M2 APE SEP - UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

2024

LA CRÉATION D'ENTREPRISES DANS L'INDUSTRIE MANUFACTURIÈRE EN FRANCE DEPUIS 2009

Une analyse et une prévision temporelle

Ecrit par :
Alexandre Brunet

SOMMAIRE

Introduction 3

Données et méthode 4

1. Analyses préalables 6

2. Stationnarité et saisonnalité 11

2.1. La non-stationnarité de la série nécessitant une différenciation 11

2.2. Des saisonnalités trimestrielle et quadrimestrielle apparentes 14

3. Choix des paramètres 16

3.1. Recherche des meilleurs paramètres selon le BIC 16

3.2. Test du modèle suggéré et adaptation 17

4. Vérification et prévisions 18

5. Comparaison avec l'intelligence artificielle : Prophet 22

Conclusion 25

Ressources 26

Table des figures 27

Table des tableaux 29

INTRODUCTION

Selon l'INSEE, la production en valeur de l'industrie manufacturière française en 2022 s'est élevée à 922 milliards d'euros sur les 5042,6 milliards d'euros qu'a produits l'ensemble des secteurs d'activité français (en prix courants)[1]. Ce résultat témoigne de la contribution non-négligeable de cette industrie à l'économie française. En effet, avec une part de 18,3% du PIB en 2022, elle se positionne comme un pilier essentiel de la richesse nationale. Cette proportion souligne la dépendance de la prospérité française à l'égard de la production manufacturière, représentant ainsi un moteur fondamental de croissance économique.

Dans ce contexte, une analyse de la création d'entreprises au sein de l'industrie manufacturière devient encore plus importante. Comprendre les tendances et la saisonnalité de ce domaine contribuera à éclairer les perspectives de cette industrie clé. De cette manière, cette étude vise à fournir des informations sur la dynamique entrepreneuriale au sein de l'industrie manufacturière en France depuis janvier 2000 à décembre 2021 ainsi qu'une vision de l'avenir à court terme. Dans le même temps, avec l'apparition de l'intelligence artificielle, nous comparerons nos prévisions avec un algorithme produit par l'intelligence artificielle (et dont nous ne connaissons pas vraiment le fonctionnement dans le détail).

[1] Pour avoir les données, suivre le lien suivant : https://www.insee.fr/fr/statistiques/6793598?sommaire=6793644#consulter puis télécharger le tableau 6.101-103 - Production par branche (38 postes).

DONNEES ET METHODE

Les données sur lesquelles nous allons travailler proviennent de l'INSEE qui enregistre l'état-civil de toutes les entreprises et leurs établissements situés en France métropolitaine et d'outre-mer, indépendamment de leur forme juridique et de leur secteur d'activité. Ces renseignements sont enregistrés dans le Système Informatique pour le Répertoire des Entreprises et de leurs Etablissements (SIRENE). C'est la Direction des Statistiques d'Entreprises (DSE) qui s'est chargée de récolter les données pour les diffuser, notamment grâce au Répertoire des entreprises et des établissements (REE) jusqu'en 2022. Le champ couvert par ce dispositif est constitué des unités légales[2] productives et marchandes (produisant et/ou vendant des biens et services). Ce qui signifie que le champ couvert comprend l'industrie, la construction, le commerce et les services. La DES s'est, entre autres, intéressée aux statistiques concernant la création d'entreprises et leurs établissements afin de suivre leur évolution. L'intérêt est que cela représente une information utile pour le suivi du cycle conjoncturel au niveau national (le cas qui va nous intéresser ici), mais aussi au niveau régional et départemental.

Dans notre cas, le but était initialement d'utiliser les données de création d'entreprises à partir de 2000, cependant de janvier 2000 à décembre 2021, les données ne sont pas homogènes. En effet, à partir de janvier 2009, les séries sont calculées en nomenclature agrégée « NA ». L'intérêt de ce changement était que cette nomenclature permettait de mieux suivre le secteur des services et de favoriser les comparaisons internationales. Les statistiques de création d'entreprises incluent les demandes des créations en auto-entrepreneur à compter de janvier 2009. Voilà pourquoi, pour des raisons de cohérences, nous n'utiliserons que les données à partir de janvier 2009. Outre ce problème d'homogénéité, en 2022, le REE a été remplacé par le Système d'Information de la Démographie des Entreprises (SIDE), de ce fait, les séries de créations d'entreprises ont été recalculées rétrospectivement jusqu'à 2000. Finalement, nous avons deux jeux de données : le 1er retraçant le nombre d'entreprises créées de 2000 à 2021 calculé par le REE[3] et le deuxième le nombre d'entreprises créées de 2000 à 2023 calculé (et recalculé) par le SIDE[4]. Cependant, l'INSEE appelle à la prudence dans l'analyse de ce dernier

[2] Les unités statistiques utilisées sont l'unité légale et l'établissement.

[3] https://www.insee.fr/fr/statistiques/serie/001564286#Tableau : série arrêtée de 2000 à 2021.

[4] https://www.insee.fr/fr/statistiques/serie/010755561#Tableau : nouvelle série de 2000 à 2023. Selon

jeu de données. De ce fait, nous allons conserver l'ancienne série composée de 156 observations mensuelles de janvier 2009 à décembre 2021.

Concernant notre méthode de travail, nous allons étudier la tendance et surtout la saisonnalité de la série et en la prenant en considération pour choisir le meilleur moyen (les meilleurs paramètres) de faire des prédictions les plus précises et fiables possibles. Pour cela, nous allons avoir recours au modèle $SARIMA_s(p,d,q)(P,D,Q)$ et tout au long de notre analyse, nous dénicherons tous les paramètres optimaux au fur et à mesure.

Ensuite, non allons vérifier la véracité que nos « résultats de prédiction » puis calculer différentes métriques permettant de savoir dans quelle mesure, nous sommes loin de la réalité. Enfin, nous allons comparer ces métriques (qui représentant donc les performances de nos résultats) avec les résultats obtenus par le biais de l'intelligence artificielle : l'algorithme Prophet.

l'INSEE : « Depuis le 1er janvier 2023 […]. Ce changement important fragilise temporairement le suivi mensuel des créations d'entreprises. Les évolutions des créations d'entreprises enregistrées sur les premiers mois de l'année 2023 doivent donc être interprétées avec une grande prudence. »

1. Analyses préalables

Pour commencer notre analyse, regardons comment le nombre de création d'entreprises se répartit (figure 1). On dénombre 19 observations extrêmes sur 156 observations, soit environ 12,2%, ce qui est non-négligeable. Ces observations sont les 19 mois où l'on a observé le plus de création d'entreprises, ce qui représente un problème car ces valeurs se concentrent pour la plupart en fin de série (tableau 1), ce qui va beaucoup jouer sur les prévisions. Nous verrons si les résultats seront satisfaisants en considérant ces valeurs aberrantes.

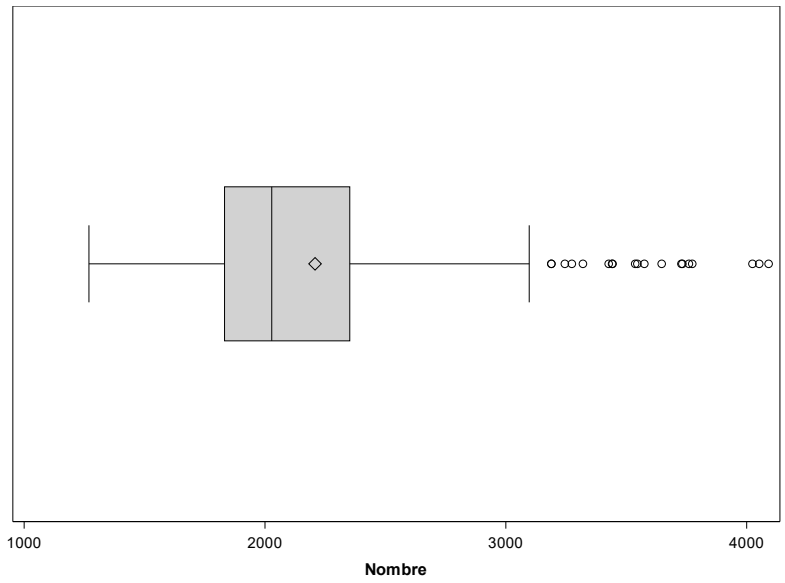

Figure 1 : Répartition du nombre de création d'entreprises au sein de l'industrie manufacturière française depuis 2009. **Lecture** : Le losange représente la moyenne et la ligne représente la médiane. **Source** : INSEE, 2022.

