8190.00	94185.00

FIGURE 1.36 – Datation des cycles des demandeurs d'emploi - série Sarthe 72

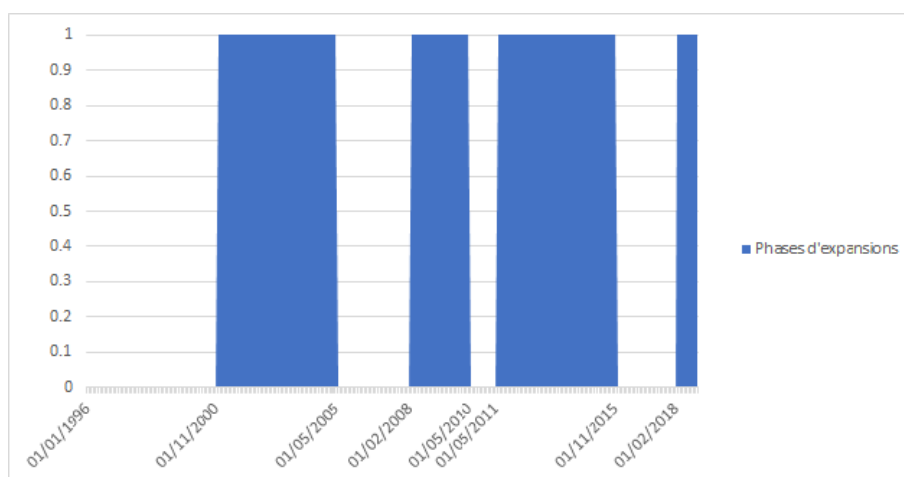


TABLE 1.28 – Datation des points de retournement du cycle des demandeurs d'emploi - série Moyenne 72

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	2000M11	NA	NA
Expansion	2000M11	2005M5	54	3680
Contraction	2005M5	2008M2	33	5030
Expansion	2008M2	2010M5	27	7210
Contraction	2010M5	2011M5	12	690
Expansion	2011M5	2015M11	54	7970
Contraction	2015M11	2018M2	27	2520
Expansion	2018M2	NA	NA	NA

TABLE 1.29 – Caractéristiques du cycle classique des demandeurs d'emploi - série Sarthe 72

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	45.00	6286.67	141450.00
Contraction	24.00	2746.67	32960.00

FIGURE 1.37 – Datation des cycles des demandeurs d'emploi - série Vendée 85

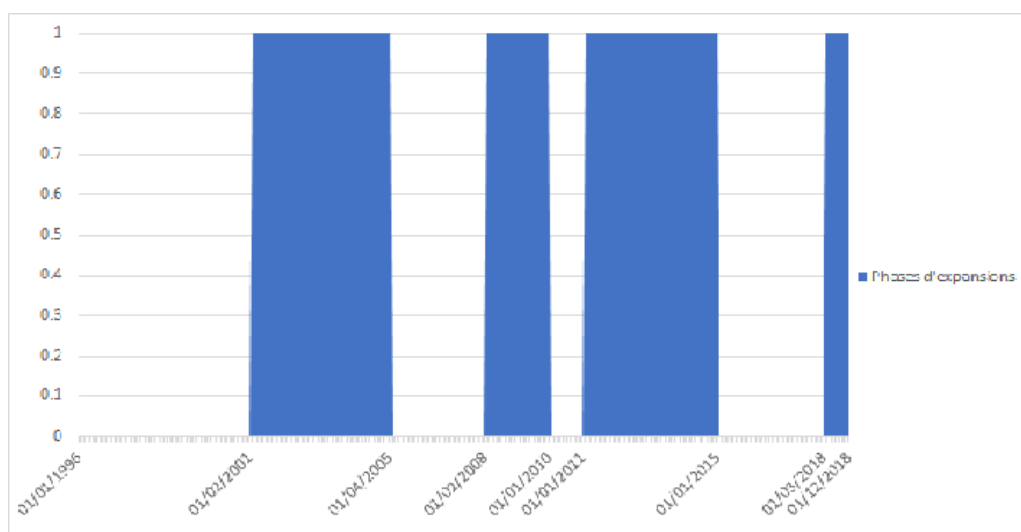


TABLE 1.30 – Datation des points de retournement du cycle des demandeurs d'emploi - série Vendée 85

Phase cycle	Début	Fin	durée en mois	Amplitude
Contraction	NA	2001M2	NA	NA
Expansion	2001M2	2005M4	50	3010
Contraction	2005M4	2008M2	34	5110
Expansion	2008M2	2010M1	23	7700
Contraction	2010M1	2011M1	12	1160
Expansion	2011M1	2015M1	48	8710
Contraction	2015M1	2018M3	38	3070
Expansion	2018M3	NA	NA	NA

TABLE 1.31 – Caractéristiques du cycle classique des demandeurs d'emploi - série Vendée 85

	Durée moyenne	Amplitude moyenne	Sévérité
Expansion	40.33	6473.33	130545.56
Contraction	28.00	3113.33	43586.67

# Chapitre 2

## Analyse du cycle de l'emploi : une étude régionale

### 2.1 Introduction

Ce chapitre 2 est en réalité un chapitre 1 bis, puisqu'il est une extension du précédent. Il explore l'analyse cyclique des régions françaises en conservant la même méthodologie. Ainsi les cycles d'emploi sont toujours définis par la définition de Burns et Mitchell (1946), c'est-à-dire une alternance de phase d'expansion et de contraction délimitées par des extrema locaux. Nous travaillons à partir des séries régionales brutes des demandeurs d'emploi de catégorie A fournies par Pôle Emploi. Les séries brutes nous permettent de procéder nous-même à la correction des valeurs aberrantes et à l'ajustement saisonnier afin de récupérer les séries dites de Tendence-Cycle, plus lisses que les séries CVS. Par conséquent, la datation des points de retournement *via* l'algorithme de Bry et Boschan (1971) est facilitée. Le découpage régional utilisé correspond à la période se situant avant la loi NOTRe<sup>1</sup>, pour des raisons de profondeur historique des données. En ce sens, nous récoltons les séries des DEFM catégorie A des 22 régions françaises entre janvier 1996 et décembre 2019. En effet, Pôle Emploi et la DARES maintiennent l'ancien découpage statistique à jour.

L'intérêt de poursuivre l'analyse réalisée dans le chapitre précédent et de l'étendre à toutes

---

1. La Loi portant nouvelle organisation territoriale de la République confie de nouvelles compétences aux régions. Elle est complétée par une loi qui modifie le découpage des régions.

TABLE 2.1 – Statistiques descriptives des données brutes régionales de la DEFM catégorie A, en millier

	Min.	Max.	Moy.	Med.	$\sigma$	Skew	Kur	J-B
Alsace	42,1	106,4	73,6	71,8	16,6	1,34	-4,06	0,00
Champagne-Ardenne	40,3	82,4	63,8	62,7	9,9	-1,12	-3,15	0,00
Lorraine	70,3	136,5	104,2	103,1	16,3	0,34	-3,18	0,00
Aquitaine	88,4	197,6	144,7	143,8	25,7	0,04	-3,03	0,01
Limousin	17,0	36,6	26,6	26,5	4,8	0,09	-3,85	0,00
Poitou-Charentes	46,0	94,6	72,6	73,2	10,9	-1,46	-2,55	0,01
Auvergne	35,1	68,7	54,1	53,2	8,0	-1,55	-3,23	0,00
Rhône-Alpes	166,4	347,2	263,3	260,8	47,2	-0,54	-3,81	0,00
Bourgogne	40,4	82,7	65,2	65,8	9,9	-2,22	-2,58	0,00
Franche-Comté	28,7	61,7	46,7	46,5	7,6	-0,74	-2,58	0,03
Bretagne	73,8	166,8	119,7	116,3	21,9	0,97	-3,29	0,00
Centre-Val-De-Loire	64,4	141,2	104,3	102,1	19,0	-0,39	-3,66	0,00
Corse	7,1	21,9	13,9	13,7	3,5	3,07	-0,95	0,00
Île-De-France	368,1	719,8	560,1	569,9	83,2	-1,68	-2,79	0,00
Languedoc-Roussillon	97,5	215,9	157,9	152,3	29,3	1,55	-2,91	0,00
Midi-Pyrénées	84,9	179,9	131,3	125,6	23,1	1,93	-3,28	0,00
Nord-Pas-De-Calais	161,9	278,9	233,8	232,1	31,6	-2,38	-3,70	0,00
Picardie	64,6	122,6	96,2	96,8	14,6	-1,38	-3,58	0,00
Basse-Normandie	38,1	77,6	60,3	60,7	8,5	-1,99	-1,40	0,06
Haute-Normandie	57,5	116,6	91,6	90,9	14,3	-2,36	-2,76	0,00
Pays-De-La-Loire	82,6	189,0	139,2	137,0	25,1	-0,22	-3,30	0,00
Provence-Alpes-Côte D'Azur	167,8	351,7	271,4	269,3	44,2	-1,32	-3,27	0,00

Source : Auteur, données Pôle Emploi. Note : Les valeurs du test de Jaque-Bera est la p-value. Si  $p < 0.5$  alors on rejette l'hypothèse nulle de normalité. et représentent les statistiques de test du skewness et du kurtosis, avec  $|v1| = \frac{(skewness-0)}{\sqrt{6/T}}$  et  $|v2| = \frac{(Kurtosis-3)}{\sqrt{24/T}}$ , avec  $T$  le nombre d'observations.

les régions françaises métropolitaines permet de poser les jalons d'une datation officielle des cycles d'emploi. Nous avons déjà exposé le paradigme auquel nous faisons face entre le cycle économique et le cycle d'emploi. Nous souhaitons mettre en avant le cycle d'emploi comme un cycle de référence au même titre que le cycle économique. Le chômage est devenu un indicateur majeur en matière de politiques publiques. C'est pourquoi l'information et la connaissance des caractéristiques des cycles d'emploi régionaux ne peuvent être qu'un atout pour l'éclairage des décisions des pouvoirs publics.

La suite de ce chapitre s'organise autour des résultats, avec la datation des points de retournement des cycles d'emploi régionaux et de leurs caractéristiques. Puis, nous verrons la datation des cycles d'accélération régionaux. Le chapitre se poursuit sur l'étude de la synchronisation et se termine avec une conclusion.

Le tableau 2.7 présente les statistiques descriptives. De manière générale, le test de Jarque-

Bera indique qu'aucune variable ne suit une loi normale. Les moyennes et les médianes sont relativement proches, ce qui indique une faible dispersion des données.

## **2.2 Résultats**

### **2.2.1 Datation des cycles d'emploi régionaux**

En suivant la logique méthodologique précédente, la datation des cycles régionaux de l'emploi de la DEFM (catégorie A) est obtenue à partir de l'algorithme de Bry-Boschan appliquée sur la tendance-cycle des séries. Conformément à la littérature, nous n'étudions pas les points de retournement présents dans les 6 premiers et les 6 derniers mois de notre chronologie, ceci afin d'éviter les faux pics, ou les faux creux qui ne possèdent pas de signification cyclique, en raison du manque de disponibilité des données avant le début et après la fin d'une série chronologique.



TABLE 2.2 – Datation des cycles d’emploi régionaux – 1996-2019

	Alsace	Champagne-Ardenne	Lorraine	Aquitaine	Limousin	Poitou-Charentes	Auvergne	Rhône-Alpes	Bourgogne	Franche-Comté
Creux				1996M10	1998M4					
Pic	1997M2	1997M3	1997M6	1997M10	1998M12	1997M5			1997M4	1997M1
Creux						1998M8				
Pic						1992M2				
Creux	2001M4	2001M5	2001M1	2001M2	2001M3	2001M4	2001M5	2001M3	2001M2	2001M1
Pic		2002M4		2003M10	2003M10	2003M10	2002M11	2003M9	2003M9	2003M12
Creux		2002M11		2004M7	2004M5	2004M7	2004M11	2004M3	2004M4	2004M10
Pic	2005M5	2003M12	2005M5	2005M1	2005M3	2005M3		2005M3	2005M3	2005M5
Creux	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2		
Pic	2010M2	2009M10	2010M9					2010M1	2010M5	2009M11
Creux	2011M3	2010M7	2011M5					2011M5	2011M3	2011M4
Pic	2015M10	2015M5	2015M5	2015M12	2016M2	2015M11	2016M2	2015M12	2016M2	2015M5

Source : Auteur, données Pôle Emploi.

TABLE 2.3 – suite : Datation des cycles d’emploi régionaux – 1996-2019

	Bretagne	C.V.D.L.	Corse	I.D.L.	Languedoc-Roussillon	Midi-Pyrénées	N.P.D.C.	Picardie	Basse-Normandie	Haute-Normandie	P.D.L.L.	P.A.C.A.	
Creux	1996M8												
Pic	1997M9	1996M12	1997M5	1997M6	1998M12	1998M12	1997M5	1997M7	1997M1	1996M12		1998M11	
Creux	1998M8						1998M8	1998M6					
Pic	1999M3						1999M2	1999M2					
Creux	2001M3	2001M4		2001M4	2001M5	2001M5	2001M4	2001M5	2001M5	2001M12		2000M11	
Pic	2003M10			2003M12	2003M1	2003M9	2002M3		2003M10			2003M10	
Creux	2004M8			2004M3	2004M3	2004M4	2002M12		2004M6			2004M7	2004M4
Pic	2005M5	2005M5	2004M11	2005M3	2005M3	2005M4	2005M5	2005M5	2005M5	2004M11		2005M5	2004M11
Creux	2008M2	2008M2	2008M4	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2	2008M2		2008M2	2008M4
Pic	2010M4	2010M5										2010M1	
Creux	2011M3	2011M3										2011M3	
Pic	2015M12	2016M2	2016M10	2016M2	2015M11	2015M11	2015M11	2015M9	2015M12	2016M2		2015M11	
Creux			2017M7	2017M1	2017M1		2016M11	2016M11		2016M11			
Pic			2018M5	2018M5	2017M10		2017M12	2017M7		2017M8			2017M10
Creux					2018M6								
Pic					2019M1								

Source : Auteur, données Pôle Emploi.

Note : C.V.D.L. = Centre-Val-de-Loire, I.D.L. = île-De-France, N.P.D.C. = Nord-Pas-De-Calais, P.D.L.L. = Pays-De-La-Loire, P.A.C.A. = Provence-Alpes-Côte D’Azur.

Les tableaux 2.2 et 2.3 synthétisent les chronologies des points de retournement du cycle de l'emploi dans l'ensemble des régions françaises. De manière générale, l'étude des cycles de l'emploi fait apparaître deux grandes phases d'expansion, toutes zones géographiques confondues. La première phase d'expansion du cycle se localise vers la fin de l'année 2000 et début 2001 et s'expliquerait par un ralentissement global de la croissance économique. En effet, la France fait face à l'onde récessive venue des États-Unis ayant pour effet une stagnation de l'activité. Cette phase de récession a été accentuée par l'attaque terroriste du 11 septembre 2001, qui a démoralisé la confiance des ménages et des entreprises notamment en France. La deuxième phase d'expansion des cycles est parfaitement identifiée en février 2008 dans toutes les séries étudiées, à quelques mois près pour certaines régions. La crise financière et économique marque ses effets dès le début 2008 dans les séries des demandeurs d'emploi. Par ailleurs, depuis peu de temps<sup>2</sup> l'Association Française de Science Economique (AFSE) a créé le Comité de Datation des Cycles de l'Economie Française (CDCEF). Les membres de ce comité ont identifiés le début de la récession économique lié à la crise financière au premier trimestre 2008, confirmant ainsi notre datation autour de cette période. Une dernière phase d'expansion marquante, commune à la plupart des séries étudiées, se situe en 2011 à la suite de la crise européenne.

### 2.2.2 Caractéristiques

Les tableaux 2.4 à 2.6 nous permettent d'observer une asymétrie entre les phases d'expansion et les phases de contraction du cycle de l'emploi à tous les niveaux géographiques étudiés. Nos résultats mettent en évidence des durées moyennes et des amplitudes moyennes plus grandes en phase d'expansion qu'en phase de contraction. L'inertie présente dans les séries du chômage pourrait être responsable d'une partie de cette asymétrie. En effet, une des caractéristiques connue du chômage est sa persistance (Sneessens, 1995 ; Allaire et al, 2000). Nous pouvons noter que les indices de l'*Excess* reportent des comportements d'inflexions conjoncturelles standards. Ainsi, les périodes d'expansions et de contractions sont pour la majorité caractérisées par des augmentations (expansions) et des diminutions (contractions) progressives. Les régions Poitou-Charentes, Bretagne, Picardie et Provence-Alpes-Côte-D'azur connaissent en moyenne des phases de contraction un peu plus brutales, c'est-à-dire une diminution du nombre de demandeurs d'emploi plus rapide en période de contraction du cycle

---

2. La création du CDCEF date du 25 octobre 2021.

de l'emploi. En revanche, seule la région Corse rencontre en moyenne des phases d'expansions plus marquées et rapides, synonyme d'augmentation rapide du nombre de demandeurs d'emploi en phase d'expansion du cycle de l'emploi.

TABLE 2.4 – Caractéristiques du cycle de l’emploi – 1996-2019

	Alsace	Champagne-Ardenne	Lorraine	Aquitaine	Limousin	Poitou-Charentes	Auvergne	Rhône-Alpes
<b>Expansion</b>								
Durée	46,00	25,50	43,67	35,25	36,25	34,25	43,00	30,00
Amplitude	26,50	10,77	27,55	26,69	5,30	11,77	10,69	55,06
<i>Excess</i>	0,25	0,20	2,95	0,42	0,28	0,69	0,86	2,23
<b>Contraction</b>								
Durée	23,00	28,00	28,00	29,67	23,00	21,25	24,00	19,00
Amplitude	14,29	10,98	19,94	26,40	4,67	10,28	7,09	26,22
<i>Excess</i>	-1,00	-0,86	-2,25	-0,84	-0,13	0,07	-0,50	-0,16

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l’amplitude moyenne est en millier de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

TABLE 2.5 – Suite : Caractéristiques du cycle de l’emploi – 1996-2019

	Bourgogne	Franche-Comté	Bretagne	Centre-Val-De-Loire	Corse	Île-De-France	Languedoc-Roussillon	Midi-Pyrénées
<b>Expansion</b>								
Durée	31,75	28,00	24,17	43,67	37,33	47,33	27,80	43,00
Amplitude	11,99	11,95	17,06	29,96	3,72	151,75	22,50	30,86
<i>Excess</i>	0,57	0,53	1,06	2,12	-0,05	9,57	1,07	1,03
<b>Contraction</b>								
Durée	24,25	27,00	17,40	33,00	43,33	36,67	20,40	24,67
Amplitude	11,78	9,39	14,31	24,20	3,34	128,73	15,55	18,11
<i>Excess</i>	-0,46	-1,23	0,45	-1,53	-0,08	-13,40	-0,34	-1,12

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l’amplitude moyenne est en millier de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

TABLE 2.6 – Suite : Caractéristiques du cycle de l’emploi – 1996-2019

	Nord-Pas-De-Calais	Picardie	Basse-Normandie	Haute-Normandie	Pays-De-La-Loire	Provence-Alpes-Côte D’Azur
<b>Expansion</b>						
Durée	30,40	38,75	44,67	50,00	31,25	60,50
Amplitude	26,83	15,84	13,08	20,36	28,23	74,96
<i>Excess</i>	3,31	2,73	1,15	2,24	1,10	10,06
<b>Contraction</b>						
Durée	19,00	21,25	31,00	32,67	18,33	53,00
Amplitude	27,46	13,40	13,64	22,58	14,55	65,69
<i>Excess</i>	-0,28	0,17	-0,77	-2,62	-0,37	4,38

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l’amplitude moyenne est en millier de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

Le tableau 2.7 met en avant les ratios moyens des durées et des amplitudes des cycles économiques des régions françaises. Un ratio supérieur à l'unité indique que la phase d'expansion de la région observée est plus longue ou plus profonde que la phase de contraction. La lecture du tableau permet une compréhension rapide et intuitive de l'asymétrie qui caractérise nos cycles régionaux d'emploi.

TABLE 2.7 – Statistiques descriptives des données brutes régionales de la DEFM catégorie A, en millier

	Ratio durée	Ratio amplitude
Alsace	1,44	1,85
Champagne-Ardenne	0,88	0,98
Lorraine	1,56	1,38
Aquitaine	1,19	1,01
Limousin	1,58	1,14
Poitou-Charentes	1,61	1,14
Auvergne	1,79	1,51
Rhône-Alpes	1,58	2,10
Bourgogne	1,31	1,02
Franche-Comté	1,04	1,27
Bretagne	1,39	1,19
Centre-Val-De-Loire	1,32	1,24
Corse	0,86	1,11
Île-De-France	1,29	1,18
Languedoc-Roussillon	1,36	1,45
Midi-Pyrénées	1,74	1,70
Nord-Pas-De-Calais	1,90	0,98
Picardie	1,82	1,18
Basse-Normandie	1,44	0,96
Haute-Normandie	1,53	0,90
Pays-De-La-Loire	1,70	1,94
Provence-Alpes-Côte D'Azur	1,14	1,14

Source : Auteur, données Pôle Emploi.

Note : Les deux ratios sont calculés avec la mesure de l'expansion au numérateur et la mesure de la contraction au dénominateur.

### 2.2.3 Datation des points de retournement des cycles d'accélération régionaux de l'emploi

Les tableaux 2.8 et 2.9 nous présentent la datation des cycles d'accélération d'emploi régionaux. Le nombre de points de retournement détecté par l'algorithme est conséquent, notamment en raison du caractère très fluctuant des taux de croissance de nos séries. Cependant, nous pouvons identifier plusieurs périodes importantes parmi l'ensemble des régions. La première période est située autour de l'année 2000, où l'on observe une accélération généralisée dans l'ensemble des régions étudiées. La seconde période d'accélération relativement présente correspond à l'année 2004. Une troisième période d'accélération des cycles d'emploi est parfai-

tement identifiée juste avant la crise de 2008, autour des années 2006 et 2007. Pour finir, on observe deux autres périodes d'accélération généralisées en 2013 et 2015.

Le premier constat que nous avons déjà fait lors de la précédente étude, correspond à l'anticipation des points de retournement du cycle d'accélération de l'emploi par rapport au cycle d'emploi. Les périodes d'accélération évoquées au-dessus sont pour la plupart enregistrées plusieurs mois avant les grandes périodes d'expansions du cycle d'emploi. Les tableaux 2.10 et 2.11 nous permettent de confirmer notre intuition, puisque en moyenne les points de retournement des phases d'accélération, correspondant aux phases d'expansions du cycle d'emploi, sont enregistrées 10,45 mois avant pour la période de l'année 2000, 10,9 mois avant pour l'année 2004 et 16 mois avant pour l'année 2008. En revanche, les phases d'accélération identifiées dans les années 2010, 2013 et 2015 ne possèdent pas toujours leurs équivalences dans les cycles d'emplois régionaux. L'anticipation moyenne des périodes de décélération sur les phases de contraction est similaire. Ainsi, on observe une anticipation d'environ 11,53 mois pour la période de 2001, 8,53 pour la période de 2004 et 15,3 pour la période de 2008.

TABLE 2.8 – Anticipation moyenne des phases d'accélération sur les phases d'expansions

Périodes	2000-2001	2004	2008
Moyenne en mois	10,45	10,9	16

Source : Auteur, données Pôle Emploi.

TABLE 2.9 – Anticipation moyenne des phases décélération sur les phases d'expansions

Périodes	2000-2001	2004	2008
Moyenne en mois	11,53	8,53	15,3

Source : Auteur, données Pôle Emploi.

Les caractéristiques des cycles d'accélération régionaux de l'emploi présentées dans les tableaux 2.12 à 2.14 nous renseignent sur plusieurs aspects. Dans un premier temps, les cycles d'accélération connaissent plus de linéarité que les cycles d'emploi régionaux, même si une légère asymétrie en terme de durée moyenne en faveur des phases de décélération est observée. Nous observons que l'amplitude des cycles d'accélération régionaux est symétrique. L'indice de l'*Excess* détaillé pour chacune des régions permettrait de retrouver le fait stylisé mis en avant par Majetti (2012). L'auteur décrit que les pics du cycle d'accélération du PIB sont arrondis, synonyme d'une baisse progressive du taux de croissance du PIB alors que les creux sont pointus, synonyme d'une reprise rapide du taux de croissance du PIB. Cependant, les résultats ne sont pas unanimes et cela dépend de la région considérée. Nous pouvons suggérer



que les régions avec des phases d'accélération brutales "creux pointus" ont le plus souvent des durées de phases d'accélération plus longues que les phases de décélération.

TABLE 2.10 – Datation des cycles d’emploi d’accélération régionaux – 1996-2019

	Alsace	Champagne-Ardenne	Lorraine	Aquitaine	Limousin	Poitou-Charentes	Auvergne	Rhône-Alpes	Bourgogne	Franche-Comté
Pic	1996M8	1996M9		1996M11	1996M11		1996M9	1996M9		1996M9
Creux	1998M3-1998M2			1998M2	1997M7	1997M8	1997M6	1998M3		
Pic	1998M10	1998M11			1998M9	1999M2	1999M1	1999M3	1998M12	
Creux	1999M10	2000M4	1999M9	2000M10	2000M10	2000M10	2000M3	2000M10	2000M10	2001M1
Pic	2001M10	2001M11	2001M7	2001M7	2001M11		2001M10	2001M10	2001M11	2002M3
Creux			2002M5	2002M10	2002M9		2002M9			
Pic			2003M1	2003M10	2003M9	2003M2	2003M5			
Creux	2004M1	2004M1			2004M4	2003M11		2004M1	2003M10	
Pic		2004M11			2004M11	2004M12		2004M8	2004M11	
Creux		2007M8	2006M11	2006M11	2007M2	2006M9	2006M5	2006M4	2007M2	2007M2
Pic	2009M2	2008M11	2009M2	2009M4	2008M12	2008M11	2008M11	2009M2	2009M1	2009M2
Creux	2010M10	2011M8	2011M3	2010M10	2010M3	2011M1	2010M10	2010M10	2010M9	2010M10
Pic	2013M1	2012M10	2013M1	2011M10	2012M9	2013M1	2012M7	2012M9	2012M9	2012M10
Creux	2010M10	2011M8	2011M3	2010M10	2010M3	2011M1	2010M10	2010M10	2010M9	2010M10
Pic	2013M1	2012M10	2013M1	2011M10	2012M9	2013M1	2012M7	2012M9	2012M9	2012M10
Creux	2013M10	2013M7	2013M8	2013M8	2013M10	2013M10	2013M8	2013M8	2013M10	2013M10
Pic	2014M5	2014M10	2014M11	2014M11	2014M7	2014M9	2014M4	2014M11	2014M10	2014M9
Creux		2016M7	2016M4	2017M4	2016M10	2016M3	2016M3	2016M3		
Pic		2017M3		2018M12	2017M10	2018M7	2018M2	2017M3		
Creux	2017M11	2018M5						2017M9	2017M9	2017M9
Pic	2019M5	2019M3	2019M6					2019M6	2019M3	2019M1

Source : Auteur, données Pôle Emploi.

TABLE 2.11 – Suite : Datation des cycles d'accélération de l'emploi régionaux – 1996-2019

	Bretagne	CVDL	Corse	IDL	LR	MP	NPDC	Picardie	BN	HN	PDLL	FACA
Pic				1996M9	1997M3			1996M9			1996M9	1996M12
Creux	1997M12	1997M8	1997M10			1997M2		1998M6	1998M3	1999M7	1997M6	
Pic	2000M4	1998M9	1998M9			1997M9		1999M1	1998M9	2000M5	1998M8	
Creux	2000M10	1999M7	2000M2	2000M11	2000M9	2000M10	1999M7	1999M7	1999M7	2001M9	2000M10	2000M10
Pic	2001M4	2001M11		2002M4	2001M10		2002M3	2001M9	2001M10	2002M10	2001M10	
Creux	2002M10	2002M9						2002M6			2003M9	
Pic	2003M7							2002M12				2003M2
Creux	2004M1				2004M1		2002M9	2004M1	2004M1		2003M11	2004M1
Pic	2005M4	2004M7	2003M6		2004M11	2003M6	2003M9	2004M11	2004M7		2004M7	2004M8
Creux	2006M5	2007M2	2005M10	2006M3	2006M5							2006M4
Pic				2006M10								2006M11
Creux				2007M6		2007M1	2007M3	2007M1	2007M10	2007M4	2007M4	2007M11
Pic	2009M2	2009M2	2009M4	2009M2	2009M2	2008M12	2009M1	2009M1	2009M1	2008M12	2008M11	2009M1
Creux	2010M10	2010M10	2011M10	2011M3	2010M9	2011M3	2011M1	2010M10	2010M6	2011M3	2011M3	2010M10
Pic	2012M9	2012M9	2012M10	2013M1	2012M10	2013M1	2013M1	2011M9	2012M8	2012M9	2012M9	2011M6
Creux	2013M8	2013M8	2013M11	2013M8	2013M5	2013M8	2013M8	2013M10		2013M8	2013M8	2012M2
Pic	2014M9	2014M9	2014M9	2014M9	2014M9	2014M11	2014M10	2015M4			2014M9	2014M9
Creux	2016M3	2016M10	2015M9	2016M4	2016M4		2016M3	2016M3	2016M3		2016M3	2016M3
Pic	2017M3		2017M7	2017M2	2017M10		2017M2	2017M5	2016M12	2016M2	2016M12	2017M3
Creux	2017M9		2018M4	2018M11					2018M10	2018M6	2017M9	
Pic	2018M12	2019M6		2019M7					2019M6		2019M3	

Source : Auteur, données Pôle Emploi.

Note : CVDL = Centre-Val-De-loire, IDL = Île De France, LR= Languedoc-Roussillon, MP = Midi-Pyrénées, NPDC = Nord-Pas-De-Calais, BN = Basse-Normandie, HN= Haute-Normandie, PDLL= Pays-de-La-Loire, PACA = Provence-Alpes-Côte D'Azur.

TABLE 2.12 – Caractéristiques des cycles d’accélération de l’emploi – 1996-2019

	Alsace	Champagne-Ardenne	Lorraine	Aquitaine	Limousin	Poitou-Charentes	Auvergne	Rhône-Alpes
<b>Durée</b>	24	12,5	21,7	16,2	14	21,3	18	18,1
<b>Amplitude</b>	0,04	0,05	0,03	0,04	0,04	0,03	0,03	0,04
<b>Excess</b>	-0,002	0,005	-0,002	0,00	0,001	0,004	0,002	0,00
<b>Durée</b>	21,5	21,25	21,4	28	16	17,2	18,7	16
<b>Amplitude</b>	0,04	0,06	0,03	0,04	0,04	0,03	0,03	0,04
<b>Excess</b>	0,00	-0,003	0,002	0,00	-0,003	0,00	-0,003	0,001

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l’amplitude moyenne est en millier de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\pm 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

TABLE 2.13 – Suite : Caractéristiques des cycles d’accélération de l’emploi – 1996-2019

	Bourgogne	Franche-Comté	Bretagne	Centre-Val-De-Loire	Corse	Île-De-France	Languedoc-Roussillon	Midi-Pyrénées
<b>Durée</b>	16	17,8	17,1	22	22,8	13,9	16,4	19,8
<b>Amplitude</b>	0,04	0,06	0,04	0,03	0,05	0,03	0,03	0,04
<b>Excess</b>	0,005	0,006	0,001	0,00	-0,007	0,001	-0,002	0,002
<b>Durée</b>	23,3	35,8	12,25	18	18,2	25,3	28,5	18,9
<b>Amplitude</b>	0,05	0,06	0,04	0,03	0,04	0,03	0,03	0,04
<b>Excess</b>	-0,005	0,001	-0,003	0,00	0,006	-0,001	0,001	-0,002

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l’amplitude moyenne est en millier de demandeurs d’emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\pm 10\%$  alors dispersion importante). L’*Excess* n’a pas d’unité, seul le signe est commenté ici. L’amplitude de l’HUR est en point de pourcentage.

TABLE 2.14 – Suite : Caractéristiques des cycles d'accélération de l'emploi – 1996-2019

	Nord-Pas-De-Calais	Picardie	Basse-Normandie	Haute-Normandie	Pays-De-La-Loire	Provence-Alpes-Côte D'Azur
<b>Expansion</b>						
Durée	19,2	14,5	13,9	16	13,9	15,3
Amplitude	0,03	0,03	0,03	0,05	0,04	0,03
<i>Excess</i>	-0,001	-0,002	0,00	0,002	0,001	0,002
<b>Contraction</b>						
Durée	19,2	16,5	26,3	21,8	19,9	19,4
Amplitude	0,03	0,03	0,03	0,05	0,04	0,03
<i>Excess</i>	0,002	0,00	-0,001	0,004	-0,001	-0,001

Source : Auteur, données Pôle Emploi

Note : La durée moyenne est en mois, l'amplitude moyenne est en millier de demandeurs d'emploi, le coefficient de variation est en pourcentage et relatif à la Tendence-Cycle (si  $\geq 10\%$  alors dispersion importante). L'*Excess* n'a pas d'unité, seul le signe est commenté ici. L'amplitude de l'HUR est en point de pourcentage.

## 2.2.4 Synchronisation cyclique régionale

Nous allons mesurer le niveau de synchronisation entre les 22 régions françaises étudiées. L'objectif est de déterminer si les cycles d'emploi sont semblables dans toutes les régions. La mesure de l'indice de concordance nous renseignera sur la convergence des cycles. Une valeur élevée de l'indice de concordance indiquera que les cycles d'emploi des régions observées seront très synchronisés. En d'autres termes, les deux cycles connaîtront des phases d'expansions et de contractions similaires en matière de durée et de datation.

### Indices de concordances régionales

Les indices de concordance du cycle de l'emploi au niveau régional et départemental sont élevés (Tableau 2.15). L'indice moyen est de 0.83, ce qui indique une forte synchronisation entre les régions. La valeur de l'indice minimum est de 0.59 (entre la région Bretagne et Provence-Alpes-Côte d'Azur). Quant à l'indice maximum, on l'observe entre la région Alsace et Centre-Val-De-Loire avec une valeur de 0.95. Nous constatons qu'en raison du cycle de l'emploi de la région PACA très atypique, les indices de concordances avec cette région comme référence sont parmi les plus faibles de notre échantillon. L'indice de concordance nous apporte un éclairage important sur la synchronisation des cycles de l'emploi entre les régions. Néanmoins, nous pouvons améliorer l'analyse des co-mouvements des séries en étudiant l'indice de Moran modifié.

