Master 1 Économétrie et Statistiques, parcours Économétrie Appliquée

Étude du taux de chômage français de 2004 à nos jours

L'outil 'Google Trends' peut-il aider à prévoir le chômage ?

Auteur Diane THIERRY

Enseignant Mme D.Girard

Année universitaire 2019-2020

SOMMAIRE :

I- Introduction	**3**
II- Analyse économique du sujet	**7**
III- Analyse des 4 variables	**10**
IV- Présentation de la méthodologie ARIMA	**21**
V- Choix du modèle ARIMA pour Yt	**27**
VI- Test de cointégration selon Engle-Granger	**38**
VII- Conclusion	**47**
VIII- Discussion	**48**
IX- Bibliographie	**49**
X- Annexes	**52**
XI- Table des matières	**60**

Résumé :

Dans cette analyse nous cherchons à voir si l'utilisation des données de recherche Google est utile pour prévoir le taux de chômage en France par rapport à d'autres variables prédictives plus traditionnelles. Nous étudions ainsi le taux de chômage de 2004 à nos jours en prenant des variables classiques et une variable issue de l'outil Google Trends. Nous utilisons la méthodologie Box-Jenkins pour définir un bon modèle de notre Y_t qu'est le taux de chômage ; nous trouvons ainsi un ARIMA[1,1,0] qui est à la fois pertinent et valide, et nous permet de réaliser des prévisions pour un trimestre. La cointégration des 4 variables explicatives avec le taux de chômage est concluante pour deux d'entre elles ; la popularité du mot 'emploi' et la production industrielle. Ainsi, la variable issue de l'outil Google Trends converge vers une même tendance que le taux de chômage à long terme et la cointégration de ces 2 dernières nous permet de conclure positivement quant à l'apport de Google Trends sur la prévision de l'emploi et du chômage.

I- Introduction

L'ère de la technologie s'est installée depuis un demi siècle déjà sur la Terre. Nos modes de vie, nos habitudes, nos goûts ont été révolutionnés par cette transformation numérique qui vient changer notre société en profondeur. Le monde connecté dans lequel nous vivons aujourd'hui est caractérisé par le développement de l'intelligence artificielle, la mise en commun des ressources grâce aux réseaux, la connexion constante aux technologies. Nous vivons à présent dans un monde où tout est intelligent : les voitures peuvent avancer et se diriger seules, certaines lunettes permettent de faire des recherches en temps réel sur internet par un simple clin d'oeil, des robots s'apparentent tellement aux humains qu'il devient difficile de les distinguer. Tout cela fascine certains, effraie d'autres et fait naître en eux des sentiments d'insécurité face aux avancées technologiques.

Ainsi, l'époque industrielle a fait place à une nouvelle génération de données et d'applications où chaque jour environ 2,5 trillions d'octets de données sont créés. Ce phénomène de Big Data, c'est à dire le stockage de milliers d'informations sur une base numérique, est apparu dès 1997 et ouvre aujourd'hui un champ presque infini de possibilité d'analyses et de recherches.[1]

De cette manière, depuis quelques années, certains économistes se concentrent sur le traitement de telles données, et plus particulièrement les **tendances de recherche** des usagers, pour améliorer et préciser leurs prévisions économiques. Ces données permettent de visualiser le comportement des consommateurs en temps réel : les tendances d'un terme ou d'un autre reflètent les besoins, les désirs des utilisateurs d'internet. L'étude de ces informations dans le but de faire de la prédiction est réalisée au travers d'enquêtes conjoncturistes, dont l'objectif selon l'INSEE est de "suivre la situation économique du moment et de prévoir les évolutions à court terme".[2] Aussi, ces enquêtes permettent de diagnostiquer les évolutions récentes de l'économie, sa situation actuelle, prévoir ses futurs variations, elles sont utiles pour orienter judicieusement les politiques (qu'elles soient menées par le gouvernement ou par la banque centrale) mais aussi pour tenir les ménages informés de l'activité économique.

[1] https://www.lebigdata.fr/definition-big-data
[2] https://www.insee.fr/fr/metadonnees/definition/c1422

En ce sens, de nombreux outils de mesure des tendances de recherche ont fait leur apparition tels que “Google Trends”, “Buzzsumo”, “Reddit”, “Social Share” etc. Dans ce dossier nous nous pencherons plus particulièrement sur ce premier outil mis en place par le géant du Web ‘Google’.

Google Trends a pour but d’identifier la popularité de certains termes dans les recherches internet des individus à une période donnée. Les données de recherche mises à disposition ne sont pas brutes, elles sont ajustées pour faciliter la comparaison entres les termes - l’outil ne donne donc pas le **volume des recherches** (quantité) mais bien la **popularité relative** d’un terme (intérêt public) par rapport au nombre total de recherches effectuées dans la même zone géographique et à la même période. Il est néanmoins possible de visualiser le volume de recherches d’un mot grâce à un autre outil de la plateforme Google : ‘Ngram Viewer’. Celui-ci indique la fréquence de recherche d’un ou plusieurs mots clefs par les internautes dans une période bien définie, il vient donc compléter les informations disponibles sous Google Trends.[3]

Aussi appelé ‘Google Tendances de recherches’ ou simplement ‘Google Tendances’, l’outil a fait son apparition en 2006, la même année que le rachat de la plateforme mondialement connue “Youtube” par Google, et représente aujourd’hui une mine d’informations très précieuse pour les professionnels du marketing et de la publicité. En effet Google Trends offre la possibilité de connaître (presque) en temps réel les tendances de recherches qui traduisent les goûts, les envies, les préférences des internautes - ce qui permet aux professionnels du webmarketing de relever, suivre et analyser les tendances du web pour adapter directement leur offre aux demandes du marché.

Les nombreuses fonctionnalités de Google Trends permettent de voir les tendances actuelles ou passées de recherche par mots-clefs, de comparer les tendances de plusieurs mots, d’affiner les recherches lorsqu’il s’agit d’un mot avec plusieurs sens, ou encore d’estimer l'intérêt suscité par un terme donné dans le futur grâce à l'outil de prévision de trafic. La plateforme permet aussi de visualiser la popularité de certains termes par des graphiques d'évolution dans le temps, ou des cartes animées de fréquences de recherche par région, par ville et par pays. [4]

[3] https://www.ya-graphic.com/google-trends-popularite-volume-de-recherche/
[4] https://www.abondance.com/20090819-10009-google-insights-for-search-disponible-en-francais.html

Les valeurs de ces fréquences sont calibrées entre 0 et 100 (où 100 correspond au taux d'utilisation le plus élevé), de manière à faciliter la lecture et l'interprétation. Pour cela, Google Trends calcule le nombre de recherches d'un terme en particulier par rapport au nombre total de recherches effectuées sur le moteur 'Google' sur une période définie, ce qui permet de se faire une idée générale des tendances sur le Web. Les données sont mises à jour **quotidiennement** dans le monde entier et dans toutes les langues gérées par Google.[5]

Lors de recherches sur Google Trends il est possible de saisir un "terme de recherche" ou un "sujet", la différence est subtile mais intéressante à connaître puisque les résultats seront différents. Un 'terme de recherche' est un mot-clef dont la recherche englobe tous les autres mots-clefs contenant ce terme dans la langue de recherche - ainsi, en recherchant le mot "gymnastique" les résultats affichés prendront en compte "cours de gymnastique", "championnats de France gymnastique 2019" etc. Le 'sujet', quant à lui, va inclure tous les autres termes qui correspondent au même concept, dans toutes les langues. Donc en cherchant le même mot, cette fois les résultats incluront aussi "agrès de sport" ou "éducation physique" par exemple.[6]

De même, il est possible de filtrer les recherches en ne choisissant que les tendances des mots entrés dans certains moteurs de recherche Google tels que "Google Shopping", "Google Actualités", "YouTube" etc. Cette fonctionnalité permet de mieux cerner les plateformes de recherches utilisées en fonction du sujet, de l'époque et de la région.

La gratuité et l'accessibilité ont fait de Google Tendances un outil incontournable dans l'analyse des séries temporelles notamment, grâce à la disponibilité de l'information en *open data* qui facilite les recherches des utilisateurs. En 2019 en France il y avait 53,1 millions d'internautes, soit près de 85% de la population de 2 ans et plus.[7] Sachant que 95% des recherches internet sont effectuées sur le moteur 'Google' (tous appareils confondus),[8] il semble que les données fournies par Google Trends soient assez représentatives de la réalité.

En évolution constante nous pouvons nous demander si cet outil puissant est réellement un bon indicateur des tendances du Web et si son attractivité est à la hauteur de

[5] https://www.anthedesign.fr/referencement/google-trends/
[6] https://www.latranchee.com/comment-passer-de-lidee-a-la-strategie-grace-a-google-trends/
[7] https://www.journaldunet.com/ebusiness/le-net/1071394-nombre-d-internautes-en-france/
[8] https://www.inwin.fr/blog/un-outil-puissant-et-gratuit-mais-meconnu-google-trends/

son efficacité. Pour cela regardons dans un premier temps une étude qui s'intéresse aux apports de Google Trends pour prévoir la consommation des ménages, dirigée par Jean-Luc Tavernier (directeur général de l'INSEE) et publiée en mars 2015, qui s'intitule "Note de conjoncture".

Pour évaluer cela, plusieurs modèles ont été élaborés en ajoutant les tendances de recherche sur Google pour voir si la prévision des dépenses mensuelles des ménages est de meilleure qualité ou non. Il en ressort que Google trends n'améliore pas significativement les prévisions à cause de la forte hétérogénéité des évolutions par produit, et la pérennité incertaine des séries obtenues sur cette plateforme. En effet, les séries fournies sont construites à partir du décompte des requêtes réalisées sur l'historique d'un échantillon d'utilisateurs - dans certains cas les échantillons sont différents et cela conduit à des séries non-homogènes voir incohérentes et instables.[9] En revanche, les résultats de la modélisation sont encourageants pour prévoir les dépenses en habillement et en équipement des ménages puisque les prévisions sont légèrement améliorées - étant le cas seulement pour ces 2 biens les auteurs de l'étude se sont interrogés sur l'éventualité d'un facteur "chance" qui prouve que les faiblesses de l'outil Google sont encore trop importantes pour prévoir efficacement la consommation des ménages.

Il semble donc difficile d'améliorer les prévisions de dépenses des ménages français à l'aide de l'outil Google Trends, mais nous ne pouvons savoir si cela est lié au sujet étudié ou à l'outil lui même. Nous allons donc mener notre propre étude sur son apport quant-à la prévision du chômage et de l'emploi.

Ainsi, dans ce dossier nous analyserons 5 séries temporelles équidistantes que nous modéliserons par un processus ARIMA, puis nous effectuerons des prévisions du taux de chômage à horizon h=1 grâce au modèle spécifié précédemment et finalement nous cointégrerons les séries des variables explicatives avec le taux de chômage.

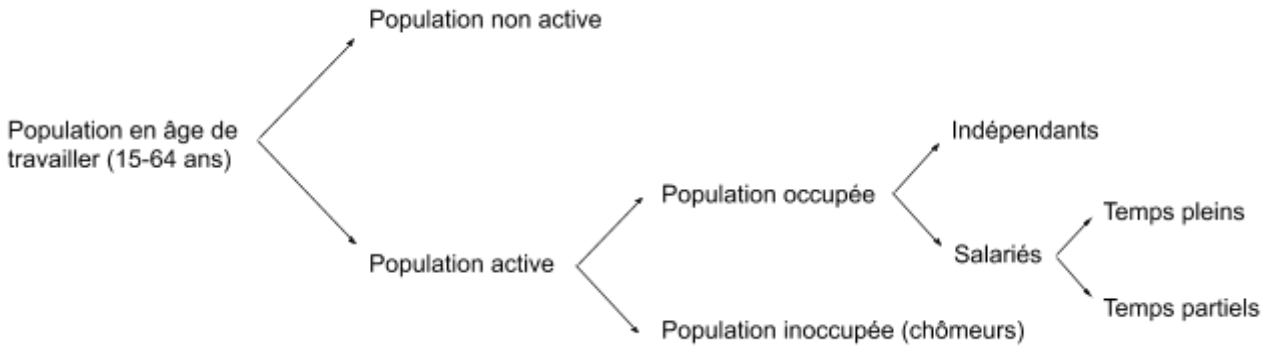
[9] Tavernier J-L., "Note de conjoncture", *INSEE*, 03/2015,pp.43-56. (consulté le 18/01/2020)

II- Analyse économique du sujet

L'année 2019 a été bénéfique d'un point de vue de l'emploi. En effet, avec 8.5% de chômage, la France a enregistré son plus faible taux depuis la crise de 2008 grâce à la création de 260.000 postes contre 188.000 l'année précédente.[10]

Ce dernier correspond selon l'INSEE au "pourcentage de chômeurs dans la population active"[11] comme visible sur le schéma suivant, aussi un chômeur est une personne n'ayant pas de travail mais en cherchant activement un.

SCHÉMA N°1 : Découpage de la population en terme d'emploi



source : élaboration personnelle à partir de l'outil "dessin" sous google drive

Dans le monde entier le chômage n'est apparu qu'en 1973 avec le premier choc pétrolier, avant cette date chaque pays était en situation de plein-emploi voire de sur-emploi (plus de demande que d'offre de travail) ce qui conduisait certains pays à faire venir des travailleurs étrangers pour pallier au manque de main d'oeuvre, faisant monter légèrement le taux de chômage. Les années 70 et 80 ont été marquées par une hausse constante du taux de chômage dans les pays d'Europe, comme aux États-Unis ou au Japon, avec une hausse plus importante pour les pays d'Europe. Puis le chômage revient à des taux plus raisonnables à partir des années 90 mais reste plus élevé en Zone Euro qu'aux États-Unis ou au Royaume-Uni par exemple.[12] Enfin, c'est avec la crise des Subprimes que les taux explosent de nouveau et qu'une scission se crée entre les pays d'Europe du Sud tels que l'Espagne, la Grèce et le Portugal, qui ont un taux de chômage très important par rapport aux pays d'Europe du Nord.