Date	Nombre
06-2021	4091
03-2021	4052
10-2021	4024
04-2021	3774
10-2019	3760
12-2021	3733
10-2020	3729
01-2021	3647
02-2021	3575
11-2021	3546
12-2020	3537
11-2020	3443

Date	Nombre
01-2020	3441
09-2021	3427
09-2020	3320
06-2020	3274
02-2020	3245
07-2020	3189
07-2021	3189

Tableau 1 : Les 19 mois où l'on a observé le plus grand nombre d'entreprises créées dans l'industrie manufacturière française depuis 2009. **Lecture** : En juin 2021, 4091 entreprises de l'industrie manufacturière ont été créés. **Source** : INSEE, 2022.

Comme expliqué dans la partie « Données et méthode », nous avons fait le choix d'exclure les données qui dataient d'avant 2009 en raison d'une modification de la comptabilité de création d'entreprises. Cela se traduit notamment par un « bond » dans le nombre de création d'entreprises (figure 2). En plus de la non-cohérence des données, ce bond peut potentiellement perturber la façon dont les prévisions sont faites et donc de faire des prévisions erronées.

En excluant ces données non-actualisées, nous pouvons voir correctement l'évolution du nombre de création d'entreprises dans l'industrie manufacturière française depuis 2009 (figure 3). À partir de là, nous pouvons voir qu'il y a une saisonnalité évidente, car il y a un même schéma (plus ou moins qui se répète), sauf à certains moments où le nombre d'entreprises créées n'est pas expliqué par la saisonnalité, notamment pendant la crise de la Covid-19 où très peu de personnes ont été incités à créer une entreprise et ceux qui l'ont fait, n'avaient pas d'autres choix que d'en créer.

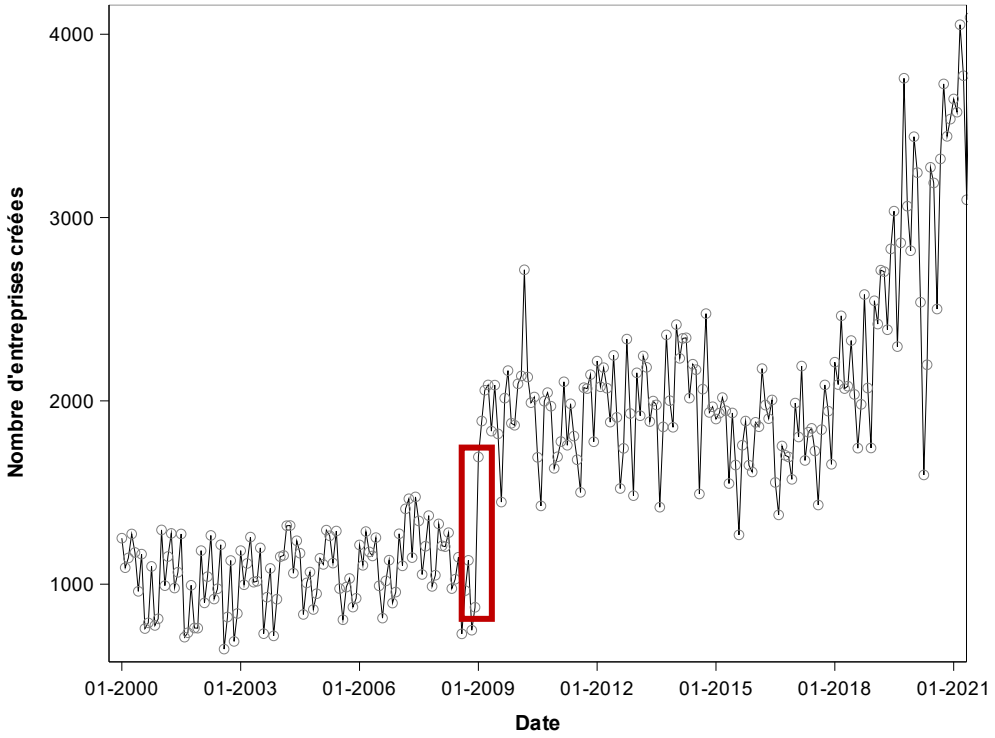

Figure 2 : Evolution du nombre de création d'entreprises dans l'industrie manufacturière française de janvier 2000 à décembre 2021. **Lecture** : Le bond de 2009 est représenté par le rectangle rouge. **Source** : INSEE, 2022.

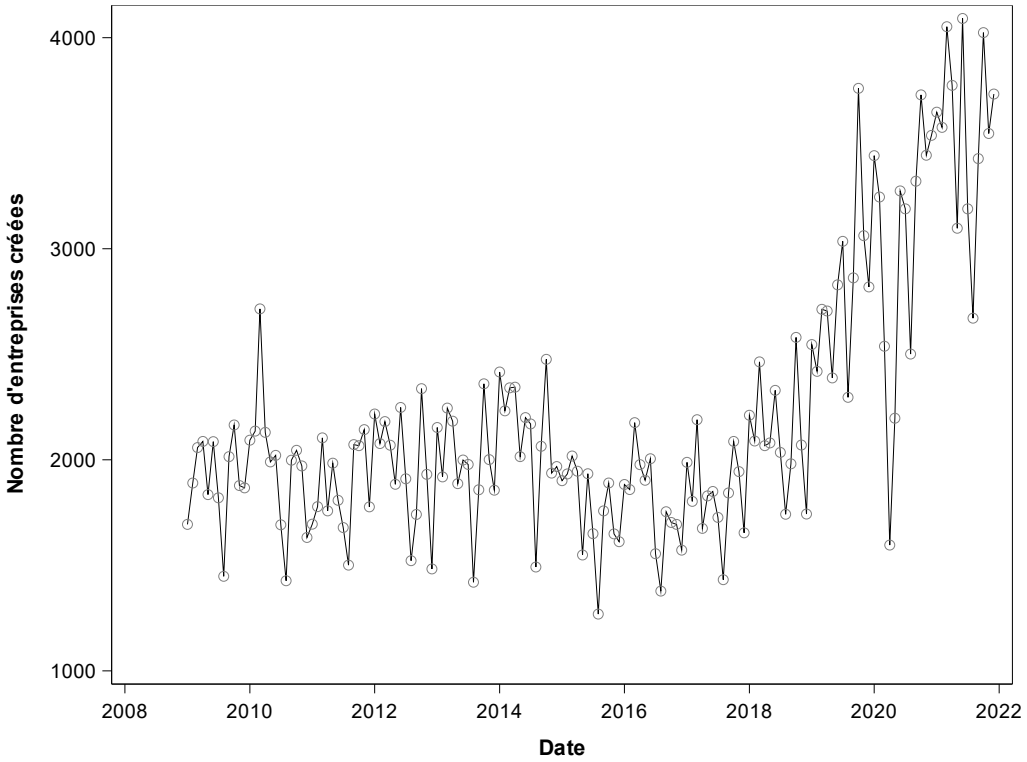

Figure 3 : Evolution du nombre de création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. **Source** : INSEE, 2022.

Pour visualiser la série sans l'effet de la saisonnalité, nous pouvons créer une série corrigée des variations saisonnières (CVS). L'intérêt d'effectuer cette tâche c'est de voir les observations influencées ou non par la saisonnalité que « subit » notre série. Dans notre cas, les observations qui sont concernées par la série CVS sont celles que nous avons mentionné précédemment : cf. période de la Covid-19 entre autres (figure 4). Autrement dit, ces observations, même si visuellement nous pouvons croire qu'elles suivent la saisonnalité, ne la suivent pas. C'est également le cas par l'observation datant de 2010 : une partie du nombre d'entreprises créées à ce moment-là tient de la saisonnalité, mais une autre partie tient à un événement extérieur au fonctionnement habituel de l'industrie manufacturière. De manière globale, nous pouvons voir que même en omettant les variations saisonnières, la création d'entreprises varie assez dans le temps.