TABLE 2.15 – Suite : Datation des cycles d'accélération de l'emploi régionaux – 1996-2019

	AL	CA	LO	AQ	LI	PC	AU	RA	BO	FC	BR	CVDL	CO	IDF	LR	MP	NPDC	PI	BN	HN	FDLL	PACA
AL	1	0,85	0,93	0,84	0,83	0,88	0,84	0,90	0,92	0,92	0,86	0,95	0,78	0,80	0,75	0,83	0,84	0,88	0,91	0,88	0,90	0,65
CA		1	0,83	0,83	0,78	0,86	0,84	0,80	0,82	0,90	0,78	0,83	0,83	0,83	0,71	0,78	0,81	0,80	0,87	0,82	0,80	0,68
LO			1	0,84	0,81	0,88	0,81	0,85	0,91	0,91	0,84	0,90	0,81	0,81	0,76	0,83	0,84	0,89	0,89	0,86	0,85	0,81
AQ				1	0,90	0,91	0,88	0,87	0,88	0,81	0,90	0,85	0,76	0,85	0,80	0,88	0,80	0,84	0,92	0,84	0,87	0,72
LI					1	0,90	0,89	0,90	0,86	0,80	0,86	0,86	0,71	0,80	0,79	0,86	0,80	0,83	0,90	0,85	0,89	0,70
PC						1	0,83	0,84	0,91	0,85	0,89	0,88	0,79	0,86	0,83	0,91	0,89	0,91	0,95	0,87	0,84	0,72
AU							1	0,86	0,82	0,76	0,78	0,88	0,82	0,80	0,71	0,78	0,83	0,81	0,87	0,89	0,84	0,71
RA								1	0,90	0,87	0,88	0,90	0,70	0,75	0,72	0,79	0,75	0,78	0,87	0,83	0,95	0,62
BO									1	0,91	0,89	0,94	0,76	0,85	0,78	0,86	0,81	0,80	0,91	0,87	0,88	0,59
FC										1	0,85	0,88	0,74	0,79	0,69	0,78	0,76	0,80	0,88	0,82	0,90	0,88
BR											1	0,85	0,74	0,78	0,80	0,79	0,83	0,82	0,80	0,84	0,77	0,83
CVDL												1	0,78	0,83	0,76	0,83	0,83	0,86	0,92	0,91	0,88	0,68
CO													1	0,79	0,75	0,76	0,85	0,84	0,79	0,84	0,68	0,83
IDF														1	0,78	0,79	0,84	0,84	0,86	0,89	0,75	0,74
LR															1	0,90	0,84	0,84	0,81	0,79	0,69	0,84
MP																1	0,81	0,88	0,89	0,81	0,78	0,81
NPDC																	1	0,93	0,85	0,88	0,75	0,77
PI																		1	0,89	0,90	0,77	0,74
BN																			1	0,90	0,88	0,71
HN																				1	0,81	0,74
FDLL																					1	0,59
PACA																						1

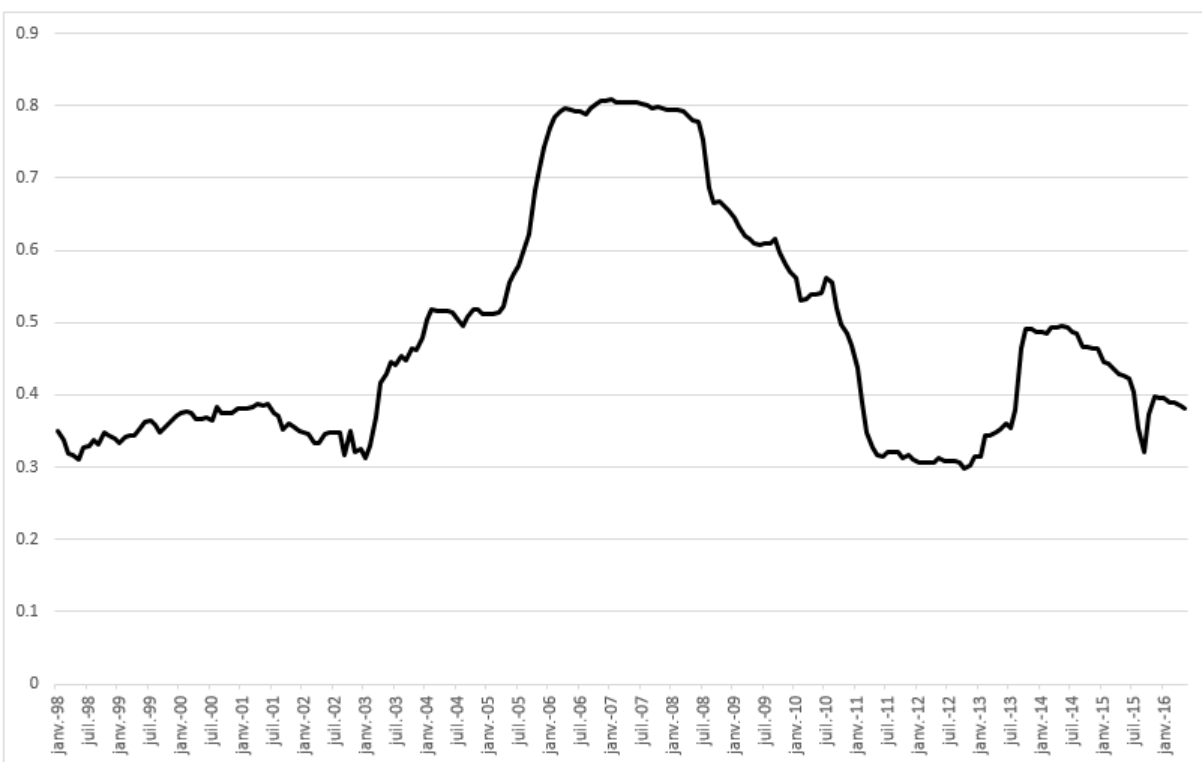
Source : Auteur, données Pôle Emploi.

Note :

## Indice de Moran modifié des cycles régionaux

L'indice de Moran modifié est calculé sur l'ensemble des régions française donnant un résultat sous la forme d'une série unique représenté avec la figure 2.1. A l'instar d'un indice de Moran classique qui mesure l'autocorrélation spatiale d'une variable, la version modifiée de cet indice s'intéresse uniquement à l'étude des co-mouvements des séries en écartant le rôle de la distance entre chaque variable, mais en s'intéressant davantage à l'évolution temporelle. L'indice permet de mesurer la synchronisation des co-mouvements dans un intervalle de 0 à 1. Une valeur de 1 signifie que les séries sont parfaitement synchronisées alors qu'une valeur de 0 indique une totale désynchronisation. Cette mesure est complémentaire à l'indice de concordance puisqu'elle ne prend pas uniquement en compte la datation des cycles de l'emploi mais capte aussi leur intensité.

FIGURE 2.1 – *Indice de Moran modifié des cycles régionaux de l'emploi,  $k=61$*



Nous pouvons constater à l'aide de la figure 2.1 que la synchronisation entre les cycles régionaux d'emploi est relativement faible entre 1998 et 2002. L'indice oscille dans un intervalle allant de 0,3 à 0,4. La datation des points de retournements du cycle d'emploi nous indique que plusieurs phases de contractions et d'expansions se succèdent durant cette période. L'alternance des phases en matière de durée et d'amplitude empêcherait le processus de syn-



chronisation. On observe donc une stagnation de la synchronisation jusqu'en 2003, date à partir de laquelle une phase de contraction des cycles est enregistrée dans la quasi-totalité des régions, qui se traduit par une hausse de la synchronisation. Beyer et Stemmer (2016) évoquent l'introduction de l'Euro (1er janvier 2002) comme source de synchronisation entre chômage régionaux des pays européens. L'indice mesure une nouvelle stagnation entre mars 2004 et mars 2005 synonyme d'une période d'expansion des cycles d'emploi avant l'embellie économique qu'a connue la France jusqu'à la récession économique de 2008. C'est ainsi que nous remarquons que l'indice de Moran modifié présente des valeurs élevées de synchronisation régionale. De plus, la valeur maximum (0.81) de l'indice est observée en janvier 2007, soit 13 mois avant l'expansion généralisée des cycles d'emploi, associée à la récession économique de 2008. Pendant cette phase post crise de désynchronisation des cycles, nous pouvons observer un léger rebond de l'indice entre mars 2010 et août 2010. Ce rebond est caractérisé par le fait que certaines régions ont recouvré une courte phase de contraction du cycle d'emploi. L'indice continue sa chute avant de se stabiliser en 2013 et de connaître un rebond fin 2013, début 2014. La phase de contraction généralisée suivante des cycles d'emploi est enregistrée autour de l'année 2015, ce qui pourrait indiquer que dès 2013, les cycles régionaux se synchronisent même en phase d'expansion du cycle d'emploi. Nous pouvons penser que les phases d'expansions des cycles connaissent une décélération à l'approche du point de retournement, comme nous l'indiquent l'*Excess* et la datation des cycles d'accélération de ces périodes, ce qui aurait permis un retour plus rapide de la synchronisation entre les régions.

Un constat très marquant peut être repéré dans cette étude de l'indice de Moran modifié. La synchronisation des cycles régionaux d'emploi se produit essentiellement pendant les phases de contractions et se détériore durant les phases d'expansions. La période 2013 à 2015 est propice à une analyse inverse où l'indice capte de la synchronisation entre régions alors que celles-ci connaissent des périodes d'expansions du cycle d'emploi. Lors de notre précédente analyse, nous avons évoqué l'impact potentiel de la fusion ANPE et Assedic fin 2008, qui avait eu pour hypothétique effet d'améliorer la synchronisation entre les cycles d'emploi départementaux de la région Pays de la Loire. Force est de constater que cette hypothèse est plus difficile à argumenter sur la synchronisation des cycles régionaux français. Une analogie très visuelle peut être réalisée en reprenant un principe de mécanique classique emprunté à la physique. Nous parlons du phénomène de synchronisation des métronomes. Imaginons 22 métronomes, représentant nos 22 régions françaises, tous reliés par un support commun

(canaux de transmission de l'information, mobilité des individus, diffusion des politiques publiques d'emploi, etc...). La démonstration physique nous apprend qu'une synchronisation entre les métronomes (régions) est opérante si et seulement si il y a un phénomène de couplage, c'est-à-dire si les informations des balancements des métronomes sont communiquées aux autres *via* le support commun. Ainsi, nous pouvons avancer que durant les phases de contraction du cycle de l'emploi (expansion économique), le phénomène de couplage entre régions est opérant. L'information circule librement, la mobilité des individus est plus présente et la diffusion des politiques publiques de l'emploi est plus fluide. En revanche, durant les périodes d'expansions du cycle de l'emploi (récession économique) le couplage entre régions est entravé, à l'image des métronomes, la synchronisation n'est plus observée. Cette analogie permet d'illustrer, selon nous, un principe majeur dans l'étude des cycles d'emploi : Le rôle joué par la région sur elle-même ainsi que sur les autres régions voisines. L'étude précédente sur les cycles départementaux permet de mettre en lumière l'importance de l'interaction des cycles qui composent un ensemble plus grand (ici la région Pays de la Loire). Nous avons remarqué que chaque département possédait ses propres caractéristiques influençant sa capacité à connaître des phases de contraction du cycle d'emploi plus rapidement que d'autres. Cependant, nous avons aussi remarqué que la synchronisation mesurée par l'indice de Moran modifié des cycles d'emploi départementaux correspondait à la datation du cycle régional des Pays de la Loire. En effet, l'hypothèse que les cycles d'emploi des plus proches voisins s'influencent entre eux n'est pas difficile à poser, sachant que le taux de chômage régionales est caractérisé par une dépendance spatiale (Patacchini et Zenou, 2007). Un commentaire similaire peut être réalisé avec la comparaison de l'indice de Moran modifié régional, qui semble suivre la datation des points de retournement du cycle d'emploi national.

## 2.3 Conclusion

Dans ce chapitre 1 bis, nous avons choisi d'étendre l'analyse des cycles d'emploi à l'ensemble des régions françaises. En suivant la même méthodologie que précédemment, nous avons estimé les cycles d'emploi et les cycles d'accélération de l'emploi des 22 régions françaises (découpage statistique antérieur à la loi NOTRe).

Plusieurs résultats sont dans la lignée du premier chapitre. Les cycles d'emploi régionaux sont asymétriques. Ils connaissent des phases d'expansion plus longues que les phases de

contraction. Par ailleurs, les phases d'expansion sont, pour l'extrême majorité, plus profondes que les phases de contraction. Nous retrouvons ici le phénomène de persistance dans les séries du chômage, caractérisées par de longues phases d'expansion. La datation des points de retournement des cycles d'accélération de l'emploi nous renseigne sur la nature symétrique de ces cycles. Après avoir identifié des périodes d'expansion généralisée dans l'ensemble des régions, nous avons mesuré la durée moyenne d'anticipation des points de retournements du cycle d'accélération et celui du cycle d'emploi. Nous avons constaté que ceux-ci se produisent en moyenne 10 mois avant, permettant ainsi de capter une information précoce sur les points de retournement du cycle de l'emploi.

L'étude de la synchronisation des cycles d'emploi nous permet d'observer un résultat important sur la particularité de cette synchronisation. En effet, les périodes d'expansion du cycle d'emploi sont identifiées pendant des phases de synchronisation selon l'indice de Moran modifié. Cependant, les périodes de contraction du cycle d'emploi sont davantage identifiées pendant les phases de désynchronisation. Beyer et Stemmer (2016) attribuent la désynchronisation observée après la crise de 2008 à des facteurs spécifiques aux régions ainsi qu'à des facteurs pays. Ce résultat met en avant le rôle que les régions peuvent avoir sur la convergence ou la divergence des cycles, qu'ils soient économiques ou d'emploi. A l'image du taux de croissance du PIB et du taux de chômage, le cycle économique et le cycle d'emploi sont négativement corrélés. On pourrait penser à une possible synchronisation des cycles pendant les phases de contraction et pendant les phases d'expansion. Mais, en réalité, on observe davantage de synchronisation quand l'économie connaît une période de stabilité ou de croissance économique et beaucoup plus de désynchronisation quand l'économie est perturbée par une crise ou un choc économique.

L'indice de concordance correspondant à l'autre mesure de synchronisation des cycles que nous utilisons délivre des résultats assez homogènes. En d'autres termes, les régions rencontrent leurs phases de contraction et d'expansion en même temps, avec quelques écarts dans la datation des points de retournement. Ce résultat apparaît moins important que l'indice de Moran modifié, car nous supposons une limite en lien avec le processus générateur des données. L'analyse des variations cycliques entre régions est probablement biaisée. La DEFM ne se forme pas au niveau régional, car les marchés du travail sont plutôt au niveau du département. Ainsi, les variations intra-régionales gommant une partie des variations

inter-régionales.

Lors d'une recherche future, il serait très intéressant de mesurer la synchronisation des cycles d'emploi départementaux. Le chapitre 1 permet d'avoir un aperçu de ce type d'analyse. On souhaiterait dégager des profils cycliques par départements comme nous l'avons réalisé pour la région des Pays De La Loire.

# Bibliographie

- [1] Anas J, Billio M, Ferrara L, Mazzi G-L (2008) A system for dating and detecting turning points in the euro area. *The Manchester School*, 76 (5), 549–577.
- [2] Anas J, Ferrara L (2004) Detecting cyclical turning points : The ABCD approach and two probabilistic indicators. *Journal of Business Cycle Measurement and Analysis*, 1 (2), 193–225.
- [3] Artis M, Krolzig H-M, Toro J (2004) The European business cycle. *Oxford Economic Papers*, 56 (1), 1–44.
- [4] Artis M-J, Kontolemis Z-G, Osborn D-R (1997) Business cycles for G7 and European countries. *The Journal of Business*, 70 (2), 249–279.
- [5] Bandrés E, Gadea Rivas M-D, Gómez-Loscos A (2017) Regional business cycles across Europe (Technical Report No 1702). Banco de España.
- [6] Box G-E, Jenkins G-M, Reinsel G (1970) *Time Series Analysis : Forecasting and Control*. Holden Day San Francisco.
- [7] Box G-E, Tiao G-C (1975) Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70 (349), 70–79.
- [8] Bry G, Boschan C (1971) Programmed selection of cyclical turning points. In *Cyclical Analysis of Time Series : Selected Procedures and Computer Programs* (pp. 7–63). NBER.
- [9] Burns A-F, Mitchell W-C (1946) *Measuring Business Cycles*. NBER.
- [10] Burman, J-P, Otto M-C (1988) *Outliers in Time Series*. US Bureau of the Census.
- [11] Buys-Ballot C H D (1847) *Les Changements Périodiques de Temperature*. Utrecht : Kemink et Fils.

- [12] Camacho M, Perez-Quiros, G, Saiz L (2006) Are European business cycles close enough to be just one. *Journal of Economic Dynamics and Control*, 30 (9-10), 1687–1706.
- [13] Chang I, Tiao G-C (1983) Effect of exogenous interventions on the estimation of time series parameters. In *Proceedings of the American Statistical Association, Business and Economics Statistics Section* (pp. 532–537).
- [14] Chang I, Tiao G-C, Chen C (1988) Estimation of time series parameters in the presence of outliers. *Technometrics*, 30 (2), 193–204.
- [15] Chen, C., & Liu, L. M. (1993). Forecasting time series with outliers. *Journal of forecasting*, 12(1), 13-35.
- [16] Comunale M. (2017) Synchronicity of real and financial cycles and structural characteristics in EU countries (Technical Report No 15). Bank of Lithuania.
- [17] Darné O, Ferrara L (2011) Identification of slowdowns and accelerations for the Euro area economy. *Oxford Bulletin of Economics and Statistics*, 73 (3), 335–364.
- [18] Ferrara L (2009) Caractérisation et datation des cycles économiques en zone euro. *Revue Économique*, 60 (3), 703–712.
- [19] Findley D-F (2005) Some recent developments and directions in seasonal adjustment. *Journal of Official Statistics*, 21 (2), 343.
- [20] Fox A-J (1972) Outliers in time series. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34 (3), 350–363.
- [21] Gómez V (2001) The use of butterworth filters for trend and cycle estimation in economic time series. *Journal of Business & Economic Statistics*, 19 (3), 365–373.
- [22] Gómez V, Maravall A (1996) Programs SEATS and TRAMO : Instructions for the user.
- [23] Gómez V, Maravall A (2001) *Automatic Modeling Methods for Univariate Series*. Wiley Online Library.
- [24] Hannan E-J, Rissanen J (1982) Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, 69 (1), 81–94.
- [25] Harding D (2004) Non-parametric turning point detection, dating rules and the construction of the euro-zone chronology. Paper presented at the Eurostat. In *Colloquium on modern tools for business cycle analysis*.

- [26] Harding D, Pagan A (2001) Extracting, analysing and using cyclical information. Mimeo, Australian National University.
- [27] Harding D, Pagan A (2002) Dissecting the cycle : A methodological investigation. *Journal of Monetary Economics*, 49 (2), 365–381.
- [28] Harding D, Pagan A (2006) Synchronization of cycles. *Journal of Econometrics*, 132 (1), 59–79.
- [29] Hillmer S-C, Bell W-R, Tiao G-C (1983) Modeling considerations in the seasonal adjustment of economic time series. *Applied Time Series Analysis of Economic Data*, 74–100.
- [30] Hodrick R-J, Prescott E-C (1997) Postwar US business cycles : An empirical investigation. *Journal of Money, Credit, and Banking*, 29 (1), 1–16.
- [31] Kaiser R, Maravall A (1999, 02) Seasonal Outliers in Time Series. *Estadística*, 53.
- [32] Juglar C, (1862) *Des Crises Commerciales et de leur retour périodique en France, en Angleterre et aux États-Unis*, Paris : Guillaumin.
- [33] Labondance F, (2013) Regional business cycle synchronisation in the euro area. *Revue d'économie régionale et urbaine*, (2), 269-296.
- Lee J, Li S (2017) Extending Moran's index for measuring spatiotemporal clustering of geographic events. *Geographical Analysis*, 49, 36-57.
- [34] Majetti, R. (2012). Caractérisation et datation des cycles économiques : le cas français. *Revue d'économie politique*, 365-403.
- [35] Maravall A, Gómez V, Caporello G (2015) Statistical and econometrics software : TRAMO and SEATS. *Statistical and Econometrics Software*.
- [36] Mazzi G, Savio G (2007) *Growth and Cycle in the Eurozone*. Palgrave-MacMillan.
- [37] Mehrhoff J (2018) Outlier detection and correction. In G. L. Mazzi D. Ladiray (Eds.), *Handbook on Seasonal Adjustment* (p. 139-167). Eurostat.
- [38] Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- [39] Otto M-C, Bell W-R (1990) Two issues in time series outlier detection using indicator variables. In *Proceedings of the American Statistical Association, Business and Economic Statistics Section* (pp. 182–187).

- [40] McKay A, Reis R (2008) The brevity and violence of contractions and expansions. *Journal of Monetary Economics*, 55 (4), 738–751.
- [41] Mink M, Jacobs J-P, de Haan J (2012) Measuring coherence of output gaps with an application to the euro area. *Oxford Economic Papers*, 64 (2), 217–236.
- [42] Mintz I, (1969) Dating postwar business cycles : Methods and their application to Western Germany, 1950–67. NBER.
- Mitchell, W-C (1913) *Business Cycles*. Berkeley : University of California Press (pt. III reprinted as “Business cycles and their causes”, 1941).
- Moore H-L, (1914) *Economic cycles : Their law and cause*. New York : The Macmillan Company.
- [43] Mundell R-A (1961) A theory of optimum currency areas. *The American Economic Review*, 51 (4), 657–665.
- [44] Peiró A, Belaide-Franch J, Gonzalo M-T (2012) Unemployment, cycle and gender. *Journal of Macroeconomics*, 34 (4), 1167–1175.
- Persons W-M, (1919) An index of general business conditions. *Review of Economic Statistics*, 1(1), 111-205.
- [45] Samarina A, Zhang L, Bezemer D (2015) Mortgages and credit cycle divergence in eurozone economies. SOM Research Report.
- Schumpeter J-A, (1939) *Business cycles. A theoretical, historical, and statistical analysis of the capitalist process*, New York : McGraw Hill.
- [46] Stock J-H, Watson M-W (2008) The evolution of national and regional factors in us housing construction. In Russel. J. Bollerslev T. Watson M. (Eds.), *Volatility and Time Series Econometrics : Essays in Honor of Robert Engle*. Oxford University Press.
- [47] Shiskin J, Young A, Musgrave J (1967) The X-11 variant of the Census method II seasonal adjustment program (Rapport technique No 15). Bureau of the Census, US Department of Commerce Washington, DC.
- [48] Tsay R-S (1986) Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, 81 (393), 132–141.
- [49] Zarnowitz V, Ozyildirim A (2006) Time series decomposition and measurement of business cycles, trends and growth cycles. *Journal of Monetary Economics*, 53 (7), 1717–1739.



# Forecasting french jobseeker

# Resume

**Forecasting French jobseekers with Google Trends** : Keywords : nowcasting ; shrinkage methods ; variable selection.

JEL Classification : C53 ; C55 ; E01 ; E17 ; O40.

Over the last decade, data from Google Trends platform has become more and more prominent. There are several reasons for this phenomenon, the availability of data, the relevance and the expanding possibilities of the search engine. In this article, we develop a literature review focusing on unemployment forecasting using Google Trends data. We perform an empirical application on the forecasting of French job seekers. The results are in agreement with those of literature and highlight the performance of Google Trends data and in particular the major interest of its availability for the forecaster.

**Nowcasting French jobseekers using data reduction methods** : Keywords : nowcasting ; shrinkage methods ; variable selection.

JEL Classification : C53 ; C55 ; E01 ; E17 ; O40.

This paper compares penalized regressions, random forest and dynamic factor models to nowcast the monthly number of French jobseekers, using a large data set of Google Trend, survey and financial data. A rolling forecast study is carried out to assess their nowcasting performance which is evaluated using the modified Diebold-Mariano test and the model confidence set approach. The results show that, the Lasso, Elastic-Net and adaptive Elastic-Net models exhibit a good nowcasting performance, and are thus relevant for nowcasting jobseeker growth rate. Their accuracy gains are large during the Covid-19 crisis whereas they provide similar performance with respects to the dynamic factor model during the non-crisis

period. Among the predictors for nowcasting jobseeker growth it appears that the variation of Google Trend variable and Finished-goods inventory in the manufacturing sector from the PMI survey are important predictors due to the fact that they are selected by all the approaches.

## Chapitre 3

# Forecasting French jobseekers with Google Trends

## 3.1 Introduction

We develop a non-exhaustive literature review on unemployment forecasting based on Google Trends data. This exercise takes place at a time when Google queries have increased year after year and have become a useful source of information and a major explanatory variable. Mainly, for economic purposes and especially, in our case, for a job search. Today, Google’s dominance in the search engine market is well established. Moreover, job searches are increasingly conducted *via* the Internet.

Google data induced a revival of interest to forecast literature, and not only in the field of economics. To the best of our knowledge, one of the first articles using Google data attempted estimating weekly influenza activity in the US by using influenza-related Google search queries (Ginsberg et al., 2009). Google econometrics became widely used by economic forecasters in various economic environment, such as private consumption (Kholodin et al., 2010), housing market (...), energy prices (Afkhami et al., 2017; economic activity (Woloszko, 2020; Ferrara and Simoni, 2020), among others. These studies have yielded encouraging results which could be an advantage in forecasting our variable namely, “Demandeurs d’emploi en fin de mois” (DEFM).

After studying what was done in the literature, we set up an empirical analysis which consists of verifying if the data from Google Trends allows us to improve the prediction of our variable of interest, the jobseekers at the end of the month (category A<sup>1</sup>) collected by the French national employment agency. We work with monthly frequency on stationary and seasonally adjusted series<sup>2</sup>. We will nevertheless study the forecasting performance of the non-stationary Google variable.

We compare the performance of our model with a benchmark AR model, commonly used in the literature. The paper is organised as follows : Section 2 describes the literature review of Google Trends, Section 3 presents the methodology behind the forecast, while Section 4

---

1. Unemployed person, obliged to perform positive acts of job search, looking for a job whatever the type of contract (permanent, fixed-term, full-time, part-time, temporary or seasonal).

2. There is no consensus in the literature as to the stationarity of Google Trends data since this depends on the keyword chosen. Choi and Varian (2012), D’Amuri and Marcucci (2017) do not perform any differencing of the series. Moreover, Fondeur and Karamé (2013) do not make any changes to the raw variable and considers seasonal adjustment to be unhelpful as it would lead to a loss of information. Nagao et al. (2019) use three different methods to seasonally adjust the series. Borup and Schütte (2020) employ a sequential testing strategy to test whether or not Google Trends data is best characterized by stationarity or a unit root.

describes the data used. Section 5 presents the results of the estimates and forecasts and Section 6 concludes.

## 3.2 Literature Review

Choi and Varian (2012) used Google data to nowcast several indicators in the US, namely retail sales, home sales and travel. They also used Google data from January 2004 to June 2009 to predict "initial claims for unemployment benefits" (Choi and Varian, 2009). The authors used an Autoregressive model with Google data namely, the search related to "Jobs" and "Welfare & Unemployment", to nowcast the weekly initial unemployment claims. They demonstrated a positive correlation between initial claims and the two search terms. Moreover, the model with the Google Trends data predicts better than the model without Google Trends data. The out-of-sample Mean-Absolute-Error (MAE) decreased by 15.74% between the two models.

This shows the beginning of the literature review of the use of Google's data in unemployment forecasting. The unemployment rate is one of the most monitored economic indicators. It is a leading indicator of economic activity and nowadays it is an indicator in its own right. Following the article by Choi and Varian (2009, 2012), many studies have emerged. Askitas and Zimmermann (2009) studied four different keywords collect from Google Trends platform, using a time period from January 2004 to April 2009. The keywords concerned are "Unemployment office or agency", "Unemployment rate", "Personnel consultant" and most popular job search engines in Germany. They collected, for each keyword, the weeks 3 and 4 for the last month and weeks 1 and 2 for the following month in order to predict the monthly German unemployment rate by using an error-correction model. Their results showed a strong correlation between keyword searches and unemployment rate and more precisely, weeks 3 and 4 of the last month could help to forecast the current month's unemployment rate. These results suggest that well-chosen keywords could provide useful information about the unemployment rate.

For Israel, Suhoj (2009) demonstrated that the Google Trends data including the leading query category of human resources from February 2004 to February 2009 (recruiting and staffing) was highly correlated with the Israeli unemployment rate. Following the same methodology as in Choi and Varian (2012), Suhoj (2009) showed that the best job of prediction can be done with Google's index (human resources category), according to the Granger causality tests. Suhoj (2009) extended the method of Choi and Varian to Israeli data. The results show, Google-based models have good results which are a confirmation of their performance.

D'Amuri (2009) was the first to predict Italian quarterly unemployment rate using Google Trends keyword ("Job offers") along with two other macroeconomic variables (Industrial Production and Employment expectation survey). He worked on a small sample (2004 :1 to 2009 :1) containing fewer than 30 observations. The author chose the structure of the ARIMAX(1,1,0) model to estimate and forecast 39 different models. Thus, modifying the explanatory variables used or the sample size. He showed that Google Trends data performs fairly well in forecasting of the unemployment rate. Moreover, the addition of the Google index in the best model produces a means square error (MSE) 33% lower than the best model without Google's index. These results imply that Google data are complementary to more traditional forecasting data.

D'Amuri and Marcucci (2017) continued this research on US data. They evaluated and compared to around 500 models of US unemployment rate. They used the keyword "jobs" and four additional explanatory variables, namely, initial claims, employment expectation, the current six-month-ahead consumer expectation and the index of economic policy uncertainty proposed by Baker, Bloom and Davis (2016). They conducted one- to twelve-step-ahead forecast of the US unemployment rate ( $h = 1, 2, \dots, 12$ ) from February 2004 to February 2014. The authors compared bivariate models which differ in the aforementioned variables. As a robustness check, many other model forms and lag structures were conducted. Their results show that Google data significantly improves the explanatory power of unemployment forecasting models. For the best Google-based models, the RMSE of the one-step and twelve-step forecast was improved by 18% and by 32%, respectively. An interesting comparison at the

state level shows that Google-based models are useful to forecast monthly unemployment rate. By the way, they discovered a strong correlation between the unemployment rate and the Google index when a greater percentage of the jobless are using the internet for their job search. As D'Amuri (2009), results suggest that models which combine Google data and soft data, like survey, would be efficient.

As a first contribution from an emerging economy, Chadwick and Şengül (2012) employed an approach using data for Turkey from January 2005 to December 2011 to predict the unemployment rate. They used a model based on a Bayesian Model Averaging (BMA) framework supplemented by Google Trends data. The authors selected 20 keywords from the Google Trends platform and two other explanatory variables namely, industrial production and initial unemployment claims. Using the BMA procedure, 45 models were selected to be compared to a benchmark AR model. Their results showed that the Google-based models outperformed the benchmark in terms of RMSE at any forecast horizon ( $h= 1, 2, 3$ ).

Google data are also evaluated for forecasting the unemployment rate in the Czech Republic, Hungary, Poland and Slovakia by Pavlicek and Kristoukef (2015). They used three AR models with different lag structures<sup>3</sup> augmented by Google Trends data from January 2004 to December 2013. The keywords "job" and "work" are selected by the authors and translated into the respective languages. Their results showed that Google-based models improved the adjusted  $R^2$ . Furthermore, two of the three models outperformed the benchmark models in terms of forecast accuracy in a statistically meaningful way. Google based-models work across many countries which confirm the validity of the previous results.

More recently, Maas (2019) used a Mixed-Data Sampling (MIDAS) regression model supplemented with Google Trends data from the keyword "unemployment". The author used several explanatory variables that were divided into six panels. The baseline variables are the initial claims and a large database published by FRED-MD including 134 monthly time se-

---

3. AR(3), AR(6), AR(12).



ries from January 2004 to December 2017. The author used the diffusion index methodology proposed by Stock and Watson (2002) in order to forecast the US unemployment rate over a horizon of one to twelve months. As a result, the best short-term prediction was made by a Google-based models, but this model did not exceed the benchmark of D'Amuri and Marcucci (2017) in terms of equality predictive accuracy (EPA, Diebold and Mariano test). Here, the author induces the famous dimensional reduction method of Stock and Watson (2002) to deal with a huge database. Thus, they reduce the numerous baseline variables into few factors which he introduces in the model augmented by Google data. Their results suggest a good performance of these models. A machine-learning technique could be a good way to undertake better unemployment forecasting.

Nagao et al. (2019) studied how the search intensity data obtained from Google Trends contributes to nowcasting of the US unemployment rate. They used two kinds of keywords namely, "jobs" and "job offer", using a time period from January 2004 to June 2018. A simple AR model augmented by Google Trends data was compared to the conventional AR model by following D'Amuri and Marcucci (2017). Their results challenge the predictive power of Google data. The model with Google's index predicts the level of unemployment rates more accurately than the benchmark. Nevertheless, they showed that adding the Google index to the AR model brings almost no significant improvement in forecast accuracy. Furthermore, the authors explain that their Google-based model is sensitive to the frequency of the Google index and the seasonal adjustment method chosen. Finally, they showed that initial claims better predict unemployment rate than the Google index. These results could induce a warning about the Google data usefulness. We assume that Google data remains a good complement in the short term but is not sufficient on its own.

Let's go back to the past with papers focusing on unemployment forecasting with a different dependent variable of the classical unemployment rate. Smith (2016) provided two empirical studies, one as part of a monthly forecast of 3-month change of International Labor Office (ILO) unemployment augmented with Google Trends data and traditional survey variables. The author constructed a Google index from a "root" keyword which is "redundancy". As

an example, "redundancy payments", "redundancy law", or "redundancy uk" among others, a total of 21 terms make up the Google "Redundancy" Index (GRI). He used two other Google search terms, namely "JSA" (Jobseeker's Allowance) and just "Redundancy" to be compared as a benchmark. In addition, eight explanatory variables were used as well as two factors derived from a principal component analysis, one without the GRI and one with it. Results showed a relative gain over the benchmark model in the one-step-ahead forecasting comparisons. However, factor-based models produced better results than individual models.

In a second step, the author elaborated a weekly model for predicting unemployment. He used a MIDAS regression to develop a model for updating unemployment forecasts. A weekly forecast was made thanks to the frequency of the Google data, then the monthly data was integrated once available. Google-derived indicators tended to record, relative to competing survey variables, a reduction in nowcasting errors in weeks 3 and 4 of an 8-week unemployment nowcast window. As in the previous works, summarizing variables information could be a good way to improve forecasting accuracy. Moreover, MIDAS regression seems to be useful to deal with different frequencies.

Borup and Schütte (2020) predicted US employment growth with Google Trends data from January 2001 to December 2018. The authors used a panel of 173 keywords derived from Google Trends platform. A method of targeted predictors was used, namely Elastic Net. Furthermore, Bagging, complete subset and Random Forests provided a large methodology and inference. Following, Bai and Ng (2008), Inoue and Kilian (2008), Elliott et al. (2013) and Breiman (2001), they established that internet search contains useful information about future employment growth. Moreover, Google-based models get accurate real-time information on future employment growth and labor market. However, the combination of the Google Trends variables tends to give better forecasting results. These results suggest that aggregated Google data has a predictive power in terms of employment growth forecasting.

As of now, there has been one study about unemployment forecasting in France. Fondeur

and Karamé (2013) examined the claimant count, categories A, B and C for continental France provided by the national employment agency, namely Pôle Emploi. They used a single keyword, "emploi" ("job" in English), from Google Trends in order to predict the claimant count as a proxy of French youth unemployment. The authors chose an unobserved component approach in according to the nature of the data (non-stationary, multiple frequencies). Out-of-sample exercise design induces a weekly calibration. They performed three types of prediction, the two-weeks-ahead, the one-week-ahead and a final nowcasts. They operated on the basis that the information of Google Trends and the number of jobseekers for the last month become available after the 24th of the current month. Results are consistent with the literature with an improvement in the accuracy of unemployment forecasts, the RMSE decreases by 40% on average. Model update method would be a great feature to take into account the availability of data step-by-step. Furthermore, the use of several explanatory variables with different machine-learning methods could be a good way to explore the predictive power of the number of jobseekers forecasting models.

We can draw some conclusions from this literature review (Tables 3.1 to 3.4). Firstly, Google Trends provide very interesting data because of its timing. In addition, Google data is available on a monthly and weekly basis, although only five years are available on a weekly basis. Furthermore, the representativeness of the queries from the Google search engine is no longer in question, as Google is the market leader in Internet search engines. However, we accept the idea that young people use Google more for their job search than older people, but this phenomenon tends to disappear as yesterday's young people become today's old.

Secondly, the research on unemployment forecasting demonstrated the good performance of the Google index. However, even if Google data performs well, it is still complementary data, in other words, Google data is important but not sufficient. Thus, combining Google data with other data is often a good idea and gives significant results. We will test the forecasting performance of Google by comparing it to the AR reference model commonly used in the literature. We used an ARX model with the number of end-of-month jobseekers collected by Pôle Emploi as the dependent variable. Then, as part of a more general research, we will observe whether Google data will be retained in variable selection methods for forecasting purposes.

### 3.3 Methodology

Our objective is short-term forecasting. In the literature, linear regression models give very good results in terms of mean squared errors. Moreover, OLS estimation should provide good forecasts. We place ourselves in a rolling window scheme in order to conduct our out-of-sample analysis. The rolling window analysis is the method recommended in the literature when the data show a potential structural break. It consists of dividing the sample into two subsets, one to train the model and the other to test and evaluate our forecasts. Each forecast is calculated with the same number of observations. Then we move to the window until we have covered our test period.