[10].https://www.ouest-france.fr/economie/emploi/chomage/le-taux-de-chomage-au-plus-bas-depuis-dix-ans-6671745 (consulté le 24/01/2020)
[11] https://www.insee.fr/fr/metadonnees/definition/c1687 (consulté le 24/01/2020)
[12] https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Dossier (consulté le 23/01/2020)

Aujourd'hui, il existe une certaine dualité du marché de l'emploi où l'on retrouve d'un côté les *insiders* que sont les travailleurs qualifiés, rémunérés correctement avec une sécurité de l'emploi et des temps pleins, et d'un autre côté les *outsiders* qui n'ont pas ou peu de qualifications et sont à cause de cela mal rémunérés, ont des conditions de travail plus difficiles et des coûts de licenciement extrêmement bas, ce qui les met dans une situation précaire.

De manière générale le taux de chômage est une variable qui connaît beaucoup de fluctuations : les périodes de récession se traduisent souvent par une hausse des licenciements donc du nombre de chômeurs, alors que les périodes d'expansion sont synonymes de création d'emplois et baisse du taux de chômage. En revanche, généralement celui-ci a tendance à baisser quand la population active diminue, ou quand il y a beaucoup d'emplois à temps partiels car cela implique un temps de travail moins important donc des facilités à embaucher et une réduction du niveau de chômage. [13]

Aussi, nous pouvons prendre l'exemple de la France et de l'Allemagne en 2018 pour comparer les niveaux de chômage en prenant en compte la structure de leur marché du travail et son impact sur l'emploi en général, comme expliqué précédemment (démographie, heures travaillées, qualification….).

TABLEAU N°1 : Comparaison de l'emploi en 2018 : France vs Allemagne

	France	**Allemagne**	**Ratios de comparaison**
Taux de chômage (% de la population active)	9,1	3,4	$\frac{9,1}{3,4} = 2,67$
Nombre d'heures travaillées (par an)	1.520	1.363	$\frac{1520}{1363} = 1,12$
Taux d'emplois à temps partiels (% de l'emploi)	14	22	$\frac{22}{14} = 1,57$
Taux de fécondité (nb d'enfants/femme)	1,87	1,46	$\frac{1,87}{1,46} = 1,28$

[13] https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Dossier (consulté le 23/01/2020)

Le tableau n°1 ci-dessus, dont les données sont extraites du site internet de l'OCDE[14], nous montre que le taux de chômage français est 2,7 fois plus élevé que le chômage allemand avec un taux à 9,1% en 2018. Seulement, il intéressant de noter qu'en moyenne les français travaillent davantage que les allemands (157 heures de différence, soit 1,12 fois plus) et que le pourcentage d'emplois à temps partiel est largement supérieur en Allemagne : 22% contre 14% en France. Cela s'explique par le fait qu'il y ait une sorte de 'pression sociale' sur les femmes pour qu'elles restent dans leur foyer s'occuper des enfants, cela justifie le taux 1,57 fois supérieur au taux français. Cependant, les temps partiels conduisent parfois à la précarité et à la hausse des inégalités. Enfin, l'accroissement démographique est plus important en France puisque le nombre moyen d'enfants par femme est de 1,87 contre 1,46 en Allemagne ; il y a ainsi un phénomène de vieillissement de la population allemande qui conduit à une baisse de la population active réduisant le taux de chômage ($= \frac{chômeurs}{population\ active}$). Par conséquent, l'Allemagne a vu son niveau de chômage se réduire grâce aux nombreux emplois à temps partiels (impliquant un nombre d'heures de travail plus faible), et à une baisse de la population active.

Le taux de chômage est donc une variable qui doit s'étudier dans un contexte, c'est à dire en prenant en compte d'autres facteurs qui caractérisent le marché du travail et permettent de mieux comprendre le niveau de chômage. Aussi, nous analyserons les fluctuations de 4 variables en France de 2004 à nos jours afin de prédire au mieux le taux de chômage, notre Y_t.

[14] https://data.oecd.org/fr/ (consulté le 24/01/2020)

III- Analyse des 4 variables

Le but de cette étude est de voir si l'utilisation de 'Google Trends' peut améliorer les prévisions du chômage et de l'emploi, c'est pourquoi nous prendrons en compte une variable explicative issue de cet outil, mais aussi des variables dites plus "classiques" qui, selon la littérature, sont de bons indicateurs du taux de chômage.

Dans cette partie nous justifierons l'utilisation des différentes variables explicatives, en montrant leur impact sur le taux de chômage, puis nous visualiserons graphiquement la relation qu'il existe entre chaque variable avec Y_t. Ainsi nous verrons s'il s'agit d'une corrélation positive qui implique une hausse du taux de chômage lorsque la variable étudiée augmente, ou d'une corrélation négative qui suppose un évolution opposée des 2 variables.

1- Présentation des données

TABLEAU N°2 : Base de données

	Taux de chômage	Popularité du mot 'emploi'	Taux d'intérêt	Production industrielle	Population active
T1 2004	8,52	41,00	4,11	110,77	27015,61
T2 2004	8,38	39,33	4,31	110,97	27033,46
T3 2004	8,47	45,00	4,16	110,63	27184,71
T4 2004	8,50	38,00	3,83	111,26	27178,72
T1 2005	8,26	38,67	3,64	110,97	27199,43
T2 2005	8,44	36,00	3,37	110,88	27324,99
T3 2005	8,62	39,00	3,23	110,62	27362,27
T4 2005	8,65	32,33	3,39	111,63	27326,01
T1 2006	8,72	33,67	3,51	111,59	27396,38
T2 2006	8,57	29,67	3,99	113,42	27396,62

Voici en tableau n°2 les 10 premières observations de la base de données élaborée personnellement pour cette étude sur le taux de chômage en France. Les observations sont trimestrielles et s'étendent de début 2004 au troisième trimestre de 2019, pour un total de 63 observations. Nous chercherons ainsi à prédire le taux de chômage du dernier trimestre de 2019 dont les données sont parues le jeudi 13 février 2019. Les données du taux de chômage,

du taux d'intérêt, de la production industrielle et de la population active sont issues du site internet de l'OCDE.[15] Les données de popularité du mot 'emploi' quant à elles, sont issues du site Google Trends.[16] Le taux de chômage est exprimé en pourcentage de la population active, le taux d'intérêt en pourcentage par année, la production industrielle est indexée base 100 en 2015, la population active en nombre de personnes, et la popularité du mot 'emploi' est calibrée entre 0 et 100 sur la période étudiée.

TABLEAU N°3 : Statistiques descriptives des 5 variables

```
> summary(df[,3:7])
  Taux_chomage     Popularite_emploi  Taux_interet      Production_indus_manu Population_active
 Min.   : 6.854   Min.   :24.67     Min.   :-0.2300   Min.   : 93.57        Min.   :27016
 1st Qu.: 8.509   1st Qu.:38.84     1st Qu.: 0.8933   1st Qu.: 99.64        1st Qu.:27800
 Median : 8.817   Median :65.33     Median : 3.0094   Median :102.64        Median :28325
 Mean   : 8.981   Mean   :61.39     Mean   : 2.5004   Mean   :104.13        Mean   :28531
 3rd Qu.: 9.786   3rd Qu.:82.50     3rd Qu.: 3.7142   3rd Qu.:110.70        3rd Qu.:29503
 Max.   :10.477   Max.   :94.67     Max.   : 4.4847   Max.   :115.38        Max.   :29867
```

```
> round(sapply(df[,3:7],mean),2)
        Taux_chomage     Popularite_emploi          Taux_interet Production_indus_manu     Population_active
                8.98                 61.39                  2.50                104.13              28531.30
> round(sapply(df[,3:7],sd),2)
        Taux_chomage     Popularite_emploi          Taux_interet Production_indus_manu     Population_active
                0.92                 22.83                  1.44                  5.96               907.36
```

Source : Logiciel R studio

D'après le tableau n°3 on voit que le taux de chômage varie entre 6,85% et 10,48% sur la période allant de 2004 à 2019, la médiane étant très proche de la moyenne cela montre qu'il n'y a pas de valeurs aberrantes ou atypiques qui tirent la moyenne à la hausse ou à la baisse, malgré ce que l'on aurait pu supposer avec la crise mondiale des Subprimes en 2008. L'écart type du taux de chômage est de 0,92 ce qui correspond à 10% de sa moyenne. La popularité du terme 'emploi' sous Google Trends fluctue entre 25 et 95%, elle n'atteint pas 100% parce que les données extraites de l'outil internet étaient initialement mensuelles, nous avons fait la moyenne pour chaque 3 mois pour les passer en trimestrielles. Cependant, les recherches atteignent leur maximum en septembre 2014 (la région du Limousin rassemble le plus grand nombre de recherches parmi les régions de France). On note aussi que toutes les variables sont homogènes puisque les écart-types sont largement inférieurs aux moyennes, cela se confirme par les diagrammes en moustache disponibles en annexe n°1 où l'on voit qu'aucune valeur n'est atypique. Enfin, il est intéressant de noter que les variations du taux d'intérêt

[15] https://data.oecd.org/fr/ (consulté le 26/01/2020)
[16] https://trends.google.fr/trends/?geo=FR (consulté le 23/01/2020)

s'étendent de -0,23% à 4,48%, constat sur lequel nous reviendrons dans l'analyse économique de cette variable.

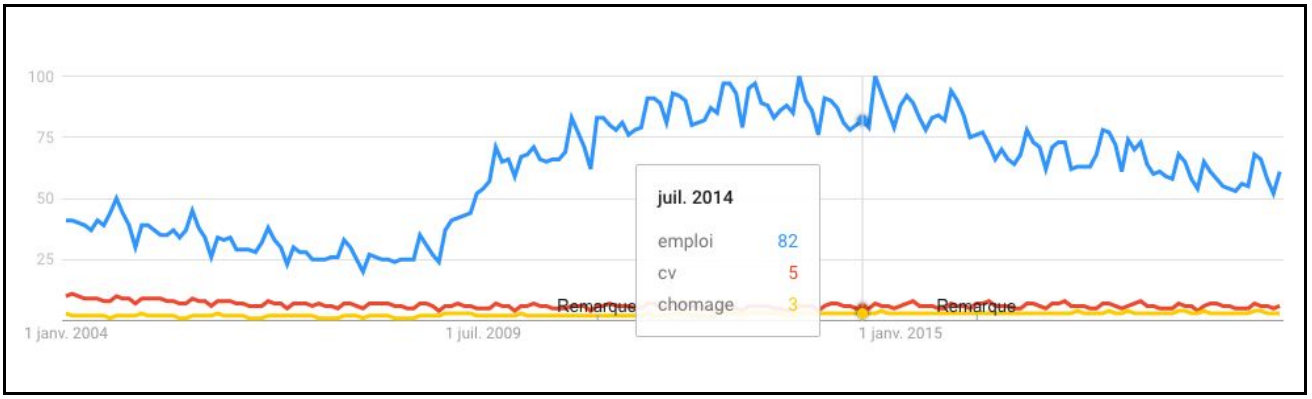
2- Google Trends : X_1

L'accès à internet n'a cessé de croître dans les foyers français depuis son apparition dans les années 70, aussi, en 2018 près de 9 ménages sur 10 y avaient accès.[17] Ce système de télécommunications informatiques permet à tout usager d'accéder à une immensité d'informations de tous types tels que ; articles, images, vidéos et d'autres encore. De plus en plus, les offres d'emploi se font en lignes : les demandeurs d'emploi (c'est à dire les entreprises) postent des annonces décrivant les emplois vacants et le profil type recherché, et les offreurs d'emploi (les particuliers) accèdent à ces requêtes en ligne et peuvent ainsi consulter et répondre aux offres. Aujourd'hui, ce sont 88% des offreurs d'emploi qui utilisent internet pour effectuer leurs démarches de recherche. [18]

Ainsi, le développement d'outils permettant la visualisation des mots utilisés lors des recherches internet est particulièrement intéressant du point de vue de l'emploi. De plus, les données sur le chômage et l'emploi mettent 1 à 2 mois à sortir, en ce sens l'utilisation d'outils disponible en instantané peut être utile pour combler cet écart et prédire facilement le niveau d'emploi français. Aussi, pour voir l'importance de Google Trends dans la prédiction du chômage il convient de choisir un mot s'y référant et de regarder ses tendances de recherche (calibrées entre 0 et 100 comme expliqué en introduction). Pour choisir le mot adéquat nous comparons la popularité de recherches des mots "emploi", "CV" et "chômage" sur la période de 2004 à ce jour. Nous cherchons la popularité relative de chacun des mots entrés comme 'terme de recherche' et non comme 'sujet' dont la différence a été expliquée en introduction. Il est effectivement plus intéressant de saisir ici un terme de recherche puisque les résultats prendront en compte toute recherche effectuée avec ce mot : pour 'emploi' par exemple seront aussi incluses des recherches telles que "le pôle emploi", "le bon coin emploi", "offres d'emploi" etc.