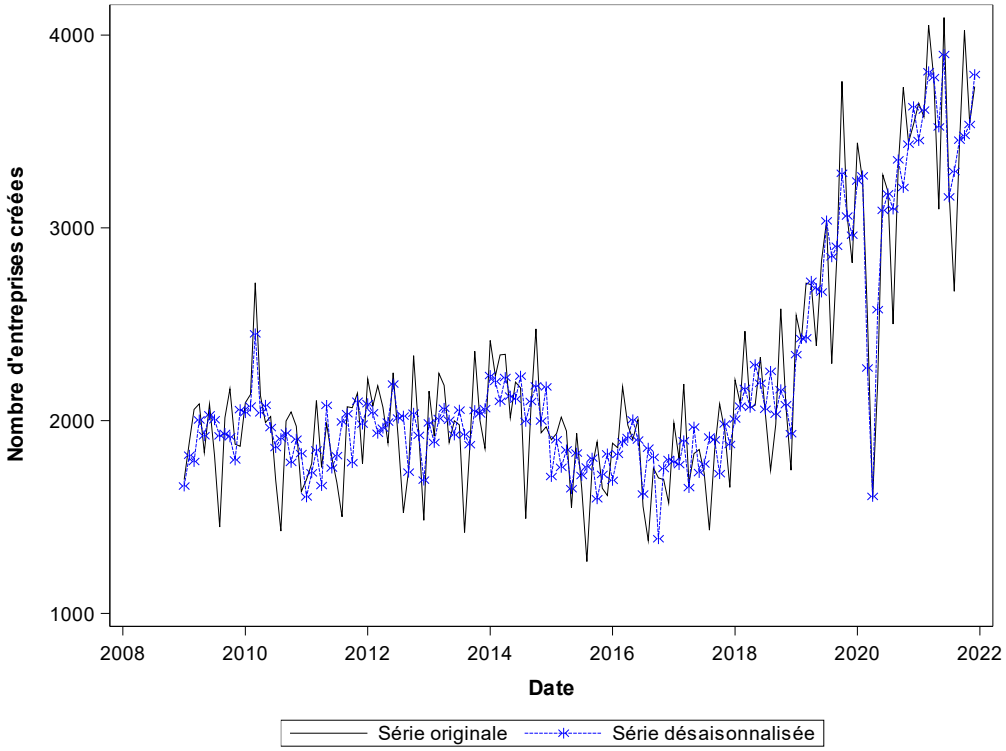

Figure 4 : Série CVS de la création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. **Source** : INSEE, 2022.

Enfin, nous pouvons visualiser la tendance générale de cette série. De 2009 à fin-2018, le nombre de création d'entreprises a varié légèrement à la hausse et à la baisse en restant à peu près au même niveau (figure 5). Mais à partir de 2019, ce nombre a fortement augmenté, malgré la crise de la Covid-19 qui a paralysé l'économie française pendant plusieurs mois. À partir de cette tendance, nous pouvons en déduire que les prévisions vont se caractériser par des

variations à la hausse et à la baisse, mais étant donné que la fin de notre série démontre une forte croissance du nombre de création d'entreprises (donc série non-stationnaire), nos prévisions sera également sûrement caractérisée par une tendance à la hausse. Pour finir, revenons sur les données aberrantes (cf. figure 1). En réalité, ce sont de fausses valeurs aberrantes. En effet, elles sont considérées ainsi étant donné les autres valeurs. Mais nous voyons que début 2019 plus d'entreprises ont été créés donc c'est normal que quelques valeurs soient considérées comme aberrantes étant donné la tendance qui a « explosé » après être resté au même point pendant plusieurs années (figure 5). Finalement, elles ne perturberont pas tant que ça les prédictions.

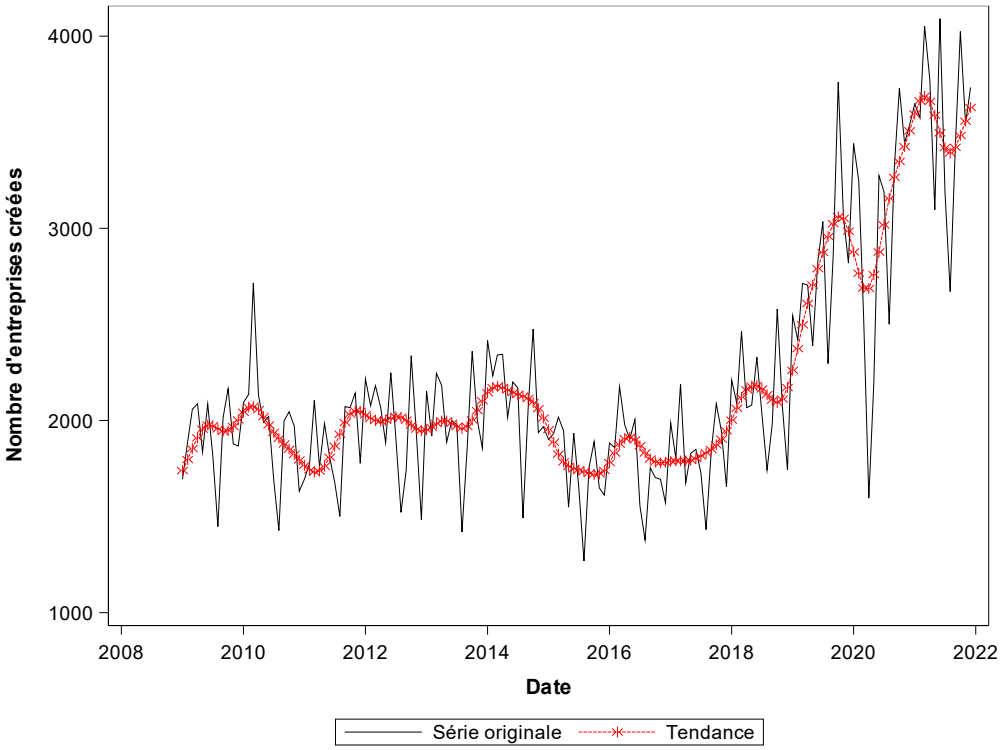

Figure 5 : Tendance de la création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. Source : **INSEE**, 2022.

2. Stationnarité et saisonnalité

2.1. La non-stationnarité de la série nécessitant une différenciation

La première chose que nous devons faire, c'est de stationnariser la série, sans cela, nous n'avons aucun moyen de faire des prédictions. Pour voir si notre série est initialement stationnaire, le test de Dickey-Fuller augmenté est un excellent moyen de confirmer ou non la nécessité de transformer les données. Dans notre cas, la série n'est clairement pas stationnaire (tableau 2) : le test rejette quasiment à chaque fois la stationnarité. De plus, les autocorrélations diminuent lentement à chaque retard, c'est une preuve de non-stationnarité (figure 6). Sans compter le fait que nous avons déjà repéré cela avec les différentes visualisations dans la partie précédente.

Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Moyenne zéro	0	-1.2581	0.4297	-0.58	0.4631		
	1	-0.4309	0.5841	-0.25	0.5938		
	2	0.6056	0.8313	0.65	0.8551		
	3	0.6456	0.8407	0.77	0.8787		
	4	0.6764	0.8479	0.83	0.8889		
	5	0.6652	0.8453	0.87	0.8960		
	6	0.6866	0.8501	0.84	0.8911		
Moyenne simple	0	-29.9936	0.0012	-3.85	0.0031	7.51	0.0010
	1	-20.6884	0.0081	-2.93	0.0450	4.40	0.0638
	2	-3.4176	0.6028	-0.85	0.8009	0.75	0.8798
	3	-1.9956	0.7772	-0.54	0.8784	0.57	0.9307
	4	-1.8270	0.7972	-0.50	0.8870	0.59	0.9238
	5	-1.1719	0.8685	-0.34	0.9149	0.53	0.9440
	6	-2.0272	0.7734	-0.53	0.8817	0.62	0.9148
Tendance	0	-49.8889	0.0005	-5.30	0.0001	14.10	0.0010
	1	-40.5742	0.0005	-4.33	0.0037	9.51	0.0010
	2	-11.3072	0.3399	-2.07	0.5598	2.72	0.6333
	3	-9.3026	0.4736	-1.82	0.6890	2.47	0.6839
	4	-8.9835	0.4973	-1.74	0.7275	2.28	0.7218
	5	-8.1351	0.5633	-1.65	0.7688	2.29	0.7204
	6	-9.8208	0.4360	-1.74	0.7273	2.22	0.7345

Tableau 2 : Tests de racine unitaire de Dickey-Fuller augmentés du nombre de création d'entreprises (série brute). **Lecture** : Dans la plupart des cas la Pr < Rho est inférieure au seuil des 5%. De ce fait, on rejette l'hypothèse selon laquelle la série est stationnaire.