The literature on forecasting economic variables (Stock and Watson, 2003; D’Amuri and Marcucci, 2017; Nagao et al. 2019) adopt in most cases a standard autoregressive model with explanatory variables as follows :

$$y_t = \beta_0 + \sum_{i=1}^p \beta_{1i} y_{t-i} + \sum_{i=1}^q \beta_{2i} x_{t-i} + \epsilon_t, \quad t = 1, 2, \dots, T. \quad (3.1)$$

where  $y_{t-i}$  and  $x_{t-i}$  denote the dependent variable for the difference in value of the month  $t-i$  and the explanatory variables, respectively, while  $\epsilon_t$  denotes the error term. In our case, we implement the same model but with one difference. In contrast to D’Amuri (2009), we do not use the lags of the explanatory variables, because we think there is no selection bias, or at least that it is minimal. Indeed, the DEFM of month  $t$  is counted between the 28<sup>th</sup> of month  $t$  and the 15<sup>th</sup> of month  $t+1$ , namely the update period. Thus, the Google data of month  $t$  is relative to the queries of the unemployed in the current month. We believe that shifting the explanatory variables in the regression is not relevant. We have nevertheless tested several autoregressive models with distributed lag *via* the *auto\_ardl* function of the R package ”**ARDL**”. The result of the lag selection by minimising the BIC criterion shows no lag for the explanatory variables. This reinforces our intuition to keep a model of the form ARX.

$$y_t = \beta_0 + \sum_{i=1}^p \beta_{1i} y_{t-i} + \sum_{s=1}^S \beta_{2s} x_{t-i} + \epsilon_t, \quad t = 1, 2, \dots, T. \quad (3.2)$$

where  $x_t$  represent  $S$  covariate from 1 to  $S$ . Here, we use a single covariate in each regression. In practice, we estimate a rolling window of 149 data points and forecast 42 points ( $h=1$ ) from July 2016 to December 2019. We also make two and three-step-ahead forecasts, so we forecast 41 and 40 points, respectively. In a simple way, to predict the variable of interest in  $t+1$ , the explanatory variable in  $t+1$  must be entered in the model. Therefore, we constructed a Google data series from the weekly data. Thus, each forecast is made using information from the first two weeks of the month  $t + 1$ . This corresponds to the weeks of the update period described above. In addition, we have also used the information of the whole month in order to compare our forecasts. The benchmark model is the baseline an AR( $p$ ) defines as follows :

$$y_t = \beta_0 + \sum_{i=1}^p \beta_{1i} y_{t-i} + \epsilon_t, \quad t = 1, 2, \dots, T. \quad (3.3)$$

where the best  $p$  is chosen according to the Bayesian Information Criterion (BIC) in both models, from an interval  $1 \leq p \leq 4$ . The mean squared error (MSE) is the common measure to assess the accuracy of predictive models.

$$MSE_i = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2,$$

where  $y_t$  et  $\hat{y}_t$  are dependent variable observations and forecasts, respectively. We compare the performance of the competing models using the out-of-sample  $R^2$  (Campbell and Thompson, 2007).

$$R_{OoS}^2 = 1 - \frac{\sum_t ((y_t - \hat{y}_{m,t})^2)}{\sum_t (y_t - \hat{y}_t)^2}$$

A positive  $R_{OoS}^2$  means that the competing model is more accurate than the reference AR model. Moreover, we use the Diebold and Mariano (1995) test of equal predictive accuracy. Let the associated forecast errors be  $e_{0,t}$  and  $e_{i,t}$ . The time- $t$  loss function will be a direct

function of the forecast error  $g(e_{0,t})$  and  $g(e_{i,t})$ . In addition, let  $d_t = g(e_{0,t}) - g(e_{i,t})$  is the loss differential. The null hypothesis in DM test is  $H0 : E[dt] = 0$ , with the test statistic is :

$$DM_{o,i} = \frac{\bar{d}_t}{\sqrt{\frac{2\pi\hat{f}_d(0)}{T}}}$$

where  $\bar{d}_t = \frac{1}{T} \sum_{t=1}^T [g(e_{0,t}) - g(e_{i,t})]$  and  $\sqrt{\frac{2\pi\hat{f}_d(0)}{T}}$  is a consistent estimate of the standard deviation of  $\bar{d}_t$ . We follow the literature and choose a quadratic loss function for L, so that our tests of forecast accuracy in fact compares the MSE of the different models. Harvey, Leybourne and Newbold (1997) proposed a modified DM test (MDM test) for small sample size. The test is based on the variance of the unbiased estimator  $\bar{d}_t$ , which gives them the following modified statistics :

$$MDM_{o,i} = \left( \frac{T + 1 - 2h + T^{-1}h(h - 1)}{T} \right)^{\frac{1}{2}} DM_{o,i}$$

where  $h$  is the horizon of prediction and  $T$  denotes the total sample size. We perform this test in our results. Clark and West (2007) suggest a new test of equal predictive accuracy for nested models define as follows :

$$CW_{o,i} = \sqrt{P} \frac{\bar{d}_{o,i}}{\hat{\sigma}_{\bar{d}_{o,i}}},$$

where  $\bar{d}_{o,i} = P^{-1} \sum_{t=R}^T e_{o,t}^2 - e_{i,t}^2 + (\hat{y}_{o,t} - \hat{y}_{i,t})^2$  is the adjusted mean squared error differential, and  $\hat{\sigma}_{\bar{d}_{o,i}}$  is a consistent estimate of the standard deviation of  $\bar{d}_{o,i}$ . As our competing models nest the reference model, we will use the Clark and West (2007) statistic to test the equality of forecast accuracy.

Following Welch and Goyal (2008) and D'Amuri and Marcucci (2017) we assess the stability of predictive accuracy with the cumulative sum of squared error difference (CSSSED). This allows us to check whether the competing model beats the benchmark on the whole sample and not just on a subset of the sample.

$$CSSED_{i,m} = \sum_{i=1}^n [(y_i - \hat{y}_i)^2 - (y_i - \hat{y}_{i,m})^2],$$

A positive *CSSED* means that the alternative model is outperforming the benchmark.

### 3.4 Data

Google Trends provides a time series of the search intensity for a search term or for a topic with several search terms that are consistent with the selected topic. The search terms display matches for all the terms in our query, in the specified language. Topics are a group of terms that share the same concept in any language. Google Trends data is available in near real time and monthly frequency. Since November 2018 only the last five years are available in weekly frequency. Google Trends is a completely free service but the researcher is confronted with these changes of method and has to take into account the potential changes of method or availability of data. Google Trends provides an anonymized measure of the search intensity of a search term (or topic) by scaling to 100 the highest value for a given area and period.

$$S_{d,r} = \frac{V_{d,r}}{T_{d,r}}$$

where  $V_{d,r}$  is the number of web searches containing that keyword for a given area ( $r$ ) and day ( $d$ ).  $T_{d,r}$  is the total number of web searches performed through Google for the same area and day. The weekly index is calculated by a simple average of each day and the monthly index is also calculated by an average of each week of a given month. Last but not least, we collect the raw data on a single download because, although the index calculation differs slightly<sup>4</sup>, the time series are almost identical with a cross-correlation that is never below 0.99 (D’Amuri and Marcucci, 2017). We have two series from Google Trends from January 2004 to December 2021. The first series is related to the search term “emploi”, which was highlighted by Fondeur and Karamé (2013), particularly for French youth unemployment. The second series is related to the topic “emploi”.

The term “emploi” includes all internet searches with the keyword employment, such as

---

4. The Google index is calculated on the basis of a sample of IPs that changes over time (daily) and with the IP address.

”Pôle Emploi” but also ”Emploi du temps”. It should be noted that 24 of the 25 queries most associated with the term employment are related to job search and unemployment (appendix). The subject ”employment” is described by Google as a group of terms that share the same concept in any language. For example, the topic ”Emploi” captures searches that include the words ”job” or ”Indeed”.(Figure 3.1). In both cases, we believe that this data can provide important information to improve the forecasting of the unemployment rate. Moreover, we have recovered the last five years of our two series of weekly frequency, which will be useful to test the performance of our early forecasts. Thus, we have four forecasting schemes (Figure 3.2), the first uses the information from the first two weeks of the month to be forecast (GI 1w for the Google term set and GI 2w for the Google Topic set). A second, more traditional scheme uses all of the month’s information to make a forecast from the original monthly series (GI 3m for the Google term set and GI 4m for the Google Topic set). A third scheme uses the weekly average of the entire current month to make a forecast (GI 5f for the Google term set and GI 6f for the Google topic set); GI 9sa and GI 2sa correspond to the same data but the seasonality has been adjusted and GI 11check corresponds to a keyword unrelated to unemployment).

FIGURE 3.1 – DEFM and search intensity (seasonally adjusted) - Pôle Emploi and Google Trends

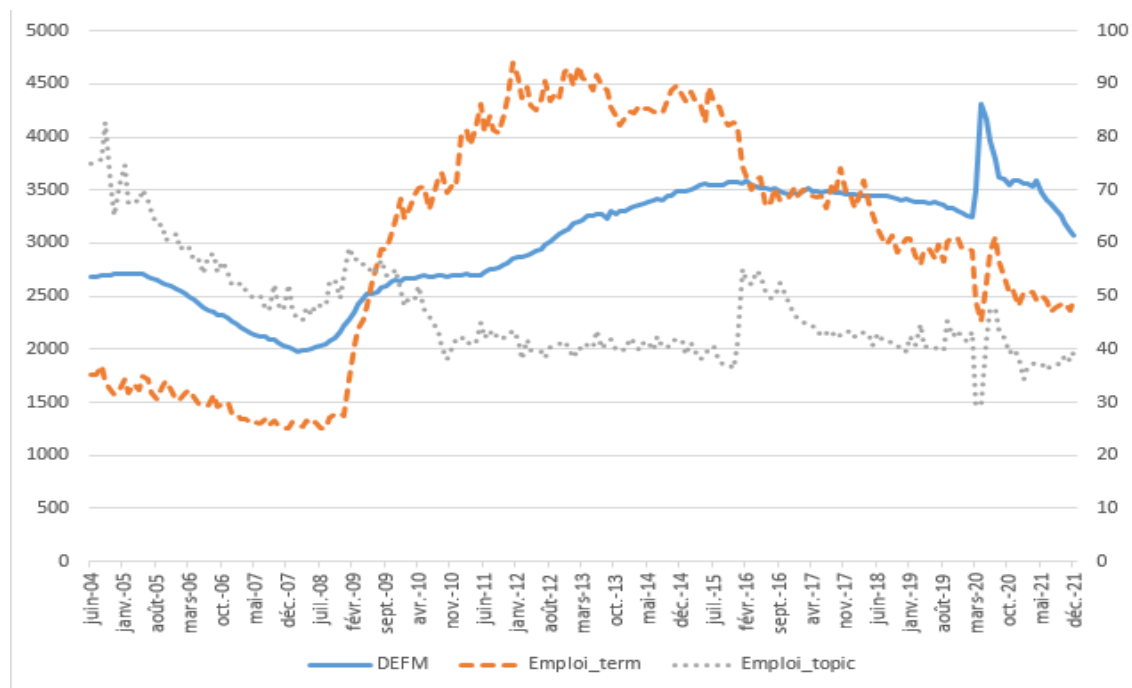
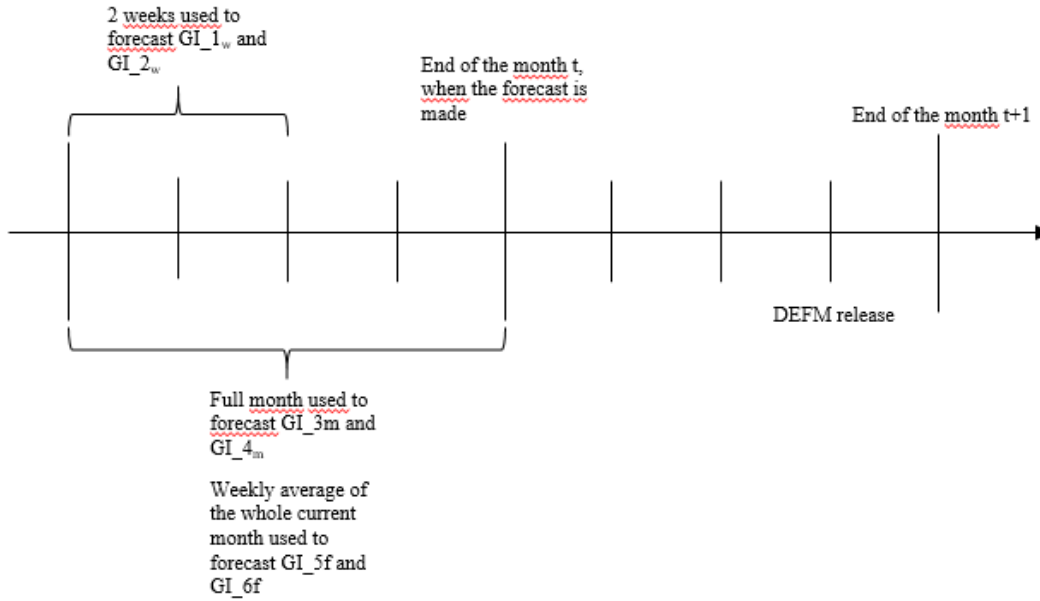




FIGURE 3.2 – Forecasting strategy - Authors



An observation of the series and its autocorrelation functions made it possible to define seasonality. The Friedman and Kruskal-Wallis tests confirmed this finding. Thus, Google Trends series for our keyword showed seasonality, which we corrected with the Tramo-Seat method for the Google index series of the search terms "emploi", the topic "emploi" as well as the weekly topic series. On the other hand, the weekly search term series was seasonally adjusted with X13-ARIMA. The choice of methods was made by minimising the AIC criterion.

To conclude this section, we perform stationary tests, namely the DF-GLS test, the KPSS test and Zivot and Andrews unit root test, which allows a break at an unknown point. Our results suggest that the Google-based data is not stationary, so we apply a first difference filter to achieve stationarity of our variables.

## 3.5 Empirical results

### 3.5.1 Nowcasting results

We compare the forecasting performance of several ARX models including Google Trends data with a benchmark AR model, traditionally used in the literature. As mentioned in the previous section, we evaluate the so-called "term" and "topic" Google data. In addition,

we test several forecasting schemes to determine whether the information readily available from Google provides an advantage in forecasting the DEFM. We use a rolling scheme of 155 observations to make 56 one-step-ahead forecasts (55 two-step-ahead and 54 three-step-ahead forecasts). Finally, we check whether there is a difference between stationary and non-stationary Google data. Results in terms of  $R_{OOS}^2$  are presented in Tables 4.10 to 4.15, with the associated CW test. The nowcasting performance is also evaluated during the non-crises (2017M5-2019M12) and crises (2020M1-2021M12) periods.

### **Full out-of-sample period (2017M5–2021M12)**

Table 5 highlights the performance of the Google variable in forecasting. The results in terms of  $R_{OOS}^2$  show contrasting performances. The "GI 4m" variable performs very well compared to the benchmark. The forecast with this variable at all three time horizons dominates all other models. Furthermore, the forecasts with the above variable are significantly different from the forecast with the reference model at 10% minimum error margin for  $h = 1$ , 10% for  $h = 2$  and 1% for  $h = 3$ . This means that Google's "topic" model performs better in predicting the DEFM. The topic variable captures a much wider range of keywords than just the term "emploi" and is therefore more useful for forecasting.

It should be noted that the other variables do not provide better forecasts than the AR model, except for one or two at the margin but the CW test indicates an equality of prediction performance between these models and the benchmark. Our robustness test, using a variable with a random keyword, namely "ampoule" is one of these exceptions. A better forecasting performance is only due to a statistical approximation of a localized outlier on the first lockdown. The following paragraph will justify this biased result.

### **Non-crisis and crisis out-of-sample period**

**Non-crisis period (2017M5–2019M12).** Only the variables "GI 3m", "GI 2w", provide MSEs smaller than the benchmark, i.e. positive ROOS, at horizon 1, 2 and 3, respectively. This time, it is the Google "term" variable (GI 3m) that becomes the most relevant for forecasting the DEFM. However, using only the information from the first two weeks of the current month (GI 2w) appears to perform better in terms of MSE. Without doubt, the test variables do not provide better forecasts than the AR model. Furthermore, our robustness

test (GI 11check) does not perform well over the whole forecast horizon.

**Crisis period (2020M1–2021M12).** The number of forecasts is a bit short but we can still present some important results. We find the "GI 4m" variable as the best variable for forecasting the DEFM at all horizons. Moreover, the test variables do not beat the benchmark, with the exception of our robustness test "GI 11check", "GI 7raw" and "GI 8raw" at the horizon  $h = 1, 2, 3$  and the horizon  $h = 2$ , respectively. This means that the March 2020 outlier disturbs the forecast somewhat. The CW test does not reject the null hypothesis of equality of prediction in the majority of models, with the exception of the GI 6f model at horizon  $h=1, 2$ .

**In conclusion, what can we learn from our results ?** In the extreme majority, stationary variables predict better than (non-stationary) test variables. The before/after decomposition of the crisis period is necessary to realise the prediction instability generated by the containment outlier. The CSSED graphs in the appendices confirm the instability of the forecasts due to the effect of lockdown and Covid-19 on the explanatory variable from google Trends. This echoes a result by Nagoa et al. (2019) criticizing the relationship of the Google Trends (keyword : "job" and "job offer" in their case) variables with the U.S. unemployment rate. The authors presented correlation coefficients that changed signs depending on the period studied.

## 3.6 Conclusion

Following the literature, we chose to use data from Google Trends to forecast French job seekers. We selected the keyword "emploi" in reference to the work of Fondeur and Karamé (2013). However, we maintain that the Google Trends data represents an increasingly older population. Therefore, our dependent variable will not be disaggregated by age. Our study therefore focus on the category A collected by Pôle Emploi as a whole. Furthermore, we chose to seasonally adjust our variables and make them stationary. The autoregressive model augmented with explanatory variables used for our analysis is commonly used in the literature. Nevertheless, we transform this model by choosing not to lag our explanatory variables.

We test the predictive power of Google data from 2004 :06 to 2021 :12. The first one-step-ahead prediction is made from May 2017 until the last one in December 2021. The empirical results show a good performance of the Google Trends data in predicting the DEFM. We noticed, that the variable of Google Trends that has the best prediction performance is the variable related to the "subject" of the keyword "emploi" (GI 4m). On the other hand, forecasting with the complete information of the current month are more accurate for forecasting of DEFM in general. We broke down the forecasts into two distinct periods (before and after containment) because this generated a lot of instability in the data.

This decomposition was useful as it highlights the robustness variables, especially (GI 11check) which does not bring any forecasting improvement during the pre-crisis period while the model becomes "better" than the benchmark during the post-crisis period. However, the CW test statistic cannot reject the null hypothesis of equal forecasts. It is difficult to determine which Google variables perform better in forecasting, between those constructed from the "Term" or the "Subject". However, one thing is for sure, the stationary and seasonally adjusted Google variables appear to perform better. It is difficult to determine which Google variables perform better in prediction, between those constructed from "term" or "topic". The variables based on the "topic" seem to generate more accurate predictions.

In parallel to this paper, we are working on forecasting DEFM *via* a large number of soft and hard variables by following a variable selection strategy with machine learning methods. Google variables are retained in most cases and have the great advantage of being available very early. The information provided in an advanced way is therefore very useful before the variables with more informative power are available.

# Bibliographie

- [1] Bai J., Ng S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146, 304-317.
- [2] Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2), 255-265.
- [3] Askitas, N., & Zimmermann, K. (2009). Google Econometrics and Unemployment Forecasting (No. 4201). Institute of Labour Economics (IZA).
- [4] Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304-317.
- [5] Boucher, C., & Maillet, B. (2012). Prévoir sans persistance. *Revue économique*, 63(3), 581-590.
- [6] Borup, D., & Schütte, E. C. M. (2020). In search of a job : Forecasting employment growth using google trends. *Journal of Business & Economic Statistics*, 1-15.
- [7] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [8] Campbell, J.Y. & Thompson, S.B. (2007). Predicting excess stock returns out of sample : Can anything beat the historical average? *The Review of Financial Studies*, 21, 1509-1531.
- [9] Chadwick, M. G., & Sengül, G. (2015). Nowcasting the unemployment rate in Turkey : Let's ask Google. *Central Bank Review*, 15(3), 15.
- [10] Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1), 291-311.
- [11] Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1, 1-5.

- [12] Choi, H., & Varian, H. (2012). Predicting the present with Google Trends, *Economic Record*, 88, 2-9.
- [13] D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search queries data. University Library of Munich, Germany.
- [14] D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801-816.
- [15] Diebold, F. X., & Mariano, R. S. (1995). Comparing forecast accuracy. *Journal of Business and Economics*.
- [16] Elliot, G., Gargano, A. & Timmermann, A. (2013). Complete subset regression. *Journal of Econometrics*, 177, 357-373.
- [17] Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117-125.
- [18] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.
- [19] Hohenstatt, R., Käsbauer, M., & Schäfers, W. (2011). "Geco" and its potential for real estate research : Evidence from the US housing market. *Journal of Real Estate Research*, 33(4), 471-506.
- [20] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464.
- [21] Harvey, D. I., Leybourne S. J. & Newbold P. (1998). Test for forecast encompassing. *Journal of Business & Economic Statistics*, 16, 254-259.
- [22] Inoue, A., & Killian L. (2008). How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. *Journal of the American Statistical Association*, 103, 511-522.
- [23] Kholodilin, K. A., Podstawski, M., & Siliverstovs, B. (2010). Do Google searches help in nowcasting private consumption? A real-time evidence for the US. *KOF Swiss Economic Institute Working Paper*, (256).
- [24] Maas, B. (2020). Short-term forecasting of the US unemployment rate. *Journal of Forecasting*, 39(3), 394-411.

- [25] Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, 30, 103-109.
- [26] Pavlicek, J., & Kristoufek, L. (2015). Nowcasting unemployment rates with google searches : Evidence from the visegrad group countries. *PloS one*, 10(5), e0127084.
- [27] Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, 3(1), 1-6.
- [28] Soybilgen, B., & Yazgan, E. (2018). Evaluating nowcasts of bridge equations with advanced combination schemes for the Turkish unemployment rate. *Economic Modelling*, 72, 99-108.
- [29] Smith, P. (2016). Google's MIDAS touch : Predicting UK unemployment with internet search data. *Journal of Forecasting*, 35(3), 263-284.
- [30] Stock, J. H., & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460), 1167-1179.
- [31] Stock, J. H., & W Watson, M. (2003). Forecasting output and inflation : The role of asset prices. *Journal of Economic Literature*, 41(3), 788-829.
- [32] Suhoy, T. (2009). Query indices and a 2008 downturn : Israeli data (No. 2009.06). Bank of Israel.
- [33] Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data : Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, 92, 132-139.
- [34] Welch, I. & Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21, 1455-1508.
- [35] Wu, L., & Brynjolfsson, E. (2015). 3. The Future of Prediction : How Google Searches Foreshadow Housing Prices and Sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.

TABLE 3.1 – Literature review

Authors	Countries	Dependent variables	Standard variables	Google Trends Keywords	Period	Method	Frequency	Forecast horizon
Choi and Varian (2009)	U.S.A.	Initial Claims		- Jobs - Welfare and Unemployment (category)	2004 :M1- 2009 :M4	ARX(1)	Weekly	One-step-ahead 2009 :M1 – 2009 :M6
Askitas and Zimmermann (2009)	Germany	Unemployment rate NSA		- Unemployment office or agency - Unemployment rate - Personnel consultant - Most popular job search engines in Germany	2004 :M1- 2009 :M4	Error-correction model	Monthly	One-step-ahead 2005 :M2 – 2009 :M4
Suhoy (2009)	Germany	Unemployment rate		- An interpolation of job openings ratio (quarterly observed variable), regressed by the query index of human resources (monthly explanatory indicator)	2004 :M1- 2009 :M3	ARIMAX(2,2)	Monthly	One-step-ahead 2007 :M12 – 2009 :M3
D'Amuri (2009)	Italy	Unemployment rate	- Industrial production index - Employment expectation survey	- Job offers	2004 :Q1- 2009 :Q1	ARIMAX(1,1,0)	Quarterly	One to three-step-ahead 2007 :Q3 – 2009 :Q1

Notes :



TABLE 3.2 – Literature review

Chadwick and Sengül (2012)	Turkey	Non-agricultural unemployment rate NSA - Industrial production - The initial unemployment claims	- cv, cv examples, looking for an employee, job, looking for a job, people, who look for a job, I am looking for a job, finding a job, employment placement agency, job ad, job ads, işkur, unemployment, unemployment insurance, career, kariyer.net, kariyer net, kariyernet, personnel hiring, secretcv	2005 :M1- 2011 :M12	Bayesian Model Averaging	Monthly	One to three-step-ahead 2010 :M11 – 2012 :M2
Fondeur and Karamé (2013)	France	Claimants count cat. A, B, C NSA	- emploi	2004 :M1- 2011 :M2	Unobserved component models	Monthly	One and Two-week-ahead, last-week-ahead 2009 :M1 – 2011 :M7
Pavlicek and Kristoufek (2015)	Czech Republic, Hungary, Poland and Slovakia	Unemployment rate	- job - work	2004 :M1- 2013 :M12	ARX(3,6,12)	Monthly	One-step-ahead 2009 :M1 – 2014 :M11

TABLE 3.3 – Literature review

Smith (2016)	United Kingdom	3-month change unemployment	<ul style="list-style-type: none"> <li>- Permanent placement, Temporary Billings, Vacancies Index, Staff Availability Index</li> <li>- Consumer expectations for unemployment</li> <li>Bank of England Agents Survey (service and manufacturing)</li> <li>F1 &amp; F2 Derived factor based on survey-based indicators and the Google redundancy and CRI variables.</li> </ul>	<ul style="list-style-type: none"> <li>- redundancy - JSA</li> <li>2004 :M1-2014 :M11</li> <li>ARX and MIDAS</li> </ul>	Monthly	One-step-ahead 2009 :M1 – 2014 :M11
D'Annuri and Marcucci (2017)	U.S.A.	Unemployment rate	<ul style="list-style-type: none"> <li>- Initial Claims</li> <li>- Employment expectation</li> <li>- The current six-month-ahead consumer expectations</li> <li>- Index of economic policy uncertainty</li> </ul>	<ul style="list-style-type: none"> <li>- jobs</li> <li>2004 :M2-2014 :M12</li> </ul>	ARDL Monthly	One-step-ahead 2007 :M3 – 2014 :M2 Twelve-step-ahead 2008 :M2-2014M2

TABLE 3.4 – Literature review

Nagao, Takeda and Tanaka (2019)	U.S.A.	Unemployment rate	- jobs - job offer	2004 :M1- 2018 :M12	ARDL	Monthly	One-step-ahead 2009 :M2 – 2018 :M12
Borup and Montes Schütte (2020)	U.S.A.	Employment rate	- - 173 internet search terms	2004 :M1- 2018 :M12	Random forest + Elastic net	Monthly	h-step-ahead (h=1,3,6,9,12) 2008 :M2 – 2018 :M12
Maas (2019)	U.S.A.	Unemployment rate	- Initial Claims - Latent factor from FRED-MD database	2004 :M1- 2017 :M12	MIDAS regression models	Monthly	h-step-ahead (h=1,2,...,12) 2008 :M3 – 2017 :M12

TABLE 3.5 – Nowcast performance during the full out-of-sample period – 2017M5-2021M12 – stated value  $R_{OOS}^2(AR)$ .

Models	h = 1	h = 2	h = 3
GI 1w	0.021**	-0.009	-0.081
GI 2w	-0.013	-0.022**	-0.028***
GI 3m	0.003	-0.028	-0.095***
GI 4m	<b>0.054*</b>	<b>0.065*</b>	<b>0.042***</b>
GI 5f	-0.070	-0.100	-0.162
GI 6f	0.011	0.016	0.000***
GI 7raw	-0.007	0.007	-0.010*
GI 8raw	-0.011	0.009	-0.008***
GI 9sa	-0.020	-0.017	-0.055
GI 10sa	-0.022	-0.015	-0.050***
GI 11check	0.003	0.017	0.011

Notes :  $R_{OOS}^2(AR)$  is the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model. \*\*\*, \*\* and \* Significant at the 1%, 5% and 10% level, respectively, for the CW test.

TABLE 3.6 – Nowcast performance during the non-crisis period – 2017M5-2019M12 – stated value  $R_{OOS}^2(AR)$ .

Models	h = 1	h = 2	h = 3
GI 1w	-1.939	-2.114	-0.967
GI 2w	0.006	<b>0.077*</b>	<b>0.429***</b>
GI 3m	<b>0.014</b>	-0.070	0.307***
GI 4m	-0.016	-0.042	0.357***
GI 5f	-1.204	-1.305	-0.505
GI 6f	-0.013	-0.034	0.366***
GI 7raw	-0.105	-0.418	0.023*
GI 8raw	-0.102	-0.327	0.122***
GI 9sa	-0.126	-0.512	-0.055
GI 10sa	-0.113	-0.386	0.071***
GI 11check	-0.156	-0.587	-0.132

Notes :  $R_{OOS}^2(AR)$  is the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model. \*\*\*, \*\* and \* Significant at the 1%, 5% and 10% level, respectively, for the CW test.

TABLE 3.7 – Nowcast performance during the crisis period – 2020M1-2021M12 – stated value  $R_{OOS}^2(AR)$ .

Models	h = 1	h = 2	h = 3
GI 1w	0.025*	0.007	-0.072
GI 2w	-0.013	-0.027	-0.030
GI 3m	0.003	-0.025	-0.092
GI 4m	<b>0.054</b>	<b>0.067</b>	<b>0.045</b>
GI 5f	-0.067	-0.093	-0.162
GI 6f	0.011*	0.013*	-0.002
GI 7raw	-0.007	0.009	-0.011
GI 8raw	-0.011	0.010	-0.009
GI 9sa	-0.020	-0.016	-0.057
GI 10sa	-0.022	-0.013	-0.052
GI 11check	0.003	0.018	0.013

Notes :  $R_{OOS}^2(AR)$  is the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model. \*\*\*, \*\* and \* Significant at the 1%, 5% and 10% level, respectively, for the CW test.

# Appendix

The graphs in the appendices represent the CSSEDs of the best models (3.3 to 3.8). We have transcribed the CSSEDs for the whole period as well as for the period without crisis. For the post-crisis period, the CSSEDs are visible on the graphs of the whole period since the best models are identical.

FIGURE 3.3 – CSSED GI 4m full period  $h = 1$  - Authors

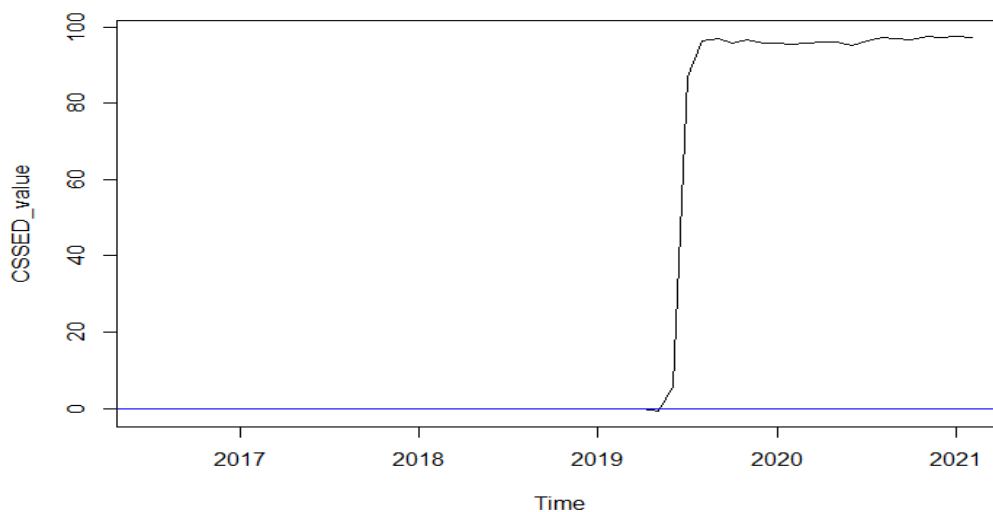


FIGURE 3.4 – CSSED GI 4m full period  $h = 2$  - Authors

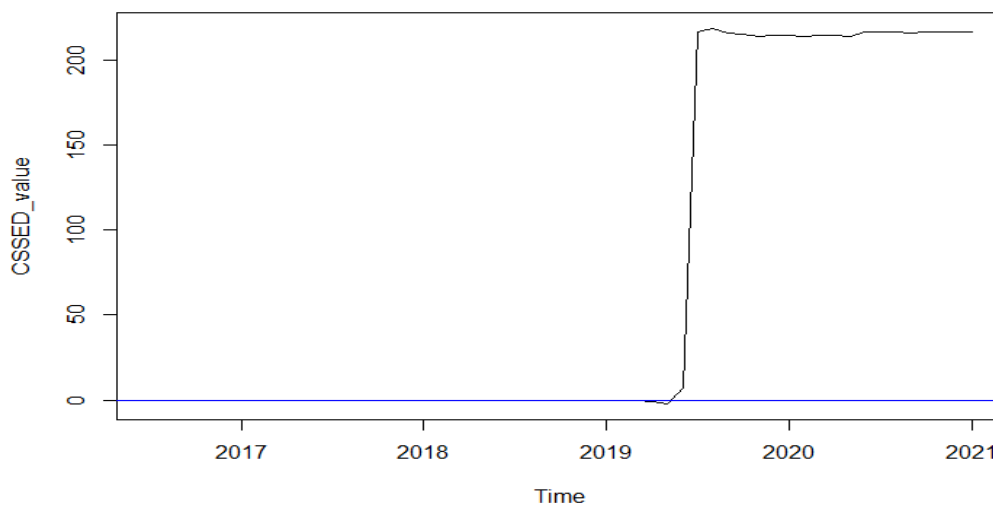


FIGURE 3.5 – CSSED GI 4m full period  $h = 3$  - Authors

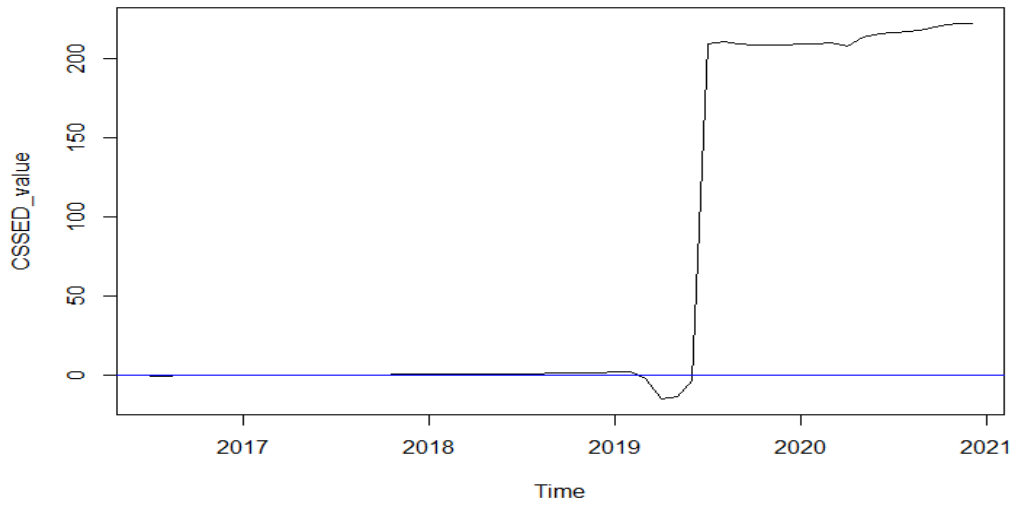


FIGURE 3.6 – CSSED GI 3m non-crisis period  $h = 1$  - Authors

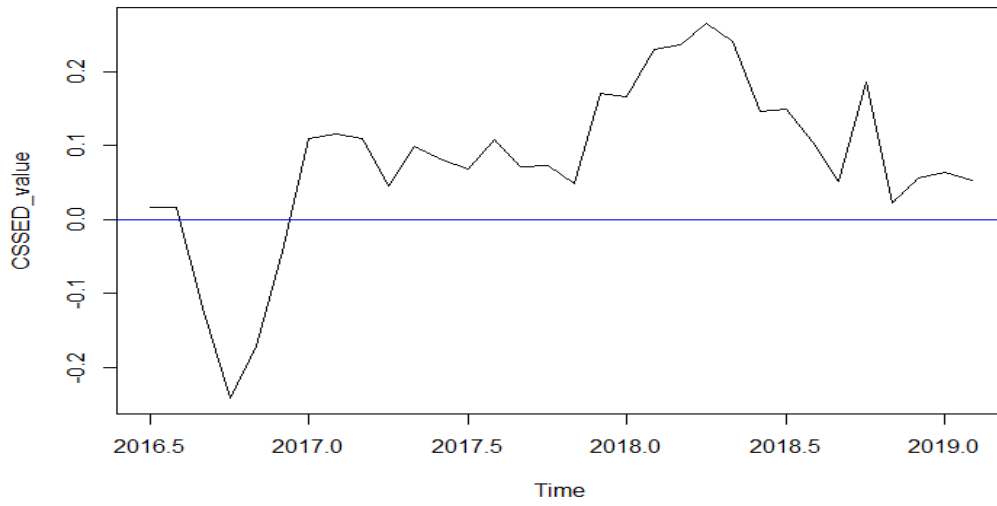


FIGURE 3.7 – CSSED GI 4m non-crisis period  $h = 2$  - Authors

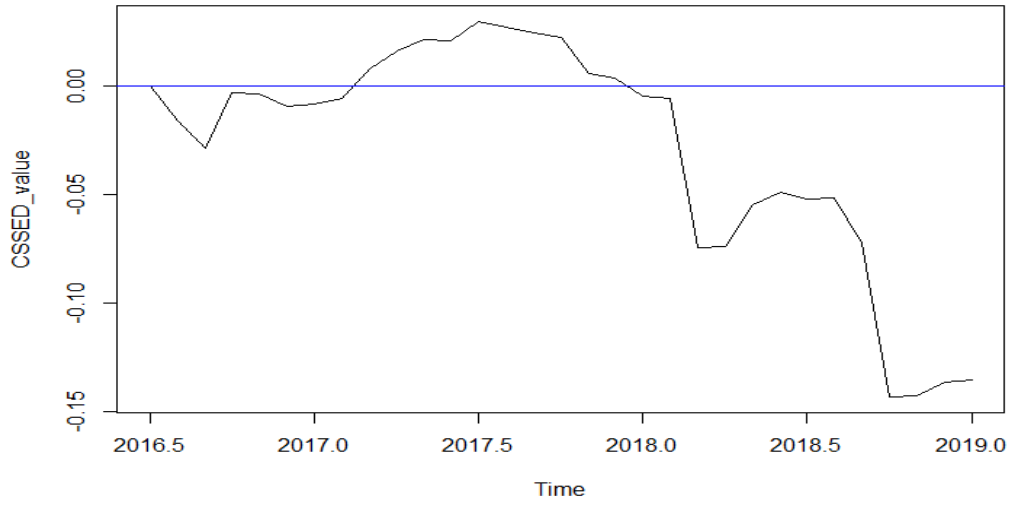
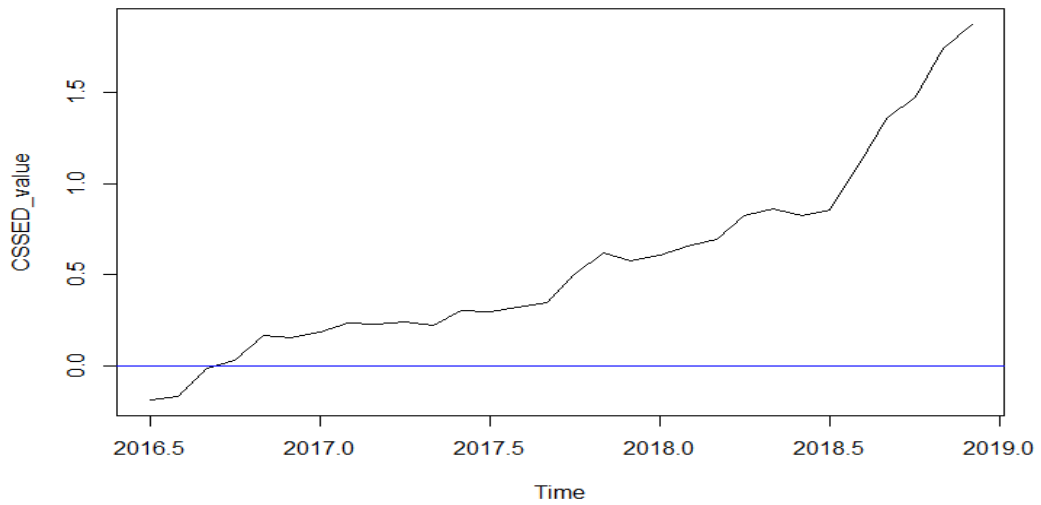


FIGURE 3.8 – CSSED GI 4m non-crisis period  $h = 3$  - Authors



# Chapitre 4

## Nowcasting French jobseekers using data reduction methods<sup>1</sup>

---

1. Ce chapitre est composé d'un travail co-écrit avec Olivier DARNE



## 4.1 Introduction

Unemployment is an important and widely analyzed indicator by policy-makers and analysts, to have an accurate and timely assessment of the state of the economy. Therefore, the accuracy of unemployment forecasts for the current month is important to provide a reliable and early analysis of the current labor market situation. However, monitoring unemployment is challenging because of significant publication delays. In France, the French national statistical institute (INSEE) publishes the unemployment rate on a quarterly basis with a time span of around 45 days after the end of the reference quarter. *Pôle Emploi*, the National Employment Agency in charge of allocating unemployment benefit and job placement, collects the number of French jobseekers at the end of the month (DEFM, *demandeurs d'emploi en fin de mois*) which is available on a monthly frequency with a lag of around 26 days after the end of the reference month.