[17] https://fr.statista.com/statistiques/509227/menage-francais-acces-internet/ (consulté le 25/01/2020)
[18].https://www.lefigaro.fr/emploi/2017/01/17/09005-20170117ARTFIG00290-les-francais-cherchent-un-emploi-sur-internet-mais-le-trouvent-grace-a-leur-reseau.php (consulté le 21/01/2020)

GRAPHIQUE N°1 : Comparaison de la popularité des termes “emploi”, “CV” et “chômage”

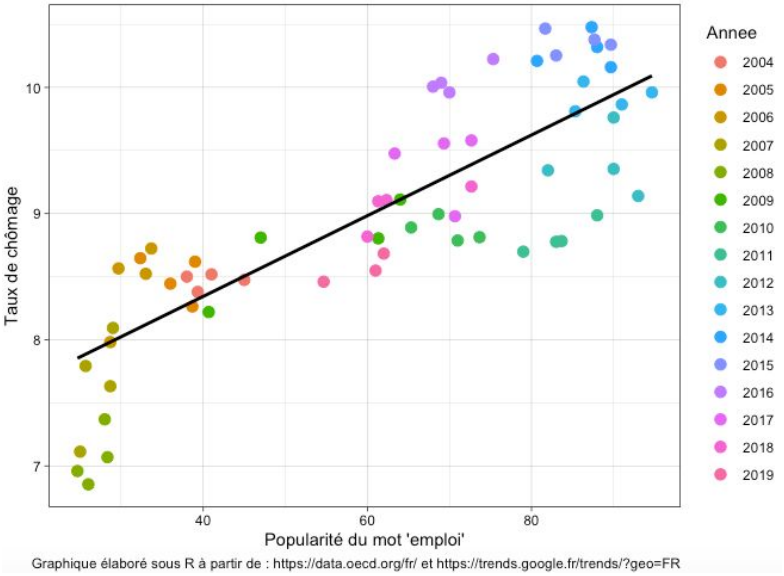


source : site internet de Google Trends

Le graphique n°1 nous donne visuellement les tendances de ces 3 mots, nous rappelons qu’il ne s’agit pas là des fréquences d’apparition mais bien de la popularité relative d’un terme par rapport au nombre total de recherches effectuées sur les mêmes périodes et régions comme c’est le cas ici. Nous constatons donc que le terme “emploi” est plus populaire que les 2 autres puisqu’il a les tendances de recherche les plus élevées sur la période étudiée, il semble ainsi mieux refléter les tendances de recherche des internautes en terme d’emploi. De plus, nous notons des pics réguliers dans la courbe d’évolution des recherches ‘emploi’, il s’agit en réalité des mois de septembre de chaque année où les recherches augmentent fortement et sont liées à la rentrée, il existe donc un phénomène de saisonnalité pour cette série. Nous décidons de retenir ce mot pour l’étude du taux de chômage en ajoutant ses fréquences d’apparition comme variable explicative, de 2004 à ce jour.

Par conséquent, nous attendons une **relation positive** entre le taux de chômage et les recherches internet du mot ‘emploi’ ; en effet sa popularité dépend de l’intérêt suscité par cette recherche. En temps d’expansion où il n’y a généralement pas ou peu de chômage, les recherches d’emplois (que ce soit via des agences spécialisées ou via internet) sont faibles puisque le pays se trouve alors dans une situation proche du plein-emploi. En revanche lorsque l’activité économique ralentit les firmes se voient parfois contraintes de licencier leurs employés car la récession implique une baisse de la demande de la part des ménages dont le pouvoir d’achat diminue, qui conduit à une baisse de l’offre proposées par les entreprises comme réponse au ralentissement économique. Ainsi, l’intérêt des recherches liées à l’emploi augmente fortement et cela peut se mesurer notamment grâce à l’outil Google Trends.

GRAPHIQUE N°2 : Relation entre le taux de chômage et la popularité du mot 'emploi'



D'après le graphique de corrélation n°2 on observe effectivement une corrélation positive entre l'intérêt de la recherche du mot 'emploi' et le taux de chômage. Ainsi sur l'échantillon étudié, c'est à dire en France de 2004 à 2019, la théorie se vérifie. On peut noter que la grande majorité des observations se situent proches de la droite de corrélation, exceptés les trimestres de 2007 et 2008 (identifiables grâce au code couleur) où le taux de chômage et donc les recherches Google Trends du mot 'emploi' sont faibles. Cette période pré-crise était en effet caractérisée par un chômage extrêmement bas dû, selon Christine Lagarde (ministre de l'Économie et de l'emploi à l'époque), à la création d'environ 340.000 emplois en 2007.[19]

3- Taux d'intérêt : X_2

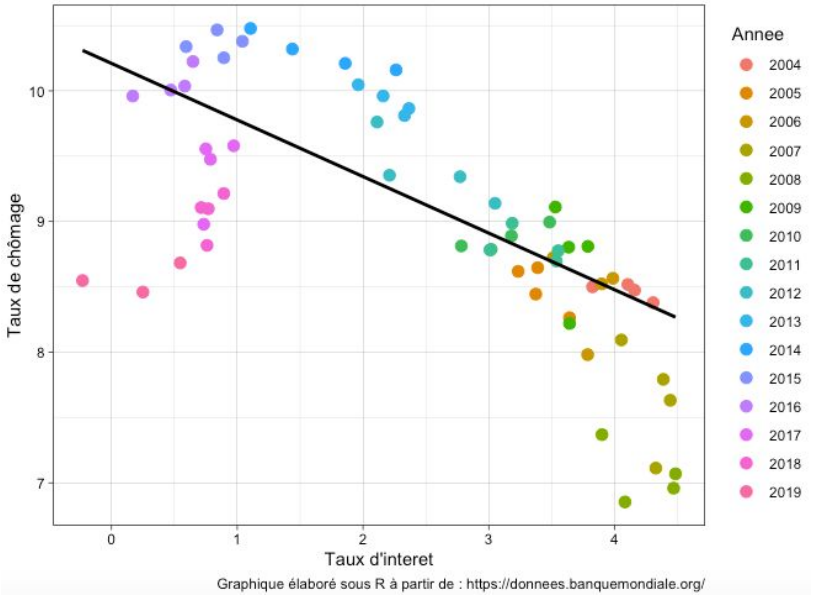
Le taux d'intérêt représente le prix qu'il faut payer pour emprunter de l'argent, prix qui rémunère le service rendu par celui qui prête l'argent (il est exprimé en pourcentage)[20]. Ainsi, les taux d'intérêt sont des déterminants importants de l'investissement des entreprises et de la consommation des ménages, principaux postes de la demande globale. En effet s'ils sont faibles, ces taux favorisent l'investissement des entreprises qui peuvent emprunter à un taux avantageu, au contraire ils freineront l'investissement s'ils sont élevés. Sachant que

[19] https://lexpansion.lexpress.fr/actualite-economique/taux-de-chomage-a-8-en-2007_470864.html (consulté le 07/02/2020)
[20] https://www.insee.fr/fr/metadonnees/definition/c1287 (consulté le 07/02/2020)

l'investissement détermine l'activité des entreprises donc la production globale et la croissance économique, il sera un bon indicateur du taux de chômage à travers l'offre d'emploi des firmes.[21] Il existe de nombreux taux d'intérêts tels que ceux de court terme, de moyen-long terme, les taux réels et nominaux… La variable que nous avons choisie est celle du taux d'intérêt français à long terme. Il correspond aux obligations d'État à échéance de 10 ans, c'est à dire les cours auxquels ces obligations s'échangent sur les marchés de capitaux.[22]

Comme expliqué précédemment le taux d'intérêt peut favoriser l'investissement lorsqu'il est faible et le freiner s'il est élevé car les coûts d'emprunts sont alors trop importants pour les entreprises. S'il le favorise (taux d'intérêt faible) il permet aux entreprises d'augmenter leur production ce qui crée un climat économique favorable à la création d'emploi et à l'embauche, et donc a une effet positif sur la santé économique globale du pays qui se traduit par une baisse du taux de chômage. [23] La relation théorique qui existe entre ces 2 variables est donc **positive**, nous allons voir maintenant si sur notre échantillon cette relation est vérifiée.

GRAPHIQUE N°3 : Relation entre le taux de chômage et le taux d'intérêt



Le lien qui apparaît sur le graphique n°3 est clair : il s'agit d'une relation négative contrairement à ce que l'on a pu supposer dans la partie économique. Cela peut être expliqué en partie par la période étudiée, en effet celle-ci englobe la situation presque utopique d'avant crise (à savoir une production forte, un taux de chômage faible etc.) puis les effets de la crise

[21] https://data.oecd.org/fr/interest/taux-d-interet-a-long-terme.htm (consulté le 07/02/2020)
[22] https://data.oecd.org/fr/interest/taux-d-interet-a-long-terme.htm (consulté le 07/02/2020)
[23] https://www.persee.fr/doc/reco_0035-2764_1999_num_50_5_410125 (consulté le 17/02/2020)

mondiale de 2008 qui est venue bouleverser l'équilibre économique des nations. On retrouve ainsi des points atypiques pour les trimestres pré-crise où le taux d'intérêt était relativement élevé (autour de 4%) pour un taux de chômage très faible (autour de 7%). En revanche sur ce graphique les valeurs les plus récentes (les 3 trimestres de 2019) semblent elles aussi atypiques ; en effet les taux d'intérêt négociés sont extrêmement bas avec un taux moyen négatif au troisième trimestre de 2019 (-0.23%). Cela est lié au taux directeur de la Banque Centrale Européenne (BCE) qui oriente les taux du marché et qui est volontairement institué à 0% (taux directeur nul) jusqu'au retour durable de l'inflation à 2%. Pour se faire, la BCE rachète de la dette publique et privée sur le marché, à hauteur de 20 milliards d'euros par mois de manière à réanimer l'activité économique via les prêts, l'investissement etc.[24] Ainsi depuis mars 2016 le taux directeur est nul (voir annexe n°2), ce qui explique les niveaux de taux faibles, ajoutés à un taux de chômage qui va diminuant, il apparaît que les points de 2019 sont inhabituels. Nous avons cependant vu qu'ils ne sont pas atypiques car comme précisé dans l'analyse des statistiques descriptives, il n'y a aucune valeur atypique pour les 5 séries.

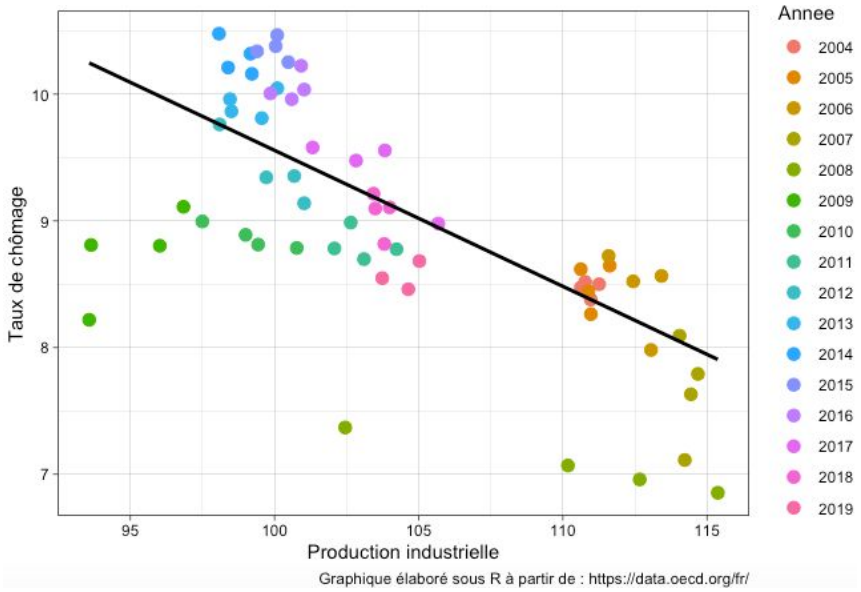
4- Production industrielle : X_3

La production industrielle manufacturière désigne la production d'entités industrielles, elle englobe plusieurs secteurs d'activité tels que l'extraction minière, les activités manufacturières, l'électricité, gaz, eau et climatisation. Exprimée sous forme d'un indice base 2015=100 dans notre cas, la production industrielle reflète les variations du volume de production du secteur secondaire sur une période donnée.[25] Représentant 16,9% du PIB français en 2018[26], la production industrielle est souvent utilisée comme indicateur de l'activité économique car elle suit de prêt ses variations. En effet le secteur secondaire est fortement lié aux cycles économiques, ainsi en phase d'expansion l'indicateur augmente et il diminue lors des périodes de récession. Les variations au sein du secteur secondaire sont souvent à l'origine des mouvements du PIB ; comme expliqué pour les taux d'intérêt, les périodes de croissance économique sont souvent synonymes de baisse du nombre de chômeurs liée à la création de nombreux emplois pour répondre à une demande en hausse. A contrario les périodes de ralentissement économique s'accompagnent souvent de

[24] https://www.lefigaro.fr/conjoncture/la-bce-maintient-ses-taux-directeurs-au-plus-bas-1-20191212 (consulté le 07/02/2020)
[25] https://data.oecd.org/fr/industry/production-industrielle.htm (consulté le 07/02/2020)
[26] https://donnees.banquemondiale.org/indicateur/NV.IND.TOTL.ZS?view=chart (consulté le 07/02/2020)

licenciements et ainsi d'un accroissement du taux de chômage dans le pays. En ce sens, la corrélation existante entre l'indice de production industrielle et le taux de chômage est **négative** : lorsque l'économie est en bonne santé (fort volume de production industrielle donc hausse de l'indice) le taux de chômage a tendance à baisser. Vérifions à présent si cette relation est vérifiée sur notre échantillon de données temporelles.

GRAPHIQUE N°4 : Relation entre le taux de chômage et la production industrielle



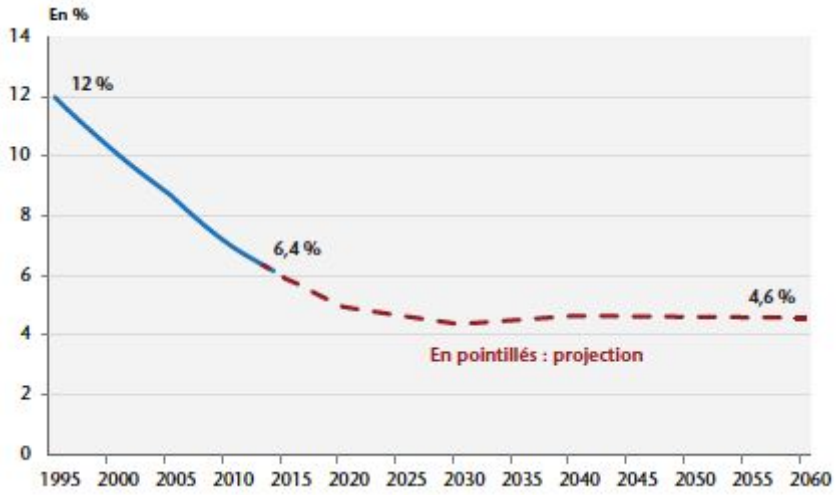
D'après le graphique n°4 on constate une corrélation négative entre ces 2 variables, ce qui confirme donc notre hypothèse. Encore une fois les trimestres de l'année 2008 sont excentrés du groupement des observations le long de la droite de corrélation. L'indice étant basé sur l'année 2015, on voit qu'avant cette période l'indice de production industrielle augmente régulièrement jusqu'en 2008 où il chute à 93 pour le premier trimestre de 2009. Les effets de la crise mondiale des Subprimes se ressentent largement sur la production nationale qui baisse progressivement en 2008 (voir points verts clairs proches de l'axe des abscisses) puis reprend de la vigueur et ré-augmente jusqu'en 2019 où elle atteint 105 points par rapport à la base 100 en 2015.