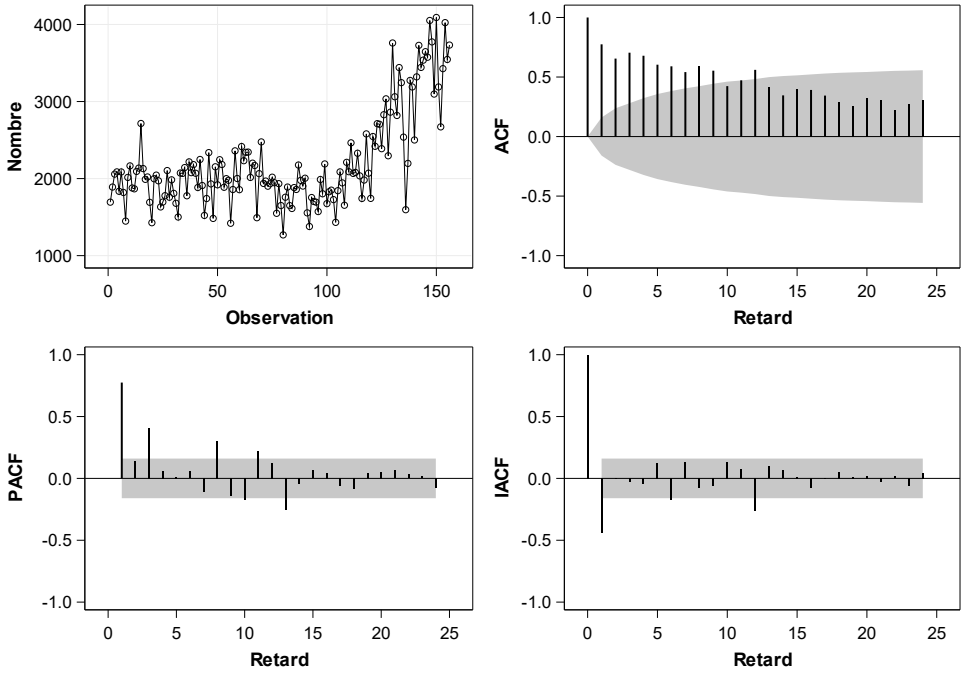

Figure 6 : Analyse des tendances et de la corrélation du nombre de création d'entreprises (série brute).

Nous venons donc à la conclusion que nous devons transformer les données pour rendre stationnaire la série. En cela, la série doit être différenciée, c'est-à-dire que $x_t = x_t - x_{t-1}$ pour x_i l'observation au temps t. Finalement, lorsque l'on a différencié une première fois, nous obtenons une série stationnaire, selon le test de Phillips-Perron (tableau 3) et nous voyons que le graphique des autocorrélations (ACF) ne fait plus apparaître une décroissante lente (figure 7).

Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau
Moyenne zéro	0	-217.492	0.0001	-19.24	<.0001
	1	-200.524	0.0001	-20.72	<.0001
	2	-171.170	0.0001	-29.10	<.0001
	3	-165.877	0.0001	-33.91	<.0001
	4	-167.708	0.0001	-31.91	<.0001
	5	-165.416	0.0001	-34.49	<.0001
	6	-163.911	0.0001	-36.69	<.0001
Moyenne simple	0	-217.495	0.0001	-19.18	<.0001
	1	-200.526	0.0001	-20.64	<.0001

Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau
	2	-171.172	0.0001	-28.98	<.0001
	3	-165.879	0.0001	-33.76	<.0001
	4	-167.711	0.0001	-31.77	<.0001
	5	-165.418	0.0001	-34.34	<.0001
	6	-163.914	0.0001	-36.52	<.0001
Tendance	0	-217.492	0.0001	-19.11	<.0001
	1	-200.523	0.0001	-20.57	<.0001
	2	-171.166	0.0001	-28.86	<.0001
	3	-165.871	0.0001	-33.62	<.0001
	4	-167.703	0.0001	-31.64	<.0001
	5	-165.412	0.0001	-34.19	<.0001
	6	-163.907	0.0001	-36.36	<.0001

Tableau 3 : Tests de la racine unitaire de Phillips-Perron. **Lecture** : La Pr < Rho est toujours bien inférieure au seuil des 5%, de ce fait on accepte l'hypothèse selon laquelle la série est stationnaire.

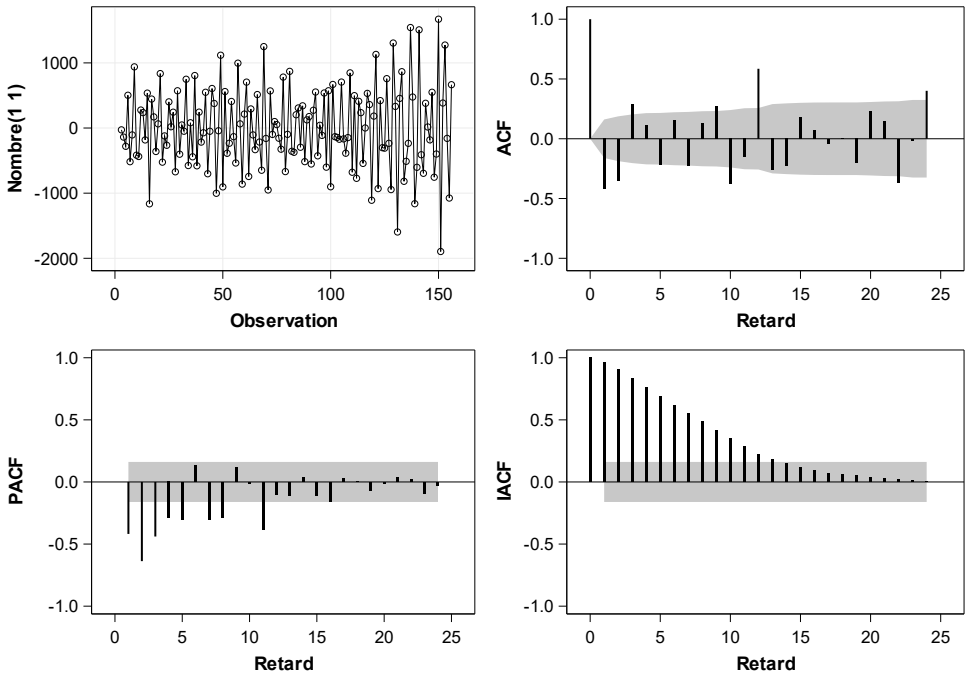

Figure 7 : Analyse des tendances et de la corrélation du nombre de création d'entreprises avec une seule différenciation.

Pour conclure sur cette partie, étant donné que nous avons différencié une seule fois ($d = 1$) les observations, nous pouvons déjà commencer à construire le modèle que nous voulons, car nous avons découvert quel sera le paramètre d : $SARIMA_s(p, 1, q)(P, D, Q)$.

2.2. Des saisonnalités trimestrielle et quadrimestrielle apparentes

Maintenant que la différenciation a été effectuée, penchons-nous sur la saisonnalité, cela nous permettra de choisir le paramètre D. Dans notre cas nous avons deux périodes saisonnières : une période trimestrielle et une période quadrimestrielle, ce qui signifie que nous sommes dans le cas de la multi-saisonnalité, ce qui rend les prévisions plus difficiles. Mais étant donné qu'une seule différenciation non-saisonnière a suffi à rendre stationnaire la série, nous allons nous contenter que d'une seule différenciation saisonnière. De plus, différencier deux fois supplémentaires fois ($D = 3$) ne supprime pas la saisonnalité existante et cause même de la sur-différenciation (cf. graphique IACF : décroissance lente) (figure 9), de même avec trois et quatre différenciations. Dans ce cas, le choix était de garder $D = 1$ pour ne pas écraser la série et de perdre de l'information. De ce fait, nous avons trouvé le paramètre s et D du modèle : $SARIMA_3(p, 1, q)(P, 1, Q)$.

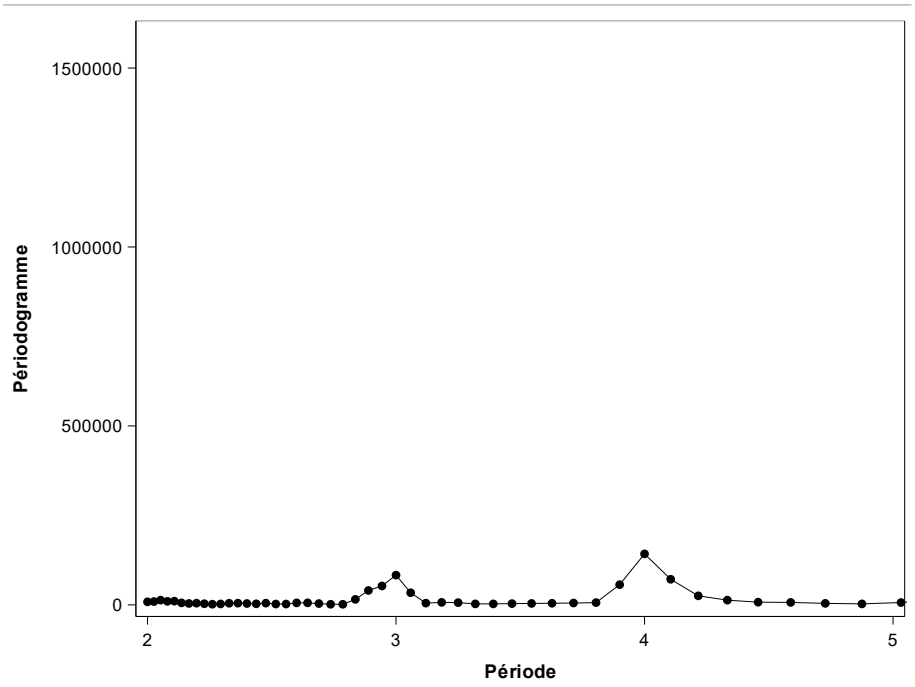

Figure 8 : Analyse de la densité spectrale du modèle.