Survey data have shown to be relevant predictors for unemployment forecasting (e.g., Abberger, 2007; Siliverstovs, 2012; Soybilgen and Yazgan (2018)). The main advantages are that they give a signal obtained directly from economic actors that reflects the short-term outlook of their own activity, sometimes of a forward-looking nature, and are characterized by very short publication lags and only very minor corrections. Recently, Google search data have shown to have predictive power for unemployment (e.g., Choi and Varian, 2009; Askitas and Zimmermann, 2009; Suhoy, 2009; D'Amuri, 2009; Chadwick and Şengül, 2012; Fondeur and Karamé, 2013; Nagao et al., 2019; Niesert et al., 2020; Simionescu and Cifuentes-Faura, 2022).

In this paper, to provide an economic interpretation, we propose a number of models designed to nowcast monthly DEFM growth, based on a large monthly dataset. The variable selection models are specified by using penalized regressions.<sup>2</sup> We focus on penalized regression methods, which is a generalization of ordinary least squares estimation, with an additional term that penalizes the size of regression coefficients. In doing so, it regularizes the model complexity, and avoids over-fitting that can cause the out-of-sample forecasting performance to deteriorate. Lasso (least absolute shrinkage and selection operator) regression method (Tib-

---

2. Considerable attention has focused on the forecasting of macroeconomic variables in a data-rich environment *via* the implementation of a variety of machine learning, variable selection and shrinkage methods (e.g., Kim and Swanson, 2014, 2018; Li and Chen, 2014; Smeekes and Wijler, 2018; Garcia, Medeiros, and Vasconcelos, 2017; Kotchoni et al., 2019; Borup et al., 2022; Goulet Coulombe et al., 2022).

shirani, 1996) is likely the most well known regularized regression method. However, Lasso regression has undesirable properties when  $N$  is greater than  $T$  or when there is a group of variables amongst which all pairwise correlations are very high (Zou and Hastie, 2005). Otherwise, Lasso can generate a selection bias between highly correlated variables, leading to arbitrary selection and under-representation of important variables. For that, we employ various penalized regressions to solve the variable selection problem of Lasso, namely adaptive Lasso (Zou, 2006), Elastic Net (EN, Zou and Hastie, 2005), SCAD (Fan and Li, 2001), adaptive EN (Zou and Zhang, 2009) and adaptive SCAD (Xiao and Xu, 2015). One of the advantages of these penalized regressions is to display a good trade-off between predictive accuracy and model interpretability.<sup>3</sup> We also use a more machine learning approach with the random forests (Breiman, 2001) which is a highly nonlinear specification. As factor models are an interesting tool for short-term forecasting of real activity by summarizing the information contained in large data sets and extracting a few common factors that can then be employed to nowcast DEFM growth, we complete this analysis by dynamic factor models. Finally, we also take into account the timing of the availability of data by proposing nowcast models according to their timing. We thus use two databases, where the first database, named small database, is available around one month before the publication of the DEFM, consisting of the Google Trend, INSEE survey and financial data. In the second database, named large database, which is available two weeks before the publication of the DEFM we add the OECD and BDF survey data.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 presents the methodology of the penalized regressions, random forest and dynamic factor models, and the nowcasting scheme is described in Section 4. Section 5 discusses the results of the variable selection and the out-of-sample nowcasting exercise. Section 6 concludes.

## 4.2 Data

The variable of interest is the monthly number of jobseekers from claimant count (“*Demandeurs d’Emploi en Fin de Mois*”, category A for continental France, DEFM hereafter) collected by the National Employment Agency (*Pôle Emploi*) in charge of unemployment

---

3. Interpretability is an advantage of penalized regressions over machine learning methods and factor models while predictive accuracy is in general more an advantage of penalized regression over OLS methods but not (systematically) over machine learning techniques and factor models.

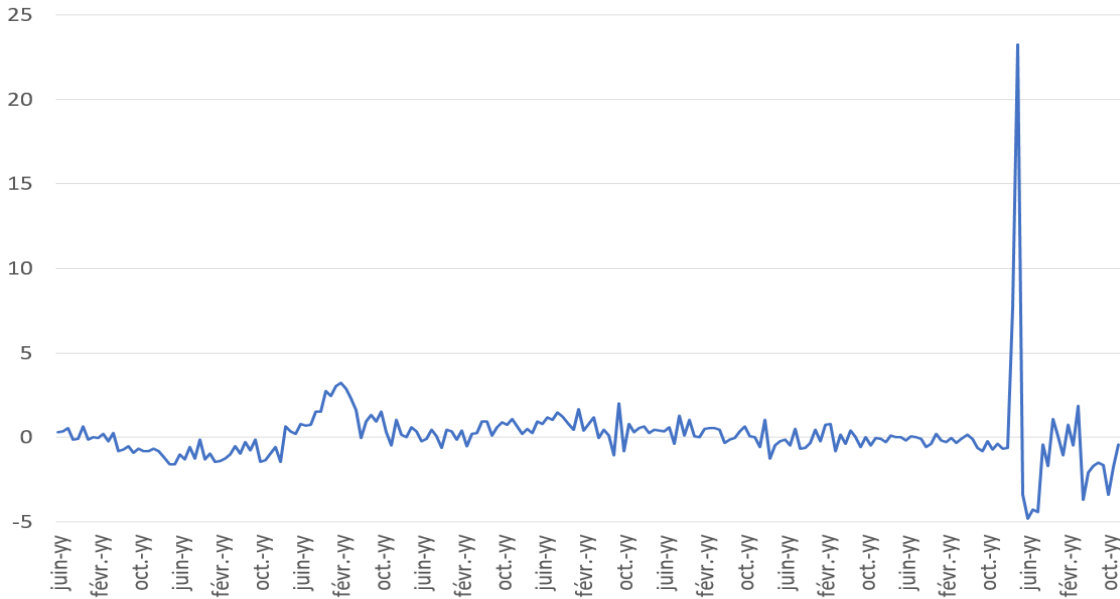
compensation and jobseeker assistance. This variable describes the inventory of unemployed people at the end of each month. The data set includes a large set of monthly indicators (108 predictors, tables 4.1 to 4.3) selected on the basis of their reliability and timeliness covering the period of 2004M6 to 2021M12. These data are based on the surveys on the manufacturing and services sectors, financial data and economic policy uncertainty indicator. The financial data concern stock market and short- and long-term interest rates and we also include the economic policy uncertainty indicator for France proposed by Baker et al. (2016). The Google Trend variables are defined as in Cariou (2022). The business survey data are mainly produced by the Banque de France (BdF), the French national statistical institute (INSEE), and the IHS Markit. They provide information which is different due to the fact, for example, they interview a different sample of businesses over a different period. The BdF and INSEE surveys are surveys of opinions, intentions or expectations, based on ten and six questions, respectively, while the PMI surveys ask questions relate to key variables such as output, new orders, prices and employment (nine questions). For the Bdf survey answers refer to the economic situation on the previous month, while for the INSEE survey answers refer to the economic situation over the recent period (usually three months) including the current month. In any given month the first available data are the business surveys published by the ISH Markit and the INSEE, around 18 and 25 days after the start of the month in question, respectively, whereas for the BdF is around 45 days. We also add survey data from the Organization for Economic Co-operation and Development (OECD) and the European Commission (EC).

We construct two different monthly databases, which we called small and large databases. The small database consists of variables available in the month of interest, namely the INSEE and ISH Markit surveys as well as the financial data. For the large database, we add the Bdf, OECD and EC surveys to the small database, which are available after the end of the reference month. We also consider four autoregressive terms. All variables are transformed to achieve stationarity. For penalized regressions and random forest variables are also standardized to have mean 0 and unit standard deviation to ensure that the shrinkage is fair and to prevent the scale of the variables affecting the estimation.

Figure 5.2 shows that the French jobseekers have known a large rise during the Covid-19 Crisis, with a sharp increase of 23.3% in April 2020. It is thus possible that some models

are more adequate during crisis periods and other models during (stable) non-crisis periods. Therefore, we will also evaluate out-of-sample performances during both periods.

FIGURE 4.1 – French jobseeker growth rate - 2004M6-2021M12.



## 4.3 Methodology

### 4.3.1 Penalized regressions

We estimate several shrinkage estimators for linear models with the estimated coefficients given by

$$\hat{\beta}_{=\beta_0, \dots, \beta_p} \left\{ \frac{1}{2} \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + P(\lambda; \beta) \right\} \quad (4.1)$$

where  $P(\lambda; \beta)$  is a penalty function that depends on the penalty parameter (or tuning parameter)  $\lambda$ . We consider different choices for the penalty functions as described below.

**Lasso regression** : Lasso (*least absolute shrinkage and selection operator*) regression, which was introduced by Tibshirani (1996), is characterized by an  $L_1$  penalty function, allowing for sparsity :

$$P^{Lasso}(\lambda; \beta) = \lambda \sum_{j=1}^p |\beta_j| \quad (4.2)$$

Lasso regressions tend to produce estimated regression coefficients that are exactly zeros, and thus can be used for variable selection, where only predictors with nonzero estimates are considered to be important (sparse solution). However, Lasso regression has undesirable properties when  $N$  is greater than  $T$  or when there is a group of variables amongst which all pairwise correlations are very high (Zou and Hastie, 2005). Otherwise, Lasso can generate a selection bias between highly correlated variables.

**Elastic Net** : Zou and Hastie (2005) develop the Elastic Net (EN) regularization method, which is a generalization including Lasso and Ridge as special cases, to avoid collinearity. The EN method simultaneously carries out automatic variable selection (as in Lasso by setting some coefficients to zero) and continuous shrinkage (as in Ridge), *via* a (convex) combination of both  $L_1$  and  $L_2$  penalty functions :

$$P^{EN}(\lambda; \beta) = \lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|) \quad (4.3)$$

where  $\alpha \in [0; 1]$ . If  $\alpha = 0$  or  $\alpha = 1$  the EN regression is equivalent to Lasso or Ridge regression, respectively. An advantage of the EN lies in its property of grouped selection, that is, the ‘grouping effect’. In other words, a group of highly correlated variables tend to have coefficients of similar magnitude and be selected in the same time.

However, the Lasso and EN methods fail to achieve the oracle property because they estimate the larger nonzero coefficients with asymptotically non-ignorable bias (Zou and Hastie, 2005). The oracle property means that when the true parameters have some zero components, the regularization parameters are estimated as 0 with probability tending to 1, and the nonzero components are estimated as well as when the correct submodel is known.<sup>4</sup>

**Adaptive Lasso regression** : Zou (2006) proposes adaptive Lasso (aLasso) to solve the variable selection problem of Lasso, showing that the inclusion of some additional information regarding the importance of each variable could considerably improve the results. aLasso uses the same penalty as Lasso with the inclusion of a weighting parameter that comes from a

---

4. Fan and Peng (2004) highlighted the importance of studying the validity of the oracle property.

first-step model that can be Lasso, Ridge or even OLS :

$$P^{aLasso}(\lambda; \beta) = \lambda \sum_{j=1}^p \widehat{\omega}_j |\beta_j| \quad (4.4)$$

where  $\widehat{\omega}_j = |\widehat{\beta}_j^*|^{-\gamma}$ , with  $\gamma > 0$  and  $\widehat{\beta}_j^*$  are the coefficients from the first-step model.

**Adaptive Elastic Net** : Zou and Zhang (2009) extend aLasso to adaptive EN (aEN) that combines the strengths of the quadratic regularization and the adaptively weighted Lasso shrinkage. Under weak regularity conditions, they establish the oracle property of the aEN.<sup>5</sup>

$$P^{aEN}(\lambda; \beta) = \lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) \widehat{\omega}_j |\beta_j|) \quad (4.5)$$

where  $\widehat{\omega}_j = |\widehat{\beta}_j|^{-\gamma}$  are the adaptive weights, with  $\gamma > 0$ , and  $\widehat{\beta}_j = \widehat{\beta}_j^{EN}$  are the EN estimators from the first-step model. If  $\alpha = 0$  then the aEN reduces to the aLasso.

**SCAD regression** : Fan and Li (2001) propose a non-concave penalty function referred to as the smoothly clipped absolute deviation (SCAD) which penalizes small coefficients more severely than Lasso and penalizes large coefficients less severely, implying more parsimonious models and bias reduction, respectively. Therefore, the SCAD estimator has the desirable properties, including unbiasedness, sparsity, and continuity as well as the oracle property (Fan and Peng, 2004).

The SCAD estimator is given by

$$\widehat{\beta}_{scad} =_{\beta_0, \dots, \beta_p} \left\{ \frac{1}{2} \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \sum_{j=1}^p P^{scad}(\lambda; |\beta_j|) \right\} \quad (4.6)$$

---

5. They show by simulations that the adaptive EN deals with the collinearity problem better than the other oracle-like methods (aLasso and SCAD), thus displaying much improved finite sample performance.

where  $p_\lambda^{scad}(\beta)$  is the penalty function defined on  $[0, \infty)$ , with  $a > 2$ , by

$$P^{SCAD}(\lambda; \beta) = \begin{cases} \lambda|\beta| & \text{si } |\beta| \leq \lambda \\ \frac{(2a\lambda|\beta|) - \beta^2 - \lambda^2}{2(a-1)} & \text{si } \lambda < |\beta| \leq a\lambda \\ \frac{\lambda^2(a^2+1)}{2} & \text{si } |\beta| > a\lambda \end{cases} \quad (4.7)$$

Xiao and Xu (2015) propose the adaptive version the SCAD regression, called aSCAD.<sup>6</sup>

### 4.3.2 Random forest

Random forest (RF) is an ensemble learning method based on classification and regression trees (Breiman, 2001).<sup>7</sup> RF allows to reduce the variance of regression trees and is based on bootstrap aggregation (bagging) of randomly constructed regression trees. This approach combines several randomized decision trees and aggregates their predictions by averaging.

To describe how a regression tree works, we follow the example discussed in Hastie et al. (2001). Consider a regression with only two predictors  $X = (X_1, X_2)$ , where each predictor  $X_i$  takes values in some given interval, and  $Y$  is the dependent variable. We first split the space into two regions at  $X_1 = s_1$ , the threshold value gives the minimum sum of squared residuals. Then the region to the left (right) of  $X_1 = s_1$  is split at  $X_2 = s_2$  ( $X_1 = s_3$ ), where  $s_2$  ( $s_3$ ) is again determined by the smallest sum of squared residuals. Finally, the region to the right of  $X_1 = s_3$  is split at  $X_2 = s_4$  (see the right plot of Figure 5.3). The final result is a partitioning into five regions  $R_k$ ,  $k = 1, \dots, 5$ , suggested to estimate the conditional expected value (prediction) of  $Y$  (see the left plot of Figure 5.3). Each region corresponds to a terminal node of the tree.

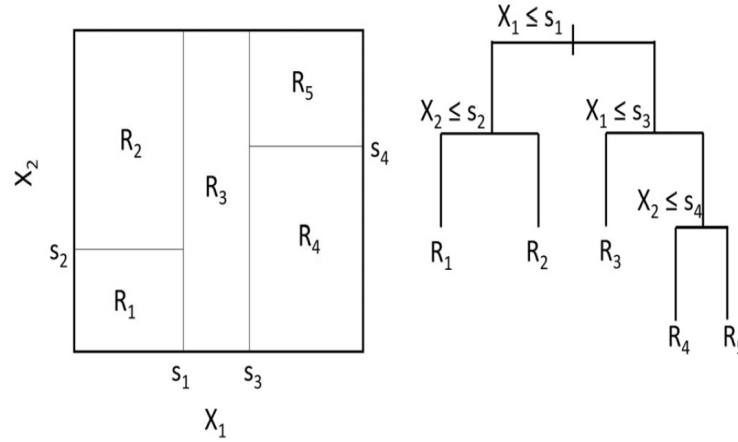
Given a dependent variable  $Y_{t+h}$ , a set of predictors  $X_t$  and a number of terminal nodes  $K$ , the splits are determined to minimize the sum of squared errors of the following regression

---

6. Xiao and Xu (2015) develop the multi-step aEN and aSCAD, a multi-step estimation algorithm built upon aEN regularization for reducing the false positives in high-dimensional variable selection while still maintaining the estimation accuracy. We applied this method and the variable selection is similar to that of aEN and aSCAD for our dataset.

7. A review of the RF algorithm can be found in Biau and Scornet (2016).

FIGURE 4.2 – Example of a regression tree (cf. Hastie et al., 2001)



model

$$Y_{t+h} = \sum_{k=1}^K c_k I_k(X_t; \theta_k),$$

where  $\theta_k$  is the set of parameters defining the  $k$ th region,  $c_k$  is a constant defined as node means, i.e.  $c_k = \sum_{j \in R_k} Y_j / N_k$ , with  $N_k$  the number of variables in the  $k$ th region, and  $I_k(X_t; \theta_k)$  is an indicator function such that

$$I_k(X_t; \theta_k) = \begin{cases} 1 & \text{if } X_t \in R_k(\theta_k) \\ 0 & \text{otherwise} \end{cases}$$

RF is a collection of regression trees, each specified in a bootstrap sample of the original data. Therefore, for each bootstrap sample  $b$ , with  $b = 1, \dots, B$ , a tree with  $K_b$  regions is estimated for a randomly selected subset of the original regressors.  $K_b$  is determined to leave a minimum number of observations in each region. The final forecast is the average of the forecasts of each tree applied to the original data

$$\hat{Y}_{t+h} = \frac{1}{B} \sum_{b=1}^B \left( \sum_{k=1}^{K_b} \hat{c}_{k,b} I_{k,b}(X_t; \hat{\theta}_{k,b}) \right)$$

It is possible to compute the importance of each variable for the RF models by using out-of-bag (OOB) samples, which are samples that are not included in the bootstrap samples. As



defined by Hastie et al. (2001) when the  $b$ th tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded. Then the values for the  $j$ th variable are randomly permuted in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable  $j$  in the random forest.

### 4.3.3 Dynamic factor model

In the factor model framework, variables  $X_t$ , are represented as the sum of two mutually orthogonal unobservable components : the common component  $\chi_t$  and the idiosyncratic component  $\xi_t$ . For a given  $t$ ,  $t = 1, \dots, T$ , the static factor model is defined by :

$$X_t = \Lambda F_t + \xi_t, \quad (4.8)$$

where  $X_t = [x_{1t}, \dots, x_{nt}]'$  is a vector of  $n$  stationary time series and it is assumed that the series have zero mean and covariance matrix  $\Gamma(0)$ ,  $\Lambda$  is the loading matrix such that  $\Lambda = [\lambda_1, \dots, \lambda_n]'$ , the common components  $\chi_t = \Lambda F_t$  are driven by a small number  $r$  of factors  $F_t$  common to all the variables in the model such that  $F_t = [F_{1t}, \dots, F_{rt}]'$ , and  $\xi_t = [\xi_{1t}, \dots, \xi_{nt}]'$  is a vector of  $n$  idiosyncratic mutually uncorrelated components, driven by variable-specific shocks.

Different approaches have been proposed to take dynamics into account in modelling the factors  $F_t$  from dynamic factor models (DFM).<sup>8</sup> Here, we use the static principal component analysis (PCA) suggested by Stock and Watson (2002) to estimate the factors  $F_t$ .<sup>9</sup> An eigenvalue decomposition of the estimated covariance matrix  $\hat{\Gamma}_0 = T^{-1} \sum_{t=1}^T X_t X_t'$  provides the  $(n \times r)$  eigenvector matrix  $\hat{S} = (\hat{S}_1, \dots, \hat{S}_r)$  containing the eigenvectors  $\hat{S}_j$  corresponding to the  $r$  largest eigenvalues for  $j = 1, \dots, r$ . The factor estimates are the first  $r$  principal components of  $X_t$  defined as  $F_t^{SW} = \hat{S}' X_t$ . To integrate dynamics in forecasting, SW propose an autoregressive model for the factors.

To take dynamics into account in modelling, another way to proceed is to model explicitly the dynamics of the factors  $F_t$ . More precisely, we assume that the DFM representation is

8. See Barhoumi, Darné, and Ferrara (2013) and Stock and Watson (2015) for a survey on DFMs.

9. We also applied other approaches of DFMs with the dynamic PCA in the frequency domain (Forni et al. (2005), the DFM based on a state-space representation (Doz et al., 2011) and the DFM estimated from quasi-maximum likelihood (Doz et al., 2012). The results have been lower than those obtained from the PCA model and are available from the authors upon request.

given by the following equation

$$X_t = A(L)F_t + \xi_t, \quad (4.9)$$

where the common components  $\chi_t = A(L)F_t$  integrate a linear dynamics where  $A(L)$  is a  $(n \times r)$  matrix describing the autoregressive form of the  $r$  factors. If we assume that there exists a  $(n \times q)$  matrix  $B(L)$  such that  $B(L) = A(L)N(L)$  with  $N(L)$  of dimension  $(r \times q)$ , then the dynamic factor is such that  $F_t = N(L)U_t$  where  $U_t$  is a  $(q \times 1)$  independent vector containing the dynamic shocks. It follows that the factor dynamics are described by

$$A(L)F_t = B(L)U_t \quad (4.10)$$

Equation (5.7) specifies a VAR( $p$ ) model for the factor  $F_t$  with lag polynomial  $A(L) = \sum_{i=1}^p A_i L^i$ .  $F_t$  is thus the  $(r \times 1)$  vector of the stacked factors with  $r = q(p + 1)$ .

## 4.4 Nowcasting scheme

To evaluate the forecasting performance, we also use the metric of out-of-sample (OOS)  $R^2$  given by

$$R_{OOS}^2 = 1 - \frac{MSE_{model}}{MSE_{benchmark}} \quad (4.11)$$

where  $MSE_{model}$  and  $MSE_{benchmark}$  are the MSE of the given and benchmark models (AR model or DFM), respectively. The given model is more accurate than the benchmark if its  $R_{OOS}^2$  is positive.

Obviously, simply comparing MSE-values does not take into account the sample uncertainty underlying observed forecast differences. Therefore, we also employed the testing procedure based on equal predictive ability (EPA) proposed by Harvey et al. (1997) with the modified Diebold-Mariano (MDM) test for small sample, which allows for pairwise comparison of forecast performances across models.<sup>10</sup>

Finally, we employ the model confidence set (MCS) procedure proposed by Hansen et al.

---

10. Due to the fact that the MDM test can be biased for nested models we have also applied the Clark-McCracken (2001) encompassing test (ENC-NEW). The results are given in Appendix B and show that we obtained the same findings than those with the MDM test.

(2011) to determine the set,  $\mathcal{M}^*$ , that consists of a subset of equivalent models in terms of predictive ability which are superior to the other competing models from a collection of models,  $\mathcal{M}_0$ . The MCS procedure yields a model confidence set,  $\widehat{\mathcal{M}}^*$ , that is a set of models constructed to contain the best models with a given level of confidence. This MCS allows getting several models displaying equivalent forecasting performance and therefore gives robustness to the forecasting exercise rather than to base the forecasting analysis only on one model. The  $t$ -statistic is defined as

$$T_{\max \mathcal{M}} = \max_{i \in \mathcal{M}} t_i \quad \text{with } t_i = \frac{\bar{d}_i}{\sqrt{\widehat{var}(\bar{d}_i)}} \quad (4.12)$$

where  $\widehat{var}(\bar{d}_i)$  denotes the estimate of  $var(\bar{d}_i)$ ,  $\bar{d}_i = m^{-1} \sum_{j \in \mathcal{M}} \bar{d}_{ij}$ , and  $\bar{d}_{ij} = n^{-1} \sum_{t=1}^n d_{ij,t}$ , with  $d_{ij,t} = L_{i,t} - L_{j,t}$  for all  $i, j \in \mathcal{M}_0$ , and  $L_{i,t}$  is a loss function (here MSE).<sup>11</sup> The  $t$ -statistic is associated with the null hypothesis  $H_{0,\mathcal{M}} : E(\bar{d}_i) = 0$  for all  $i \in \mathcal{M}$ , where  $\mathcal{M} \subset \mathcal{M}_0$ . If  $H_{0,\mathcal{M}}$  is accepted at level  $\alpha$  then the MCS is the set  $\widehat{\mathcal{M}}_{1-\alpha}^*$ .<sup>12</sup>

## 4.5 Empirical results

### 4.5.1 Variable selection

In the penalized regressions the tuning parameters are selected *via* cross-validation, which is a data-driven method that is designed to maximize the expected out-of-sample predictive accuracy. For the EN approach the parameter  $\alpha$  is set to 0.5 as in Medeiros et al. (2021) and is also chosen sequentially with  $0 < \alpha < 1$ .<sup>13</sup> For the adaptive version of the Lasso, EN and SCAD methods the estimators  $\widehat{\beta}_j$  in the first-step model – used to compute the adaptive weights  $\widehat{\omega}_j$  – are the estimates from the Ridge regression, and the parameter  $\gamma$  is set to 0.5.

Here we use the RF algorithm implemented in the `randomForest` package in R and the parameter `mtry` is estimated during the validation phase by implementing the `caret` package in R. The penalized regressions are implemented in the `glmnet` package.

---

11.  $\bar{d}_{ij}$  measures the relative sample loss between the  $i$ -th and  $j$ -th models, while  $\bar{d}_i$  is the sample loss of the  $i$ -th model relative to the average across the models in  $\mathcal{M}$ .

12. The MCS  $p$ -values are calculated using bootstrap implementation with 10,000 resamples (Hansen et al., 2011).

13. We test several values of  $\alpha$ , from 0.1 to 0.9 with a step of 0.1, and choose the one that optimizes the cross-validation criterion.

**Small database.** Tables 1, 3 and 4 display the most selected variables by the various methods for the small database. The variation of Google Trends variable (diff Google” relative to the ”term) is the predictor the most selected, in average, for the whole period and the non-crisis period. The autoregressive term with two lags is also very frequently selected. The INSEE surveys on Past trend and Expected trend in production and the PMI survey on Finished-goods inventory in the manufacturing sector are among the most frequently selected variables. Moreover, some models frequently chose predictors on the employment/work with Past trend and Expected trend in the workforce industry from the INSEE survey and the Backlogs of Work Index in the manufacturing and services sectors from the PMI survey.

**Large database.** Tables 2, 5 and 6 show that the Google Trends variable (diff Google) and the two-order autoregressive term remain very relevant and are the most selected predictor. In addition, we observe that when adding the BdF survey data most of the selected predictors come from the BdF survey as respect to the INSEE survey data, with the most frequently predictor is Expected employment in the service sector. This finding suggests that the BdF surveys seem to be more informative than the INSEE survey for jobseeker nowcasting. Note that it appears that the business surveys in the industry sector are more frequently chosen than those in the service sector.

#### 4.5.2 Nowcasting results

For the nowcasting evaluation we compare the various approaches to the AR model and the DFM. Although penalized regressions generally display good forecasting properties, there is evidence that performing an OLS on the group of variables selected can improve forecasting performance (Belloni and Chernozhukov, 2013). In this paper, both the models using the coefficients derived from penalized regressions and OLS-derived coefficients are tested. Results in terms of MSE are presented in Tables 4.10 and 4.11 as well as the  $R_{OOS}^2$  computed with the benchmark models and the associated DM test, and the MCS  $p$ -values with their ranks for the small and large database, respectively. The nowcasting performance is also evaluated during the non-crisis (2016M1-2019M1) and crisis (2020M1-2021M12) periods.

## Full out-of-sample period (2016M1–2021M12)

**Small database.** The results in terms of MSE show that the models based on coefficients derived from penalized regressions produce better MSE than those based on OLS-derived coefficients, where the smallest MSE is exhibited by the EN model where the parameter  $\alpha = 0.5$  is sequentially selected. This result is confirmed by the  $R_{OOS}^2$  because these models display a positive large value for the AR(1) benchmark and by the DM tests which are significant at the 5% level. Further, these penalized regressions also exhibit positive values of the  $R_{OOS}^2$  for the DFM benchmark, but they are very slightly positive. Lasso, Elastic-Net, Elastic-Net with  $\alpha = 0.5$  and adaptive Elastic-Net regressions exhibit MCS  $p$ -values upper to 0.90, giving thus similar forecasting performance, where EN has the rank 1. The other regularized regressions without OLS are in the  $\widehat{\mathcal{M}}_{50\%}^*$  as well as the RF model and the DFM, whereas the other models are in the  $\widehat{\mathcal{M}}_{90\%}^*$ .

**Large database.** Table 4.11 confirms the main results obtained from the small database where the Lasso, EN, EN0.5 and aEN models give the lowest MSE, positive values of the two  $R_{OOS}^2$  and being in the  $\widehat{\mathcal{M}}_{50\%}^*$ . The differences are that the aLasso is now included in the range of models displaying thus similar good forecasting performance, and the aEN model provides the smallest MSE. We can not observe a consensus on the fact that the MSE of all the models obtained from the small database are smallest than those using the large database. This result suggests that adding the data from the BdF survey seems interesting for nowcasting the French DEFM.<sup>14</sup>

To better see whether adding the BdF surveys to the small database can improve the nowcast performance, in Figure 4.5, we plot the cumulative sum of squared forecast error (CUMSFE) loss differential of the nowcasts provided by the EN and aEN models from the large database against those obtained by the EN and aEN models from the small database. The CUMSFE is given by

$$\text{CUMSFE}_{t,t+k} = \sum_{i=t}^{t+k} e_{i,M_1}^2 - e_{i,M_2}^2$$

where  $M_1$  is the benchmark, in our case the EN or aEN model estimated from the small

---

14. Note that the EN model seems to be improved when the parameter  $\alpha$  is chosen sequentially than to be set to 0.5 but not in a significant matter.

database, and  $M_2$  is the EN or aEN model estimated from the large database. A positive value of  $\text{CUMSFE}_{t,t+k}$  means that the benchmark displays larger squared forecast errors compared to the model estimated from the large database up to  $t+k$ , and negative values imply the opposite. We see that both models estimated from the small and large databases do not display significant difference until the Covid-19 crisis since their CUMSFE are very close to 0. We observe largest losses of additional BdF survey data during the Covid19 crisis. This can be explained by the fact that the BdF surveys was conducted after the lockdown measures came in 16 March (Banque de France, 2020) , whereas most responses of the INSEE surveys were collected before the lockdown (INSEE, 2020), implying that the BdF surveys in March could not well take into account the effects of the lockdown measures. The aftermath of the Covid-19 crisis is different for both models since the aEN model exhibits positive values suggesting that the nowcasts from the large database outperform those from the small database, whereas it is the opposite for the EN model.

We also provide time-series plots of the cumulative squared prediction error (CSPE) for the penalized models to assess their nowcasting performance over time. Figures 4.3 and 4.4 for the small and large databases, respectively, show that the CSPE of the different methods displays similar behavior, with a strong jump associated with the Covid-19 crisis. Now we analyze whether some models are more adequate during crisis periods and other models during non-crisis periods.

### **Non-crisis and crisis out-of-sample period**

**Non-crisis period (2016M1–2019M12).** For the small database (Table 4.12) the MSE produced by most of the models are very close, with values ranging between 0.233 and 0.325, except for SCAD, aSCAD, Lasso+OLS, EN+OLS and EN05+OLS. The best models display positive values of  $R_{OOS}^2$  for both benchmarks but without having significant DM statistics. These results are generally confirmed on the large database (Table 4.14), with a set of models having similar nowcasting performance, including the AR(1) model and the DFM. An interesting finding is that the RF model appears to be the more accuracy model for nowcasting the DEFM during the non-crisis periods, with both the small and large database. Finally, our results also suggest that the penalized regressions do not improve standard models, namely AR(1) model and DEFM, during non-crisis periods. Figure 4.6 confirms that the

behavior of the different models is similar during the non-crisis period for the small database.