5- Population active : X_4

"Chaque année, en France, 800.000 jeunes entrent sur le marché du travail tandis que 670.000 seniors partent à la retraite. La population active progresse donc de 130.000 personnes, ce qui signifie qu'il faut créer de très nombreux postes pour que le chômage stagne

ou ne progresse pas."[27] Cette explication de Éric Heyer (docteur en Sciences Économiques) illustre le fait que la population active peut croître du fait de la démographie, entraînant ainsi une hausse de la population en âge de travailler pour laquelle il est parfois difficile de trouver un emploi car les offres d'emploi (des individus) deviennent supérieures aux demandes d'emploi (des entreprises). La population active correspond au nombre de personnes en âge de travailler c'est à dire entre 15 et 65 ans étant actifs, c'est donc la population en âge de travailler à laquelle on retire les inactifs que sont les étudiants, les hommes/femmes au foyer et les retraités. Depuis quelques années, la France comme de nombreux autres pays, assiste à un vieillissement de sa population. Effectivement, le taux de fécondité français en 2017 était de 1,86[28] enfants par femme alors que le taux qui permet de garder une population active stable (sans vieillissement de la population) est de 2,1.

GRAPHIQUE N°5 : Ratio "actifs âgés de 15-54 ans" sur "actifs de 55 ans et plus"



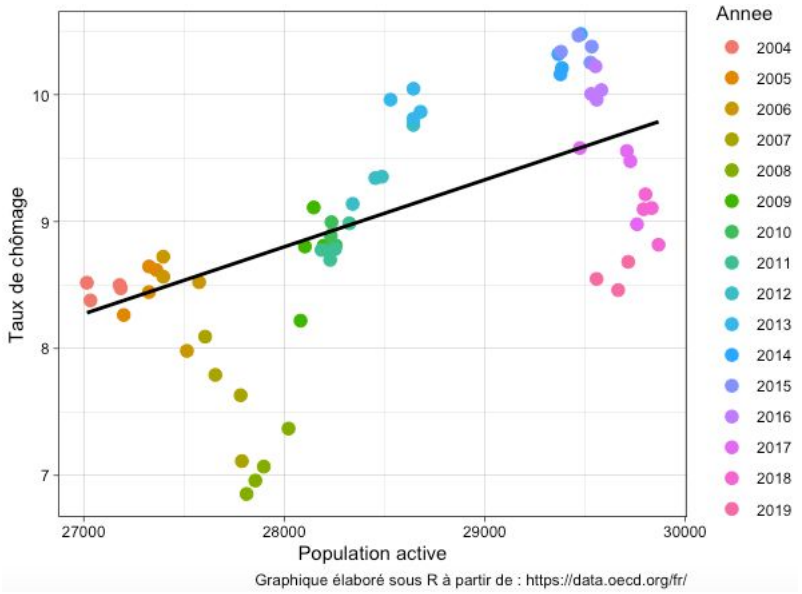
Source : https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm (consulté le 09/02/2020)

Aussi, comme visible sur le graphique n°5, au cours des 30 dernières années la population active française a beaucoup vieilli ; en 1995 il y avait 12 fois plus d'actifs de 15 à 54 ans, que d'actifs de 55 ans et plus alors que ce ratio tend à se stabiliser autour de 4,5% à partir de 2030[29]. Cette rapide décroissance du ratio s'explique par l'arrivée de nombreux baby-boomers à l'âge de la retraite dès les années 2000. Le pic de natalité d'après guerre expliqué par l'optimisme général, le redémarrage de l'économie et l'amélioration du niveau de vie global, a conduit à un accroissement démographique important allant de 1942 à 1973.

[27] https://www.brief.eco/a/2018/10/31/on-fait-le-point/les-determinants-du-chomage/ (consulté le 23/01)
[28] https://data.oecd.org/fr/pop/taux-de-fecondite.htm (consulté le 09/02/2020)
[29] https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm (consulté le 09/02/2020)

Or, un fort taux démographique entraîne une population active plus nombreuse nécessitant ainsi d'importantes créations d'emplois pour garder constant le taux d'emploi au sein du pays. Seulement, en temps de crises ou de ralentissement économique les emplois se font rares et très peu sont créés, provoquant ainsi une hausse soudaine du nombre de chômeurs. Aussi, comme expliqué précédemment[30] le taux de chômage a tendance à diminuer quand la population active diminue puisqu'il y a alors un moins grand écart entre l'offre et la demande d'emplois : la corrélation entre taux de chômage et population active est donc **positive**.

GRAPHIQUE N°6 : Relation entre le taux de chômage et la population active



D'après le graphique n°6 on observe une relation positive comme supposée théoriquement, en revanche cette relation n'apparaît pas très clairement car on peut voir que les points sont regroupés dans 3 zones différentes. Le niveau de chômage faible ces 2 dernières années (2018 et 2019) se distingue ici encore sur le graphique puisque les points correspondants sont plus éloignés de la courbe que les autres. Il en va de même pour les valeurs de 2007 et 2008 distinguables sur chaque graphique de corrélation par leur atypicité.

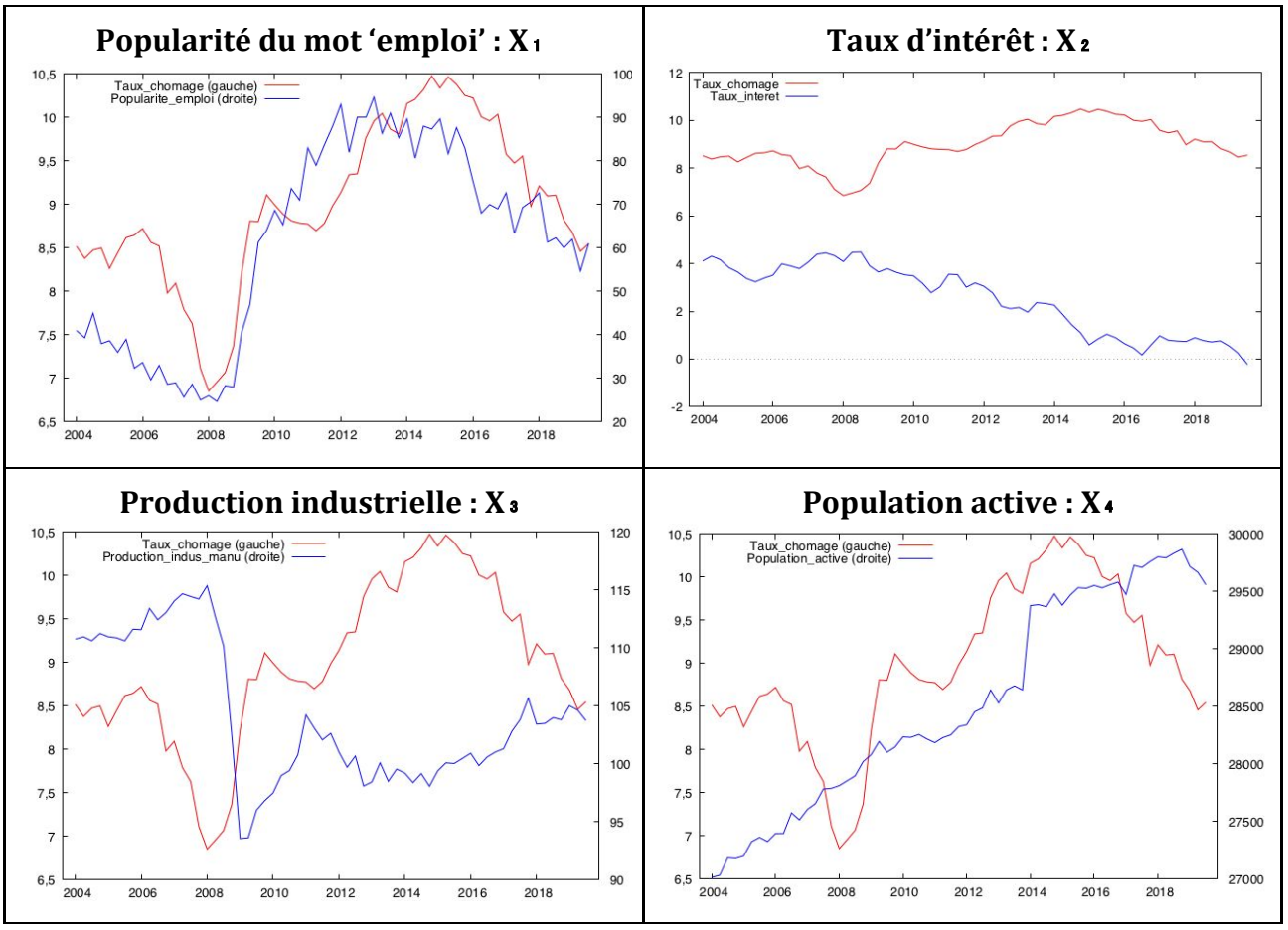
6- Graphiques d'évolution des variables explicatives

Nous pouvons terminer cette partie en regardant l'évolution dans le temps de chaque variable explicative mise sur un même graphique que celui du taux de chômage (Yt), de manière à voir

[30] Voir partie II ; analyse économique du taux de chômage page 7

si elles ont des évolutions similaires et ainsi pressentir si elles pourront être cointégrées ou non.

SCHÉMAS N°2 : Graphiques d'évolution des variables explicatives



Sur les schémas n°2 on voit en rouge le taux de chômage et en bleu les 4 variables explicatives. On voit d'emblée que les variables X_1 et X_3 suivent approximativement les fluctuations du taux de chômage, on peut donc facilement supposer qu'elles seront cointégrées au taux de chômage car on voit qu'elles convergent à long terme. Elles se rejoignent effectivement aux alentours de 2019. En revanche, les variables X_2 et X_4 semblent évoluer d'une manière totalement indépendante au taux de chômage, elles n'ont pas les mêmes variations sur la période étudiée, il apparaît ainsi qu'elles seront difficilement co-intégrables. De plus on note que les fluctuations du taux d'intérêt sont exactement opposées à celles du taux de chômage, il y a comme une symétrie de leurs évolutions ce qui confirme le relation négative que l'on a pu observée entre ces 2 variables sur le graphique de corrélation n°3.

IV- Présentation de la méthodologie ARIMA

L'économétrie est la mesure de l'économie qui se fait grâce la création et l'estimation de modèles ayant pour but d'identifier la relation entre des variables Y et X. Seulement, lors de la modélisation de cette relation, de nombreuses sources peuvent détériorer la qualité du modèle et l'estimation des paramètres. Cela peut être lié à la spécification de la forme du modèle, à la qualité d'estimation de ses paramètres ou encore à la partie aléatoire d'une variable qui baisse la qualité de prévision d'un tel modèle. C'est pourquoi il existe des méthodes d'estimation alternatives fondées sur les variations d'une série Y de manière à minimiser les erreurs en se basant sur les éléments connus d'une série ; c'est l'étude des séries temporelles.

Ainsi, une série temporelle ou 'processus temporel' est une suite d'observations numériques d'une variable donnée dans le temps qui est caractérisée par une tendance (à la hausse ou à la baisse), des cycles, une saisonnalité et une composante accidentelle. Les différentes méthodes d'analyse de séries temporelles diffèrent selon l'importance accordée à cette dernière, il en existe 3 approches : fréquentielle, globale et temporelle.

C'est en 1927 que G.U.Yule introduit les modèles autorégressifs AR dans son article "*On the method of investigating periodicities in distributed series with special reference to Wolfer's sunspot numbers*", et que E.Slutsky introduit les modèles moyennes-mobiles MA dans son article "*The summation of random causes as the source of cyclical processes*".[31] Puis en 1938, Wold propose un modèle ARMA combinant les modèles proposés par Yule et Slutsky, en un modèle linéaire basé sur la notion d'un processus infini de chocs aléatoires. Enfin, en 1954, il propose un processus ARMA stationnaire où :

➢ la partie autorégressive (AR) est constituée d'une combinaison linéaire finie de valeurs passées du processus et du terme aléatoire. Ce processus permet de modéliser les observations actuelles qui dépendent des observations antérieures.

➢ la partie moyenne mobile (MA) est constituée d'une combinaison linéaire finie de valeurs passées d'une variable aléatoire appelée "bruit blanc". Ce dernier est un processus qui contient des variables aléatoires indépendantes avec une variance constante et une espérance mathématique nulle, il aide à représenter les effets d'un choc dans un futur proche.