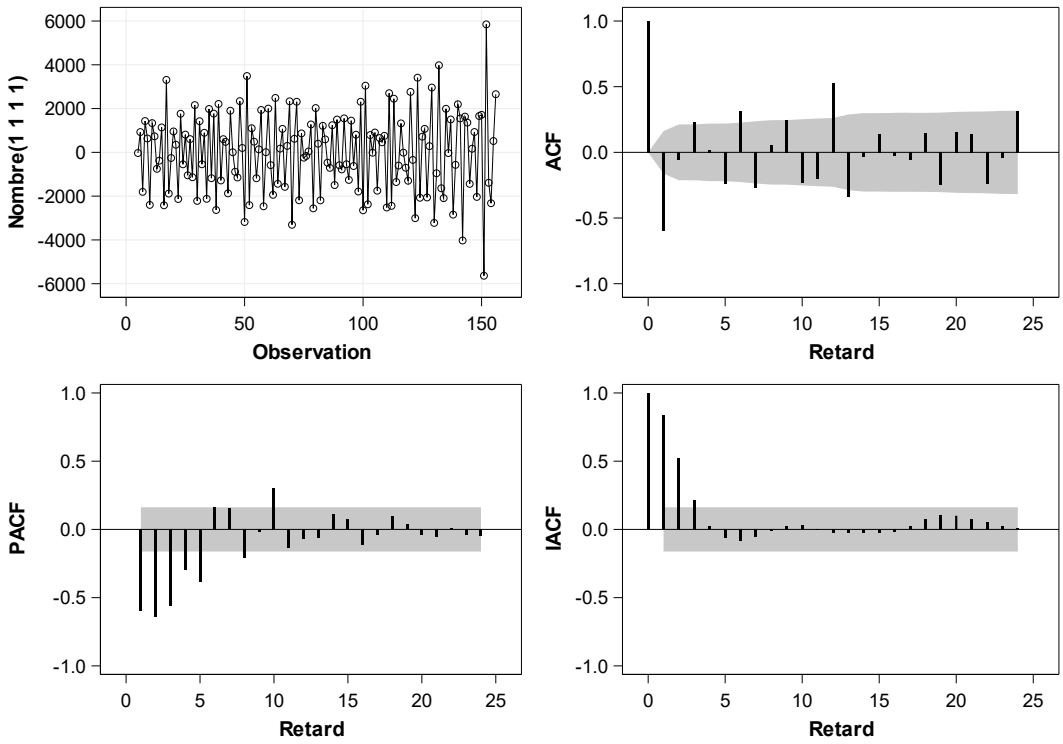

Figure 9 : Analyse des tendances et de la corrélation du nombre de création d'entreprises avec trois différenciations saisonnières.

3. Choix des paramètres

3.1. Recherche des meilleurs paramètres selon le BIC

Maintenant, il nous faut trouver les ordres autorégressifs (p et P) et les ordres de moyenne-mobile (q et Q). Pour cela, nous allons déployer une recherche par le BIC en utilisant trois méthodes : SCAN, ESACF et MINIC. Chacune de ces méthodes recherche les meilleures combinaisons de p et q (elles recherchent en fait le meilleur ARMA). Cela se fait en minimisant le BIC. Lorsque ces trois méthodes ont défini le meilleur modèle (au sens du BIC), nous allons comparer ces trois modèles et nous allons choisir l'ARMA qui minimise de nouveau le BIC. Finalement, nous avons $ARMA(13,5)$ proposé par la méthode ESACF ayant un BIC de 11,38 ; ensuite le modèle $ARMA(11,1)$ proposé par la méthode SCAN ayant un BIC de 11,29 puis la méthode MINIC propose un $ARMA(16,0)$ qui obtient un BIC de 11,26. C'est donc ce modèle que nous allons choisir, ce qui nous fait un modèle $SARIMA_3(16,1,0)(P,1,Q)$.

Modèle de série incorrect : AR(40)
Valeur de table minimale : BIC(16,0) = 11.2637

ARMA(p+d,q) Tests de sélection d'ordre provisoire

SCAN			ESACF		
p+d	**q**	**BIC**	**p+d**	**q**	**BIC**
5	5	11.44404	5	6	11.42763
11	1	11.28827	14	3	11.39709
4	12	11.50804	13	5	11.37535
2	13	11.76879	12	6	11.37922
19	0	11.31555	8	8	11.39887
			9	8	11.40703
			17	4	11.40212
			18	3	11.404
			4	12	11.50804
			0	13	12.19652
			2	13	11.76879
			3	13	11.76221
			20	4	11.40278

Tableau 4 : Recherche du modèle $ARMA(p,q)$ qui minimise le BIC en employant les méthodes SCAN, ESACF et MINIC. **Lecture** : La méthode SCAN trouve un $BIC = 11{,}32$ pour un $ARMA(19,0)$.

3.2. Test du modèle suggéré et adaptation

Selon la suggestion vue précédemment, nous allons faire plusieurs tests de modèles. L'idée est que si le modèle testé n'est pas caractérisé par une autocorrélation des résidus, alors il est conservé pour faire des prédictions, sinon nous allons tester un autre modèle.

Premièrement, nous allons tester le modèle $SARIMA_3(16,1,0)(1,1,1)$, cependant il est caractérisé par une forte autocorrélation des résidus (tableau 5). Dans cette optique et en gardant à l'esprit que le modèle associé à notre contexte, aime particulièrement les ordres autorégressifs, contrairement aux ordres de moyennes mobiles (dit très grossièrement), on retire l'ordre de moyenne mobile saisonnier et on ajoute un ordre autorégressif saisonnier. Cela nous donne le modèle $SARIMA_3(16,1,0)(2,1,0)$ qui semble convenir à première vue car il satisfait notre critère de sélection : les résidus ne sont pas autocorrélés. C'est donc ce modèle qu'on utilisera pour faire des prédictions.

Jusqu'au retard	Khi-2	DDL	Pr > khi-2	Autocorrélations					
6	.	0	.	-0.328	-0.129	-0.027	0.250	-0.143	0.079
12	.	0	.	-0.107	0.126	0.066	-0.153	-0.092	0.334
18	.	0	.	-0.127	-0.145	0.051	0.160	-0.022	-0.068
24	110.82	5	<.0001	-0.088	0.221	0.054	-0.224	0.023	0.289
30	116.41	11	<.0001	-0.083	-0.116	0.070	0.052	-0.000	-0.042

Tableau 5 : Vérification de l'autocorrélation des résidus du modèle $SARIMA_3(16,1,0)(1,1,1)$.
Lecture : La Pr > khi-2 étant bien inférieure au seuil des 5%, alors nous rejetons l'hypothèse selon laquelle les résidus ne sont pas autocorrélés.

Jusqu'au retard	Khi-2	DDL	Pr > khi-2	Autocorrélations					
6	.	0	.	-0.004	0.003	-0.007	-0.007	-0.004	0.010
12	.	0	.	-0.034	-0.079	0.013	0.050	-0.019	-0.021
18	.	0	.	0.008	0.080	0.030	0.012	0.046	0.022
24	9.77	6	0.1347	-0.014	0.082	0.019	-0.057	0.099	0.116
30	11.88	12	0.4552	0.012	0.010	-0.003	-0.082	0.027	-0.057

Tableau 6 : Vérification de l'autocorrélation des résidus du modèle $SARIMA_3(16,1,0)(2,1,0)$.
Lecture : La Pr > khi-2 étant supérieure au seuil des 5%, alors nous conservons l'hypothèse selon laquelle les résidus ne sont pas autocorrélés.

4. Vérification et prévisions

Avant de faire des prévisions, nous devons au préalable voir si le modèle que nous avons déterminé respecte les hypothèses les plus importantes :

- Stationnarité du modèle ;
- Non-saisonnalité de la série ;
- Absence d'autocorrélation des résidus ;
- Normalité des résidus.

Comme nous avons vu dans les parties précédentes, nous avons déjà stationnarisé la série (donc pas de tendance), mais nous n'avons pas pu supprimer la saisonnalité, sans doute à cause de la multi-saisonnalité de cette série. Concernant les résidus, nous avons réussi à trouver un modèle qui supprime leur autocorrélation. Il ne manque plus qu'à vérifier leur normalité. Pour cela, il existe différents tests de normalité plus ou moins contraignant et trois d'entre eux certifient la présence de la normalité des résidus au seuil des 5% (tableau 7). Cependant, le test le plus contraignant (Shapiro-Wilk) rejette la normalité des résidus à ce seuil mais l'accepte si nous sommes un peu plus stricts (en réduisant la marge d'erreur à 1%). Dans cette optique, l'hypothèse du modèle est vérifiée. Cependant, le risque de ce genre de manœuvre c'est d'augmenter le risque d'accepter l'hypothèse nulle sachant qu'elle est fausse. Mais au vu des trois autres tests, cela semble peu probable. Au final, nous avons trois hypothèses sur quatre qui sont vérifiées, ce qui est un résultat assez satisfaisant.