**Crisis period (2020M1–2021M12).** During the Covid-19 crisis the Lasso, EN, EN05 and aEN models outperform the other models, especially both benchmarks and the models based on OLS-derived coefficients, being the only models in the  $\widehat{\mathcal{M}}_{10\%}^*$  and  $\widehat{\mathcal{M}}_{50\%}^*$  for the small (Table 4.13) and large (Table 4.15) databases, respectively. Figure 4.7 show that the four penalized models display better nowcasts than the two benchmarks during the crisis period, especially compared to the AR(1) model and in a least level to the DFM for the small database. This result suggests that these penalized regressions are clearly appropriate during this crisis period for nowcasting the French jobseekers.

## 4.6 Conclusion

In this paper we compared nowcasting performance of penalized regressions, random forest and dynamic factor model for nowcasting the monthly number of French jobseekers, using a large dataset. A rolling nowcast study was carried out to assess their nowcasting performance. The results showed that, the Lasso, Elastic-Net and adaptive Elastic-Net models exhibit a good nowcasting performance, and are thus relevant for nowcasting jobseeker growth rate. Their accuracy gains are large during the Covid-19 crisis whereas they provide similar performance with respects to the dynamic factor model during the non-crisis period. Among the predictors for nowcasting jobseeker growth it appears that the variation of Google Trend variable, Finished-goods inventory in the manufacturing sector from the PMI survey and the autoregressive term with two lags are important predictors due to the fact that they are selected by all the approaches. We also observe that when adding the BdF survey data most of the selected predictors come from the BdF survey as respect to the INSEE survey data, suggesting that the BdF surveys seem to be more informative for jobseeker nowcasting.

For practitioners, our findings suggest that when factor models and penalized regressions give similar performances the latter methods should be preferred since they also allow for interpretability and ability to identify the main drivers of the models.

This paper is focused on nowcasting the French jobseeker growth rate. Further research would be to design the models to forecast the number of jobseekers to a longer horizon.

Another extension could be to nowcast/forecast the jobseeker growth by restricting to the lags of the regressors instead of using the contemporaneous values. Further research can be focused on youth jobseekers for the age group 15–24 years in order to fully exploit the Google Trend variable because internet use is most likely affected by a generation bias (D’Amuri, 2009; Fondeur and Karamé, 2013). For the machine learning approaches, we used penalized regressions. We could also consider other approaches which allow nonlinear specification in future research, such as decision trees or support vector machine as suggested by Ahmad et al. (2021) and Gogas et al. (2022) for forecasting unemployment. Finally, in this data-rich environment, it would be interesting to compare our approaches with that of Schiavoni et al. (2021) based on a combination of PCA and state space models.



# Bibliographie

- [1] Ahmad M., Khan Y.A., Jiang C., Kazmi S.J.H., Abbas S.Z. (2021). The impact of COVID-19 on unemployment rate : An intelligent based unemployment rate prediction in selected countries of Europe. *International Journal of Finance and Economics*, 1-16.
- [2] Askitas N., Zimmermann K. (2009). *Google Econometrics and Unemployment Forecasting*. Working paper No. 4201, Institute of Labour Economics.
- [3] Bai J., Ng S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146, 304-317.
- [4] Banque de France (2020). Update on business conditions in France at the end of March 2020. April 2020.
- [5] Belloni A., Chernozhukov V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19, 527-547.
- [6] Bessec M. (2013). Short-term forecasts of French GDP : A dynamic factor model with targeted predictors. *Journal of Forecasting*, 32, 500-511.
- [7] Borup D., Christensen B.J., Mühlbach N.S., Nielsen M.S. (2020). Targeting predictors in random forest regression. Available at SSRN : <https://ssrn.com/abstract=3551557>.
- [8] Borup D., Schütte E.C.M. (2020). In search of a job : Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, forthcoming.
- [9] Breiman L. (2001). Random forests. *Machine Learning*, 45 (1), 5-32.
- [10] Chadwick M.G., Sengül G. (2015). Nowcasting the unemployment rate in Turkey : Let's ask Google. *Central Bank Review*, 15(3), 15.
- [11] Choi H., Varian H. (2009). Predicting initial claims for unemployment benefits. *Google Inc*, 1, 1-5.

- [12] D'Amuri F. (2009). Predicting unemployment in short samples with internet job search queries data. University Library of Munich, Germany.
- [13] D'Amuri F., Marcucci J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801-816.
- [14] Fan J., Li R. (2001). Variable selection *via* nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.
- [15] Fan J., Lv J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Serie B*, 70, 849-911.
- [16] Ferrara L., Simoni A. (2019). When are Google data useful to nowcast GDP? An approach *via* preselection and shrinkage. Working paper No. 717, Banque de France.
- [17] Fondeur Y., Karamé F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30, 117-125.
- [18] Fortin-Gagnon O., Leroux M., Stevanović D., Surprenant S. (2020). A large Canadian database for macroeconomic analysis. Working Paper No 2018S-25, CIRANO.
- [19] Gentzkow M., Kelly B., Taddy M. (2019). Text as data. *Journal of Economic Literature*, 57, 535-74.
- [20] Giannone D., Lenza M. and Primiceri G. (2021). Economic predictions with big data : The illusion of sparsity. *Econometrica*, forthcoming.
- [21] Gogas P., Papadimitriou T., Sofianos E. (2022). *Journal of Forecasting*, 41, 551-566.
- [22] Goulet-Coulombe P. (2020). To Bag is to Prune. arXiv preprint arXiv :2008.07063.
- [23] Goulet-Coulombe P. (2021). The macroeconomy as a random forest. arXiv preprint arXiv :2006.12724.
- [24] Goulet-Coulombe P., Leroux M., Stevanovic D., Surprenant S. (2019). How is machine learning useful for macroeconomic forecasting? Working Papers No 2019s-22, CIRANO.
- [25] Goulet-Coulombe P., Marcellino M., Stevanović D. (2021). Can machine learning catch the Covid-19 recession? *National Institute Economic Review*, 256, 71-109.
- [26] Goulet-Coulombe P., Leroux M., Stevanović D., Surprenant S. (2021). Macroeconomic data transformations matter. *International Journal of Forecasting*, 37, 1338-1354.
- [27] Hastie T., Tibshirami R., Friedman J.H (2001). *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, New York : Springer.

- [28] INSEE (2020). Economic outlook. April 2020.
- [29] Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2), 339-354.
- [30] Kock A.B., Medeiros M.C., Vasconcelos G. (2020). Penalized Time Series Regression. In Fukely P. (ed.) *Macroeconomic Forecasting in the Era of Big Data*, Springer.
- [31] Kotchoni R., Leroux M., Stevanovic D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34, 1050-1072.
- [32] Li R., Zhong W. and Zhu L. (2012). Feature screening *via* distance correlation learning. *Journal of the American Statistical Association*, 107, 1129-1139.
- [33] Liu J., Zhong W. and Li R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58, 1-22.
- [34] Masini R.P., Mendes E.F., Medeiros M.C. (2021). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, forthcoming
- [35] Maas B. (2020). Short-term forecasting of the US unemployment rate. *Journal of Forecasting*, 39(3), 394-411.
- [36] McCracken M.W., Ng S. (2016). FRED-MD : A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34, 574-589.
- [37] Medeiros M.C., Vasconcelos G.F.R., Veiga A., Zilberman E. (2021). Forecasting inflation in a data-rich environment : The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39, 98-119.
- [38] Nagao S., Takeda F., Tanaka R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, 30, 103-109.
- [39] Niesert R.F., Oorschot J.A., Veldhuisen C.P., Brons K., Lange R.-J. (2020). *International Journal of Forecasting* 36, 1163-1172.
- [40] Pan W., Wang X., Xiao W., Zhu H. (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114, 928-937.
- [41] Schiavoni C., Palm F., Smeekes S., van den Brakel J. (2021). A dynamic factor model approach to incorporate Big Data in state space models for official statistics. *Journal of the Royal Statistical Society, Series A*, 184, 324-353.

- [42] Schumacher C. (2010). Factor forecasting using international targeted predictors : The case of German GDP. *Economics Letters*, 107(2), 95-98.
- [43] Simionescu M., Cifuentes-Faura J. (2022). Forecasting National and Regional Youth Unemployment in Spain Using Google Trends. *Social Indicators Research*, forthcoming.
- [44] Smith P. (2016). Google's MIDAS touch : Predicting UK unemployment with internet search data. *Journal of Forecasting*, 35(3), 263-284.
- [45] Tibshirani R. (1996). Regression shrinkage and selection *via* the Lasso. *Journal of the Royal Statistical Society. Series B*, 58, 267-288.
- [46] Wager S., Athey S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228-1242.
- [47] Wager S., Walther G. (2016). Adaptive concentration of regression trees, with application to random forests. Working Paper, Stanford University. Available at arXiv : <https://arxiv.org/abs/1503.06388>
- [48] Zou H., Hastie T. (2005). Regularization and variable selection *via* the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320.
- [49] Zou H., Zhang H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37, 1733-1751.

TABLE 4.1 – Variables names and abbreviations - 1/3

Variables names	Abbreviations	Source
Demandeurs d'Emploi en Fin de Mois (growth rate)	DEFM	Pôle Emploi/DARES
Composite Leading Indicator	CLI OECD	OECD
Business Confidence Index	BCI OECD	OECD
Consumer Confidence Index	CCI OECD	OECD
Economic Sentiment Indicator	ESI DGECFIN	DGECFIN
Economic Sentiment Indicator Construction	IDC const DGECFIN	DGECFIN
Economic Sentiment Indicator Industry	IDC indus DGECFIN	DGECFIN
Economic Sentiment Indicator Retail	IDC com DGECFIN	DGECFIN
Economic Sentiment Indicator Consumers	IDC conso DGECFIN	DGECFIN
Economic Sentiment Indicator Services	IDC serv DGECFIN	DGECFIN
Monthly business sentiment indicator in market services	IMCA serv BDF	Banque de France
Monthly business sentiment indicator in manufacturing industry	IMCA indus BDF	Banque de France
Change in employment, compared with previous month	Evol eff ind BDF	Banque de France
Expected employment for the coming month	Prev eff ind BDF	Banque de France
Change in employment over next month	Evol eff serv BDF	Banque de France
Expected employment, compared with previous month	Prev eff serv BDF	Banque de France
Change in foreign orders, compared with previous month	Evol cmde indus BDF	Banque de France
Change in overall level of new orders, compared with previous month	$Evol_{cmd_i}ndus_BDF$	Banque de France
Change in deliveries, compared with previous month	Evol liv indus BDF	Banque de France
Change in commodity prices, compared with previous month	Evol pmp indus BDF	Banque de France
Current position in inventories of commodities	Stock mp indus BDF	Banque de France
Change in industrial producer prices, compared with previous month	$Evol_{pprod_i}ndus_BDF$	Banque de France
Expected industrial prices for coming month	Prev pprod indus BDF	Banque de France
Change in output, compared with previous month	Evol prod indus BDF	Banque de France
Expected production for the coming month	$Prev_{prod_i}ndus_BDF$	Banque de France
Change in inventories of final goods, compared with previous month	$Evol_{stock_i}ndus_BDF$	Banque de France
Current position in inventories of final goods	$Stock_i ndus_BDF$	Banque de France
Expected inventories of final goods for the coming month	$Prev_{stock_i}ndus_BDF$	Banque de France
Cash positions at end of month	Tresor indus BDF	Banque de France
End-of-month cash position SME	Tresor PME indus BDF	Banque de France
Average capacity utilisation rate	$TUC_i ndus_BDF$	Banque de France
Expected overall activity over next month	Prev act serv BDF	Banque de France
Change in foreign orders, compared with previous month	Evol dmndee serv BDF	Banque de France
Forecasts on foreign orders over next month	Prev dmndee serv BDF	Banque de France
Change in aggregate demand, compared with previous month	Evol dmnde serv BDF	Banque de France
Expected aggregate demand over next month	Prev dmnde serv BDF	Banque de France
Change in prices, compared with previous month	Evol prix serv BDF	Banque de France
Expected prices over next month	Prev prix serv BDF	Banque de France
Cash positions at end of month	Tresor serv BDF	Banque de France
Cash positions of SMEs at end of month	Tresor PME serv BDF	Banque de France

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.

TABLE 4.2 – Variables names and abbreviations 2/3

Variables names	Abbreviation	Source
Business climate summary indicator	ICA INSEE	INSEE
Summary indicator of overall economic situation reversal	IRC INSEE	INSEE
Employment climate summary indicator	ICE INSEE	INSEE
Business climate summary indicator industry	ICA indus INSEE	INSEE
Business climate summary indicator services	ICA serv INSEE	INSEE
Business climate summary indicator construction	ICA bat INSEE	INSEE
Business climate summary indicator retail	ICA com INSEE	INSEE
Expected trend in the workforce industry	Tendprev eff indus INSEE	INSEE
Past trend in the workshop industry	Tendpass eff indus INSEE	INSEE
Level of stocks of manufactured products industry	Stock prodfin INSEE	INSEE
Past trend in production industry	Tendpass prod INSEE	INSEE
Expected trend in production industry	Tendprev prod INSEE	INSEE
Probable trend in industrial production volume industry	Tendvol prod indus INSEE	INSEE
Probable trend in sale prices industry	Evol prix INSEE	INSEE
Probable trend in industrial price general level industry	Tend prix indus INSEE	INSEE
Past trend of activity services	Tendpass acti serv INSEE	INSEE
Expected trend of activity services	Tendprev act serv INSEE	INSEE
Expected trend of demand services	Tendprev dem serv INSEE	INSEE
Past trend of prices services	Tendpass prix serv INSEE	INSEE
Expected trend of prices services	Tendprev prix serv INSEE	INSEE
General prospects on activity services	Persp act serv INSEE	INSEE
Change in the situation of the enterprise over the last 3 months services	Evol entrep serv INSEE	INSEE
Expected trend of employment services	Tendprev eff serv INSEE	INSEE
Past trend of employment services	Tendpass eff serv INSEE	INSEE
Expected trend of employment construction	Tendprev eff bat INSEE	INSEE
Past trend of employment construction	Tendpass eff bat INSEE	INSEE
Change in the number of persons employed in the next 3 months	Evol eff com1 INSEE	INSEE
Change in the number of persons employed in the last 3 months	Evol eff com2 INSEE	INSEE
Level of global order books	Niv car indus INSEE	INSEE
Level of foreign order books	Niv care indus INSEE	INSEE
Current order books	Niv car cmd indus BDF	Banque de France
Business development (sales) in the next 3 months	Evol vente com INSEE	INSEE
Intents for orders in the next 3 months	Int cmd com INSEE	INSEE
Opinion on the future trend of unemployment	Op men chom INSEE	INSEE
Consumer price index year-on-year change	IPC gliss annuel INSEE	INSEE
CAC40 (rate)	rcac40	Banque de France
Long-term interest rate	taux long	Banque de France
Short-term interest rate	taux short	Banque de France
Economic Policy Uncertainty	EPU	Baker et al. (2016)
Keyword "emploi"	log Google	Google Trends
Keyword "emploi"	diff Google	Google Trends
Keyword "emploi"	log Google topic	Google Trends
Keyword "emploi"	Diff Google topic	Google Trends
Spread Long-Short	Spread LS	BdF - author

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.

TABLE 4.3 – Variables names and abbreviations -3/3

Variables names	Abbreviation	Source
Composite - Output Index	PMI <sub>FR</sub> COM <sub>OB</sub>	IHS Markit
Composite - New Orders Index	PMI <sub>FR</sub> COM <sub>NI</sub>	IHS Markit
Composite - Employment Index	PMI <sub>FR</sub> COM <sub>EI</sub>	IHS Markit
Composite - Backlogs of Work Index	PMI <sub>FR</sub> COM <sub>BW</sub>	IHS Markit
Composite - Input Prices Index	PMI <sub>FR</sub> COM <sub>IP</sub>	IHS Markit
Composite - Output Prices Index	PMI <sub>FR</sub> COM <sub>OP</sub>	IHS Markit
Manufacturing PMI Index	PMI <sub>FR</sub> MAN <sub>PM</sub>	IHS Markit
Manufacturing - Output Index	PMI <sub>FR</sub> MAN <sub>OB</sub>	IHS Markit
Manufacturing - New Orders Index	PMI <sub>FR</sub> MAN <sub>NI</sub>	IHS Markit
Manufacturing - New Export Orders	Index PMI <sub>FR</sub> MAN <sub>NE</sub>	IHS Markit
Manufacturing - Employment Index	PMI <sub>FR</sub> MAN <sub>EI</sub>	IHS Markit
Manufacturing - Backlogs of Work Index	PMI <sub>FR</sub> MAN <sub>BW</sub>	IHS Markit
Manufacturing - Quantity of Purchases Index	PMI <sub>FR</sub> MAN <sub>QP</sub>	IHS Markit
Manufacturing - Stocks of Purchases Index	PMI <sub>FR</sub> MAN <sub>SP</sub>	IHS Markit
Manufacturing - Stocks of Finished Goods Index	PMI <sub>FR</sub> MAN <sub>SF</sub>	IHS Markit
Manufacturing - Suppliers' Delivery Times Index	PMI <sub>FR</sub> MAN <sub>SD</sub>	IHS Markit
Manufacturing - Input Prices Index	PMI <sub>FR</sub> MAN <sub>IP</sub>	IHS Markit
Manufacturing - Output Prices Index	PMI <sub>FR</sub> MAN <sub>OP</sub>	IHS Markit
Service - Output Index	PMI <sub>FR</sub> SER <sub>OB</sub>	IHS Markit
Service - New Orders Index	PMI <sub>FR</sub> SER <sub>NI</sub>	IHS Markit
Service - Employment Index	PMI <sub>FR</sub> SER <sub>EI</sub>	IHS Markit
Service - Backlogs of Work Index	PMI <sub>FR</sub> SER <sub>BW</sub>	IHS Markit
Service - Input Prices Index	PMI <sub>FR</sub> SER <sub>IP</sub>	IHS Markit
Service - Output Prices Index	PMI <sub>FR</sub> SER <sub>OP</sub>	IHS Markit
Service - Future Activity Index	PMI <sub>FR</sub> SER <sub>FB</sub>	IHS Markit

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.

TABLE 4.4 – The most important variables during the full out-of-sample period – 2016M1-2021M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-2)	0.99	0.97	0.99	0.99	0.65	-	0.94	0.36
DEFM(-4)	0.76	-	0.93	-	0.53	0.78	0.81	-
diff Google	1	0.99	0.99	1	0.88	0.96	0.76	-
Tendprev prod INSEE	0.99	0.99	0.99	1	0.81	0.92	0.71	-
PMI FR COM BW	0.90	0.92	0.99	0.97	-	0.76	-	0.67
PMI FR MAN SF	0.76	-	0.96	0.89	0.47	0.85	0.97	-
Tendpass prod INSEE	0.92	0.81	0.96	0.97	-	-	-	0.72
Evol eff com2 INSEE	0.79	0.81	-	0.83	0.58	0.76	-	-
Evol vente com INSEE	0.88	0.82	0.93	0.92	-	-	-	-
log Google topic	-	0.75	-	-	0.61	0.74	0.96	-
PMI FR MAN NE	0.79	0.78	-	0.92	-	-	-	-
Persp act serv INSEE	-	-	0.93	0.88	-	-	-	0.47
Tendpass eff indus INSEE	-	0.76	-	-	0.56	0.82	-	-
rcac40	-	-	-	-	-	0.79	0.99	-
PMI FR MAN OP	-	-	0.92	-	-	-	-	-
PMI FR SER IP	-	-	-	-	-	-	0.85	-
PMI FR MAN PM	-	-	-	-	-	-	-	0.85
log Google	-	-	-	-	-	-	0.82	-
PMI FR SER EI	-	-	-	-	-	0.78	-	-
IPC gliss annuel INSEE	-	-	-	-	-	-	0.74	-
PMI FR MAN QP	-	-	-	-	-	-	-	0.71
PMI FR MAN BW	-	-	-	-	-	-	-	0.67
EPU	-	-	-	-	0.54	-	-	-
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.53
Diff Google Topic	-	-	-	-	0.47	-	-	-
Tendprev act serv INSEE	-	-	-	-	-	-	-	0.43
PMI FR MAN OB	-	-	-	-	-	-	-	0.40

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples. The RF column shows the frequency of selection among the ten most important variables. "-" means that the variables are not among the top 10 selected variables but are still selected to a lesser extent.

TABLE 4.5 – The most important variables during the full out-of-sample period – 2016M1-2021M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-1)	-	-	-	-	-	-	0.88	-
DEFM(-2)	0.99	1	1	0.99	-	-	0.97	-
DEFM(-4)	-	-	-	-	0.51	0.51	0.96	-
diff Google	0.99	0.99	0.99	0.99	1	0.97	0.92	-
PMI FR MAN SF	0.99	1	1	0.99	-	0.99	0.96	-
Prev eff serv BDF	0.99	1	1	0.99	0.56	1	0.92	-
Spread LS	0.88	0.76	0.99	0.93	0.78	0.86	-	-
Prev stock indus BDF	0.99	0.97	1	0.99	-	0.97	-	-
Evol cmde indus BDF	0.99	0.96	1	1	-	0.96	-	-
Prev prod indus BDF	0.89	0.88	-	0.93	-	0.83	0.88	-
Evol dmnde serv BDF	0.97	0.94	0.99	0.97	-	-	-	0.44
Evol vente com INSEE	0.90	0.92	-	0.93	0.60	-	-	-
log Google topic	-	-	-	-	0.75	0.78	0.93	-
IMCA indus BDF	-	-	0.97	-	0.72	-	-	0.60
Evol prix serv BDF	-	-	0.94	-	-	-	0.89	-
Evol eff com2 INSEE	-	-	-	-	0.51	0.76	-	-
rcac40	-	-	-	-	-	-	0.94	-
PMI FR MAN PM	-	-	-	-	-	-	-	0.68
Evol stock indus BDF	-	-	-	-	0.67	-	-	-
PMI FR MAN OP	-	-	-	-	0.67	-	-	-
Tendpass prod INSEE	-	-	-	-	-	-	-	0.67
PMI FR MAN BW	-	-	-	-	-	-	-	0.61
PMI FR MAN QP	-	-	-	-	-	-	-	0.57
PMI FR COM BW	-	-	-	-	-	-	-	0.51
Prev act serv BDF	-	-	-	-	-	-	-	0.51
Tendprev act serv BDF	-	-	-	-	-	-	-	0.32
Prev dmnde serv BDF	-	-	-	-	-	-	-	0.31

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples. The RF column shows the frequency of selection among the ten most important variables. "-" means that the variables are not among the top 10 selected variables but are still selected to a lesser extent.

TABLE 4.6 – The most important variables during the non-crisis period – 2016M1-2019M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-2)	1	1	1	1	0.65	-	1	0.54
DEFM(-4)	-	-	0.90	-	-	0.75	1	-
diff Google	1	1	1	1	0.85	1	1	-
Tendprev prod INSEE	1	1	1	1	0.79	0.94	-	-
Tendpass prod INSEE	0.94	0.90	1	1	-	0.77	-	1
PMI FR COM BW	0.85	0.90	0.98	1	-	0.77	-	0.85
Evol eff com2 INSEE	0.75	0.79	0.90	0.77	0.44	0.83	0.83	-
Evol vente com INSEE	0.90	0.90	0.98	0.96	-	-	-	-
PMI FR MAN SF	-	-	0.94	0.83	-	0.77	1	-
PMI FR MAN NE	0.77	0.77	-	0.94	0.44	-	-	-
Niv care indus INSEE	0.73	0.77	-	-	0.33	0.79	-	-
PMI FR SER EI	0.67	0.71	-	-	-	0.75	-	-
log Google topic	-	-	-	-	0.60	-	0.96	-
ICA bat INSEE	-	-	0.90	-	-	-	-	0.50
Persp act serv INSEE	-	-	-	0.81	-	-	-	0.58
Tendpass eff indus INSEE	-	-	-	-	0.40	0.77	-	-
rcac40	-	-	-	-	-	-	1	-
PMI FR MAN PM	-	-	-	-	-	-	-	1
PMI FR MAN QP	-	-	-	-	-	-	-	0.96
IPC gliss annuel INSEE	-	-	-	-	-	-	0.90	-
PMI FR MAN BW	-	-	-	-	-	-	-	0.88
PMI FR SER IP	-	-	-	-	-	-	0.83	-
log Google	-	-	-	-	-	-	0.75	-
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.60
Tendprev act serv INSEE	-	-	-	-	-	-	-	0.58
EPU	-	-	-	-	0.38	-	-	-
Diff Google topic	-	-	-	-	0.33	-	-	-

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.



TABLE 4.7 – The most important variables during the crisis period – 2020M1-2021M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-1)	0.96	0.96	0.92	-	0.92	-	0.83	0.58
DEFM(-2)	-	-	1	0.92	-	-	0.83	-
DEFM(-3)	-	-	0.96	0.92	-	-	-	-
DEFM(-4)	0.96	-	-	-	0.96	-	-	-
PMI FR MAN SF	1	0.96	1	1	0.92	1	0.92	-
Persp act serv INSEE	1	1	1	1	-	0.92	-	0.33
diff Google	0.96	0.96	1	0.96	0.92	-	-	-
PMI FR COM BW	1	0.96	1	0.96	-	-	-	0.38
ICA com INSEE	0.96	0.96	-	-	0.88	0.92	-	0.33
taux long	-	0.96	-	0.92	-	1	-	-
rcac40	-	0.96	-	-	-	0.92	0.96	-
EPU	0.96	-	0.92	-	0.88	-	-	-
Tendpass eff indus INSEE	-	0.96	-	-	0.88	0.92	-	-
PMI FR SER OP	-	0.96	-	-	0.88	0.92	-	-
Tendprev prod INSEE	-	-	1	1	-	-	-	-
PMI FR MAN EI	0.96	-	-	-	-	0.92	-	-
Tendprev eff serv INSEE	-	-	-	-	0.88	0.92	-	-
log Google topic	-	-	-	-	-	-	0.96	0.33
PMI FR SER IP	-	-	-	-	-	-	0.88	0.38
PMI FR MAN BW	-	-	-	-	-	-	0.83	0.38
Evol prix INSEE	0.96	-	-	-	-	-	-	-
log Google	-	-	-	-	-	-	0.96	-
PMI FR MAN NE	-	-	-	0.96	-	-	-	-
Tendprev dem serv INSEE	-	-	-	0.96	-	-	-	-
Tendprev eff indus INSEE	-	-	-	-	-	0.92	-	-
Diff Google topic	-	-	0.92	-	-	-	-	-
Evol eff com2 INSEE	-	-	-	-	0.88	-	-	-
taux short	-	-	-	-	-	-	0.88	-
PMI FR MAN SD	-	-	-	-	-	-	0.83	-
PMI FR SER NI	-	-	-	-	-	-	-	0.50
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.46
PMI FR COM EI	-	-	-	-	-	-	-	0.38

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 4.8 – The most important variables during the non-crisis period – 2016M1-2019M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-1)	-	-	-	-	-	-	0.96	-
DEFM(-2)	1	1	1	1	-	-	1	-
DEFM(-4)	-	-	-	-	0.31	-	1	-
diff Google	1	1	1	1	1	1	1	-
Prev eff serv BDF	1	1	1	1	0.40	1	1	-
PMI FR MAN SF	1	1	1	1	-	0.98	1	-
Evol dmnde serv BDF	1	1	1	1	-	0.90	-	0.63
Prev prod indus BDF	1	1	1	1	-	1	-	-
Evol cmde indus BDF	1	0.96	1	1	-	0.98	-	-
Prev stock indus BDF	1	0.96	1	1	-	0.96	-	-
Evol vente com INSEE	0.96	0.98	0.96	0.98	0.48	-	-	-
IMCA indus BDF	-	-	0.96	-	0.92	-	-	0.77
log Google topic	-	-	-	-	0.83	0.75	1	-
Spread LS	-	-	-	0.90	0.79	0.81	-	-
Prev act serv BDF	0.83	0.81	-	-	-	-	-	0.73
Evol stock indus BDF	-	-	-	-	0.58	-	0.96	-
rcac40	-	-	-	-	-	-	0.98	-
Evol prix serv BDF	-	-	-	-	-	-	0.96	-
Tendpass prod INSEE	-	-	-	-	-	-	-	0.94
PMI FR MAN PM	-	-	-	-	-	-	-	0.85
PMI FR MAN BW	-	-	-	-	-	-	-	0.75
IDC com DGEFCFIN	-	-	-	-	-	0.73	-	-
PMI FR MAN QP	-	-	-	-	-	-	-	0.67
PMI FR COM BW	-	-	-	-	-	-	-	0.65
PMI FR MAN OP	-	-	-	-	0.69	-	-	-
Tendprev act serv BDF	-	-	-	-	-	-	-	0.42
Prev dmnde serv BDF	-	-	-	-	-	-	-	0.40
Evol eff com2 INSEE	-	-	-	-	0.33	-	-	-

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 4.9 – The most important variables during the crisis period – 2020M1-2021M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
DEFM(-1)	-	-	-	-	0.92	-	-	0.38
DEFM(-2)	1	1	1	1	0.88	0.92	0.92	-
DEFM(-4)	0.96	0.96	-	-	0.88	-	0.88	-
Evol cmde indus BDF	0.96	0.96	1	1	0.83	0.92	0.88	0.25
diff Google	1	1	1	1	0.92	0.92	0.75	-
PMI FR MAN SF	1	1	1	1	-	1	0.88	-
Evol prix serv BDF	1	1	1	1	0.92	0.96	-	-
Prev stock indus BDF	1	1	1	1	-	1	-	0.63
Prev eff serv BDF	1	1	1	1	-	1	-	0.29
IDC com DGEFCFIN	0.96	1	1	-	0.92	0.96	-	-
Persp act serv INSEE	-	0.96	1	0.96	-	-	-	-
IMCA indus BDF	-	-	-	1	-	0.92	-	0.33
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.29
PMI FR MAN OP	-	-	1	1	-	-	-	-
Spread LS	-	-	-	-	-	0.96	0.79	-
Evol dmnde serv INSEE	0.96	-	-	-	-	-	-	-
Evol eff com2 INSEE	-	-	-	-	0.92	-	-	-
CLI OECD	-	-	-	-	-	-	0.92	-
Tendpass eff indus INSEE	-	-	-	-	0.88	-	-	-
log Google	-	-	-	-	-	-	0.88	-
rcac40	-	-	-	-	-	-	0.88	-
Evol prix INSEE	-	-	-	-	0.83	-	-	-
log Google topic	-	-	-	-	-	-	0.79	-
Prev prod indus BDF	-	-	-	-	-	-	-	0.67
PMI FR COM BW	-	-	-	-	-	-	-	0.33
PMI FR MAN QP	-	-	-	-	-	-	-	0.29
PMI FR MAN OB	-	-	-	-	-	-	-	0.25

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 4.10 – Nowcast performance during the full out-of-sample period – 2016M1-2021M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	6.433	0.759*	0.291	0.9897	2
aLasso	7.991	0.700*	0.119	0.7387	5
EN	<b>6.345</b>	0.762*	0.301	1.0000	1
EN0.5	6.475	0.757*	0.286	0.9897	2
SCAD	8.834	0.669**	0.026	0.5551	8
aEN	6.577	0.753*	0.275	0.9897	2
aSCAD	8.953	0.664**	0.013	0.5420	9
Lasso+OLS	12.268	0.540**	-0.352	0.2928	14
aLasso+OLS	11.119	0.583**	-0.225	0.3756	11
EN+OLS	12.956	0.537**	-0.359	0.2928	14
EN0.5+OLS	12.956	0.514*	-0.428	0.2928	14
SCAD+OLS	11.335	0.575**	-0.249	0.2982	13
aEN+OLS	11.645	0.563**	-0.283	0.2928	14
aSCAD+OLS	10.464	0.607**	-0.153	0.5420	9
RF	9.378	0.648**	-0.034	0.7387	5
DFM	9.073	0.660**	—	0.7387	5
AR(1)	26.650	—	-1.937	0.2891	12

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 4.11 – Nowcast performance during the full out-of-sample period - 2016M1-2021M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	7.264	0.727*	0.148	0.5628	2
aLasso	7.520	0.718*	0.118	0.5628	2
EN	6.998	0.737*	0.179	0.5628	2
EN0.5	7.331	0.725*	0.140	0.5628	2
SCAD	9.157	0.656**	-0.074	0.4273	6
aEN	<b>6.238</b>	0.766*	0.269	1.0000	1
aSCAD	10.178	0.618**	-0.194	0.3992	8
Lasso+OLS	14.548	0.454	-0.706	0.0892	15
aLasso+OLS	13.392	0.497	-0.570	0.1927	14
EN+OLS	29.106	-0.092	-2.413	0.0471	17
EN0.5+OLS	23.845	0.105	-1.796	0.0556	16
SCAD+OLS	11.576	0.566**	-0.358	0.3992	8
aEN+OLS	10.930	0.590**	-0.282	0.3992	8
aSCAD+OLS	11.119	0.583**	-0.304	0.3992	8
RF	9.257	0.653**	-0.086	0.4273	6
DFM	8.528	0.680**	—	0.3992	8
AR(1)	26.650	—	-2.125	0.2541	13

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 4.12 – Nowcast performance during non-crisis period - 2016M1-2019M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	0.317	0.026	0.020	0.0445	9
aLasso	0.291	0.108	0.102	0.3497	4
EN	0.320	0.017	0.011	0.0445	9
EN0.5	0.312	0.043	0.037	0.0489	7
SCAD	0.388	-0.193	-0.201	0.0310	16
aEN	0.348	-0.070	-0.077	0.0445	9
aSCAD	0.384	-0.180	-0.187	0.0212	17
Lasso+OLS	0.337	-0.035	-0.041	0.0445	9
aLasso+OLS	0.279	0.144	0.139	0.4808	2
EN+OLS	0.341	-0.048	-0.054	0.0445	9
EN0.5+OLS	0.473	-0.452	-0.461	0.0445	9
SCAD+OLS	0.275	0.155	0.149	0.4808	2
aEN+OLS	0.312	0.041	0.035	0.0489	7
aSCAD+OLS	0.300	0.079	0.035	0.3497	4
RF	<b>0.233</b>	0.283	0.279	1.0000	1
DFM	0.324	0.006	—	0.0445	9
AR(1)	0.326	—	-0.006	0.0618	6

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 4.13 – Nowcast performance during crisis period - 2020M1-2021M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	18.67	0.765*	0.298	0.9901	2
aLasso	23.39	0.705*	0.120	0.7234	5
EN	<b>18.39</b>	0.768*	0.308	1.0000	1
EN0.5	18.80	0.763*	0.292	0.9901	2
SCAD	25.73	0.676*	0.032	0.5612	8
aEN	19.03	0.760*	0.284	0.9901	2
aSCAD	26.09	0.671*	0.018	0.5171	9
Lasso+OLS	36.13	0.544	-0.360	0.2540	12
aLasso+OLS	32.80	0.586**	-0.234	0.3147	11
EN+OLS	36.30	0.542	-0.366	0.2540	12
EN0.5+OLS	37.92	0.522	-0.427	0.2540	12
SCAD+OLS	33.46	0.578**	-0.259	0.2540	12
aEN+OLS	34.31	0.567	-0.291	0.2540	12
aSCAD+OLS	30.79	0.612**	-0.159	0.5171	9
RF	27.67	0.651**	-0.041	0.7234	5
DFM	26.57	0.665**	—	0.7234	5
AR(1)	79.30	—	-1.984*	0.254	12

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 4.14 – Nowcast performance during non-crisis period - 2016M1-2019M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	0.292	0.105	0.047	0.0642	3
aLasso	0.291	0.105	0.047	0.0642	3
EN	0.296	0.091	0.033	0.0814	2
EN0.5	0.293	0.099	0.041	0.0642	3
SCAD	0.345	-0.061	-0.129	0.0642	3
aEN	0.325	0.001	-0.063	0.0559	7
aSCAD	0.337	-0.034	-0.101	0.0559	7
Lasso+OLS	1.039	-2.192*	-2.398*	0.0559	7
aLasso+OLS	0.808	-1.482*	-1.641*	0.0559	7
EN+OLS	1.322	-3.776*	-3.615*	0.0516	17
EN0.5+OLS	1.404	-3.313*	-3.591*	0.0559	7
SCAD+OLS	0.588	-0.806*	-0.922*	0.0559	7
aEN+OLS	0.579	-0.778*	-0.893*	0.0559	7
aSCAD+OLS	0.549	-0.687	-0.796*	0.0559	7
RF	<b>0.240</b>	0.262	0.214	1.0000	1
DFM	0.306	0.060	—	0.0559	7
AR(1)	0.326	—	-0.064	0.0559	7

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 4.15 – Nowcast performance during crisis period - 2020M1-2021M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Lasso	21.21	0.733*	0.151	0.5418	2
aLasso	21.98	0.723*	0.120	0.5418	2
EN	20.40	0.743*	0.183	0.5418	2
EN0.5	21.41	0.730*	0.143	0.5418	2
SCAD	26.78	0.662**	-0.072	0.4254	6
aEN	<b>18.06</b>	0.772*	0.277	1.0000	1
aSCAD	29.86	0.623**	-0.196	0.4122	8
Lasso+OLS	41.56	0.476	-0.665	0.0754	15
aLasso+OLS	38.56	0.514	-0.544	0.1983	14
EN+OLS	64.67	0.184	-1.590*	0.0583	16
EN0.5+OLS	68.73	0.133	-1.752*	0.0583	16
SCAD+OLS	33.55	0.577	-0.344	0.4122	8
aEN+OLS	31.63	0.601**	-0.267	0.4122	8
aSCAD+OLS	32.26	0.593	-0.292	0.412	8
RF	27.29	0.656**	-0.093	0.4254	6
DFM	24.97	0.685**	—	0.4122	8
AR(1)	79.30	—	-2.176*	0.245	13

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively, with \* and \*\* meaning that the EPA test of Diebold and Mariano (1995) is significant at the 5% and 10% level, respectively. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

FIGURE 4.3 – Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the small database.