[31] https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf (consulté le 13/02/2020)

Le processus ARMA a été popularisé en 1970 lorsque les chercheurs Box et Jenkins ont publié l'ouvrage "*Time series analysis, forecasting and control*", montrant que la méthodologie ARMA pour l'étude de séries temporelles pouvait s'appliquer à de nombreux domaines. Le modèle ARMA ainsi popularisé est composé de 'p' éléments pour déterminer AR et de 'q' éléments pour déterminer MA ; tels que ARMA[p,d].[32] On a donc ;

AR(p) : $Z_t = \delta + a_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p}$

MA(q) : $Z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$

Aussi, il y a 2 conditions pour qu'un processus $[X_t]$ soit considéré comme un processus ARMA[p,d] :

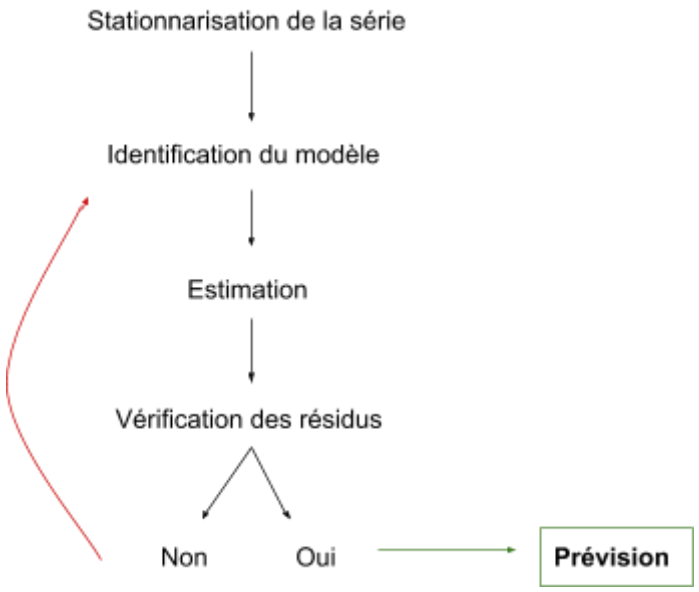
➢ **<u>Stationnarité</u>** : pour la partie AR, signifie que la série doit garder les mêmes caractéristiques dans le temps, on dit ainsi qu'elle est indépendante du temps. Ce concept de stationnarité est très important et constitue la première étape de la méthodologie ARIMA, il implique une même distribution de probabilités, une même espérance mathématique et une même variance pour chaque observations. Par conséquent, une série stationnaire tend à varier autour d'une valeur moyenne à long terme, c'est pourquoi la méthode d'estimation et de prévision 'Box Jenkins' s'applique à très court terme puisqu'à plus long terme c'est la moyenne mathématique qui dira les prédictions. Statistiquement, la condition se matérialise de la manière suivante ; pour un processus AR(1) on doit avoir $|\phi_1| < 1$ et pour un AR(2) on doit avoir $|\phi_2| < 1$, $\phi_1 + \phi_2 < 1$ et $\phi_2 - \phi_1 < 1$.[33]

➢ **<u>Inversibilité</u>** : pour la partie MA où tout est aléatoire, cette condition est indépendante de la stationnarité et est donc applicable sur un processus non stationnaire. De même que pour la stationnarité, la condition d'inversibilité se vérifie à travers la valeur des coefficients estimés pour le processus MA ; il faut $|\theta_1| < 1$ pour un MA(1), et $|\theta_2| < 1$, $\theta_1 + \theta_2 < 1$ et $\theta_2 - \theta_1 < 1$ pour un MA(2) .[34]

Les chercheurs Box et Jenkins ont développé une méthode permettant de trouver les paramètres 'p' et 'q' d'un modèle ARMA[p,q][35] et tenter de prévoir les variations d'une série sur le court terme, dont le déroulé est visible en schéma n°3.

[32] https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf (consulté le 13/02/2020)
[33] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.28-38.
[34] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.42-48.
[35] **Modèle ARMA** : Autoregressive moving average model

SCHÉMA N°3 : Méthode Box-Jenkins pour l'analyse de Y



source : élaboration personnelle à partir de l'outil "dessin" sous google drive

→ Stationnarisation de la série

Comme expliqué précédemment la première étape de la méthodologie ARIMA est celle de vérifier la stationnarité de la série de manière à pouvoir identifier le modèle pour l'estimer et, s'il est vérifié, passer à la prévision. La stationnarité s'effectue à 2 niveaux : au niveau de la variance et au niveau de la moyenne.

➢ Stationnarité autour de la variance : la série doit avoir une espérance mathématique constante ne dépendant pas du temps

➢ Stationnarité autour de la moyenne : la série doit fluctuer autour d'une valeur moyenne

Aussi, dans une série stationnaire l'autocovariance entre 2 observations dépendra uniquement du nombre de périodes qui les sépare et non du temps. Prenons l'exemple d'une série brute Y_t que l'on cherche à modéliser à l'aide d'un processus ARIMA. Imaginons que cette série ne soit pas stationnaire, on ne peut alors pas utiliser un modèle ARMA[p,q], il va falloir procéder à une transformation de la série brute Y_t. Si cette dernière n'est pas stationnaire autour de la variance, c'est à dire qu'il existe une tendance à la hausse ou à la baisse, on va alors la mettre en logarithme de manière à corriger cette *trend*, on a ainsi : y_t =log(Y_t). Il se peut aussi que cette nouvelle série ne soit visuellement pas stationnaire autour de la moyenne, c'est à dire que la dispersion autour de la moyenne augmente/diminue au cours du temps, on va alors procéder à une différenciation qui viendra corriger ces écarts, on

passe ainsi de y_t à Z_t. Celle-ci permet d'avoir une série stationnarisée, seulement il n'est plus possible d'utiliser un modèle ARMA puisque l'on a dû passer par une différenciation pour rendre Y_t stationnaire, on va alors parler de processus ARIMA[p,d,q][36] où d représente le nombre de différenciation nécessaires pour rendre la série stationnaire autour de la moyenne.

→ Identification du modèle

La deuxième étape de la méthode de Box-Jenkins consiste à identifier le type de processus, cela est possible en regardant la forme des fonctions d'autocorrélation (FAC) et d'autocorrélation partielle (FACP) d'une série stationnaire (d'où l'importance de l'étape précédente). Ces tracés sont contenus dans les corrélogrammes simples et partiels, ceux-ci ont été développés pour la première fois en 1884 par le physicien Poynting qui étudiait la relation entre le mouvement du prix du blé et les importations de coton et de soie.[37] Le processus d'identification du modèle est donc un processus visuel qui s'effectue uniquement en regardant la FAC et la FACP, 3 cas de figure sont alors possibles :

➢ Modèle classique : la ou les premières valeurs d'un des corrélogrammes sont significatives puis décroissent pour se placer autour de zéro à terme, et seules 1 ou 2 valeurs sont significatives dans l'autre corrélogramme, puis on observe une rupture. Cela signifie qu'il n'y a pas de partie saisonnière dans la série. Un modèle classique se présente sous forme d'un ARIMA[p,d,q] avec d=1 s'il a fallu différencier la série une fois pour la rendre stationnaire.

➢ Modèle saisonnier : la ou les premières valeurs des corrélogrammes ne sont pas significatives en revanche sur les périodes postérieures on observe des valeurs significatives, que ce soit dans la FAC ou dans la FACP. Ce phénomène de saisonnalité peut apparaître dans les séries temporelles qui ne sont pas annuelles, par exemple si l'on étudie les ventes de glaces mensuellement on observera un pic chaque année au moment de l'été : c'est la saisonnalité. Un modèle saisonnier se présente sous forme d'un SARIMA[P,D,Q $]_S$ [38] où 'S' correspond à la périodicité de la série, ainsi pour des données mensuelles S=12, pour des données trimestrielles S=4 etc.

➢ Modèle mixte : les premières valeurs des corrélogrammes suivent la description d'un modèle classique, mais au bout de k périodes on observe des valeurs qui sortent du seuil de significativité. Nous sommes alors dans le cas d'un modèle classique ET d'un modèle

[36] **Modèle ARIMA** : Autoregressive integrated moving average model
[37] https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf (consulté le 14/02/2020)
[38] **Modèle SARIMA** : Seasonal autoregressive moving average model

saisonnier. Aussi appelé 'modèle multiplicatif' le modèle mixte se présente ainsi : ARIMA[p,d,q]x[P,D,Q $]_S$.[39]

La spécification des processus doit suivre une règle importante : la règle de parcimonie. Celle-ci indique qu'il est mieux de choisir un modèle avec le moins de paramètres possibles, pour perdre le moins d'informations et s'éloigner au minimum de la série initiale.

→ Estimation du modèle

L'estimation du modèle identifié à l'étape précédente se fait de manière informatique, en revanche il convient de prêter attention aux résultats de cette estimation. Un modèle bien spécifié est un modèle dont les paramètres sont pertinents, c'est à dire qu'ils doivent être significatifs, et justes ; il faut que leur valeur ne soit pas trop proche de 1 pour vérifier les conditions de stationnarité et d'inversibilité (on fixe donc la limite à 0,95).

→ Vérification des résidus notés $\{\widehat{a}_t\}$

L'étape de vérification des résidus est très importante puisqu'elle permet de valider ou non le modèle identifié en deuxième étape de la méthode Box-Jenkins, permettant ainsi d'améliorer le modèle si besoin. Pour être validé, un modèle doit avoir des résidus indépendamment distribués, cela implique que ces derniers suivent un processus 'bruit blanc' (au moins pour les 2 premières valeurs du corrélogramme, à savoir K=1 et K=2) et qu'ils suivent approximativement une loi normale. En effet en observant le corrélogramme des résidus, aucune valeur ne doit (théoriquement) dépasser le seuil de significativité fixé par le test de Bartlett à $\widehat{\varphi_K}(\widehat{a}_t) = \pm 1,96\sqrt{\frac{1}{T}}$. Il y a pour cela 2 manières de vérifier la distribution des résidus :

➢ <u>En analysant chaque autocorrélation de manière indépendante</u> : la première valeur de la FAC($\widehat{a}_t$) doit être très éloignée du seuil de significativité car cela montre que la variance est faible et donc que le modèle est pertinent. Dans le cas contraire (premières valeurs de la FAC significatives) il faut retourner à l'étape d'identification du modèle, car si les résidus ne suivent pas un processus de 'bruit blanc' cela signifie que le modèle est incorrect, peu pertinent, que la saisonnalité n'a pas été prise en compte ou que la série n'a pas été stationnarisée.

[39] Tsay R. S., "Analysis of financial times series", *University of Chicago*, 2002, pp.48-53 .

➢ <u>En analysant les résidus dans leur ensemble</u> : grâce au test porte-manteau qui s'applique de la manière suivante :

(1) H_0 : $\varphi_1 = \varphi_2 = \varphi_K = 0$ (hypothèse d'indépendance des résidus)

H_1 : pas H_0

(2) $Q=T \sum_{k=1}^{K} \varphi_K^2(\widehat{a_t}) \sim \chi^2$

(3) Quand la valeur de Q augmente la p-value diminue et on refuse H_0.
Quand la valeur de Q diminue la p-value augmente et on accepte H_0.

Aussi, chaque inadéquation du modèle est signalée par l'accroissement de la valeur de Q, lorsque l'on compare plusieurs modèles on va donc choisir celui qui a la p-value la plus élevée, de manière à accepter H_0 pour que les résidus soient vérifiés.

→ Prévisions

"All models are false, but some are useful" (G.Box, 1979). Cette phrase de Box montre que les processus de modélisation et d'estimation des paramètres d'une série sont toujours approximatifs, mais que certains modèle sont tout de même utiles notamment pour faire de la prévision ; il y a donc toujours une erreur que l'on essaye de minimiser. Les prévisions se font à très court terme car à long terme la série tend vers la tendance centrale (c'est à dire sa moyenne arithmétique) si la série est stationnaire. Par conséquent, la prévision se base sur l'hypothèse de dispersion constante autour de la variance. On cherche donc à savoir si la série observée entre 1 et T, s'éloigne ou s'approche de la tendance centrale à horizon h. La prévisions d'un processus ARIMA[p,d,q]x[P,D,Q $]_S$ s'effectue de la manière suivante ; on utilise la formule $(1-\phi_1 B) Z_t = (1-\Theta B^{12})(1-\theta_1 B) a_t$ pour trouver Z_{T+1}, prévision en t+1 de notre série stationnarisée, que l'on estime par $\widehat{Z_T}(1)$ de manière à finalement trouver $Y_T(1)$.

Nous allons à présent pouvoir appliquer cette méthodologie sur notre base de données !

V- Choix du modèle ARIMA pour Y_t

L'analyse que nous effectuerons dans ce dossier suit le schéma n°2 détaillé dans la partie précédente grâce au logiciel de statistiques **'Gretl'**.

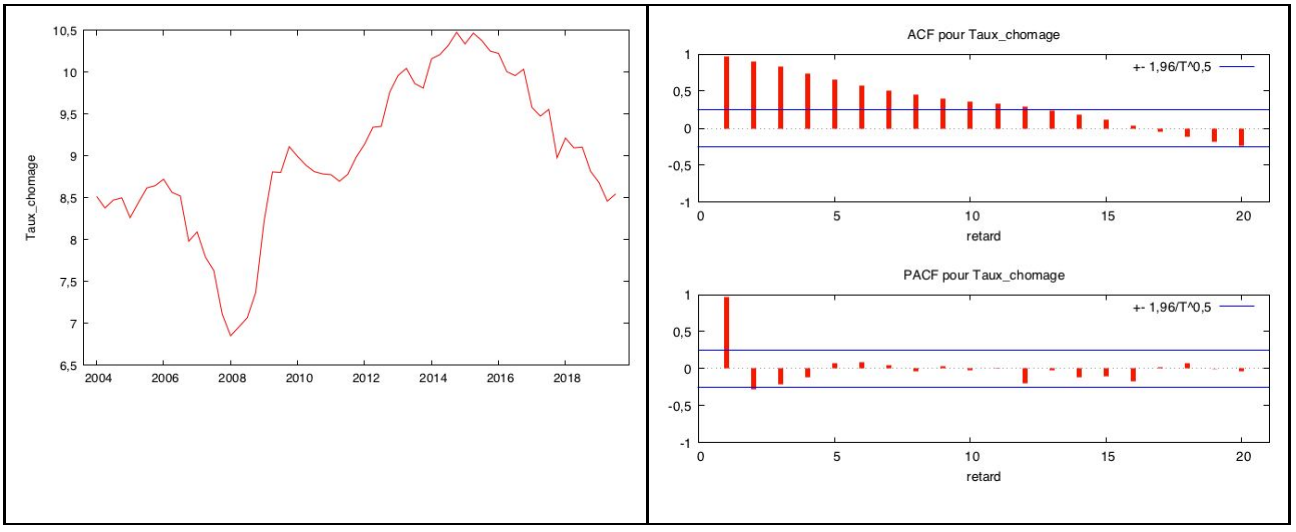
Notre série étant trimestrielle nous pouvons supposer l'existence d'une saisonnalité où l'on retrouve un même phénomène toutes les 4 périodes. De plus, les observations s'étendant de 2004 à 2019 nous pouvons imaginer que la crise mondiale de 2008 créera de fortes fluctuations dans nos séries puisqu'il s'agit d'indicateurs économiques.

Nous décidons de prendre une ordre maximum de retard de 20 périodes pour les corrélogrammes, ce qui correspond à 5 années d'observations consécutives allant de début 2004 à fin 2008. En prenant un nombre important de retards nous retenons l'impact de la crise de manière à garder un échantillon représentatif et à visualiser d'éventuelles valeurs significatives en dehors des premières périodes dans les différents corrélogrammes. Nous garderons ainsi ce même nombre de retard pour toute la partie de stationnarisation, de manière à pouvoir comparer les différents modèles estimés.