Test	Statistique		p-value	
Shapiro-Wilk	**W**	0.977917	**Pr < W**	0.0141
Kolmogorov-Smirnov	**D**	0.055313	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.058529	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.436596	**Pr > A-Sq**	>0.2500

Tableau 7 : Tests de normalité des résidus du modèle $SARIMA_3(16,1,0)(2,1,0)$. **Lecture** : Les différentes p-value accepte l'existence de la normalité des résidus du modèle étudiés au seuil des 10%.

Maintenant, il nous faut encore voir si le modèle ne produit pas de sur-ajustement, pour cela, on applique les prévisions du modèle en les superposant avec la série brute (figure 10). Il semblerait qu'il n'y ait pas de sur-ajustement du modèle aux données brutes et il semblerait également qu'il soit bon car s'aligne relativement bien à ces données. Il capte même les observations qui ne sont pas influencées par la saisonnalité (cf. mi-2020 et fin-2021).

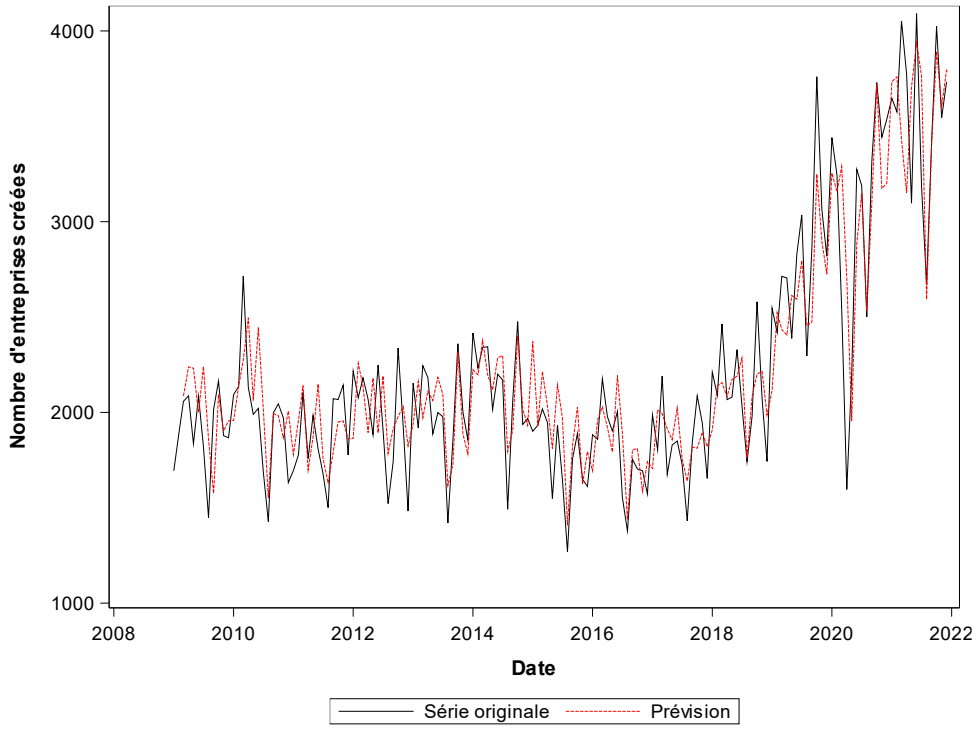

Figure 10 : Prévision du modèle $SARIMA_3(16,1,0)(2,1,0)$ sur l'ensemble des données brutes.

Notre modèle est choisi, vérifié et (presque) validé, donc vient le moment d'effectuer des prédictions. Dans notre cas, les prédictions ne sont pas farfelues, donc cohérentes. Nous avons donc un aperçu global des résultats (figure 11) un aperçu un peu plus localisé (figure 12) et les résultats précis pour chaque mois prédits (tableau 8).

Ce que nous pouvons dire sur ces prédictions, c'est que premièrement, comme nous l'avons mentionné dans l'analyse sur la tendance, c'est que les prédictions ont une tendance générale à la hausse. Mais avec l'analyse saisonnière, nous pouvons également confirmer que les prédictions se caractérisent également par des hausses et des baisses légères. Une dernière remarque (qui est plus un avertissement) : nous voyons bien que l'erreur type grandit à chaque période future prédite. En réalité, il faut s'en méfier comme les prévisions météorologiques d'une semaine à l'autre : plus c'est loin plus on a de chances de se tromper. Ainsi, prédire au-delà d'une année est peut-être un peu prétentieux. Ce que nous aurions pu faire c'est de comparer les données prédites avec les données de la nouvelle série qu'a publiée l'INSEE, mais comme expliqué précédemment, les données ne sont pas fiables à 100%.

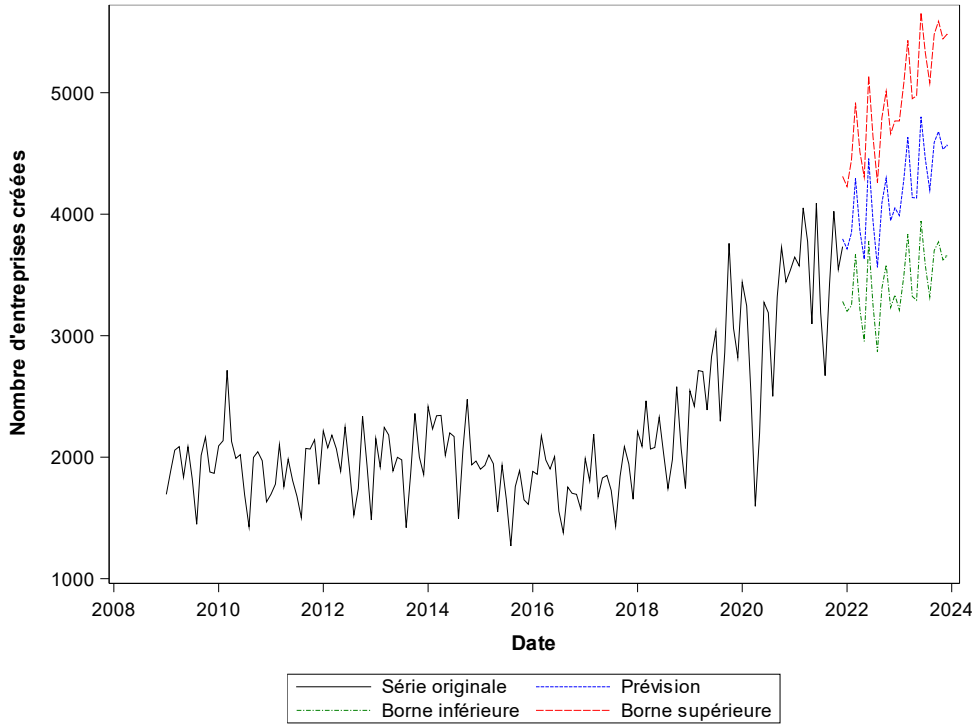

Figure 11 : Série originale et prévision du modèle $SARIMA_3(16,1,0)(2,1,0)$ sur deux ans après la dernière observation de la série originale.

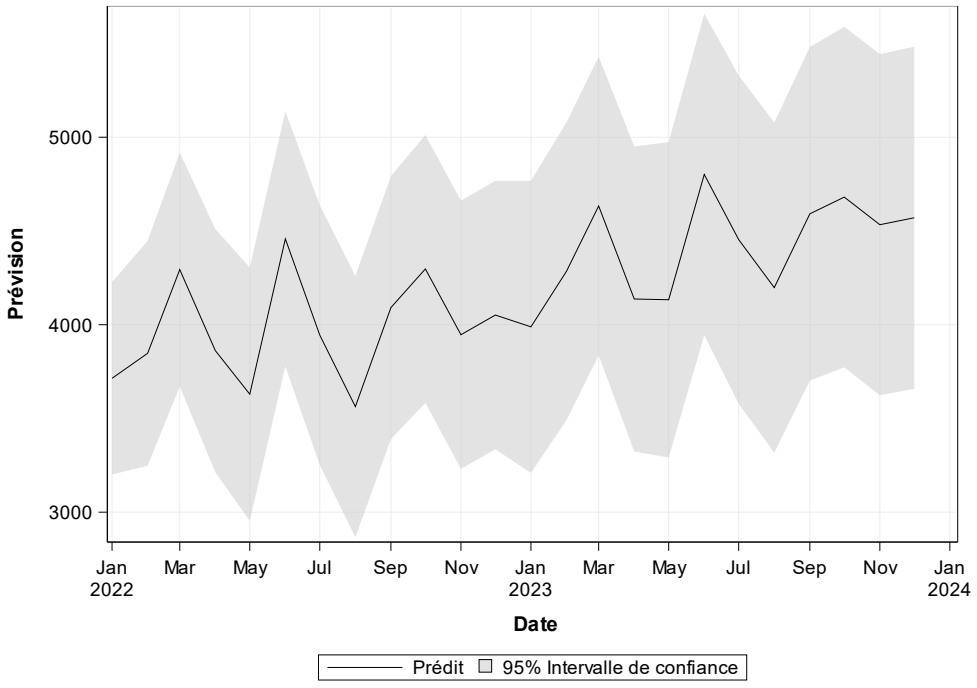

Figure 12 : Prévision du modèle $SARIMA_3(16,1,0)(2,1,0)$ sur deux ans après la dernière observation de la série originale.