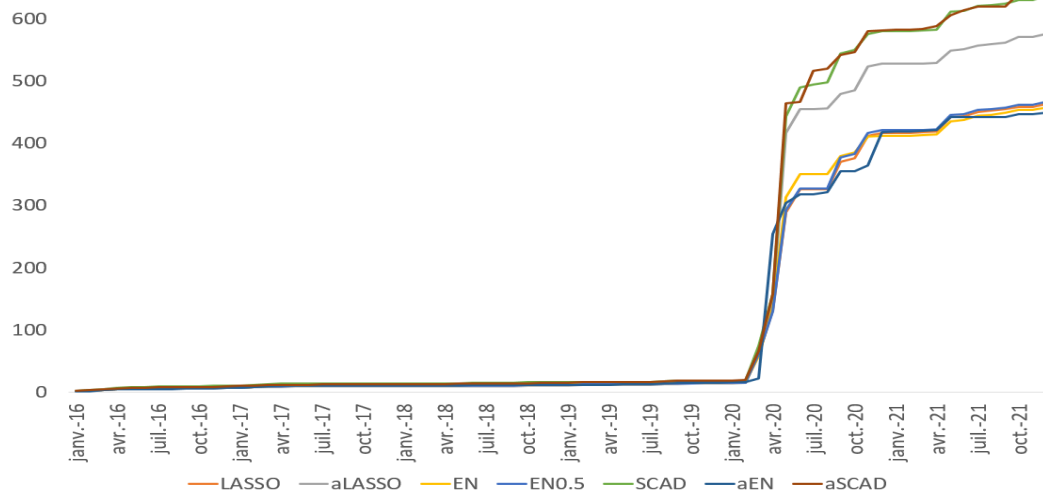


FIGURE 4.4 – Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the large database.

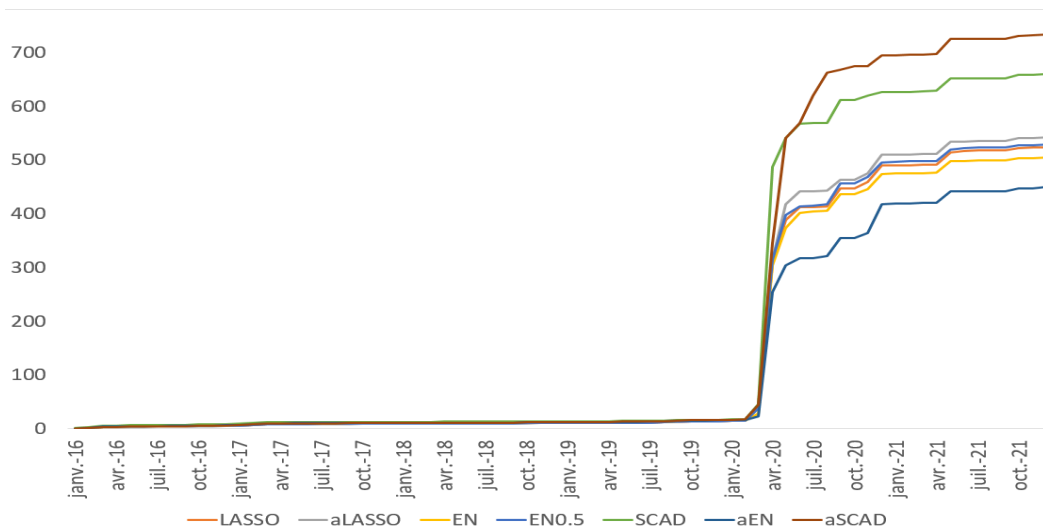
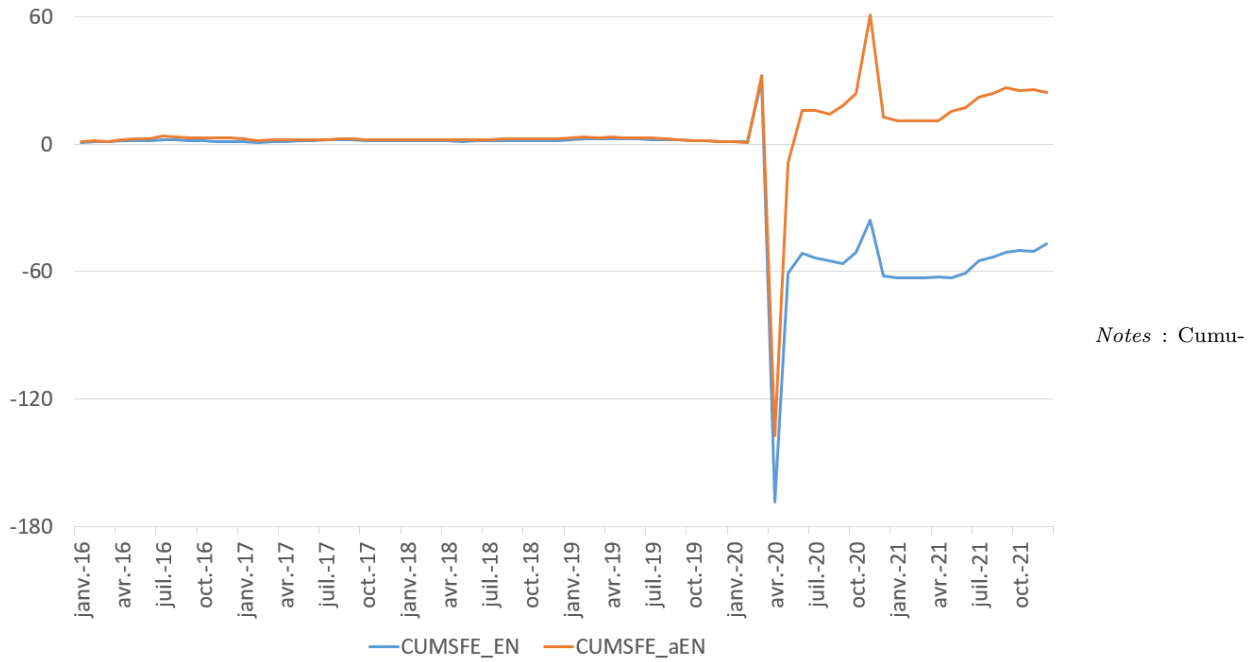


FIGURE 4.5 – Cumulative sum of squared forecast error (CUMSFE) from the small and large databases.



lative sum of squared forecast error (CUMSFE) loss differential of the nowcasts provided by the EN or aEN models estimated from the large database against those obtained from the EN or aEN models estimated from the small database.

FIGURE 4.6 – Cumulative squared prediction errors during the non-crisis period – 2016M1-2019M12 – with the small database.

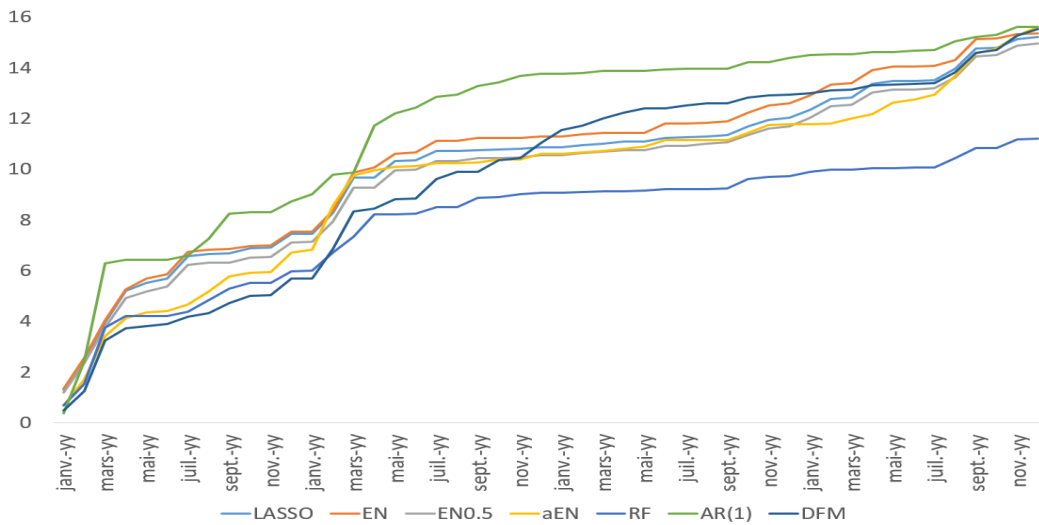
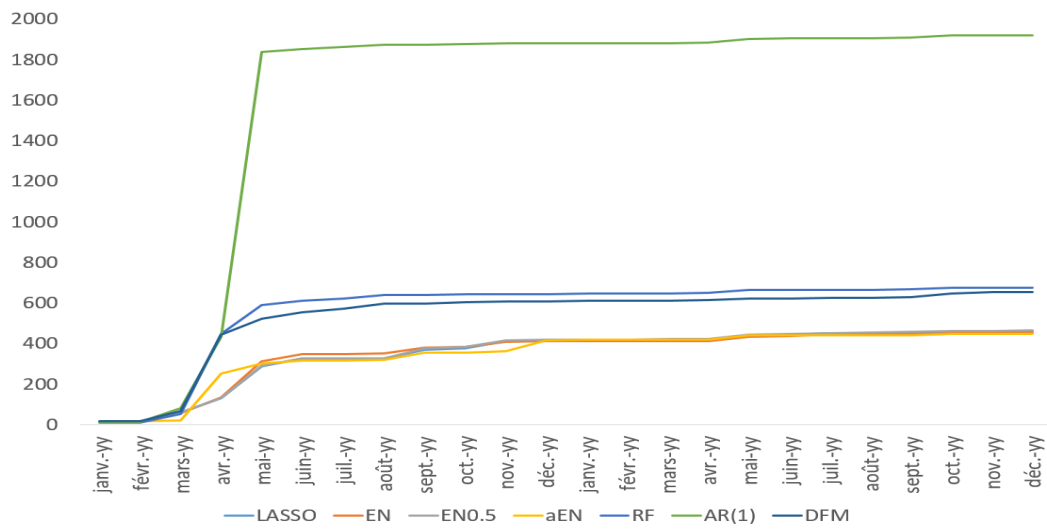


FIGURE 4.7 – Cumulative squared prediction errors during the crisis period – 2020M1-2021M12 – with the large database.





# Chapitre 5

## Are national or regional surveys useful for nowcasting jobseekers ? The case of the French region of Pays-de-la-Loire <sup>1</sup>

### 5.1 Introduction

Unemployment is a crucial issue for France. A high unemployment rate has detrimental effects both on the society and the national economy. An increasing unemployment rate reduces tax revenue (due to the loss of income by the unemployed) and increases public spending in terms of unemployment benefits and the associated retraining and placement programs. Therefore, early and accurate forecasts are required to propose the best economic and social policies at both national and regional levels. Indeed, the Regions are in charge of the coordination on their territory of all the actions in favor of the economy. For example, they are in charge of the professional training of young people and jobseekers. In France, the French national statistical institute (INSEE) publishes the unemployment rate on a quarterly basis with a significant delay of around 45 days after the end of the reference quarter, whereas *Pôle Emploi*, the National Employment Agency in charge of allocating unemployment benefit and job placement, publishes the number of jobseekers at the end of the month (DEFM, *demandeurs d'emploi en fin de mois*) on a monthly basis with a lag of around 26 days after the end of

---

1. Ce chapitre est composé d'un travail co-écrit avec Amélie CHARLES et Olivier DARNE

the reference month. In this paper we focus on nowcasting the Pays-de-la-Loire’s jobseekers, a regional dynamic economy which displays the second lowest level of unemployment for the 2022Q2 with 5.9%, largely under the national level with 7.4%.<sup>2</sup> The PdL’s employment dynamism can be explained by the structure of economic activity, the population growth, the economic performance of companies, the number of business creations (companies and micro-enterprises), and the attractiveness of the region. Figure 5.1 shows that French’s and PdL’s jobseeker growths seem to display similar behavior. Moreover, the majority of studies on unemployment forecasting are conducted at the national level and there are very few at the regional level (see, for instance, Simionescu and Cifuentes-Faura, 2022). Further, there is only one work which forecasts the jobseekers at the national level in France (Fondeur and Karamé, 2013), but there is no study at the regional level.

A number of studies have shown that traditional unemployment forecasting models can be outperformed by a variety of machine learning methods in data-rich environment; see, for example, Kim and Swanson (2018), Li and Chen (2014), Kotchoni et al. (2019), Goulet Coulombe et al. (2022), or Gogas et al. (2022). Moreover, some studies have shown the interest to target predictors from a large database of predictors, suggesting a prior screening stage; see, for example, Bai and Ng (2008), Kotchoni et al. (2019), Borup and Schütte (2022), Yousuf and Feng (2022), or Borup et al. (2022). In particular, Fan and Lv (2008) emphasize the importance of an additional screening method in high dimensions, and introduce the sure independence screening (SIS) procedure. The idea of the SIS procedure is to rank all predictors by employing an utility measure between the target and each predictor and then using the more important variables for modelling in a second stage.<sup>3</sup> In this paper, we consider three extensions of this dimension reduction technique which are model-free screening procedures and thus including linear and nonlinear models, namely the sure independent ranking and screening (SIRS) procedure of Zhu et al. (2011), the SIS procedure based on the distance correlation (DC-SIS) of Li et al. (2012) and the SIS procedure based on the martingale difference correlation of Shao and Zhang (2014). In a second step we apply various penalized regressions with Ridge, Lasso and Elastic-Net methods, as well as random forest (RF) regression

---

2. Since 2016 there are 13 regions in continental France : Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté, Bretagne, Centre-Val de Loire, Corse, Grand Est, Hauts-de-France, Île-de-France, Normandie, Nouvelle-Aquitaine, Occitanie, Pays de la Loire and Provence-Alpes-Côte d’Azur.

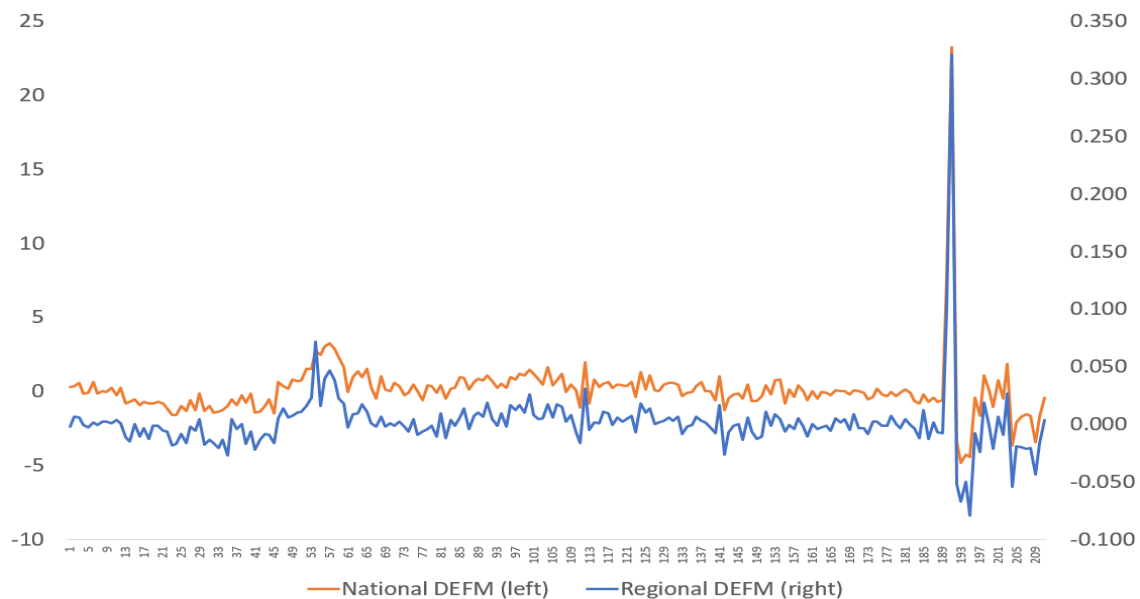
3. SIS pre-selection has been coupled with the SCAD method by Fan and Lv (2008), the adaptive EN method by Zou and Zhang (2009) or the Ridge method by Ferrara and Simoni (2022).

on the set of targeted predictors selected during the screening stage. As factor models are become traditional model for short-term forecasting of unemployment by summarizing the information contained in large databases in few unobserved common factors (Brunhes-Lesage and Darné, 2012; Costantini, 2013; Kim and Swanson, 2018; Li and Chen, 2014; Kotchoni et al., 2019) we complete this analysis by applying dynamic factor models on the targeted predictors.

The aim of our paper is twofold. First, we question whether it is more appropriate to only use regional data or combine national and regional data into various machine learning methods for nowcasting purpose. Second, we compare the nowcasting performance of penalized and random forest regressions and DFM with or without a prior dimension reduction stage to nowcast regional jobseekers growth rate. We also assess the nowcasting performance of the various methods by distinguishing crisis (Global Financial Crisis and Covid-19 crisis) and non-crisis periods.

The rest of this paper is structured as follows. Section 2 describes the data. Section 3 introduces the methodology of the penalized regressions, random forest, dynamic factor model and the SIS-type methods, as well as the nowcasting evaluation. The main results are discussed in Section 4, and Section 5 concludes.

FIGURE 5.1 – Pays-de-la-Loire and French jobseeker growth rate - 2004M6-2021M12.



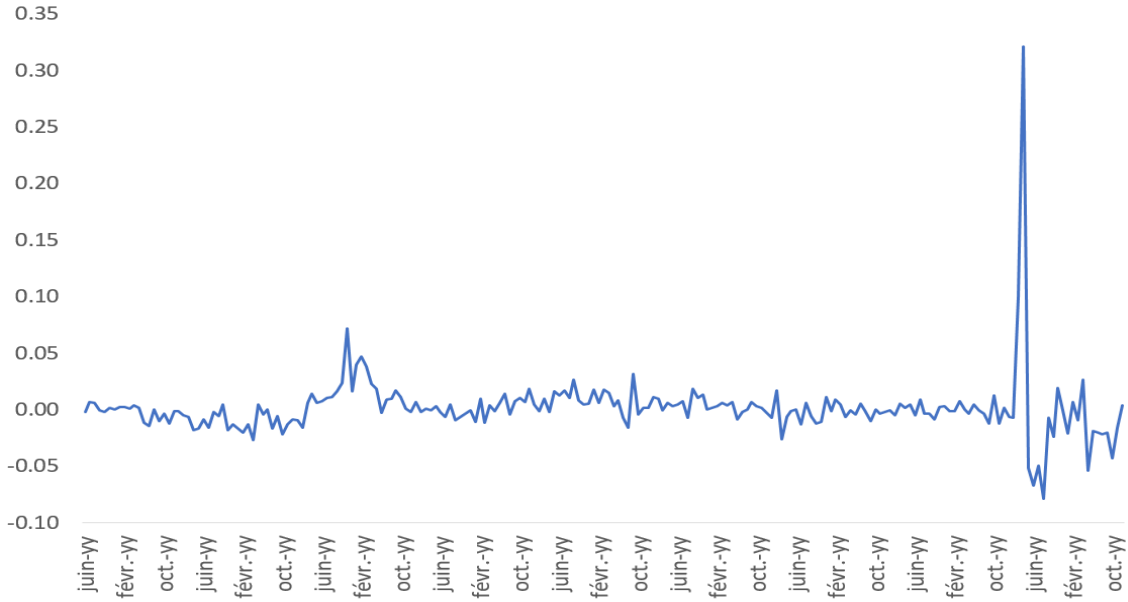
## 5.2 Data

The data set includes monthly data for the number of jobseekers from claimant count (“*Demandeurs d’Emploi en Fin de Mois*”, category A, DEFM hereafter) for the region *Pays de la Loire* (PdL, hereafter) provided by *Pôle Emploi*, the National Employment Agency. This variable describes the inventory of unemployed people at the end of each month. Figure 5.2 shows that the PdL jobseekers have known a large rise during the Covid-19 Crisis, with a large increase of 32.1% in April 2020.

We consider a large set of monthly indicators selected on the basis of their reliability and timeliness covering the period of 2004M6 to 2021M12 (Tables 5.1 to 5.4). These data are based on the surveys on the manufacturing and services sectors, financial data and economic policy uncertainty indicator. The financial data concern stock market and short- and long-term interest rates and we also include the economic policy uncertainty indicator for France proposed by Baker et al. (2016). The Google Trend variables are defined as in Cariou (2022) by focusing on the PdL region. The business survey data are mainly produced by the Banque de France (BdF), the French national statistical institute (INSEE), the Organization for Economic Co-operation and Development (OECD) and the European Commission (EC) and the IHS Markit.

We construct two different monthly databases, starting from the same data set which includes the financial data, the economic policy uncertainty indicator and the Google Trend variables. In the first database, called regional database, we add the regional surveys provided by the BdF on the manufacturing and service sectors, giving a set of 30 variables. For the second database, called large database, the national surveys provided by the BdF, the INSEE, the OECD, the EC and the IHS Markit are added to the regional database, leading to a large database of 134 predictors. We also consider four autoregressive terms. All variables are transformed to achieve stationarity. For penalized regressions and random forest variables are also standardized to have mean 0 and unit standard deviation to ensure that the shrinkage is fair and to prevent the scale of the variables affecting the estimation.

FIGURE 5.2 – PdL’s jobseeker growth rate - 2004M6-2021M12.



## 5.3 Methodology

### 5.3.1 Penalized regressions

We estimate several shrinkage estimators for linear models with the estimated coefficients given by

$$\hat{\beta}_{=\beta_0, \dots, \beta_p} \left\{ \frac{1}{2} \sum_{i=1}^N \left( y_i - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + P(\lambda; \beta) \right\} \quad (5.1)$$

where  $P(\lambda; \beta)$  is a penalty function that depends on the penalty parameter (or tuning parameter)  $\lambda$ . We consider different choices for the penalty functions as described below.

**Lasso regression** : Lasso (*least absolute shrinkage and selection operator*) regression, which was introduced by Tibshirani (1996), is characterized by an  $L_1$  penalty function, allowing for sparsity :

$$P^{Lasso}(\lambda; \beta) = \lambda \sum_{j=1}^p |\beta_j| \quad (5.2)$$

Lasso regressions tend to produce estimated regression coefficients that are exactly zeros, and thus can be used for variable selection, where only predictors with nonzero estimates

are considered to be important (sparse solution). However, Lasso regression has undesirable properties when  $N$  is greater than  $T$  or when there is a group of variables amongst which all pairwise correlations are very high (Zou and Hastie, 2005). Otherwise, Lasso can generate a selection bias between highly correlated variables.

**Elastic Net** : Zou and Hastie (2005) develop the Elastic Net (EN) regularization method, which is a generalization including Lasso and Ridge as special cases, to avoid collinearity. The EN method simultaneously carries out automatic variable selection (as in Lasso by setting some coefficients to zero) and continuous shrinkage (as in Ridge), *via* a (convex) combination of both  $L_1$  and  $L_2$  penalty functions :

$$P^{EN}(\lambda; \beta) = \lambda \sum_{j=1}^p (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|) \quad (5.3)$$

where  $\alpha \in [0; 1]$ . If  $\alpha = 0$  or  $\alpha = 1$  the EN regression is equivalent to Lasso or Ridge regression, respectively. An advantage of the EN lies in its property of grouped selection, that is, the ‘grouping effect’. In other words, a group of highly correlated variables tend to have coefficients of similar magnitude and be selected in the same time.

However, the Lasso and EN methods fail to achieve the oracle property because they estimate the larger nonzero coefficients with asymptotically non-ignorable bias (Zou and Hastie, 2005). The oracle property means that when the true parameters have some zero components, the regularization parameters are estimated as 0 with probability tending to 1, and the nonzero components are estimated as well as when the correct submodel is known.<sup>4</sup>

### 5.3.2 Random forest

Random forest (RF) is an ensemble learning method based on classification and regression trees (Breiman, 2001).<sup>5</sup> RF allows to reduce the variance of regression trees and is based on bootstrap aggregation (bagging) of randomly constructed regression trees. This approach combines several randomized decision trees and aggregates their predictions by averaging.

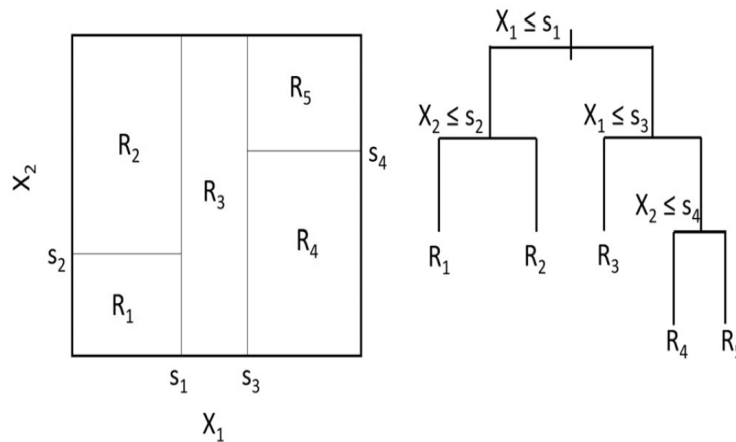
---

4. Fan and Peng (2004) highlighted the importance of studying the validity of the oracle property.

5. A review of the RF algorithm can be found in Biau and Scornet (2016).

To describe how a regression tree works, we follow the example discussed in Hastie et al. (2001). Consider a regression with only two predictors  $X = (X_1, X_2)$ , where each predictor  $X_i$  takes values in some given interval, and  $Y$  is the dependent variable. We first split the space into two regions at  $X_1 = s_1$ , the threshold value gives the minimum sum of squared residuals. Then the region to the left (right) of  $X_1 = s_1$  is split at  $X_2 = s_2$  ( $X_1 = s_3$ ), where  $s_2$  ( $s_3$ ) is again determined by the smallest sum of squared residuals. Finally, the region to the right of  $X_1 = s_3$  is split at  $X_2 = s_4$  (see the right plot of Figure 5.3). The final result is a partitioning into five regions  $R_k$ ,  $k = 1, \dots, 5$ , suggested to estimate the conditional expected value (prediction) of  $Y$  (see the left plot of Figure 5.3). Each region corresponds to a terminal node of the tree.

FIGURE 5.3 – Example of a regression tree (cf. Hastie et al., 2001)



Given a dependent variable  $Y_{t+h}$ , a set of predictors  $X_t$  and a number of terminal nodes  $K$ , the splits are determined to minimize the sum of squared errors of the following regression model

$$Y_{t+h} = \sum_{k=1}^K c_k I_k(X_t; \theta_k),$$

where  $\theta_k$  is the set of parameters defining the  $k$ th region,  $c_k$  is a constant defined as node means, i.e.  $c_k = \sum_{j \in R_k} Y_j / N_k$ , with  $N_k$  the number of variables in the  $k$ th region, and  $I_k(X_t; \theta_k)$  is an indicator function such that

$$I_k(X_t; \theta_k) = \begin{cases} 1 & \text{if } X_t \in R_k(\theta_k) \\ 0 & \text{otherwise} \end{cases}$$

RF is a collection of regression trees, each specified in a bootstrap sample of the original data. Therefore, for each bootstrap sample  $b$ , with  $b = 1, \dots, B$ , a tree with  $K_b$  regions is estimated for a randomly selected subset of the original regressors.  $K_b$  is determined to leave a minimum number of observations in each region. The final forecast is the average of the forecasts of each tree applied to the original data

$$\hat{Y}_{t+h} = \frac{1}{B} \sum_{b=1}^B \left( \sum_{k=1}^{K_b} \hat{c}_{k,b} I_{k,b}(X_t; \hat{\theta}_{k,b}) \right)$$

It is possible to compute the importance of each variable for the RF models by using out-of-bag (OOB) samples, which are samples that are not included in the bootstrap samples. As defined by Hastie et al. (2001) when the  $b$ th tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded. Then the values for the  $j$ th variable are randomly permuted in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable  $j$  in the random forest.

### 5.3.3 Dimensionality reduction approaches

We consider three dimensionality reduction approaches based on the sure screening property that "all important variables survive after applying a variable screening procedure with probability tending to 1" (Fan and Lv, p. 853, 2008). The idea of the dimensionality reduction procedure is to rank all predictors by using an utility measure between the response and each predictor and then to retain the important variables for further investigation.<sup>6</sup>

Let  $Y$  be the univariate response variable and  $X = \{X_k\}_{k=1}^p$  be the predictor variables, with  $n$  the sample size. For any marginal utility measure  $\hat{\omega}_k$  the dimensionality reduction approach selects a set of important predictors with large  $\hat{\omega}_k$  and define a submodel  $\widehat{\mathcal{M}}$  such as

$$\widehat{\mathcal{M}} = \{k : \hat{\omega}_k \text{ is among the first } d_n \text{ largest of all, for } 1 \leq k \leq p\}$$

---

6. See Liu et al. (2015) for a selective survey on screening methods.



where the submodel size  $d_n$  is predefined to be smaller than the sample size  $n$ , with  $d_n$  the tuning parameter. This reduces the full model of size  $p \gg n$  to a submodel with size  $d_n$ .

**SIS approach** : Fan and Lv (2008) propose an approach referred to as Sure Independence Screening (SIS hereafter) for linear models. They show that the Pearson correlation ranking procedure possesses the sure screening property for linear regression with Gaussian predictors and responses. They use Pearson correlation as a marginal utility for feature screening. The sample correlation between the  $k$ -th predictor and the response variable is defined as

$$\omega_k^{sis} = \frac{1}{n} \mathbf{X}_k^T \mathbf{Y}, \quad \text{for } k = 1, 2, \dots, p$$

where  $\mathbf{X}_k$  and  $\mathbf{Y}$  are marginally standardized. Fan and Lv (2008) suggest ranking all predictors according to  $|\omega_k^{sis}|$ , and they select the  $d_n$  important predictors to obtain the submodel  $\widehat{\mathcal{M}}_{SIS}$ .

**SIRS approach** : Zhu et al. (2011) propose a sure independent ranking and screening (SIRS) procedure to screen significant predictors by relaxing the model assumptions or the error distribution assumption of the SIS procedure.<sup>7</sup> The SIRS allows a general model framework, covering a wide large of parametric and semiparametric models, which can be viewed as a model-free screening method, instead of a specific linear model as in the SIS, and is robust to heavy-tailed responses.<sup>8</sup> To avoid specifying a structure for regression function, Zhu et al. (2011) consider the correlation between  $\mathbf{X}_k$  and  $1(\mathbf{Y} < y)$  to measure the importance of the  $k$ -th predictor instead of the correlation between  $\mathbf{X}_k$  and  $\mathbf{Y}$ . They advocate the estimator  $\widehat{\omega}_k^{sirs}$  as the marginal utility measure for predictor ranking given by

$$\widehat{\omega}_k^{sirs} = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{X}_{ik} \mathbf{1}(\mathbf{Y} < y) \right\}^2, \quad \text{for } k = 1, 2, \dots, p$$

---

7. They further show that under linearity condition assumption on the predictor vector, the SIRS enjoys the ranking consistency property, namely the SIRS can rank the important predictors at the top asymptotically.

8. By model-free, it means that one does not need to impose a specific model structure on regression functions to carry out a creening procedure.

Zhu et al. (2011) propose to select the top predictors according to  $\widehat{\omega}_k^{sirs}$  by defining the sub-model  $\widehat{\mathcal{M}}_{SIRS}$ .

**DC-SIS approach :** Li, Zhong and Zhu (2012) develop a SIS procedure based on the distance correlation (DC). the DC-SIS improves the SIS procedure by validating the sure screening property under more general settings, including linear and nonlinear models. Compared with the model-based screening procedure, as in the SIS approach, the DC-SIS procedure is a model-free screening procedure, which thus robust to model misspecification.<sup>9</sup> DC was introduced by Székely, Rizzo and Bakirov (2007), for measuring dependence and testing independence between two random vectors.

They first define the distance covariance between two random vectors  $\mathbf{u} \in \mathcal{R}^q$  and  $\mathbf{v} \in \mathcal{R}^p$  with finite first moments to be nonnegative number  $\text{dcov}(\mathbf{u}, \mathbf{v})$  given by

$$\text{dcov}^2(\mathbf{u}, \mathbf{v}) = \int_{\mathcal{R}^{p+q}} |\phi_{u,v}(t, s) - \phi_u(t)\phi_v(s)|^2 \omega(t, s) dt ds$$

where  $\phi_u(t)$  and  $\phi_v(s)$  are the marginal characteristic function of  $\mathbf{u}$  and  $\mathbf{v}$ , respectively,  $\phi_{u,v}(t, s)$  the joint characteristic function of  $\mathbf{u}$  and  $\mathbf{v}$ ,  $\omega(t, s) = \{c_p c_q |\mathbf{t}|_p^{1+p} |s|_q^{1+q}\}^{-1}$  the weight function, with  $c_p = \frac{\pi^{1+p/2}}{\Gamma((1+p)/2)}$ .

The DC between two random vectors  $\mathbf{u}$  and  $\mathbf{v}$  is defined as

$$\text{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\text{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\text{dcov}(\mathbf{u}, \mathbf{u}) \text{dcov}(\mathbf{v}, \mathbf{v})}}$$

The property of the DC is that  $\text{dcorr}(\mathbf{u}, \mathbf{v}) = 0$  if and only if  $\mathbf{u}$  and  $\mathbf{v}$  are independent.<sup>10</sup>

For two univariate normal random variables  $U$  and  $V$

$$\text{dcorr}(U, V) = \left\{ \frac{\rho \arcsin(\rho) + \sqrt{1 - \rho^3} - \rho \arcsin(\rho/2) - \sqrt{4 - \rho^2 + 1}}{1 + \pi/3 - \sqrt{3}} \right\}^{1/2}$$

9. Further, Li et al. (2012) show that the DC-SIS can effectively handle grouped predictors, compared with the SIRS.

10. Székely et al. (2007) and Székely and Rizzo (2009) show the relationship between the DC and the Pearson correlation coefficient  $\rho$ .

which is strictly increasing in  $|\rho|$ . For any sample DC  $\widehat{\omega}_k^{dcsis} = \widehat{dcorr}(X_k, y)^2$  the DC-SIS select a set of important predictors with large  $\widehat{\omega}_k^{dcsis}$  and define a submodel  $\widehat{\mathcal{M}}_{DC}$ .