→ Stationnarisation

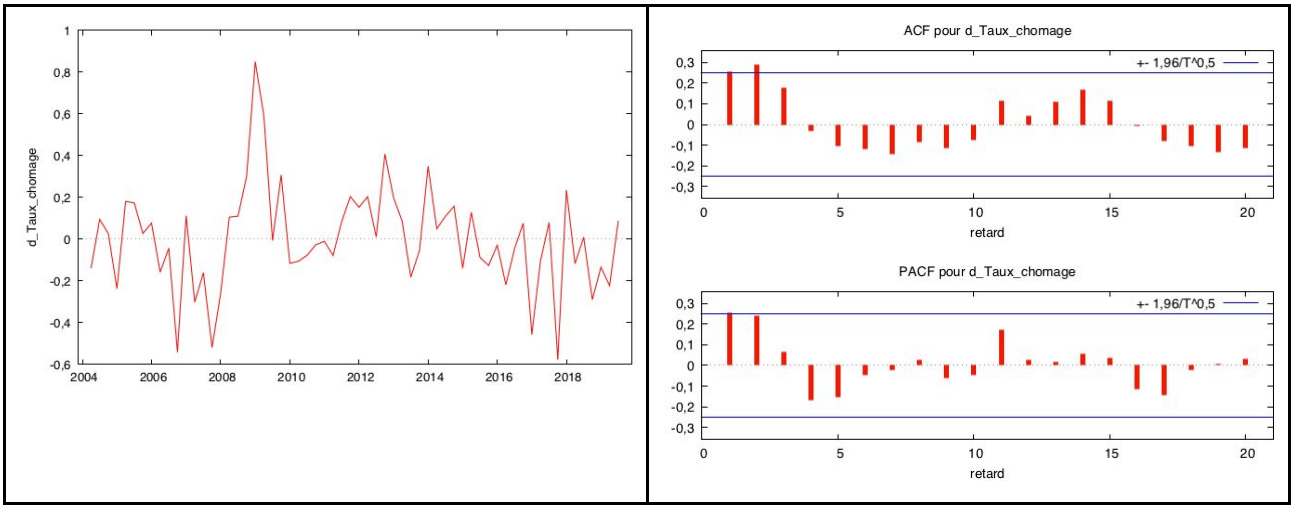
Comme expliqué précédemment, un processus est dit stationnaire si la série est stationnaire pour la variance et pour la moyenne ; c'est à dire qu'il ne doit pas y avoir de tendance à la hausse ou à la baisse, ni de coupure dans la série ou encore de processus aléatoire.

GRAPHIQUE N°7 : Évolution temporelle et corrélogramme du taux de chômage "brut" (Y_t)



Il apparaît sur le graphique n°7 que la série n'est pas stationnaire, en effet les dispersions autour de la moyenne augmentent et diminuent fortement sur la période étudiée avec un pic inhabituel lié à la crise de 2008, comme supposé précédemment. Il convient donc de passer par un retard pour corriger cette trop grande volatilité de la série, en faisant une différenciation qui nous coûtera une observation (nous n'aurons plus que 62 périodes) mais permettra de stabiliser la série. Il ne semble pas y avoir de tendance à la hausse ou à la baisse car la chute du taux de chômage jusqu'à 2008 est compensée par une forte hausse dès 2012, il n'est donc à priori pas nécessaire de passer la série en logarithme car elle est déjà stationnaire autour de la variance. Disponibles en annexe n°3, les graphiques du taux de chômage passé en logarithme confirment que cette transformation n'est pas nécessaire puisqu'il n'y a pas ou très peu de différence avec ceux de la série brute. Nous effectuons donc une différenciation sur la série initiale. De plus, les valeurs dans la fonction d'autocorrélation décroissent lentement tandis qu'on observe une réelle coupure dans la fonction d'autocorrélation partielle dès K=2 ; on peut supposer l'existence d'un modèle autorégressif AR(p). Enfin, nous constatons qu'il n'y a pas de valeurs significatives en dehors des premières valeurs, cela signifie qu'il n'y a pas de saisonnalité, contrairement à ce que l'on avait supposé avant de regarder les graphiques de Y_t.

GRAPHIQUE N°8 : Évolution temporelle et corrélogramme du taux de chômage différencié



Après une différenciation visible sur le graphique n°8, il semble que la série soit davantage stationnaire par rapport à la moyenne ; l'inertie générale de la série a été réduite, les dispersions autour de la moyenne sont plus homogènes. La transformation était donc nécessaire pour stationnariser la série, l'effet de la crise des Subprimes est toujours visible mais semble atténué puisque l'inertie post-choc est plus faible : la trend revient plus

rapidement à la moyenne. Notre série Y_t est stationnarisée grâce à une différenciation, nous la notons Z_t et l'utiliserons pour la méthode Box-Jenkins. Enfin, nous pouvons vérifier la stationnarité de Z_t en effectuant le test ADF[40] qui s'intéresse au comportement des résidus à long terme, pour cela on pose les hypothèses :

(1) H_0 : La racine unitaire existe, Z_t n'est pas stationnaire

H_1 : Z_t est stationnaire

(2) $t_{obs} \gtrless ADF_{0,05}$

(3) Quand la p-value diminue on refuse H_0 ; Z_t est stationnaire.

La p-value associé au test de la racine unitaire pour notre modèle est 0,0003 comme visible dans le tableau n°4, nous refusons H_0 ; notre série différenciée une fois est bel et bien stationnaire. Nous pouvons à présent passer à l'estimation en cherchant le modèle correspondant à notre série différenciée.

TABLEAU N°4 : Résultats du test ADF sous Gretl

```
Test de Dickey-Fuller augmenté pour d_Taux_chomage
testing down from 10 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec un retard de (1-L)d_Taux_chomage
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,559901
  statistique de test: tau_nc(1) = -3,60912
  p. critique asymptotique 0,000304
  Coeff. d'autocorrélation du 1er ordre pour e: -0,018
```

→ Identification et estimation du modèle

L'annexe n°4 indique que les 2 premières valeurs de la FAC et la première de la FACP de Z_t sont significatives au seuil de 5%. Sachant qu'aucune autre valeur n'est significative au bout de K périodes, on sait qu'il s'agit d'un modèle classique qui se présentera donc sous la forme d'un ARIMA[p,d,q] où d=1 puisque nous avons différencié une fois la série initiale pour la rendre stationnaire. On observe à partir du corrélogramme du graphique n°8 qu'il y a 2 valeurs significatives dans la FAC et une dans la FACP ce qui nous amène à modéliser un processus ARIMA[1,1,0] puisque les corrélogrammes s'apparentent à ceux d'une spécification AR(1).

[40] **Test ADF** : test de Dickey-Fuller augmenté

TABLEAU N°5 : Estimation d'un premier modèle ARIMA[1,1,0]

```
Évaluations de la fonction : 14
Évaluations du gradient : 6

Modèle 1: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z       p. critique
  ---------------------------------------------------------------
  const      0,000202342    0,0397780    0,005087     0,9959
  phi_1      0,251465       0,122154     2,059        0,0395    **

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000637
  Ec. type des innovations   0,235715
  R2                         0,064824
  R2 ajusté                  0,064824
  Log de vraisemblance       1,591349
  Critère d'Akaike           2,817303
  Critère de Schwarz         9,198706
  Hannan-Quinn               5,322803

                       Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  AR
  Racine  1           3,9767     0,0000     3,9767     0,0000
  ---------------------------------------------------------------
```

L'estimation du modèle est disponible dans la tableau n°5, on voit d'emblée que la constante n'est pas significative, ce qui signifie que la série évolue autour de 0, il faut donc réestimer le modèle en retirant la constante.

TABLEAU N°6 : Estimation du même modèle ARIMA[1,1,0] sans la constante

```
Évaluations de la fonction : 16
Évaluations du gradient : 3

Modèle 2: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z      p. critique
  ---------------------------------------------------------------
  phi_1       0,251468      0,122152     2,059     0,0395    **

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000789
  Ec. type des innovations   0,235715
  R2                         0,064827
  R2 ajusté                  0,064827
  Log de vraisemblance       1,591336
  Critère d'Akaike           0,817329
  Critère de Schwarz         5,071598
  Hannan-Quinn               2,487662

                       Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  AR
  Racine  1           3,9766     0,0000     3,9766     0,0000
  ---------------------------------------------------------------
```

On voit dans le tableau n°6 l'estimation de ce même modèle sans la constante qui n'était pas significative, on peut voir que $\widehat{\phi_1}$ =0,251 donc les conditions de stationnarité et

d'inversibilité sont respectées puisque $\widehat{\varphi_1}$ est inférieur à 1 et en est assez éloigné pour être certain de la stationnarité. S'agissant d'un processus AR(1) auquel on ajoute une différentiation, celui-ci est par définition invertible. Économiquement parlant, $\widehat{\varphi_1} = 0,251$ signifie que la série est stationnaire avec une légère persistance de 25% (sachant que 0< ϕ <1 où 0 signifie aucune persistence/retour immédiat à une situation normale, et où 100% signifie que la série ne reviendra jamais à son état initial). De plus, l'écart-type s'élève à 0,236 ce qui est peu comparé à l'écart-type de la variable brute de 0,92, et donc satisfaisant pour notre modèle.

TABLEAU N°7 : Pourcentage d'erreur de prévision du modèle ARIMA[1,1,0] sans la constante

2018:1	0,235416	-0,145123	0,380539
2018:2	-0,116902	0,059200	-0,176102
2018:3	0,009221	-0,029397	0,038618
2018:4	-0,289367	0,002319	-0,291686
2019:1	-0,134507	-0,072767	-0,061740
2019:2	-0,222715	-0,033824	-0,188891
2019:3	0,087860	-0,056006	0,143866

Note : * indique un résidu supérieur à 2,5 fois l'écart type
Statistiques d'évaluation des prédictions using 62 observations

Erreur Moyenne	0,00078902
Racine de la moyenne des erreurs au carré	0,23571
Erreur absolue moyenne	0,17439
Mean Percentage Error	120,81
Mean Absolute Percentage Error	138,38
U de Theil	0,99092

À horizon h>0 il y a plusieurs moyens de connaître le pourcentage d'erreur associé à telle ou telle prévision, définis par les 3 critères suivants[41] :

➢ L'erreur moyenne : ce critère est très peu utilisé car la moyenne est faite sur les valeurs 'brutes' donc si l'erreur de prévision pour une année est de -0,3% et celle de l'année suivante est de +0,3% alors ce critère donnera une erreur moyenne de 0 c'est à dire une qualité de prévision parfaite, alors qu'en réalité elle ne l'est pas.

➢ L'erreur absolue moyenne : contrairement au premier critère, celui-ci fait la moyenne des erreurs de prévision en valeurs absolues, il corrige donc les limites de la méthode n°1.

➢ L'erreur quadratique moyenne : ce dernier est le critère le plus largement utilisé puisqu'il prend en compte la variance associée à chaque erreur de prévision en mettant chacune d'elle au carré, c'est donc le moyen le plus précis pour connaître l'erreur de prévision.

[41] Vaté M., "Statistiques chronologiques et prévisions", *Economica*, 1993, pp.217-225.

Nous cherchons la qualité de prévision du modèle ARIMA[1,1,0] sans la constante en terme de pourcentage d'erreur. D'après le tableau n°7 on voit que les deux premiers critères présentés ci-dessus sont disponibles en pourcentage d'erreur, nous choisissons donc l'erreur **absolue** moyenne en pourcentage pour les raisons énoncées précédemment. Ainsi, le pourcentage d'erreur est de 138,38% ce qui est considérable, ceci peut être expliqué par 2 choses. Premièrement nous avons estimé un modèle ne comprenant pas de partie moyenne-mobile (MA) alors que celle-ci, comme précisé dans le développement de la méthodologie ARIMA page 21, aide à représenter les effets d'un choc en période 't' mais aussi dans un futur proche. Avoir une partie MA dans le modèle aiderait ainsi à réduire ce pourcentage d'erreur. Deuxièmement l'étendue de la variable 'taux de chômage' (notre Y_t) est assez réduite puisqu'elle va de 6,85% à 10,47%, donc dans cette analyse la qualité de la prévision se joue à des dizaines ou centaines après la virgule. Effectivement, l'échelle étant réduite, une petite erreur de prévision se ressentira directement sur le taux d'erreur qui augmentera fortement. Par exemple, si pour une variation de 0,01 le modèle prévoit 0,02 le pourcentage d'erreur sera de 100%, c'est pour cela qu'il est important de regarder l'étendue / les variations de la série initiale. Cet écart est aussi notable sur le graphique des valeurs prédites par le modèle et des valeurs observées, disponible en annexe n°5.

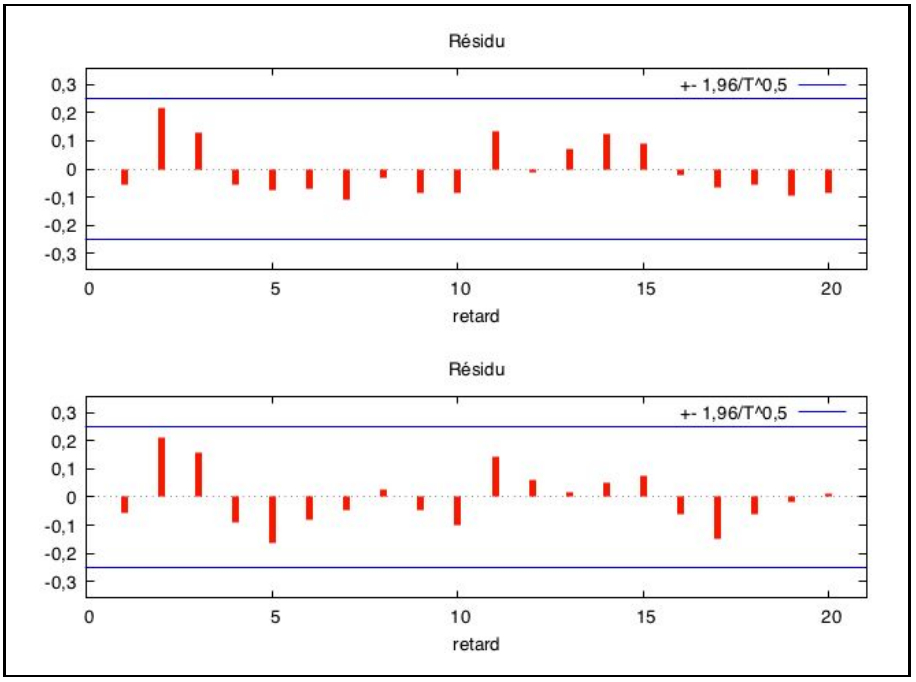
Nous pouvons passer à présent à l'étape de vérification des résidus qui viendra valider ou invalider notre modèle.