Obs.		Prévision	Erreur type	Intervalle de confiance à 95%	
2022	Jan	3714.0611	261.7281	3201.0835	4227.0386
	Feb	3847.4260	306.0998	3247.4814	4447.3705
	Mar	4294.0538	318.1246	3670.5411	4917.5665
	Apr	3861.5785	330.9337	3212.9603	4510.1967
	May	3629.5193	344.7228	2953.8751	4305.1635
	Jun	4457.8274	346.9499	3777.8180	5137.8367
	Jul	3944.5138	353.4812	3251.7034	4637.3241
	Aug	3563.0152	355.7693	2865.7202	4260.3103
	Sep	4091.4331	357.9334	3389.8965	4792.9697
	Oct	4297.5378	364.5487	3583.0356	5012.0400
	Nov	3946.5195	364.9949	3231.1427	4661.8964
	Dec	4051.4762	365.3647	3335.3744	4767.5779
2023	Jan	3988.5832	397.5412	3209.4168	4767.7496
	Feb	4285.3218	405.2129	3491.1191	5079.5246
	Mar	4633.3773	406.7148	3836.2309	5430.5237
	Apr	4137.0529	415.0799	3323.5112	4950.5945
	May	4133.2249	429.3118	3291.7893	4974.6605
	Jun	4801.4654	437.6208	3943.7444	5659.1865
	Jul	4454.1092	446.7067	3578.5802	5329.6383
	Aug	4197.8641	449.3265	3317.2005	5078.5278
	Sep	4591.4549	454.0525	3701.5283	5481.3815
	Oct	4680.6884	463.3329	3772.5726	5588.8042
	Nov	4533.3823	464.1377	3623.6893	5443.0754
	Dec	4570.5291	465.5931	3657.9834	5483.0748

Tableau 8 : Résultats des prédictions. **Lecture** : Le modèle prédit qu'en juin 2023, qu'environ 4801 entreprises seront créées au sein de l'industrie manufacturière.

5. Comparaison avec l'intelligence artificielle : Prophet

Cette dernière partie est de comparer les résultats que nous avons obtenu avec le modèle que nous avons développé avec un modèle produit par l'intelligence artificielle. Un exemple de ce type est Prophet qui peut être déployé sur R et Python (entre autres). La première chose que ce modèle peut donner, c'est la tendance générale de la série et les variations annuelles (figure 13). Sur la tendance générale, nous pouvons dire que le modèle de Prophet est assez (trop) général et ne rend pas compte des différentes variations, même si finalement la série est assez stationnaire de 2009 à 2017. C'est sûrement pour cela que les variations annuelles apparaissent avec la tendance, pour combler ce manque d'information global.

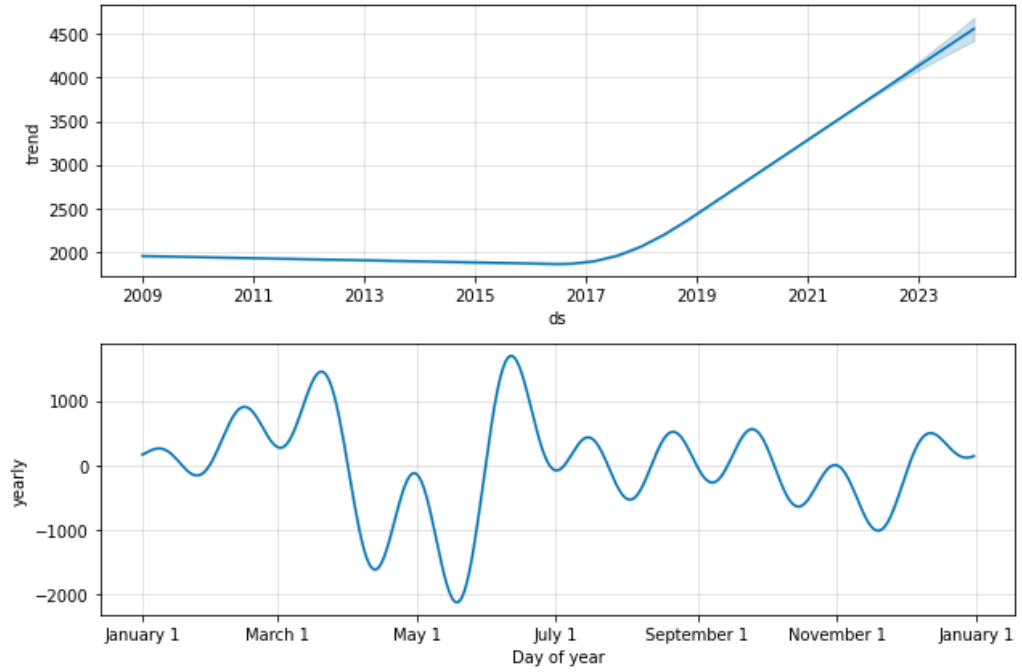

Figure 13 : Tendance de la série originale et les variations annuelles du modèle produit par Prophet.

Concernant les prédictions, le modèle de Prophet est assez grossier et ne fait que répéter plus ou moins un même schéma au fil du temps (figure 14). Contrairement à lui, notre modèle capte bien à la fois les variations saisonnières (ce qui le modèle de Prophet fait bien également) mais aussi les événements exogènes (figure 15), ce que fait mal l'autre modèle. En réalité, le modèle produit par Prophet semble être assez « fainéant » en ne captant qu'un minimum d'informations.

Il nous faut donc déterminer une bonne fois pour toutes quel est le meilleur modèle. Deux métriques peuvent être utilisées : l'erreur quadratique moyenne (RMSE) et l'erreur absolue

moyenne (MAE). Il semblerait que le modèle que nous avons développé affiche de bien meilleures performances que le modèle de Prophet (tableau 9).

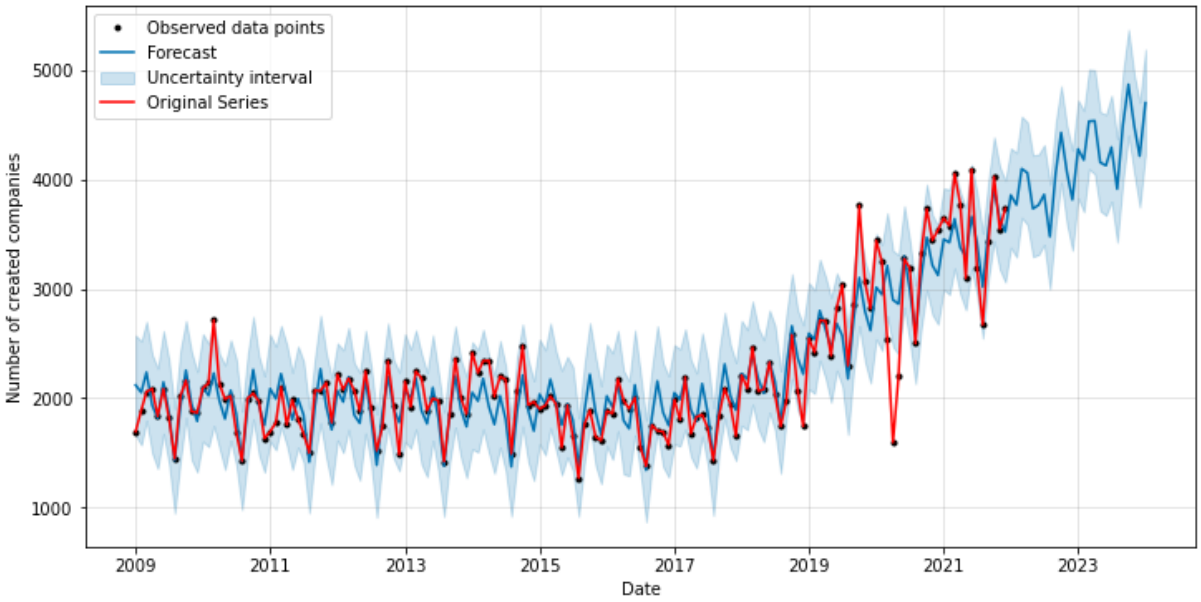

Figure 14 : Prédictions du modèle produit par Prophet.