**MDC-SIS approach** : Shao and Zhang (2014) propose the martingale difference correlation, to measure the departure of conditional mean independence between a scalar response variable  $\mathbf{v}$  and a vectore predictor variable  $\mathbf{u}$ . They define the martingale difference divergence (MDD) for  $\mathbf{u} \in \mathcal{R}^q$  and  $\mathbf{v} \in \mathcal{R}^1$  to be the nonnegative number  $MDD(\mathbf{v}|\mathbf{u})$  as follows

$$MDD(\mathbf{v}|\mathbf{u})^2 = \int_{\mathcal{R}^q} |\Psi_{v,u}(s) - \Psi_v \Psi_u(s)|^2 \omega(s) ds$$

where  $\Psi_{v,u}(s)$ ,  $\psi_v$  ans  $\Psi_u(s)$  denote the joint and marginal characteristic functions, and  $\omega(s) = \{c_q |s|_q^{1+q}\}^{-1}$ .

The martingale difference correlation (MDC) of  $\mathbf{v}$  given  $\mathbf{u}$  is given by

$$MDC(\mathbf{v}|\mathbf{u})^2 = \begin{cases} \frac{MDD(\mathbf{v}|\mathbf{u})^2}{\sqrt{var(\mathbf{v})^2 dvar(\mathbf{u})^2}} & \text{if } var(\mathbf{v})^2 dvar(\mathbf{u})^2 > 0 \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

where  $dvar(\mathbf{u}) = dcov(\mathbf{u}, \mathbf{u})$  is the distance variance of  $\mathbf{u}$ . For any sample MDC  $\widehat{\omega}_k^{mdcsis} = (MDC_n^j(y|X_k))^2$  the MDC-SIS select a set of important predictors with large  $\widehat{\omega}_k^{mdcsis}$  and define a submodel  $\widehat{\mathcal{M}}_{MDC}$ .

The MDC-SIS procedure is model free, like the DC-SIS procedure. However, unlike the DC-based procedure, which screens out the variables that are marginally independent of the distribution of the response  $Y$ , the MDC-based screening procedure focuses on a specific aspect of the conditional distribution of  $Y$  given  $X$ , say conditional mean of  $Y$  given  $X$ .<sup>11</sup>

---

11. Other approaches based on the DC is the partial DC (PDC) proposed by Székely and Rizzo (2014) for which Yousuf and Feng (2021) develop a PDC-based screening procedure (PDC-SIS), and the conditional DC (CDC) proposed by wang et al. (2015) with the CDC-SIS procedure suggested by Wen et al. (2018).

### 5.3.4 Dynamic factor model

In the factor model framework, variables  $X_t$ , are represented as the sum of two mutually orthogonal unobservable components : the common component  $\chi_t$  and the idiosyncratic component  $\xi_t$ . For a given  $t$ ,  $t = 1, \dots, T$ , the static factor model is defined by :

$$X_t = \Lambda F_t + \xi_t, \quad (5.5)$$

where  $X_t = [x_{1t}, \dots, x_{nt}]'$  is a vector of  $n$  stationary time series and it is assumed that the series have zero mean and covariance matrix  $\Gamma(0)$ ,  $\Lambda$  is the loading matrix such that  $\Lambda = [\lambda_1, \dots, \lambda_n]'$ , the common components  $\chi_t = \Lambda F_t$  are driven by a small number  $r$  of factors  $F_t$  common to all the variables in the model such that  $F_t = [F_{1t}, \dots, F_{rt}]'$ , and  $\xi_t = [\xi_{1t}, \dots, \xi_{nt}]'$  is a vector of  $n$  idiosyncratic mutually uncorrelated components, driven by variable-specific shocks.

Different approaches have been proposed to take dynamics into account in modelling the factors  $F_t$  from dynamic factor models (DFM).<sup>12</sup> Here, we use the static principal component analysis (PCA) suggested by Stock and Watson (2002) to estimate the factors  $F_t$ .<sup>13</sup> An eigenvalue decomposition of the estimated covariance matrix  $\widehat{\Gamma}_0 = T^{-1} \sum_{t=1}^T X_t X_t'$  provides the  $(n \times r)$  eigenvector matrix  $\widehat{S} = (\widehat{S}_1, \dots, \widehat{S}_r)$  containing the eigenvectors  $\widehat{S}_j$  corresponding to the  $r$  largest eigenvalues for  $j = 1, \dots, r$ . The factor estimates are the first  $r$  principal components of  $X_t$  defined as  $F_t^{SW} = \widehat{S}' X_t$ . To integrate dynamics in forecasting, SW propose an autoregressive model for the factors.

To take dynamics into account in modelling, another way to proceed is to model explicitly the dynamics of the factors  $F_t$ . More precisely, we assume that the DFM representation is given by the following equation

$$X_t = A(L)F_t + \xi_t, \quad (5.6)$$

where the common components  $\chi_t = A(L)F_t$  integrate a linear dynamics where  $A(L)$  is a  $(n \times r)$  matrix describing the autoregressive form of the  $r$  factors. If we assume that there exists a  $(n \times q)$  matrix  $B(L)$  such that  $B(L) = A(L)N(L)$  with  $N(L)$  of dimension  $(r \times q)$ ,

12. See Barhoumi, Darné, and Ferrara (2013) and Stock and Watson (2015) for a survey on DFMs.

13. We also applied other approaches of DFMs with the dynamic PCA in the frequency domain (Forni et al. (2005), the DFM based on a state-space representation (Doz et al., 2011) and the DFM estimated from quasi-maximum likelihood (Doz et al., 2012). The results have been lower than those obtained from the PCA model and are available from the authors upon request.

then the dynamic factor is such that  $F_t = N(L)U_t$  where  $U_t$  is a  $(q \times 1)$  independent vector containing the dynamic shocks. It follows that the factor dynamics are described by

$$A(L)F_t = B(L)U_t \quad (5.7)$$

Equation (5.7) specifies a VAR( $p$ ) model for the factor  $F_t$  with lag polynomial  $A(L) = \sum_{i=1}^p A_i L^i$ .  $F_t$  is thus the  $(r \times 1)$  vector of the stacked factors with  $r = q(p + 1)$ .

## 5.4 Nowcasting scheme

To evaluate the forecasting performance, we also use the metric of out-of-sample (OOS)  $R^2$  given by

$$R_{OOS}^2 = 1 - \frac{MSE_{model}}{MSE_{benchmark}} \quad (5.8)$$

where  $MSE_{model}$  and  $MSE_{benchmark}$  are the MSE of the given and benchmark models (AR model or DFM), respectively. The given model is more accurate than the benchmark if its  $R_{OOS}^2$  is positive.

Obviously, simply comparing MSE-values does not take into account the sample uncertainty underlying observed forecast differences. Therefore, we also employed the testing procedure based on equal predictive ability (EPA) proposed by Harvey et al. (1997) with the modified Diebold-Mariano (MDM) test for small sample, which allows for pairwise comparison of forecast performances across models.<sup>14</sup>

Finally, we employ the model confidence set (MCS) procedure proposed by Hansen et al. (2011) to determine the set,  $\mathcal{M}^*$ , that consists of a subset of equivalent models in terms of predictive ability which are superior to the other competing models from a collection of models,  $\mathcal{M}_0$ . The MCS procedure yields a model confidence set,  $\widehat{\mathcal{M}}^*$ , that is a set of models constructed to contain the best models with a given level of confidence. This MCS allows getting several models displaying equivalent forecasting performance and therefore gives robustness to the forecasting exercise rather than to base the forecasting analysis only

---

14. Due to the fact that the MDM test can be biased for nested models we have also applied the Clark-McCracken (2001) encompassing test (ENC-NEW). The results are given in Appendix B and show that we obtained the same findings than those with the MDM test.

on one model. The  $t$ -statistic is defined as

$$T_{\max \mathcal{M}} = \max_{i \in \mathcal{M}} t_i \quad \text{with } t_i = \frac{\bar{d}_i}{\sqrt{\widehat{var}(\bar{d}_i)}} \quad (5.9)$$

where  $\widehat{var}(\bar{d}_i)$  denotes the estimate of  $var(\bar{d}_i)$ ,  $\bar{d}_i = m^{-1} \sum_{j \in \mathcal{M}} \bar{d}_{ij}$ , and  $\bar{d}_{ij} = n^{-1} \sum_{t=1}^n d_{ij,t}$ , with  $d_{ij,t} = L_{i,t} - L_{j,t}$  for all  $i, j \in \mathcal{M}_0$ , and  $L_{i,t}$  is a loss function (here MSE).<sup>15</sup> The  $t$ -statistic is associated with the null hypothesis  $H_{0,\mathcal{M}} : E(\bar{d}_i) = 0$  for all  $i \in \mathcal{M}$ , where  $\mathcal{M} \subset \mathcal{M}_0$ . If  $H_{0,\mathcal{M}}$  is accepted at level  $\alpha$  then the MCS is the set  $\widehat{\mathcal{M}}_{1-\alpha}^*$ .<sup>16</sup>

## 5.5 Empirical results

### 5.5.1 Variable selection

In the penalized regressions the tuning parameters are selected *via* cross-validation, which is a data-driven method that is designed to maximize the expected out-of-sample predictive accuracy. For the EN approach the parameter  $\alpha$  is set to 0.5 as in Medeiros et al. (2021) and is also chosen sequentially with  $0 < \alpha < 1$ .<sup>17</sup> For the adaptive version of the Lasso, EN and SCAD methods the estimators  $\widehat{\beta}_j$  in the first-step model – used to compute the adaptive weights  $\widehat{\omega}_j$  – are the estimates from the Ridge regression, and the parameter  $\gamma$  is set to 0.5.

Here we use the RF algorithm implemented in the `randomForest` package in R and the parameter `mtry` is estimated during the validation phase by implementing the `caret` package in R. The penalized regressions are implemented in the `glmnet` package.

**Small / regional database.** Tables 5.5, 5.7 and 5.8 display the most selected variables by the various methods for the small database. The small database corresponds to the regional database only. We find that "Evol act serv BDF reg" - evolution of the activity in the services - is the most selected, in average, for the whole period, the non-crisis period and for the crisis period. the level of the order book in the industry is also often selected.

15.  $\bar{d}_{ij}$  measures the relative sample loss between the  $i$ -th and  $j$ -th models, while  $\bar{d}_i$  is the sample loss of the  $i$ -th model relative to the average across the models in  $\mathcal{M}$ .

16. The MCS  $p$ -values are calculated using bootstrap implementation with 10,000 resamples (Hansen et al., 2011).

17. We test several values of  $\alpha$ , from 0.1 to 0.9 with a step of 0.1, and choose the one that optimizes the cross-validation criterion.

**Large database.** The large base corresponds to the base with the national data that completes the regional data. It can be seen that the national data outperform the regional data and seems more useful for forecasting the growth rate of the DEFMs. In addition, we find the variation of Google Trends variable (diff Google” relative to the ”term) as a very often selected predictor. Except for the business survey in the services on the evolution of the activity. We note that it appears that the business surveys in the industry sector are more frequently chosen than those in the service sector.

### 5.5.2 Nowcasting results

For the nowcasting evaluation we compare the various machine learning approaches to the AR model and the DFM. Results in terms of MSE are presented in Tables 5.11 and 5.12 as well as the  $R_{OOS}^2$  computed with the benchmark models and the associated DM test, and the MCS  $p$ -values with their ranks for the regional and large database, respectively. The nowcasting performance is also evaluated during the non-crisis (2016M1-2019M1) and crisis (2020M1-2021M12) periods.

#### Full out-of-sample period (2016M1–2021M12)

**Regional database.** The results in terms of MSE show that the DFM produces the smallest MSE, with a large positive value of the  $R_{OOS}^2$  and a significant DM test against the AR(1) model, and being the only model a MCS  $p$ -value upper to 0.90. This finding suggests that the DFM seems to be appropriate for nowcasting the regional DEFM using the regional database. Penalized regressions are also interesting alternatives since they exhibit MSE relatively close to that of the DFM, significant DM tests against the AR(1) benchmark and are in the  $\widehat{\mathcal{M}}_{30\%}^*$ , giving thus similar nowcasts as respects to the DFM. The findings are confirmed by Figure 5.4 that provides time-series plots of the cumulative squared prediction error (CSPE) for the penalized models and DFM to assess their nowcasting performance over time. Note that two models giving the lowest MSE are models which take into account all the variables of the regional database without selecting some of them, namely the Ridge model and the DFM.

**Large database.** For the nowcasts based on the large database during the full period (Table 5.12), the results in terms of MSE show that the Elastic-Net model with the DCSIS and MDC-

SIS pre-selections produce the lowest MSEs. These models outperform the AR(1) benchmark, with large positive values of  $R_{OOS}^2$ , significant DM tests and are the only models in the  $\widehat{\mathcal{M}}_{10\%}^*$ . All the other models display similar nowcast performance since they are in the  $\widehat{\mathcal{M}}_{50\%}^*$ , except the AR(1) model. Another interesting result is that the Lasso, Elastic-Net and RF models based on a screening procedure exhibit better MSE than their counterpart without pre-selection. This finding confirms those obtained by Borup and Schütte (2022) and Borup et al. (2022) showing that pre-selecting predictors improve RF model. However, it is not the case for the Ridge regression. Finally, when adding the national surveys to the regional database it appears that the nowcast performance is improved for the penalized regressions whereas it is not the case for the DFM.

To better see whether adding the national surveys to the regional database can improve the nowcast performance, Figure 5.5 plots the cumulative sum of squared forecast error (CUMSFE) loss differential of the nowcasts provided by the EN model and DFM from the large database against those obtained by the EN model and DFM from the regional database. The CUMSFE is given by

$$\text{CUMSFE}_{t,t+k} = \sum_{i=t}^{t+k} e_{i,M_1}^2 - e_{i,M_2}^2$$

where  $M_1$  is the benchmark, in our case the EN model or the DFM estimated from the regional database, and  $M_2$  is the EN model or the DFM estimated from the large database. A positive value of  $\text{CUMSFE}_{t,t+k}$  means that the benchmark displays larger squared forecast errors compared to the model estimated from the large database up to  $t+k$ , and negative values imply the opposite. We see that both models estimated from the regional and large databases do not display significant difference until the Covid-19 crisis since their CUMSFE are very close to 0. We observe largest gains of additional national survey data at the beginning of the Covid19 crisis. However, during the Covid-19 crisis both models show opposite behavior since the EN model exhibits positive values suggesting that the nowcasts from the large database outperform those from the small database, whereas it is the opposite for the DFM for which the CUMSFE decreases.

Figures 5.4 and 5.5 show a strong jump associated with the Covid-19 crisis. Thus, we now



analyze whether some models are more adequate during crisis periods and other models during non-crisis periods.

### Non-crisis and crisis out-of-sample period

**Non-crisis period (2016M1–2019M12).** During the non-crisis period all the models produce close MSE, except the AR(1) model, for the regional (Table 5.13) and large (Table 5.15) databases since most of the models are in the  $\widehat{\mathcal{M}}_{50\%}^*$ . The lowest MSE are given by the EN model with or without sequentially selecting the  $\alpha$  parameter. Note that the two-step procedures with a screening approach do not seem to improve the accuracy during this period.

**Crisis period (2020M1–2021M12).** The results from the regional database show that the DFM outperforms the other models since it is the only model in the  $\widehat{\mathcal{M}}_{50\%}^*$ . For the large database most of the models display similar nowcast performance with close MSE and being in the  $\widehat{\mathcal{M}}_{10\%}^*$ . An interesting finding is that adding the national surveys to the regional database seems to have benefits on the nowcast accuracy for the penalized regressions. This could be explained by the fact that the national surveys have better taken into account the Covid-19 crisis than the regional surveys.

## 5.6 Conclusion

This paper developed nowcasting models for the jobseekers of PdL by using penalized regressions, random forest and dynamic factor model applied on a broad set of regional and national predictors. The results showed that when adding the national surveys to the regional database it appears that the nowcast performance is improved for the penalized regressions whereas it is not the case for the dynamic factor model. In particular, the Elastic-Net model with the DCSIS and MDCSIS pre-selections display accuracy gains during the Covid-19 crisis in adding national surveys, suggesting that national surveys seem to be more informative during this crisis period.

For practitioners, our findings suggest that when factor models and penalized regressions give similar performances the latter methods should be preferred since they also allow for interpretability and ability to identify the main drivers of the models.

Variables selection reveals the usefulness of national variables for forecasting the growth rate of French jobseekers over regional variables. Another point highlights the relevance of the Google Trends variable also at the national level.

This paper is focused on nowcasting the PdL's jobseeker growth rate. Further research would be to nowcast the jobseekers for the other French regions. For the machine learning approaches, we used penalized regressions. We could also consider other approaches which allow nonlinear specification in future research, such as decision trees or support vector machine as suggested by Ahmad et al. (2021) and Gogas et al. (2022) for forecasting unemployment.

# Bibliographie

- [1] Bai J., Ng S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146, 304-317.
- [2] Belloni A., Chernozhukov V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19, 527-547.
- [3] Bessec M. (2013). Short-term forecasts of French GDP : A dynamic factor model with targeted predictors. *Journal of Forecasting*, 32, 500-511.
- [4] Borup D., Christensen B.J., Mühlbach N.S., Nielsen M.S. (2020). Targeting predictors in random forest regression. Available at SSRN : <https://ssrn.com/abstract=3551557>.
- [5] Borup D., Schütte E.C.M. (2020). In search of a job : Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, forthcoming.
- [6] Breiman L. (2001). Random forests. *Machine Learning*, 45 (1), 5-32.
- [7] Fan J., Li R. (2001). Variable selection *via* nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.
- [8] Fan J., Lv J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Serie B*, 70, 849-911.
- [9] Fondeur Y., Karamé F. (2013). Can Google data help predict French youth unemployment ? *Economic Modelling*, 30, 117-125.
- [10] Ferrara L., Simoni A. (2019). When are Google data useful to nowcast GDP ? An approach *via* preselection and shrinkage. Working paper No. 717, Banque de France.
- [11] Fortin-Gagnon O., Leroux M., Stevanović D., Surprenant S. (2020). A large Canadian database for macroeconomic analysis. Working Paper No 2018S-25, CIRANO.
- [12] Gentzkow M., Kelly B., Taddy M. (2019). Text as data. *Journal of Economic Literature*, 57, 535-74.

- [13] Giannone D., Lenza M. and Primiceri G. (2021). Economic predictions with big data : The illusion of sparsity. *Econometrica*, forthcoming.
- [14] Gogas P., Papadimitriou T., Sofianos E. (2022). *Journal of Forecasting*, 41, 551-566.
- [15] Goulet-Coulombe P. (2020). To Bag is to Prune. arXiv preprint arXiv :2008.07063.
- [16] Goulet-Coulombe P. (2021). The macroeconomy as a random forest. arXiv preprint arXiv :2006.12724.
- [17] Goulet-Coulombe P., Leroux M., Stevanovic D., Surprenant S. (2019). How is machine learning useful for macroeconomic forecasting ? Working Papers No 2019s-22, CIRANO.
- [18] Goulet-Coulombe P., Marcellino M., Stevanović D. (2021). Can machine learning catch the Covid-19 recession ? *National Institute Economic Review*, 256, 71-109.
- [19] Goulet-Coulombe P., Leroux M., Stevanović D., Surprenant S. (2021). Macroeconomic data transformations matter. *International Journal of Forecasting*, 37, 1338-1354.
- [20] Hastie T., Tibshirami R., Friedman J.H (2001). *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, New York : Springer.
- [21] Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2), 339-354.
- [22] Kock A.B., Medeiros M.C., Vasconcelos G. (2020). Penalized Time Series Regression. In Fukely P. (ed.) *Macroeconomic Forecasting in the Era of Big Data*, Springer.
- [23] Kotchoni R., Leroux M., Stevanovic D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34, 1050-1072.
- [24] Li R., Zhong W. and Zhu L. (2012). Feature screening *via* distance correlation learning. *Journal of the American Statistical Association*, 107, 1129-1139.
- [25] Liu J., Zhong W. and Li R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58, 1-22.
- [26] Masini R.P., Mendes E.F., Medeiros M.C. (2021). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, forthcoming
- [27] McCracken M.W., Ng S. (2016). FRED-MD : A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34, 574-589.

- [28] Medeiros M.C., Vasconcelos G.F.R., Veiga A., Zilberman E. (2021). Forecasting inflation in a data-rich environment : The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39, 98-119.
- [29] Pan W., Wang X., Xiao W., Zhu H. (2019). A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114, 928-937.
- [30] Schumacher C. (2010). Factor forecasting using international targeted predictors : The case of German GDP. *Economics Letters*, 107(2), 95-98.
- [31] Simionescu M., Cifuentes-Faura J. (2022). Forecasting National and Regional Youth Unemployment in Spain Using Google Trends. *Social Indicators Research*, forthcoming.
- [32] Székely G.J., Rizzo M.L. (2009). Brownian distance covariance. *Annals of Applied Statistics*, 3, 1233-1303.
- [33] Székely G.J., Rizzo M.L. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42, 2382–2412.
- [34] Székely G.J., Rizzo M.L., Bakirov N.K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35, 2769-2794.
- [35] Tibshirani R. (1996). Regression shrinkage and selection *via* the Lasso. *Journal of the Royal Statistical Society. Series B*, 58, 267-288.
- [36] Wager S., Athey S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228-1242.
- [37] Wager S., Walther G. (2016). Adaptive concentration of regression trees, with application to random forests. Working Paper, Stanford University. Available at arXiv : <https://arxiv.org/abs/1503.06388>
- [38] Wang X., Pan W., Hu W., Tian Y., Zhang H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110, 1726-1734.
- [39] Wen C., Pan W., Huang M. and Wang X. (2018). Sure independence screening adjusted for confounding covariates with ultrahigh dimensional data. *Statistica Sinica*, 28, 293-317.
- [40] Yousug K. and Feng Y. (2021). Targeting predictors *via* partial distance correlation with applications to financial forecasting. *Journal of Business & Economic Statistics*, forthcoming.

- [41] Zhu L.P., Li L., Li R., Zhu L.X. (2011). Model-free feature screening for ultrahigh dimensional data. *Journal of the American Statistical Association*, 106, 1464-1475.
- [42] Zou H., Hastie T. (2005). Regularization and variable selection *via* the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320.
- [43] Zou H., Zhang H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37, 1733-1751.

TABLE 5.1 – Variables names and abbreviations - 1/3

Variables names	Abbreviations	Source
Demandeurs d'Emploi en Fin de Mois (growth rate)	DEFM	Pôle Emploi/DARES
Composite Leading Indicator	CLI OECD	OECD
Business Confidence Index	BCI OECD	OECD
Consumer Confidence Index	CCI OECD	OECD
Economic Sentiment Indicator	ESI DGEFCIN	DGEFCIN
Economic Sentiment Indicator Construction	IDC const DGEFCIN	DGEFCIN
Economic Sentiment Indicator Industry	IDC indus DGEFCIN	DGEFCIN
Economic Sentiment Indicator Retail	IDC com DGEFCIN	DGEFCIN
Economic Sentiment Indicator Consumers	IDC conso DGEFCIN	DGEFCIN
Economic Sentiment Indicator Services	IDC serv DGEFCIN	DGEFCIN
Monthly business sentiment indicator in market services	IMCA serv BDF	Banque de France
Monthly business sentiment indicator in manufacturing industry	IMCA indus BDF	Banque de France
Change in employment, compared with previous month	Evol eff ind BDF	Banque de France
Expected employment for the coming month	Prev eff ind BDF	Banque de France
Change in employment over next month	Evol eff serv BDF	Banque de France
Expected employment, compared with previous month	Prev eff serv BDF	Banque de France
Change in foreign orders, compared with previous month	Evol cmde indus BDF	Banque de France
Change in overall level of new orders, compared with previous month	Evol <sub>cmd</sub> indus <sub>BDF</sub>	Banque de France
Change in deliveries, compared with previous month	Evol liv indus BDF	Banque de France
Change in commodity prices, compared with previous month	Evol pmp indus BDF	Banque de France
Current position in inventories of commodities	Stock mp indus BDF	Banque de France
Change in industrial producer prices, compared with previous month	Evol <sub>p</sub> prod <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Expected industrial prices for coming month	Prev pprod indus BDF	Banque de France
Change in output, compared with previous month	Evol prod indus BDF	Banque de France
Expected production for the coming month	Prev <sub>p</sub> prod <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Change in inventories of final goods, compared with previous month	Evol <sub>s</sub> stock <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Current position in inventories of final goods	Stock <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Expected inventories of final goods for the coming month	Prev <sub>s</sub> stock <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Cash positions at end of month	Tresor indus BDF	Banque de France
End-of-month cash position SME	Tresor PME indus BDF	Banque de France
Average capacity utilisation rate	TUC <sub>i</sub> indus <sub>BDF</sub>	Banque de France
Expected overall activity over next month	Prev act serv BDF	Banque de France
Change in foreign orders, compared with previous month	Evol dmndee serv BDF	Banque de France
Forecasts on foreign orders over next month	Prev dmndee serv BDF	Banque de France
Change in aggregate demand, compared with previous month	Evol dmnde serv BDF	Banque de France
Expected aggregate demand over next month	Prev dmnde serv BDF	Banque de France
Change in prices, compared with previous month	Evol prix serv BDF	Banque de France
Expected prices over next month	Prev prix serv BDF	Banque de France
Cash positions at end of month	Tresor serv BDF	Banque de France
Cash positions of SMEs at end of month	Tresor PME serv BDF	Banque de France

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.

TABLE 5.2 – Variables names and abbreviations 2/3

Variables names	Abbreviation	Source
Business climate summary indicator	ICA INSEE	INSEE
Summary indicator of overall economic situation reversal	IRC INSEE	INSEE
Employment climate summary indicator	ICE INSEE	INSEE
Business climate summary indicator industry	ICA indus INSEE	INSEE
Business climate summary indicator services	ICA serv INSEE	INSEE
Business climate summary indicator construction	ICA bat INSEE	INSEE
Business climate summary indicator retail	ICA com INSEE	INSEE
Expected trend in the workforce industry	Tendprev eff indus INSEE	INSEE
Past trend in the workshop industry	Tendpass eff indus INSEE	INSEE
Level of stocks of manufactured products industry	Stock prodfin INSEE	INSEE
Past trend in production industry	Tendpass prod INSEE	INSEE
Expected trend in production industry	Tendprev prod INSEE	INSEE
Probable trend in industrial production volume industry	Tendvol prod indus INSEE	INSEE
Probable trend in sale prices industry	Evol prix INSEE	INSEE
Probable trend in industrial price general level industry	Tend prix indus INSEE	INSEE
Past trend of activity services	Tendpass acti serv INSEE	INSEE
Expected trend of activity services	Tendprev act serv INSEE	INSEE
Expected trend of demand services	Tendprev dem serv INSEE	INSEE
Past trend of prices services	Tendpass prix serv INSEE	INSEE
Expected trend of prices services	Tendprev prix serv INSEE	INSEE
General prospects on activity services	Persp act serv INSEE	INSEE
Change in the situation of the enterprise over the last 3 months services	Evol entrep serv INSEE	INSEE
Expected trend of employment services	Tendprev eff serv INSEE	INSEE
Past trend of employment services	Tendpass eff serv INSEE	INSEE
Expected trend of employment construction	Tendprev eff bat INSEE	INSEE
Past trend of employment construction	Tendpass eff bat INSEE	INSEE
Change in the number of persons employed in the next 3 months	Evol eff com1 INSEE	INSEE
Change in the number of persons employed in the last 3 months	Evol eff com2 INSEE	INSEE
Level of global order books	Niv car indus INSEE	INSEE
Level of foreign order books	Niv care indus INSEE	INSEE
Current order books	Niv car cmd indus BDF	Banque de France
Business development (sales) in the next 3 months	Evol vente com INSEE	INSEE
Intents for orders in the next 3 months	Int cmd com INSEE	INSEE
Opinion on the future trend of unemployment	Op men chom INSEE	INSEE
Consumer price index year-on-year change	IPC gliss annuel INSEE	INSEE
CAC40 (rate)	rcac40	Banque de France
Long-term interest rate	taux long	Banque de France
Short-term interest rate	taux short	Banque de France
Economic Policy Uncertainty	EPU	Baker et al. (2016)
Keyword "emploi"	log Google	Google Trends
Keyword "emploi"	diff Google	Google Trends
Keyword "emploi"	log Google topic	Google Trends
Keyword "emploi"	Diff Google topic	Google Trends
Spread Long-Short	Spread LS	BdF - author

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.



TABLE 5.3 – Variables names and abbreviations -3/3

Variables names	Abbreviation	Source
Composite - Output Index	PMI <sub>FR</sub> COM <sub>OB</sub>	IHS Markit
Composite - New Orders Index	PMI <sub>FR</sub> COM <sub>NI</sub>	IHS Markit
Composite - Employment Index	PMI <sub>FR</sub> COM <sub>EI</sub>	IHS Markit
Composite - Backlogs of Work Index	PMI <sub>FR</sub> COM <sub>BW</sub>	IHS Markit
Composite - Input Prices Index	PMI <sub>FR</sub> COM <sub>IP</sub>	IHS Markit
Composite - Output Prices Index	PMI <sub>FR</sub> COM <sub>OP</sub>	IHS Markit
Manufacturing PMI Index	PMI <sub>FR</sub> MAN <sub>PM</sub>	IHS Markit
Manufacturing - Output Index	PMI <sub>FR</sub> MAN <sub>OB</sub>	IHS Markit
Manufacturing - New Orders Index	PMI <sub>FR</sub> MAN <sub>NI</sub>	IHS Markit
Manufacturing - New Export Orders	Index PMI <sub>FR</sub> MAN <sub>NE</sub>	IHS Markit
Manufacturing - Employment Index	PMI <sub>FR</sub> MAN <sub>EI</sub>	IHS Markit
Manufacturing - Backlogs of Work Index	PMI <sub>FR</sub> MAN <sub>BW</sub>	IHS Markit
Manufacturing - Quantity of Purchases Index	PMI <sub>FR</sub> MAN <sub>QP</sub>	IHS Markit
Manufacturing - Stocks of Purchases Index	PMI <sub>FR</sub> MAN <sub>SP</sub>	IHS Markit
Manufacturing - Stocks of Finished Goods Index	PMI <sub>FR</sub> MAN <sub>SF</sub>	IHS Markit
Manufacturing - Suppliers' Delivery Times Index	PMI <sub>FR</sub> MAN <sub>SD</sub>	IHS Markit
Manufacturing - Input Prices Index	PMI <sub>FR</sub> MAN <sub>IP</sub>	IHS Markit
Manufacturing - Output Prices Index	PMI <sub>FR</sub> MAN <sub>OP</sub>	IHS Markit
Service - Output Index	PMI <sub>FR</sub> SER <sub>OB</sub>	IHS Markit
Service - New Orders Index	PMI <sub>FR</sub> SER <sub>NI</sub>	IHS Markit
Service - Employment Index	PMI <sub>FR</sub> SER <sub>EI</sub>	IHS Markit
Service - Backlogs of Work Index	PMI <sub>FR</sub> SER <sub>BW</sub>	IHS Markit
Service - Input Prices Index	PMI <sub>FR</sub> SER <sub>IP</sub>	IHS Markit
Service - Output Prices Index	PMI <sub>FR</sub> SER <sub>OP</sub>	IHS Markit
Service - Future Activity Index	PMI <sub>FR</sub> SER <sub>FB</sub>	IHS Markit

Notes : Each variable is at the national level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them.

TABLE 5.4 – Variables names and abbreviations - regional level

Variables names	Abbreviations	Source
Reg 52-Monthly business sentiment indicator in market services	IMCA serv BDF reg	Banque de France
Reg 52-Monthly business sentiment indicator in manufacturing industry	IMCA indus BDF reg	Banque de France
Reg 52-Change in employment, compared with previous month	Evol eff ind BDF reg	Banque de France
Reg 52-Expected employment for the coming month	Prev eff ind BDF reg	Banque de France
Reg 52-Change in employment over next month	Evol eff serv BDF reg	Banque de France
Reg 52-Expected employment, compared with previous month	Prev eff serv BDF reg	Banque de France
Reg 52-Change in foreign orders, compared with previous month	Evol cmde indus BDF reg	Banque de France
Reg 52-Change in overall level of new orders, compared with previous month	Evol cmd indus BDF reg	Banque de France
Reg 52-Change in commodity prices, compared with previous month	Evol pmp indus BDF reg	Banque de France
Reg 52-Change in industrial producer prices, compared with previous month	Evol pprod indus BDF reg	Banque de France
Reg 52-Expected industrial prices for coming month	Prev pprod indus BDF reg	Banque de France
Reg 52-Change in output, compared with previous month	Evol prod indus BDF reg	Banque de France
Reg 52-Expected production for the coming month	Prev prod indus BDF reg	Banque de France
Reg 52-Change in inventories of final goods, compared with previous month	Evol stock indus BDF reg	Banque de France
Reg 52-Current position in inventories of final goods	Stock indus BDF reg	Banque de France
Reg 52-Average capacity utilisation rate	TUC indus BDF reg	Banque de France
Reg 52-MoM change in activity	Evol act serv BDF reg	Banque de France
Reg 52-Expected overall activity over next month	Prev act serv BDF reg	Banque de France
Reg 52-Change in prices, compared with previous month	Evol prix serv BDF reg	Banque de France
Reg 52-Expected prices over next month	Prev prix serv BDF reg	Banque de France
Reg 52-Cash positions at end of month	Tresor serv BDF reg	Banque de France
Reg 52-Cash positions of SMEs at end of month	Niv car cmd indus BDF reg	Banque de France
Keyword "emploi"	log Google reg	Google Trends
Keyword "emploi"	Diff Google reg	Google Trends
Keyword "emploi"	log Google topic reg	Google Trends
Keyword "emploi"	Diff Google topic reg	Google Trends

Notes : Each variable is at the regional level. The variables are all seasonally adjusted, otherwise TRAMO-SEATS or X-13ARIMA can adjust them. Reg 52 corresponds to the code of the region Pays de la Loire.

TABLE 5.5 – The most important variables during the full out-of-sample period – 2016M1-2021M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
Evol act serv BDF reg	1	1	1	1	0.86	0.81	0.36	0.90
Niv car cmd indus BDF reg	1	1	1	1	0.33	0.19	-	0.83
Prev act serv BDF reg	0.92	1	1	0.92	0.44	0.53	0.33	0.72
Evol cmd indus BDF ref	0.90	0.94	0.94	0.89	-	0.26	-	0.72
Evol cmde indus BDF reg	0.74	0.74	-	0.74	-	-	-	0.57
Stock indus BDF reg	0.69	0.79	1	0.71	0.65	0.49	0.26	-
Evol pmp indus BDF reg	0.71	0.76	0.99	0.72	-	-	-	0.88
Evol eff ind BDF reg	0.66	0.75	0.92	0.69	0.58	0.53	0.26	0.79
Prev eff serv BDF reg	0.72	-	-	0.72	0.36	0.5	-	0.71
Prev pprod indus BDF reg	0.63	0.79	1	0.64	-	-	-	-
Prev prod indus BDF reg	-	0.85	0.93	-	-	-	-	-
log Google topic reg	-	-	-	-	0.28	0.26	0.67	-
DEFM2 reg	-	-	-	-	-	0.61	0.74	-
IMCA serv BDF reg	-	-	-	-	0.28	-	-	0.76
Evol stock indus BDF reg	-	-	0.91	-	-	-	-	-
IMCA indus BDF reg	-	-	-	-	-	-	-	0.89
rcac40	-	-	-	-	-	-	0.74	-
DEFM3 reg	-	-	-	-	-	-	0.72	-
DEFM1 reg	-	-	-	-	-	-	0.68	-
DEFM4 reg	-	-	-	-	-	-	0.54	-
Evol prix serv BDF reg	-	-	-	-	0.46	-	-	-
Diff Google topic reg	-	-	-	-	0.25	-	-	-
Evol eff serv BDF reg	-	-	-	-	-	0.22	-	-

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples. The RF column shows the frequency of selection among the ten most important variables. "-" means that the variables are not among the top 10 selected variables but are still selected to a lesser extent.