→ Vérification et choix du modèle

Comme précisé en partie précédente, il existe 2 approches pour vérifier que les résidus suivent un processus de bruit blanc. Nous commencerons par l'approche individuelle puis nous étudierons les résidus dans leur ensemble en regardant la valeur de la statistique Q, enfin nous regarderons la normalité des résidus grâce aux histogrammes et aux graphiques Q-Q, en privilégiant la valeur du Q-stat.

➢ <u>Analyse individuelle</u> : D'après le graphique n°9 on voit d'emblée que la première valeur du corrélogramme des résidus est très éloignée du seuil de significativité, elle se situe à -0,05 environ alors que le seuil, lui, est de 1,96 $\sqrt{\frac{1}{62}}$ =0,249. Cela indique que la variance est faible, donc notre modèle est pertinent. De plus, aucune valeur ne dépasse le seuil de significativité [-0,249 ; +0,249] donc les résidus suivent bien un processus 'bruit blanc', ils sont donc stationnaires et stochastiques.

GRAPHIQUE N°9 : Corrélogramme des résidus du modèle ARIMA[1,1,0] sans la constante



➢ Analyse globale : D'après le tableau n°8 on constate que la statistique Q du 20è retard s'élève à 13,39 pour une p-value de 0,818> $\alpha = 0,05$ ce qui signifie que l'on accepte largement l'hypothèse nulle selon laquelle les résidus sont indépendamment distribués. Cette nouvelle approche nous permet encore une fois de valider le modèle ARIMA[1,1,0] puisque ses résidus ont été vérifiés. Enfin, nous pouvons vérifier la normalité de ces derniers en sachant que c'est la valeur Q-stat qui prévaut sur toutes les approches possibles, et que celle-ci a déjà été vérifiée.

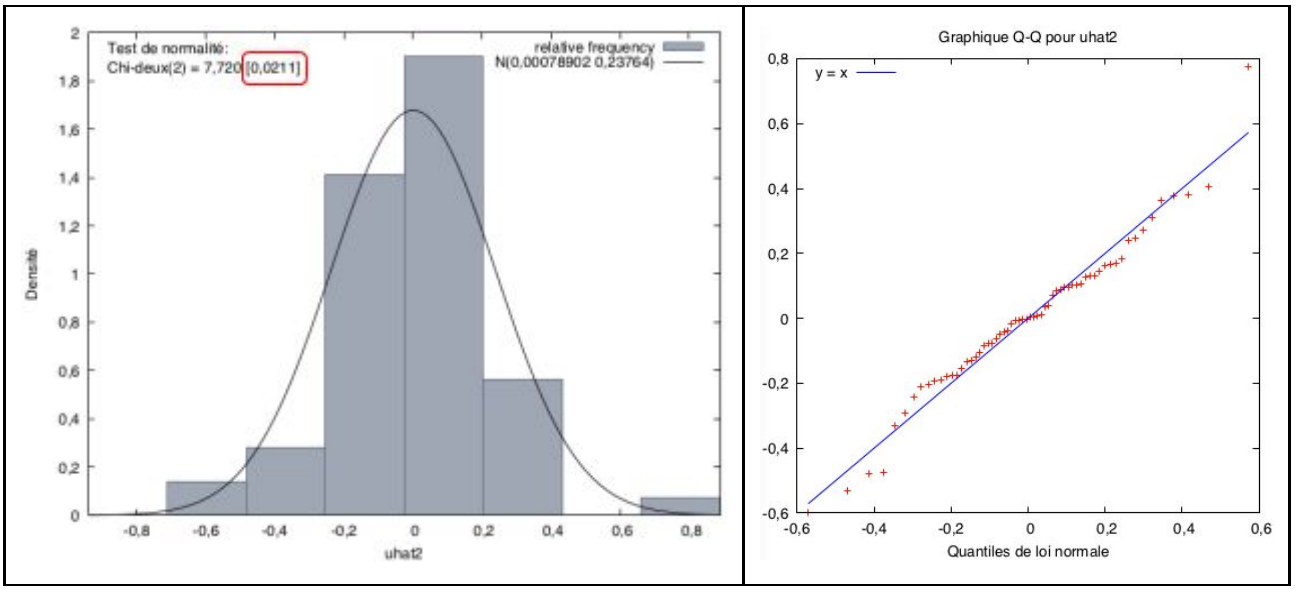
TABLEAU N°8 : Test porte-manteau des résidus du modèle ARIMA[1,1,0] sans la constante

Fonction d'auto-corrélation résiduelle
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5

RETARD	ACF		PACF		Q	[p. crit.]
1	-0,0578		-0,0578			
2	0,2145	*	0,2118	*	3,2593	[0,071]
3	0,1304		0,1601		4,4025	[0,111]
4	-0,0567		-0,0906		4,6225	[0,202]
5	-0,0767		-0,1626		5,0325	[0,284]
6	-0,0684		-0,0807		5,3643	[0,373]
7	-0,1081		-0,0461		6,2068	[0,400]
8	-0,0316		0,0255		6,2803	[0,507]
9	-0,0867		-0,0454		6,8437	[0,554]
10	-0,0850		-0,1003		7,3945	[0,596]
11	0,1342		0,1422		8,7954	[0,552]
12	-0,0139		0,0587		8,8106	[0,639]
13	0,0703		0,0154		9,2108	[0,685]
14	0,1257		0,0514		10,5159	[0,651]
15	0,0886		0,0767		11,1791	[0,672]
16	-0,0201		-0,0604		11,2140	[0,737]
17	-0,0643		-0,1510		11,5782	[0,772]
18	-0,0579		-0,0631		11,8808	[0,807]
19	-0,0944		-0,0160		12,7035	[0,809]
20	-0,0849		0,0106		13,3850	[0,818]

➢ Normalité des résidus :

GRAPHIQUE N°10 : Histogramme et QQ plot des résidus



Le test de normalité des résidus est fondé sur l'hypothèse nulle H_0 : les résidus sont distribués selon une loi normale. La p-value associé à ce test est de 0,021 comme visible sur le graphique n°10, cela signifie que nous rejetons H_0 au seuil de risque de 0,05% mais nous remarquons que sa valeur en est proche donc que la distribution des résidus est relativement proche d'une loi normale.

Cela se confirme sur le diagramme QQ-plot où l'on voit que les valeurs se situent autour de la droite d'Henry, sans être vraiment alignées sur cette dernière. Comme expliqué précédemment, malgré le fait que les résidus ne suivent pas un loi normale (H_0 refusée) c'est la valeur de la Q-stat qui est la plus importante ; la conclusion de cette partie sur la vérification des résidus est donc qu'ils sont 'bruit blanc', c'est à dire qu'aucune valeur n'est significative dans les fonctions d'autocorrélation, et qu'ils ne suivent pas une loi normale mais s'en approchent fortement.

Ainsi, le modèle ARIMA[1,1,0] est un modèle pertinent que l'on valide. Cependant, le taux d'erreur de prévision étant très élevé et les résidus ne suivant pas une loi normale, nous estimons 2 nouveaux modèles dont les informations se trouvent en annexes 6 et 7 pour voir s'il est possible d'améliorer sa pertinence afin d'avoir les meilleurs prévisions possibles. L'information des 3 modèles est résumée dans le tableau n°9.

TABLEAU N°9 : Récapitulatif des modèles estimés

	A : ARIMA(1,1,0)	**B : ARIMA(0,1,1)**	**C : ARIMA(1,1,1)**
Nombre de coefficients significatifs*	1/1	0/1	1/2
Valeur des coefficients estimés	$\widehat{\phi}_1$ =0,251	$\widehat{\theta}_1$ =0,175	$\widehat{\phi}_1$ =0,655 $\widehat{\theta}_1$ =−0,411
Ecart-type	0,236	0,238	0,231
P-value des résidus pour K=20	0,818	0,614	0,795
Prévision (% d'erreur)	138%	118%	170%
P-value test normalité	0,021<0,05	0,011<0,05	0,031<0,05
Conclusion	✓	✗	✗

* : au seuil de risque de 5% tel que α =0,05

On voit ainsi qu'il n'y a pas de meilleur modèle que le modèle A en se basant sur la règle de parcimonie, car c'est celui pour lequel tous les coefficients sont significatifs et la p-value des résidus est la plus élevée. Comme supposé précédemment, on voit que le modèle B fait de meilleurs prévisions puisque son taux d'erreur est de 118%, cela est dû en fait qu'il y ait une partie moyenne-mobile qui permet de mieux prévoir les effets d'un choc. En revanche les différences d'écart-type entre les modèles sont minimes donc on considère que du point de vue de ce critère, les 3 modèles sont pertinents. Finalement, les résidus qui s'approchent le plus de la loi normale sont ceux du modèle C, mais comme ce critère n'est pas essentiel à la validation d'un modèle nous ne le retenons pas dans le choix du meilleur modèle.

Ainsi, nous décidons de retenir le modèle A qui est le plus pertinent ; il vérifie les principales conditions, c'est donc celui que nous utiliserons dans la prévision du taux de chômage en t+1 soit à un horizon h=1.

→ Prévisions à une période

La prévisions s'effectue en plusieurs étapes, il convient dans un premier temps de définir Z_{T+1} à partir du modèle d'estimation de Zt (notre série Yt stationnarisée), puis de trouver $\widehat{Z_T}(1)$ pour finalement obtenir $Y_T(1)$ comme étant la prévision de Yt en t+1. La période que nous étudions s'étend du premier trimestre de 2004 au troisième de 2019, la prévision que nous effectuerons sera donc pour le dernier trimestre de 2019. Pour rappel, le modèle final que nous choisisson est de forme ARIMA[1,1,0], on a donc ; $Z_t = \phi_1 Z_{t-1} + a_t$.

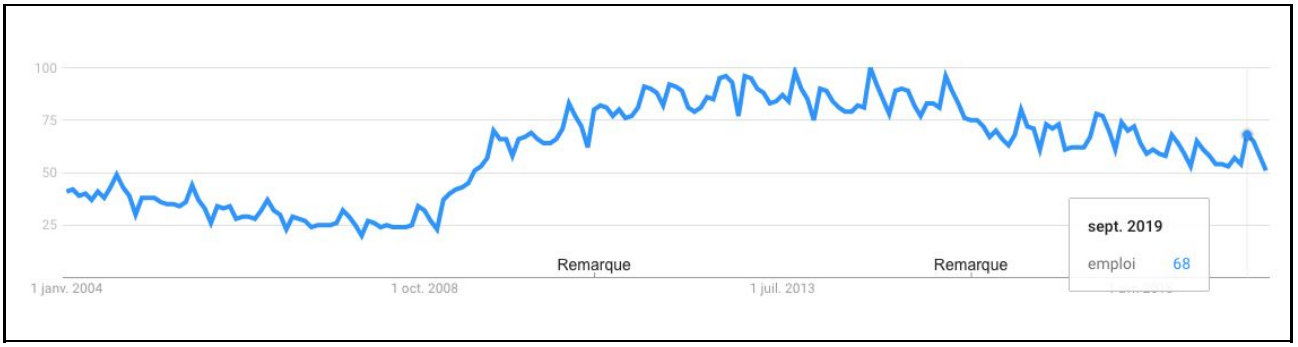
Nous cherchons les prévisions de notre série stationnarisée, pour cela nous avançons d'une période dans le temps pour trouver $Z_{T+1} = \phi_1 Z_T + a_{T+1}$. Sachant que les résidus $\{\widehat{a_t}\}$ se comportent comme un bruit blanc leur valeur est nulle, ils n'apparaitront donc pas dans l'estimation de Zt : $\widehat{Z_T}(1) = \widehat{\phi_1} * \widehat{Z_T}$. Les valeurs de $\widehat{Z_T}$ et Y_T sont disponibles en annexe n°8, on trouve ainsi $\widehat{Z_T}(1)$=0,2515*0,0879=0,0221. Nous avons donc la prévision de la série Zt, c'est à dire du taux de chômage différencié une fois, pour trouver l'estimation du 'vrai' taux de chômage il suffit d'additionner $\widehat{Z_T}(1)$ à sa dernière observation disponible, tel que $\widehat{Y_T}(1) = \widehat{Z_T}(1) + Y_T \Leftrightarrow \widehat{Y_T}(1) = 0,0221 + 8,54806 = 8,526\%$.

Ainsi à partir du modèle ARIMA[1,1,0] le taux de chômage s'élèverait à 8,526% au quatrième trimestre de 2019. En réalité ce dernier se situait à 8,1% selon l'INSEE[42], il a connu une forte de baisse entre le troisième et le quatrième trimestres de 2019, diminution qu'il était difficile de prévoir sachant que pour les trois premiers trimestres de 2019, le taux de chômage était compris entre 8,46 et 8,68%. Par une simple règle de trois nous trouvons un taux d'erreur de prévision de 105,26% ce qui est légèrement mieux que le pourcentage d'erreur proposé par le logiciel Gretl qui était de 138%.

Cette baisse soudaine du taux de chômage de 0,45% entre le troisième et quatrième trimestres de 2019 étaient difficiles à prévoir. Étant donné que nous nous intéressons à l'apport de Google Trends dans la prévision de l'emploi, nous pouvons regarder le graphique d'évolution de la popularité du mot 'emploi' de 2004 à 2019 en incluant cette fois le dernier trimestre de 2019, c'est à dire les 3 derniers mois de l'année.