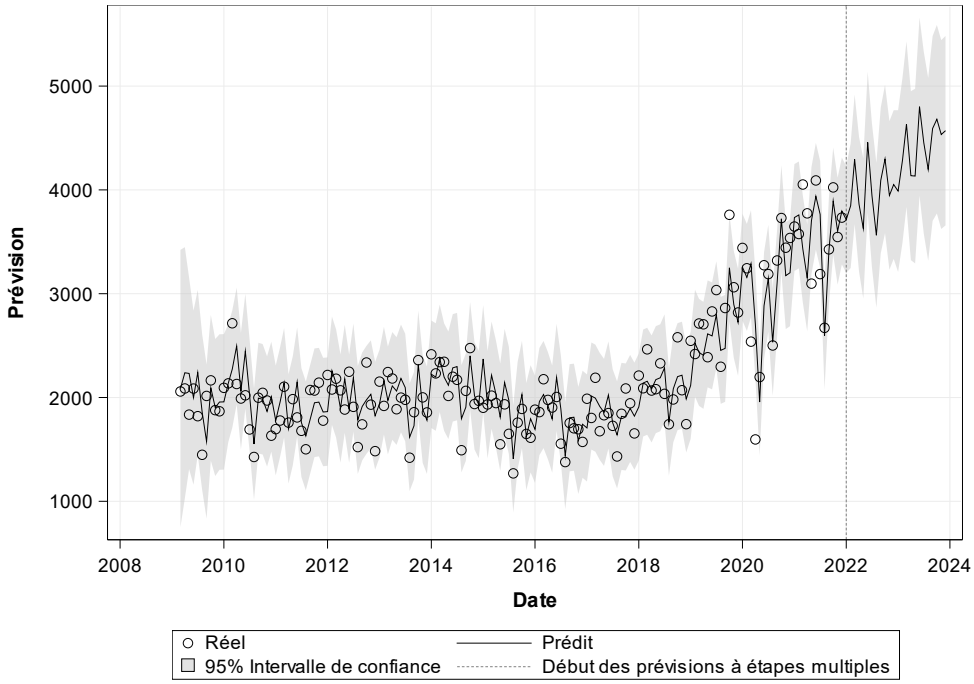

Figure 15 : Prédictions du modèle $SARIMA_3(16,1,0)(2,1,0)$.

Modèles	RMSE	MAE
$SARIMA_3(16,1,0)(2,1,0)$	215.2	151.4
PROPHET	239.1	176.7

Tableau 9 : Comparaison finale entre les deux modèles. **Lecture** : Les prédictions du modèle produit par Prophet s'écartent en moyenne de 239,1 unités par rapport aux valeurs réelles.

CONCLUSION

Le nombre de création d'entreprises de l'industrie française a fortement augmenté depuis 2019, malgré les événements paralysant l'économie. Notre objectif était de faire des prédictions sur deux ans pour voir comment se caractériserait l'évolution de ce nombre après 2021. Finalement, nous avons opté pour le modèle $SARIMA_3(16,1,0)(2,1,0)$ qui valide bien toutes les hypothèses statistiques, à l'exception de la saisonnalité. Cela s'explique notamment par le fait que la série étudiée est marquée par une multi-saisonnalité (une saisonnalité trimestrielle et quadrimestrielle). Malgré cet inconvénient, le modèle effectue de bonnes prédictions et ces prédictions sont caractérisées par une tendance globale à la hausse, mais avec quand même des variation saisonnières qui s'illustrent par des hausses et des baisses légères du nombre de création d'entreprises.

Après tout cela, nous avons comparé les résultats du modèle que nous avons développé avec un modèle produit par l'intelligence artificielle (Prophet) qui semble assez grossier dans l'ensemble en ne captant pas toutes les informations. Pour évaluer les performances des deux modèles, nous avons utilisé l'erreur quadratique moyenne et l'erreur absolue moyenne et ces deux métriques attestent que le modèle $SARIMA_3(16,1,0)(2,1,0)$ est le meilleur.

RESSOURCES

« Système d'information sur la démographie d'entreprises | Insee », consulté le 11 février 2024, https://www.insee.fr/fr/metadonnees/source/serie/s2120.

« Production et consommation intermédiaire en 2022 − Les comptes de la Nation en 2022 | Insee », consulté le 10 février 2024, https://www.insee.fr/fr/statistiques/6793598?sommaire=6793644#consulter.

« Présentation statistique − Système d'information sur la démographie d'entreprises | Insee », consulté le 11 février 2024, https://www.insee.fr/fr/metadonnees/source/serie/s2120/presentation.

« Présentation statistique − Créations d'entreprises | Insee », consulté le 11 février 2024, https://www.insee.fr/fr/metadonnees/source/indicateur/p1631/presentation.

« Nombre de créations d'entreprises - Industrie manufacturière - Ensemble - France - Données mensuelles brutes | Insee », consulté le 10 février 2024, https://www.insee.fr/fr/statistiques/serie/010755561#Tableau.

« Nombre de créations d'entreprises - Industrie manufacturière - Ensemble - France - Données mensuelles brutes - Série arrêtée | Insee », consulté le 10 février 2024, https://www.insee.fr/fr/statistiques/serie/001564286#Tableau.

« Indicateurs macroéconomiques de l'industrie manufacturière | Insee », consulté le 10 février 2024, https://www.insee.fr/fr/statistiques/2123180.

« Documentation sur la méthodologie − Créations d'entreprises | Insee », consulté le 11 février 2024, https://www.insee.fr/fr/metadonnees/source/indicateur/p1631/documentation-methodologique.

« Description − Créations d'entreprises | Insee », consulté le 11 février 2024, https://www.insee.fr/fr/metadonnees/source/indicateur/p1631/description.

TABLE DES FIGURES

Figure 1 : Répartition du nombre de création d'entreprises au sein de l'industrie manufacturière française depuis 2009. 6

Figure 2 : Evolution du nombre de création d'entreprises dans l'industrie manufacturière française de janvier 2000 à décembre 2021. 8

Figure 3 : Evolution du nombre de création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. 8

Figure 4 : Série CVS de la création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. 9

Figure 5 : Tendance de la création d'entreprises dans l'industrie manufacturière française de janvier 2009 à décembre 2021. 10

Figure 6 : Analyse des tendances et de la corrélation du nombre de création d'entreprises (série brute). 12

Figure 7 : Analyse des tendances et de la corrélation du nombre de création d'entreprises avec une seule différenciation. 13

Figure 8 : Analyse de la densité spectrale du modèle. 14

Figure 9 : Analyse des tendances et de la corrélation du nombre de création d'entreprises avec trois différenciations saisonnières. 15

Figure 10 : Prévision du modèle $SARIMA3(16,1,0)(2,1,0)$ sur l'ensemble des données brutes. 19

Figure 11 : Série originale et prévision du modèle $SARIMA3(16,1,0)(2,1,0)$ sur deux ans après la dernière observation de la série originale. 20

Figure 12 : Prévision du modèle $SARIMA3(16,1,0)(2,1,0)$ sur deux ans après la dernière observation de la série originale. 20

Figure 13 : Tendance de la série originale et les variations annuelles du modèle produit par Prophet. 22

Figure 14 : Prédictions du modèle produit par Prophet. 23

Figure 15 : Prédictions du modèle $SARIMA3(16,1,0)(2,1,0)$. 23

TABLE DES TABLEAUX

Tableau 1 : Les 19 mois où l'on a observé le plus grand nombre d'entreprises créées dans l'industrie manufacturière française depuis 2009. 7

Tableau 2 : Tests de racine unitaire de Dickey-Fuller augmentés du nombre de création d'entreprises (série brute). 12

Tableau 3 : Tests de la racine unitaire de Phillips-Perron. 13

Tableau 4 : Recherche du modèle $ARMA(p, q)$ qui minimise le BIC en employant les méthodes SCAN, ESACF et MINIC. 16

Tableau 5 : Vérification de l'autocorrélation des résidus du modèle $SARIMA3(16,1,0)(1,1,1)$. 17

Tableau 6 : Vérification de l'autocorrélation des résidus du modèle $SARIMA3(16,1,0)(2,1,0)$. 17

Tableau 7 : Tests de normalité des résidus du modèle $SARIMA3(16,1,0)(2,1,0)$..... 18

Tableau 8 : Résultats des prédictions. 21

Tableau 9 : Comparaison finale entre les deux modèles. 24