TABLE 5.6 – The most important variables during the full out-of-sample period – 2016M1-2021M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
Evol cmde indus BDF	0.96	1	1	1	-	0.61	-	-
diff Google	0.85	0.96	1	0.86	0.67	0.72	0.69	-
Evol cmd indus BDF	0.94	-	-	1	0.36	0.57	-	-
Evol act serv BDF reg	0.90	0.99	-	0.92	0.64	0.72	-	-
Prev prod indus BDF	0.90	-	1	0.89	0.47	0.72	-	-
Evol dmnde serv BDF	0.89	0.96	-	0.97	-	0.60	-	0.5
Prev eff serv BDF reg	0.72	-	-	0.72	0.5	0.71	-	-
PMI FR MAN SF	0.78	-	1	0.82	0.38	0.68	-	-
Prev act serv BDF	0.43	-	-	0.58	-	-	-	0.44
Tendprev prod INSEE	-	1	1	-	-	-	-	-
Tendpass acti serv INSEE	-	0.99	1	-	-	-	-	-
IDC com DGECFIN	-	0.97	1	-	-	-	-	-
Evol dmndee serv BDF	-	0.94	1	-	-	-	-	-
log Google topic reg	-	-	-	-	0.38	0.22	0.67	-
Evol eff ind BDF reg	0.39	-	-	0.5	-	-	-	-
Niv car cmd indus BDF	-	-	1	-	-	-	-	-
Prev prod indus BDF reg	-	-	1	-	-	-	-	-
Tendprev eff serv INSEE	-	-	1	-	-	-	-	-
Evol stock indus BDF	-	-	-	-	0.53	-	0.51	-
IMCA indus BDF	-	-	-	-	0.5	-	-	0.46
Evol cmde indus BDF reg	-	0.95	-	-	-	-	-	-
EPU	-	0.97	-	-	-	-	-	-
DEFM1 reg	-	-	-	-	-	-	0.75	-
DEFM2 reg	-	-	-	-	-	-	0.75	-
Tendpass prod INSEE	-	-	-	-	-	-	-	0.67
rcac40 reg	-	-	-	-	-	-	0.68	-
DEFM3 reg reg	-	-	-	-	-	-	0.65	-
Prev stock indus BDF	-	-	-	-	-	-	0.57	-
DEFM4 reg	-	-	-	-	-	-	0.54	-
PMI FR COM BW	-	-	-	-	-	-	-	0.54
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.54
PMI FR MAN OB	-	-	-	-	-	-	-	0.51
PMI FR MAN PM	-	-	-	-	-	-	-	0.49
IMCA serv BDF reg	-	-	-	-	-	-	-	0.44
PMI FR MAN BW	-	-	-	-	-	-	-	0.44
Prev pprod indus BDF reg	-	-	-	-	0.33	-	-	-
Diff Google topic reg	-	-	-	-	-	0.24	-	-

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples. The RF column shows the frequency of selection among the ten most important variables. "-" means that the variables are not among the top 10 selected variables but are still selected to a lesser extent.

TABLE 5.7 – The most important variables during the non-crisis period – 2016M1-2019M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
Evol act serv BDF reg	1	1	1	1	1	0.71	-	1
Niv car cmd indus BDF reg	1	1	1	1	-	-	-	1
Prev act serv BDF ref	1	-	-	1	0.63	0.71	-	1
Evol cmde indus BDF reg	1	-	-	1	-	-	-	1
Prev eff serv BDF reg	1	-	-	1	-	0.69	-	1
Evol eff ind BDF reg	-	1	1	-	0.79	0.71	-	-
Evol pmp indus BDF reg	-	1	1	-	-	-	-	-
Evol prix serv BDF reg	-	1	1	-	0.60	-	-	-
DEFM2 reg	-	-	-	-	-	0.85	1	-
Stock indus BDF reg	-	-	-	-	0.85	-	-	-
DEFM3 reg	-	-	-	-	-	-	0.98	-
DEFM1 reg	-	-	-	-	-	-	0.98	-
log Google topic reg	-	-	-	-	-	-	0.90	-
rcac40	-	-	-	-	-	-	0.77	-
IMCA serv BDF reg	-	-	-	-	-	-	-	1

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 5.8 – The most important variables during the crisis period – 2020M1-2021M12 – with the small database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
Evol act serv BDF reg	1	1	1	1	1	1	-	-
Evol cmd indus BDF reg	0.92	1	1	0.92	0.67	0.79	-	-
Niv car cmd indus BDF reg	1	1	1	1	-	-	-	-
Prev act serv BDF reg	0.71	1	1	0.71	-	-	-	-
Prev prod indus BDF reg	0.67	1	-	0.67	-	-	-	0.92
DEFM1 reg	-	-	-	-	-	-	0.21	0.92
log Google topic reg	-	-	-	-	-	-	0.21	0.83
DEFM3 reg	-	-	-	-	-	0.17	0.21	-
Evol eff ind BDF reg	-	-	-	-	0.17	0.17	-	-
Diff Google topic reg	-	-	-	-	0.17	0.17	-	-
Prep pprod indus BDF reg	-	-	1	-	-	-	-	-
IMCA indus BDF reg	-	-	-	-	-	-	-	0.88
TUC indus BDF reg	-	-	-	-	-	-	-	0.83
DEFM2 reg	-	-	-	-	-	-	0.21	-
DEFM4 reg	-	-	-	-	-	-	0.21	-
Stock indus BDF reg	-	-	-	-	0.21	-	-	-

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 5.9 – The most important variables during the non-crisis period – 2016M1-2019M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
diff Google	1	1	1	1	0.75	1	0.98	-
Evol act serv BDF reg	1	1	1	1	0.85	1	-	-
PMI FR MAN SF	1	-	-	1	-	0.98	-	-
Diff Google reg	-	1	1	-	-	-	-	-
Prev eff serv BDF reg	1	-	-	-	0.56	0.98	-	-
Evol cmde indus BDF	-	1	1	1	-	-	-	-
Evol cmd indus BDF	1	-	-	1	-	-	-	-
Log Google topic reg	-	-	-	-	0.52	-	0.88	-
Evol cmde indus BDF reg	-	1	-	-	-	-	-	-
Prev prod indus BDF	-	-	-	-	-	1	-	-
EPU	-	-	1	-	-	-	-	-
DEFM1 reg	-	-	-	-	-	-	1	-
DEFM2 reg	-	-	-	-	-	-	0.98	-
Tendpass prod INSEE	-	-	-	-	-	-	-	0.94
rcac40	-	-	-	-	-	-	0.92	-
IMCA indus BDF	-	-	-	-	0.75	-	-	-
Evol dmnde serv BDF	-	-	-	-	-	-	-	0.71
Tendvol prod indus INSEE	-	-	-	-	-	-	-	0.71
PMI FR COM BW	-	-	-	-	-	-	-	0.69
Prev avt serv BDF	-	-	-	-	-	-	-	0.67

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 5.10 – The most important variables during the crisis period – 2020M1-2021M12 – with the large database.

Variables	Lasso	aLasso	EN	EN05	SCAD	aEN	aSCAD	RF
Evol cmd indus BDF	0.79	1	1	1	0.79	0.25	-	-
Evol cmde indus BDF	0.92	1	1	1	-	-	-	-
Evol dmnde serv BDF	0.75	-	-	0.92	-	-	-	-
Evol act serv BDF reg	0.71	-	1	0.75	-	0.17	-	-
Prev prod indus BDF	0.71	-	-	-	-	-	-	0.63
Tendprev prod INSEE	-	1	-	0.79	0.67	-	-	-
EPU	-	0.96	1	-	-	-	-	-
Evol liv indus BDF	-	1	-	-	-	-	-	-
diff Google	-	-	1	-	1	1	-	-
DEFM2 reg	-	-	-	-	0.71	-	0.29	-
DEFM4 reg	-	-	-	-	0.71	-	0.25	-
Diff Google topic reg	-	-	-	-	-	0.17	-	-
PMI FR COM BW	-	-	-	-	-	0.17	-	-
DEFM1 reg	-	-	-	-	-	-	0.25	0.29
DEFM3 reg	-	-	-	-	-	-	0.25	-
log Google topic reg	-	-	-	-	-	-	0.25	-
IMCA indus BDF reg	-	-	-	-	-	-	-	0.29
Prev stock indus BDF	-	-	-	-	-	-	-	0.63
Prev prod indus BDF reg	-	-	-	-	-	-	-	0.54

Notes : The table reports the frequency of selection of the most important variables for each method. The importance of each variable for the RF models is measured by using out-of-bag (OOB) samples.

TABLE 5.11 – Nowcast performance during the full out-of-sample period – 2016M1-2021M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Ridge	0.165	0.632	-0.154	0.3836	2
Lasso	0.219	0.512	-0.532	0.3778	4
EN	0.188	0.579	-0.321	0.3778	4
EN0.5	0.211	0.529	-0.478	0.3778	4
RF	0.174	0.611	-0.222	0.3836	2
DFM	0.143	0.681	0.000	1.0000	1
AR(1)	0.448	0.000	-2.139	0.2474	7

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 5.12 – Nowcast performance during the full out-of-sample period - 2016M1-2021M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank	
Ridge	0.155	0.653	0.996	0.985	0.8528	3
Lasso	0.173	0.615	0.978	0.984	0.6909	18
EN	0.170	0.620	0.981	0.984	0.7506	16
EN0.5	0.176	0.608	0.975	0.983	0.6459	19
RF	0.173	0.614	0.978	0.984	0.6459	19
dcsis+en05	0.160	0.642	0.991	0.985	0.7788	5
dcsis+en	0.145	0.676	1.006	0.986	0.9012	2
dcsis+lasso	0.160	0.643	0.991	0.985	0.7788	5
dcsis+ridge	0.159	0.644	0.992	0.985	0.7788	5
dcsis+rf	0.160	0.643	0.991	0.985	0.7788	5
sirs+en05	0.165	0.631	0.986	0.984	0.7788	5
sirs+en	0.158	0.648	0.993	0.985	0.7788	5
sirs+lasso	0.168	0.626	0.983	0.984	0.7506	16
sirs+ridge	0.163	0.635	0.988	0.984	0.7788	5
sirs+rf	0.173	0.613	0.978	0.983	0.6306	21
mdcsis+en05	0.162	0.638	0.989	0.985	0.7788	5
mdcsis+en	0.145	0.677	1.006	0.986	1	1
mdcsis+lasso	0.160	0.643	0.991	0.985	0.7788	5
mdcsis+ridge	0.159	0.644	0.992	0.985	0.7788	5
mdcsis+rf	0.163	0.637	0.988	0.984	0.7788	5
DFM	0.151	0.663	1.000	0.986	0.8528	3
AR(1)	0.448	0.000	0.703	0.957	0.2153	22

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 5.13 – Nowcast performance during non-crisis period - 2016M1-2019M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Ridge	0.0070	0.273	0.148	1	1
Lasso	0.0072	0.247	0.119	0.6771	5
EN	0.0071	0.258	0.131	0.7686	2
EN0.5	0.0071	0.261	0.134	0.7686	2
RF	0.0072	0.248	0.119	0.6838	4
DFM	0.0082	0.146	0.000	0.3352	6
AR(1)	0.0096	0.000	-0.171	0.1457	7

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 5.14 – Nowcast performance during crisis period - 2020M1-2021M12 – with the small database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank
Ridge	0.480	0.637	-0.166	0.3679	2
Lasso	0.641	0.516	-0.558	0.3679	2
EN	0.551	0.584	-0.339	0.3679	2
EN0.5	0.618	0.533	-0.503	0.3679	2
RF	0.508	0.616	-0.235	0.3679	2
DFM	0.412	0.689	0.000	1.0000	1
AR(1)	1.324	0.000	-2.217	0.2445	7

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 5.15 – Nowcast performance during non-crisis period - 2016M1-2019M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank	
Ridge	0.0071	-0.024	1.001	0.005	0.9885	6
Lasso	0.0070	-0.009	1.001	0.020	0.9938	4
EN	0.0069	0.000	1.001	0.028	1	1
EN0.5	0.0069	0.000	1.001	0.028	1	1
RF	0.0069	0.000	1.001	0.028	1	1
dcsis+en05	0.0071	-0.031	1.001	-0.002	0.9883	7
dcsis+en	0.0078	-0.132	1.000	-0.099	0.9370	18
dcsis+lasso	0.0072	-0.047	1.001	-0.017	0.9852	10
dcsis+ridge	0.0071	-0.021	1.001	0.008	0.9888	5
dcsis+rf	0.0077	-0.120	1.000	-0.088	0.9067	19
sirs+en05	0.0072	-0.046	1.001	-0.016	0.9839	12
sirs+en	0.0077	-0.117	1.000	-0.086	0.9567	17
sirs+lasso	0.0072	-0.035	1.001	-0.006	0.9880	8
sirs+ridge	0.0072	-0.048	1.001	-0.018	0.9843	11
sirs+rf	0.0074	-0.069	1.001	-0.039	0.9790	14
mdcsis+en05	0.0074	-0.068	1.001	-0.037	0.9670	16
mdcsis+en	0.0086	-0.242	0.999	-0.207	0.4433	22
mdcsis+lasso	0.0073	-0.051	1.001	-0.021	0.9839	12
mdcsis+ridge	0.0072	-0.042	1.001	-0.013	0.9880	8
mdcsis+rf	0.0076	-0.096	1.000	-0.064	0.9697	15
DFM	0.0080	-0.153	1.000	-0.120	0.8938	20
AR(1)	0.0096	-0.387	0.998	-0.348	0.5366	21

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .

TABLE 5.16 – Nowcast performance during crisis period - 2020M1-2021M12 – with the large database.

Models	MSE(%)	$R_{OOS}^2(AR)$	$R_{OOS}^2(DFM)$	MCS	rank	
Ridge	0.452	0.659	-0.034	0.295	0.8301	2
Lasso	0.504	0.620	-0.153	0.214	0.701	16
EN	0.497	0.624	-0.138	0.224	0.6371	18
EN0.5	0.513	0.612	-0.175	0.200	0.5372	19
RF	0.505	0.619	-0.156	0.212	0.5372	19
dcsis+en05	0.466	0.648	-0.067	0.273	0.7443	5
dcsis+en	0.419	0.683	0.040	0.346	0.8301	2
dcsis+lasso	0.465	0.649	-0.065	0.274	0.7443	5
dcsis+ridge	0.463	0.650	-0.061	0.277	0.7443	5
dcsis+rf	0.464	0.649	-0.063	0.276	0.7443	5
sirs+en05	0.481	0.637	-0.101	0.250	0.7443	5
sirs+en	0.457	0.655	-0.046	0.287	0.7443	5
sirs+lasso	0.488	0.631	-0.118	0.238	0.701	16
sirs+ridge	0.475	0.641	-0.088	0.259	0.7443	5
sirs+rf	0.505	0.619	-0.156	0.213	0.5152	21
mdcsis+en05	0.471	0.644	-0.078	0.265	0.7443	5
mdcsis+en	0.417	0.685	0.046	0.350	1	1
mdcsis+lasso	0.465	0.649	-0.063	0.275	0.7443	5
mdcsis+ridge	0.464	0.650	-0.062	0.277	0.7443	5
mdcsis+rf	0.473	0.643	-0.082	0.263	0.7443	5
DFM	0.437	0.670	0.000	0.319	0.8301	2
AR(1)	1.324	0.000	-2.031	-1.065	0.1583	22

Notes :  $R_{OOS}^2(AR)$  and  $R_{OOS}^2(DFM)$  are the  $R_{OOS}^2 = 1 - (MSE_{model}/MSE_{benchmark})$  where the benchmark model is the AR(1) model and the DFM model, respectively. \* and \*\* Significant at the 5% and 10% level, respectively, for the SPA<sub>c</sub> test, the null of the Hansen test for superior predictive ability. MCS denotes the  $p$ -value of the statistic  $T_{\max \mathcal{M}}$  of Hansen et al. (2011) based on the MSE loss function. Rank gives model ranking position based on the  $\widehat{\mathcal{M}}_{90\%}^*$ .



FIGURE 5.4 – Cumulative squared prediction errors during the full out-of-sample period – 2016M1-2021M12 – with the regional database.

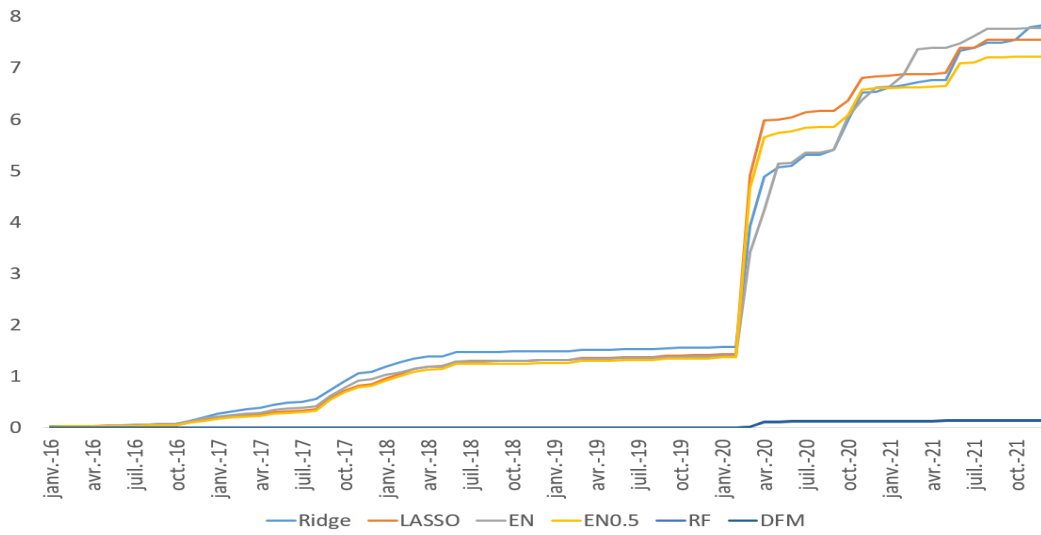
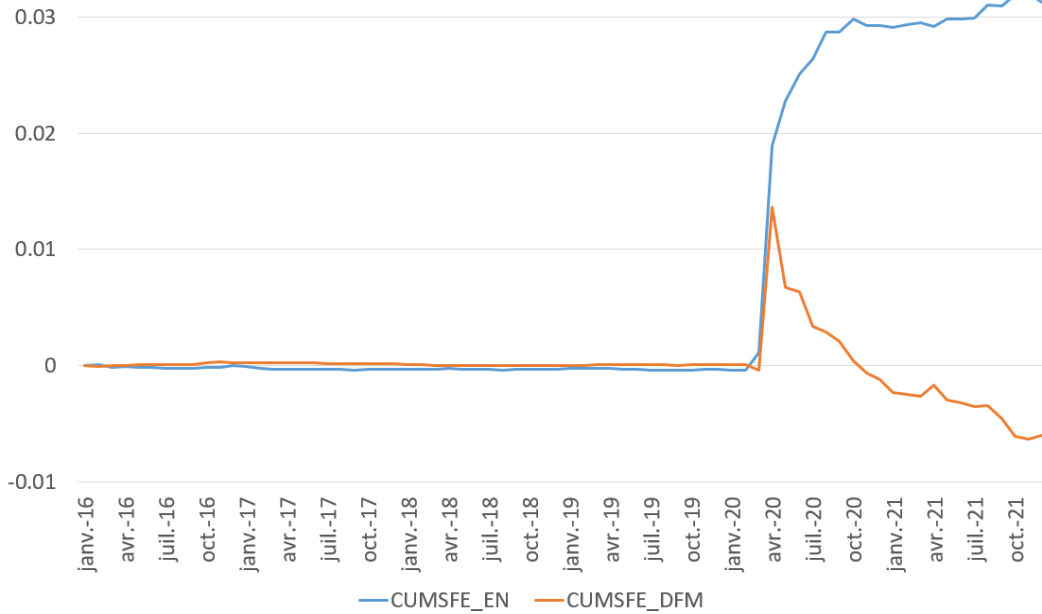


FIGURE 5.5 – Cumulative sum of squared forecast error (CUMSFE) from the regional and large databases.



*Notes* : Cumulative sum of squared forecast error (CUMSFE) loss differential of the nowcasts provided by the EN model or the DFM estimated from the large database against those obtained from the EN model or the DFM estimated from the regional database.

# Conclusion générale

L'enjeu principal de cette thèse était d'analyser de manière efficace la conjoncture au niveau du chômage. Nous espérons avoir contribué au développement scientifique de cette littérature. L'analyse conjoncturelle des cycles classiques est un phénomène largement étudié au niveau international, mais également à l'échelle nationale. Cependant, l'analyse conjoncturelle du chômage souffrait d'un manque d'études se consacrant à l'estimation de son cycle et au développement d'une datation de référence. Nous continuons de soutenir qu'une variable d'intérêt telle que le chômage, aussi discuté dans les médias et dans la sphère politico-économique, doit être analysée de la même manière que l'activité économique. Dans un premier temps, nous nous sommes attelés à l'étude des cycles d'emploi, ainsi que de leurs caractéristiques. Par ailleurs, nous avons souhaité mettre en avant une analyse de la synchronisation permettant de comprendre le contexte temporel dans lequel s'inscrivaient nos cycles. Afin de parfaire cette étude, nous l'avons généralisée à l'ensemble des régions françaises.

Dans un second temps, et après avoir analysé le contexte des cycles, notamment régionaux, nous avons cherché à développer un modèle de prévision mensuelle en plusieurs étapes. La première étape répondait à la littérature grandissante sur les *webdata*. La seconde étape constituait le cœur de cette seconde partie, puisqu'elle s'attachait à démontrer l'utilité des méthodes de réduction de dimensionnalité, au service de la prévision des demandeurs d'emploi. Le **Machine Learning** et le raisonnement économique nous ont permis d'aboutir à un double résultat. Le premier concerne l'identification des variables pertinentes à la prévision des DEFM, et le second est la performance de certains modèles sur le modèle de référence.

La première partie s'est d'abord intéressée à la désaisonnalisation des séries temporelles, ainsi qu'à l'extraction de la Tendence-Cycle. Nous renseignons aussi une étude de la détection des *outliers* en expliquant la raison de leurs apparitions. Ce travail préliminaire nous a permis de mesurer un premier cycle d'emploi national, ses caractéristiques et la synchronisation des cycles départementaux de la région des Pays de la Loire.

Dans le chapitre 1, nous nous concentrons sur les demandeurs d'emploi en catégorie A collectée par Pôle Emploi au niveau national, puis au niveau régional, avec un focus sur la région des Pays de la Loire et ses départements, à savoir la Loire-Atlantique, le Maine-et-Loire, la Mayenne, la Sarthe et la Vendée. Cette première analyse nous permet de dater un premier cycle d'emploi national, en utilisant l'algorithme de Bry-Boschan. Nous avons réalisé une comparaison du cycle d'emploi et du cycle de l'activité. Ce faisant, nous avons identifié le

caractère contracyclique du cycle d'emploi sur le cycle des affaires. Cependant, il émerge de ce résultat, que le cycle d'emploi renvoie à une forte inertie de l'expansion du cycle d'emploi suite à une période de récession de l'activité économique.

Nous avons cherché par la suite à mesurer les caractéristiques des cycles d'emploi respectivement : la durée, l'amplitude et la sévérité. Nos résultats ont mis en évidence un cycle asymétrique, avec des phases d'expansion plus longues et plus amples que les phases de contraction. Un constat faisant écho à ce que nous avons appelé précédemment, l'inertie du chômage. En outre, les trajectoires d'inflexion conjoncturelle, quantifiées par l'indice de sévérité, nous apprennent que les phases d'expansion et de contraction sont plutôt progressives, c'est-à-dire qu'en début de période, juste après un point de retournement, l'augmentation ou la diminution est douce.

Une analyse présente dans ce chapitre concerne l'étude du cycle d'accélération de l'emploi. Les résultats sont équivoques à la littérature. Le cycle d'accélération de l'emploi anticipe de quelques mois les points de retournement du cycle de l'emploi. L'intérêt de cette comparaison est d'identifier à l'avance les potentiels points de retournement du cycle d'emploi grâce aux points de retournement du cycle d'accélération de l'emploi. On retrouve ici l'idée de capter une information précoce sur l'arrivée d'un point de retournement.

Le dernier résultat important de ce chapitre, découle de l'analyse de la synchronisation des cycles départementaux de la région des Pays de la Loire. Dans un premier temps, nous avons utilisé l'indice de concordance et fait état d'une forte synchronisation cyclique au niveau des départements de la région étudiée. Le maillage du tissu économique étant plutôt la zone d'emploi, nous pouvons considérer que cette analyse manque de pertinence, puisqu'elle démontre uniquement que les départements d'une même région connaissent une évolution cyclique identique. Ainsi, nous avons cherché à améliorer cette étude de la synchronisation cyclique au travers d'une analyse novatrice, utilisant l'indice de Moran Modifié.

Dans un second temps, cet indice est calculé sur les cinq départements de la région des Pays de la Loire afin de conserver le même périmètre géographique. Nous ressortons de cette analyse un résultat important. L'indice moyen avant la crise de 2008 est plus faible que l'indice moyen post-crise. Ce résultat démontre une meilleure synchronisation des cycles d'emploi départementaux sur la période de l'après-crise. L'hypothèse soutenue, qui sera nettement

nuancée dans le chapitre 2, est l'impact de la fusion de l'ANPE et des Assedic fin 2008, qui aurait contribué à une meilleure gestion des demandeurs d'emploi. Nous pensions à une meilleure diffusion des politiques publics d'emploi ainsi qu'à une meilleure formation des agents de Pôle Emploi. Depuis cette fusion, les conseillers sont de plus en plus formés aux deux missions de Pôle Emploi, à savoir le placement et l'indemnisation.

Dans le Chapitre 2, qui est en réalité un chapitre 1 bis, nous réalisons une extension du premier. Nous mesurons les cycles d'emploi de toutes les régions françaises avec l'ancien découpage statistique précédente la loi NOTRe promulguée en 2015. Le grand intérêt de cette étude réside dans la comparaison des cycles d'emploi, entre les différentes régions françaises afin d'identifier les différences notables, en matière de datation cyclique. Les résultats observés sont similaires à la précédente étude. Les cycles d'emploi sont asymétriques en faveur des phases d'expansion. Certaines régions sortent du lot avec une asymétrie en faveur des phases de contraction, tel que la Corse et la Champagne-Ardenne, mais l'extrême majorité possèdent des phases d'expansion plus longues et plus ample que les phases de contraction. Le fait intéressant de cette étude apparaît dans l'analyse de la synchronisation avec l'indice de Moran Modifié. Grâce aux données de l'ensemble des régions françaises, nous avons pu calculer cet indice en captant toute l'information régionale à l'échelle de la France. Les résultats sont interprétés de manière descriptive.

Par ailleurs, l'hypothèse présentée auparavant est remise en cause, car nous n'observons plus de différence majeure entre la période avant la crise de 2008 et la période après cet crise. Cependant, puisque que notre indice est calculé sur la base de tous les cycles d'emploi régionaux, nous avons pu identifier un effet global. Nous avons constaté que les périodes d'expansion des cycles d'emploi étaient identifiées, pendant les diminutions du niveau de l'indice. À l'inverse, les périodes de contraction du cycle de l'emploi s'observent, au moment où le niveau de l'indice augmente ou que ce niveau soit déjà élevé. En le résumant simplement, quand tout va bien, les cycles régionaux se synchronisent, alors que quand tout va mal, les cycles se désynchronisent et forment une sorte de cacophonie. Nous avons nommé ce phénomène, l'effet de couplage. Selon nous, cet effet est opérant pendant que la situation économique d'un pays est stable, puis il s'atténue quand la situation devient moins favorable. La construction de l'indice étant réalisée en début de thèse, nous n'avons pas pu capter la période du confinement lié à l'épidémie du Covid-19. Néanmoins, il aurait été très instructif

d’observer la synchronisation cyclique d’emploi des régions pendant cette période caractérisée par une forte instabilité.

La deuxième partie tente de calibrer un modèle de prévision mensuelle des DEFM. Les enquêtes de conjoncture françaises ont été mises à l’honneur ainsi que les données issues de Google Trend. Nous procéderons au même schéma que pour la première partie en réalisant un modèle de prévision sur la variable nationale avant de nous concentrer sur la région des Pays de la Loire.

Dans le chapitre 3, nous proposons une revue de la littérature des travaux qui se sont intéressés à la prévision du chômage ou de l’emploi en utilisant les *webdata*, et plus particulièrement les données fournies par Google Trends. La littérature commence par des résultats optimistes en matière de performance des prévisions. En revanche, certaines études pointent du doigt la relation entre la variable construite avec Google et leurs variables d’intérêt. Nous nous retrouvons dans cette dualité de résultat.

En premier lieu, nous constatons un fort pouvoir prédictif de notre variable Google construite à partir du mot-clé ”Emploi”. Nous obtenons ainsi des résultats encourageants permettant d’établir la pertinence des *webdata* dans une optique de prévision à court terme du taux de croissance des DEFM. Nos modèles de prévision basés sur la variable explicative de Google Trends sont plus précis que le modèle de référence, et ils possèdent des erreurs quadratiques moyennes plus faibles. En outre, nous avons observé une certaine instabilité dans nos résultats. En effet, la variable Google et nos prévisions s’avèrent altérées par le pic lié au confinement suite à l’épidémie de Covid-19. Ce pic est présent dans notre variable dépendante et dans notre variable explicative, perturbant ainsi la cohérence de nos prévisions. Mis à part ce point de nuance, la donnée issue de Google Trends semble pertinente pour la prévision des DEFM. Cela démontre, d’une certaine manière, l’information captée par cette mesure d’intensité de recherche. Certes, le chercheur est dépendant du bon vouloir de Google qui se réserve le droit de modifier l’information qu’il transmet, mais la disponibilité et la fréquence des données reste un atout majeur pour un exercice de prévision, comme celui que nous avons réalisé.

Dans le chapitre 4, nous avons souhaité construire un modèle de prévision mensuelle des DEFM performant. Pour ce faire, nous avons sélectionné un grand nombre d’enquêtes de

conjoncture, ainsi que d'autres variables explicatives reliées au marché du travail et au chômage. Notre objectif était d'utiliser le délai, plus précoce, de parution des enquêtes de conjoncture, afin de fournir une prévision anticipée du chômage selon Pôle Emploi. La grande quantité de variables sélectionnées nous impose l'utilisation des méthodes de réduction de dimensionnalité. Les résultats peuvent être présentés en deux points.

Le premier point concerne la fréquence de sélection des variables jugées utiles à la prévision par les méthodes de régression pénalisée et plus généralement de *Machine Learning*. Les variables issues de la base de données la plus rapidement disponible sont : la variable de Google Trends, et les enquêtes de conjoncture dans l'industrie relatives à la tendance passée et prévue de la production publiées par l'Insee. De plus, les variables les plus sélectionnées depuis la base de données plus tardive, c'est-à-dire les données qui sont publiées moins rapidement, essentiellement les enquêtes de conjoncture de la Banque de France, sont : la variable de Google Trends et les enquêtes de conjoncture de la Banque de France qui se substituent aux enquêtes de conjoncture de l'Insee. Ce résultat suggère que les enquêtes BDF possèdent davantage d'informations que les enquêtes Insee pour un objectif de prévision des DEFM à court terme. Mais, les résultats indiquent aussi une bonne performance de la variable de Google Trends, confirmant les résultats présentés dans le chapitre 3. Par ailleurs, le fait de retrouver les variables autorégressives dans les plus sélectionnées confirme, que le plus souvent, les variables retardées sont très utiles pour la prévision, surtout que la variable dépendante est caractérisée par une forte inertie.

Le second point est destiné à la mesure de la performance des modèles de prévision sous forme d'une mise en compétition ou pour utiliser le terme anglais *horse-race*. Nous décomposons notre compétition en trois échantillons. La première relative à l'ensemble de notre période disponible ; La seconde à la période post-crise du Covid-19 ; et la troisième à la période pré-crise du Covid-19. Les résultats montrent que les modèles Lasso, Elastic-Net et Elastic-Net adaptatif présentent une bonne performance de prévision à très court terme, et sont donc pertinents pour la prévision "immédiate" du taux de croissance des demandeurs d'emploi. Leurs gains de précision sont importants pendant la crise du Covid-19, alors qu'ils fournissent des performances similaires à celles du modèle de facteur dynamique pendant la période sans crise. Ce constat envoie un bon signal aux régressions pénalisées par rapport aux modèles à facteurs puisque cela peut signifier, que les régressions pénalisées délivrent des performances

identiques en matière de prévision que les modèles à facteurs et de surcroît permettent une meilleure identification des variables pertinentes et donc de leur interprétation.

Dans le chapitre 5, nous nous intéressons à la prévision du taux de croissance de la DEFM de la région des Pays de la Loire. L'aspect régional est le second fil rouge de notre thèse, en plus de notre variable dépendante. Nous cherchons à construire un modèle de prévision mensuelle en conservant la méthodologie précédemment utilisée. Nous avons constaté que les enquêtes de conjoncture nationales apportaient une information non-négligeable à la prévision du taux de croissance des DEFM. De manière similaire à la précédente étude, les données issues de Google Trends font partie des données les plus sélectionnées parmi les différentes méthodes de sélection. Les performances de prévision à court terme semblent s'améliorer, à mesure que les enquêtes nationales sont intégrées, pour les régressions pénalisées. Un résultat que l'on observe pas pour le modèle factoriel dynamique. Par ailleurs, le modèle Elastic-Net avec les présélections DCSIS et MDCSIS montre des gains de précision pendant la crise Covid-19 en ajoutant les enquêtes nationales, ce qui suggère que les enquêtes nationales semble apporter une information supplémentaire pendant cette période de crise.

### **Perspectives de recherches futures**

Beaucoup de perspectives de recherches émergent de ce travail. La première est en lien avec les conclusions du chapitre 1. La création de la datation du cycle d'emploi de référence amène une réflexion autour de la nécessité d'estimer un indicateur de retournement cyclique. En effet, la création d'un indicateur de retournement doit s'appuyer sur une chronologie de référence des points de retournement des cycles à partir de laquelle on pourra juger de la qualité de l'indicateur.

Une deuxième idée serait d'étudier la synchronisation cyclique au niveau des zones d'emploi. En effet, réitérer cette analyse sur le périmètre des zones d'emploi semble pertinent à plusieurs égards. La zone d'emploi permet de capter l'information pertinente à l'échelle d'un territoire. Nous avons déjà mentionné que les variations intra-régionales pouvaient masquer une partie des variations inter-régionales. La zone d'emploi apparaît utile aux diagnostics conjoncturels locaux.

Une dernière idée vient de la partie deux, relative aux modèles de prévision mensuelle du taux de croissance des demandeurs d'emploi. La réplication de ce même travail sur les données



du chômage Insee pourrait être un nouveau point de départ. Bien que nous ayons défendu les données de Pôle Emploi, disponible plus rapidement et avec une fréquence mensuelle, la mesure de l'Insee permettrait de réaliser une comparaison des performances du modèle de prévision à l'échelle internationale. Par ailleurs, nous pourrions creuser la prévision *via* les données issues de Google Trends, en élargissant de manière conséquente le nombre de mots-clés. Les techniques de *Machine Learning* seraient un outil formidable pour réaliser cette extension. Nous pourrions aussi nous concentrer sur les demandeurs d'emploi appartenant à la catégorie des jeunes plutôt que sur l'ensemble de la population. La variable Google serait probablement encore plus performante.

---

**Titre :** Analyse et prévision du chômage

**Mots clés :** cycle d'emploi ; demandeurs d'emploi ; *Machine Learning* ; prévision ; série temporelle

Plus la connaissance de l'évolution de la situation de l'économie à venir est rapide, plus les décideurs politiques et économiques pourront mettre en place des stratégies efficaces au niveau national et régional. Ce premier postulat annonce une problématique qui se traduit comme suit ; comment analyser de manière efficace la conjoncture au niveau du chômage ?

Cette thèse s'emploie à créer une datation de référence du cycle d'emploi des demandeurs d'emploi français. Les caractéristiques du cycle sont mesurées ainsi que la synchronisation identifiant les comportements idiosyncratiques. Puis, nous élaborons un modèle de prévision mensuelle à partir des données des demandeurs d'emploi collectées par Pôle Emploi afin d'éclairer le décideur public.

---

**Title :** Analysis and forecasting of unemployment

**Keywords :** employment cycle ; jobseekers ; Machine Learning ; forecasting ; time series

*The faster the knowledge of the evolution of the future economic situation, the better the political and economic decision-makers will be able to set up efficient strategies at the national and regional level. This first postulate announces a problem that can be translated as follows: how to effectively analyze the economic situation in terms of unemployment?*

*This thesis seeks to create a baseline dating of the employment cycle of French job seekers. The characteristics of the cycle are measured as well as the timing identifying idiosyncratic behaviors. Then, we develop a monthly forecasting model based on job seeker data collected by Pôle Emploi in order to inform the public policy maker.*