[42] https://www.insee.fr/fr/statistiques/4309346 (consulté le 16/02/2020)

GRAPHIQUE N°11 : Évolution de la popularité du terme ‘emploi’ de 2004 à fin 2019



Sur le graphique n°11 on voit une nette chute des recherches de ce terme à partir de septembre 2019, cela laisse donc supposer que les recherches internet aident effectivement à prévoir le niveau d’emploi puisque la la baisse du taux de chômage au quatrième trimestre de l’année 2019 se voit dans la popularité de ses recherches qui diminuent elles aussi. En revanche il faut rester prudents quant à ce constat car comme précisé précédemment, il existe une certaine saisonnalité des recherches internet pour ce mot qui augmentent à chaque rentrée scolaire puis baissent à nouveau. Nous ne pouvons donc pas savoir s’il s’agit du caractère saisonnier de la variable Google Trends ou du réel reflet du marché du travail.

VI- Test de cointégration selon Engle-Granger

Le principe de cointégration apparaît dans les années 1980 grâce à l'économètre Clive William John Granger, puis la méthode se développe dans les années 90 avec la cointégration de plusieurs séries. Il s'agit de voir si celles-ci évoluent dans le même sens, ainsi, elles sont cointégrées si elles présentent les mêmes évolutions sur toute la période. Rappelons que chaque série temporelle est caractérisée par une composante déterministe et une composante aléatoire, le but de la cointégration est de voir au bout de combien de différenciations la composante déterministe disparaît pour qu'il ne reste que la partie stochastique, afin de voir si celle-ci peut être représentée par un modèle ARIMA (ou ARMA si la série est déjà stationnaire). Par conséquent on cherche à savoir si les erreurs peuvent être corrigées à court terme pour que les 2 séries trouvent une tendance commune à long terme, pour cela on regarde s'il y a une réduction ou un ajustement des erreurs par rapport à une tendance centrale.

Il y a 2 conditions à vérifier pour parler de séries cointégrées :

➢ <u>Même ordre d'intégration des 2 séries</u> : pour pouvoir cointégrer 2 séries il est impératif qu'elles soient intégrées du même ordre, c'est à dire qu'elles aient besoin du même nombre de différenciations pour être stationnaire. Pour vérifier la **stationnarité** d'une série il faut valider 3 conditions ; l'espérance de la série est indépendante du temps, la constante est finie et indépendante du temps, et la covariance entre Yt et Yt-k est une fonction finie de K périodes que l'on peut utiliser au niveau des prévisions et doit être elle aussi indépendante du temps.

➢ <u>La relation n'est pas fallacieuse</u> : il faut vérifier que la relation entre les 2 séries est pertinente, pour cela on utilise le test de causalité d'Engle-Granger. Une **relation fallacieuse** est caractérisée par un R^2 élevé, par une valeur de la statistique 't' élevée, enfin, la régression peut être acceptée d'un point de vue statistique mais pas d'un point de vue économique car les résidus ne sont pas stationnaires, donc la relation ne peut pas être interprétée.

Aussi, une relation n'est pas fallacieuse lorsque les variables sont cointégrées, pour tester la cointégration on doit être capable d'exprimer Y_t en fonction de X_t tel que $\widehat{Y}_t - \widehat{\alpha} - \widehat{\beta} X_t = \widehat{\varepsilon}_t$ où les résidus, notés $\widehat{\varepsilon}_t$, doivent être intégrés d'ordre 0 : $\varepsilon_t \sim I(0)$. En effet l'écart entre les 2 séries doit être constant à travers le temps pour parler de cointégration, on revient ainsi à la définition de la cointégration qui est la combinaison linéaire de 2 séries intégrées d'ordre inférieur.

Pour voir si l'on peut construire une combinaison linéaire entre 2 variables, c'est à dire si on peut les cointégrer, on procède par étape selon la méthodologie d'Engle-Granger :

1) On commence par vérifier que les 2 séries étudiées soient <u>intégrées du même ordre</u>, pour cela on a recours au test ADF dont la méthodologie a été développée en début de partie, ou au test KPSS dont l'hypothèse nulle est l'inverse de celle du test ADF c'est à dire H_0 : La variable est stationnaire (la racine unitaire n'existe pas). Dans notre cas, Y_t en niveau n'était pas stationnaire donc nous avons dû différencier la variable une fois. Par conséquent les tests appliqués à une de nos 4 variables explicatives doivent dire que les séries brutes sont non stationnaires, en revanche sur les séries différenciées 1 fois on doit retrouver une p-value< α =0,05 pour ADF et p-value> α =0,05 pour KPSS, de manière à valider la stationnarité de la variable X_t avec d=1.
2) Ensuite on effectue une <u>régression linéaire simple</u> pour vérifier la stationnarité des résidus de long terme, et sachant qu'à terme les séries évoluent autour d'une tendance centrale, on prend les variables brutes pour cette régression. On commence alors à vérifier la deuxième condition de la cointégration ; il s'agit de voir si la relation est fallacieuse ou non.
3) On cherche ainsi à <u>vérifier la relation</u> énoncée précédemment à savoir $\widehat{Y}_t - \widehat{\alpha} - \widehat{\beta} X_t = \widehat{\varepsilon}_t \sim I(0)$ c'est à dire voir si les résidus sont intégrés d'ordre 0. Pour cela on effectue le test de Dickey-Fuller Augmenté ou le test de KPSS sur les résidus sauvegardés de la régression linéaire simple estimée à l'étape précédente. Pour rappel si les séries en niveau sont stationnaires, c'est à dire qu'elles ne nécessitent pas de différenciation, alors les résidus peuvent être intégrés du même ordre à savoir d=0.
4) Enfin, pour vérifier la pertinence de la relation entre Y_t et X_t on s'intéresse à la relation de court terme en construisant un <u>modèle à correction d'erreur</u> (MCE). Celui-ci sert à voir l'influence des observations les unes par rapport aux autres, c'est pour cela qu'il

prend en compte les variables différenciées (car de toutes façons à long terme les séries se stabilisent autour de la moyenne). Pour se faire on utilise les résidus de la relation de long terme que l'on décale afin de voir s'il existe une correction entre les observations actuelles et celles qui précèdent. On obtient une relation de la forme suivante : $\Delta Y_t = \alpha_1 \Delta X + \delta(\widehat{Y_{t-1}} - Y_{t-1}) + \upsilon_t$ où δ correspond au coefficient de correction. Celui-ci doit être impérativement significatif et négatif pour qu'il y ait un retour de Y_t à sa valeur d'équilibre de long terme qui est déterminée par $\beta X_{t-1} + \alpha$. En effet, lorsque Y_{t-1} est supérieur à cette expression d'équilibre, il y aura une force de rappel vers l'équilibre de long terme uniquement si $\delta < 0$. Ainsi, le modèle à correction d'erreur permet de modéliser conjointement les dynamiques de court terme représentées par les variables différenciées une seule fois, et une dynamique de long terme représentée par les variables brutes. Enfin, le coefficient estimé de δ donne le pourcentage de déséquilibre entre les 2 séries qui sera corrigé chaque période (chaque trimestre dans notre cas), nous pouvons ainsi voir au bout de combien de trimestres les variables s'apparentent à 100%, c'est à dire où l'on retrouve une tendance commune.

Nous allons à présent appliquer cette méthode de cointégration sur nos 4 variables en commençant par la popularité du terme 'emploi' sous Google Trends.

→ Cointégration du taux de chômage avec la popularité du mot 'emploi' (X_1)

La première étape du processus vise à vérifier que l'ordre de différenciation est le même pour les 2 variables, si ce n'est pas le cas notre analyse s'arrête directement puisqu'il n'est pas possible de cointégrer des variables qui n'ont pas le même degré de différenciation. Nous procédons donc au test ADF sur la variable X_1 en niveau de manière à voir si elle est stationnaire.

TABLEAU N°10 : Résultats du test ADF pour la série X_1 brute

```
Test de Dickey-Fuller augmenté pour Popularite_emploi
testing down from 20 lags, criterion AIC
taille de l'échantillon 53
hypothèse nulle de racine unitaire : a = 1

  test avec constante
  avec 9 retards de (1-L)Popularite_emploi
  modèle: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0664661
  statistique de test: tau_c(1) = -2,44633
  p. critique asymptotique 0,1291
  Coeff. d'autocorrélation du 1er ordre pour e: -0,009
  différences retardées: F(9, 42) = 8,679 [0,0000]
```

On voit d'après le tableau n°10 que la p-value associée au test ADF est supérieure à 0,05 donc nous acceptons l'hypothèse nulle selon laquelle la série X_1 n'est pas stationnaire. Tout comme le taux de chômage, la variable issue de Google Trends n'est initialement pas stationnaire. Voyons à présent si elle l'est avec une différenciation, de manière à pouvoir cointégrer les 2 variables.

TABLEAU N°11 : Résultats du test KPSS pour la série X_1 différenciée 1 fois

```
Test KPSS pour d_Popularite_emploi

T = 62
Paramètre du délai de troncation = 3
Statistique de test = 0,280612

                        10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique > .10
```

D'après le tableau n°11 on remarque que la série différenciée une fois est effectivement stationnaire puisque la p-value du test KPSS (attention H_0 : X_1 est stationnaire, contrairement au test ADF) est supérieure au seuil de risque de 10% et donc celui de 5%, donc on accepte H_0. Nous pouvons ainsi continuer la cointégration entre ces 2 variables car elles respectent la première condition du processus de cointégration. Pour cela nous construisons la relation de long terme entre ces 2 variables en prenant les séries brutes, relation de laquelle nous sauvegardons les résidus de manière à vérifier leur stationnarité, première condition d'une relation non fallacieuse.

TABLEAU N°12 : Modèle de long terme et test ADF des résidus entre Y_t et X_1

```
Modèle 4: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                    coefficient   erreur std.   t de Student   p. critique
  const             6,92890       0,186245      37,20          1,24e-43   ***
  Popularite_emploi 0,0334344     0,00284635    11,75          2,65e-17   ***

Moy. var. dép.        8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus  15,97688   Éc. type de régression  0,511777
R2                    0,693433   R2 ajusté               0,688407
F(1, 61)              137,9778   p. critique (F)         2,65e-17
Log de vraisemblance -46,17539   Critère d'Akaike        96,35077
Critère de Schwarz    100,6370   Hannan-Quinn            98,03658
rho                   0,867049   Durbin-Watson           0,271032
```

```
Test de Dickey-Fuller augmenté pour uhat4
testing down from 10 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 2 retards de (1-L)uhat4
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,147837
  statistique de test: tau_nc(1) = -2,17041
  p. critique asymptotique 0,02886
  Coeff. d'autocorrélation du 1er ordre pour e: -0,009
  différences retardées: F(2, 57) = 2,948 [0,0605]
```

Ainsi, d'après le tableau n°12 on constate que les résidus de long terme sont bien stationnaires car la p-value du test ADF s'élève à 0,029 ce qui est inférieur à 0,05 et nous permet de rejeter l'hypothèse nulle de non stationnarité des résidus. Nous avons donc

$\widehat{\varepsilon_t} \sim I(0)$ ce qui nous permet de continuer notre démarche de cointégration de la variable issue de l'outil 'Google Trends' avec le taux de chômage français, en estimant cette fois la relation de court terme, c'est à dire avec les variables différenciées une fois auxquelles on ajoute les résidus de long terme retardés d'une période.

TABLEAU N°13 : Modèle de court terme (MCE) entre Y_t et X_1

```
Modèle 5: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                      coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------------
  const               -0,00205472   0,0287255      -0,07153       0,9432
  d_Popularite_emp~    0,0110799    0,00495600      2,236         0,0292     **
  uhat4_1             -0,153064     0,0569901      -2,686         0,0094     ***

Moy. var. dép.          0,000482   Éc. type var. dép.       0,245735
Somme carrés résidus    3,008157   Éc. type de régression   0,225800
R2                      0,183348   R2 ajusté                0,155665
F(2, 59)                6,623114   p. critique (F)          0,002541
Log de vraisemblance    5,825819   Critère d'Akaike        -5,651637
Critère de Schwarz      0,729766   Hannan-Quinn            -3,146137
rho                     0,199650   Durbin-Watson            1,598166
```

À partir du tableau n°13 qui nous donne les informations du modèle à correction d'erreur, on voit que le coefficient de correction est significatif et négatif puisque $\widehat{\delta}$ =-0,153. Ainsi la relation qui lit la popularité du mot 'emploi' au taux de chômage n'est pas fallacieuse ; en effet les résidus de long terme sont stationnaires et Y_t retourne bien à son équilibre de long terme. Ainsi, la relation de court terme des variables 'taux de chômage' et 'popularité du mot emploi' nous informe que chaque trimestre, les 2 séries convergent de 15% environ -elles s'apparenteront donc à 100% au bout de 7 périodes soit près de 2 ans.

Il est possible d'interpréter la relation de long terme qui lit Y_t et X_1 donnée en tableau n°12. Aussi nous voyons que le coefficient estimé de X_1 est significatif, l'outil de Google Trends est donc pertinent pour expliquer les variations du taux de chômage, et est positif ce qui confirme la relation théorique qui lie ces 2 variables et la relation empirique que nous avons pu analyser en partie III à l'aide des graphiques de corrélation. Par conséquent, lorsque la popularité du mot 'emploi' augmente d'un point le taux de chômage augmentera de 0,033%. La cointégration des 2 premières variables a été concluante, voyons ce qu'il en est des 3 autres variables explicatives : le taux d'intérêt, la production industrielle et la population active.

→ Cointégration du taux de chômage avec le taux d'intérêt (X_2)

Nous procédons à la même analyse pour la deuxième variable explicative du taux de chômage ; le taux d'intérêt. Les résultats du test ADF disponibles en tableau n°12 ci-dessous indiquent que la série brute n'est pas stationnaire, en revanche elle l'est après une différenciation puisque la p-value est de 0 ce qui nous permet de rejeter H_0 ; la série X_2 est donc stationnaire après une différenciation.

TABLEAU N°14 : Résultats du test ADF pour la série X_2 brute et différenciée 1 fois

```
Test de Dickey-Fuller augmenté pour Taux_interet
testing down from 20 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 2 retards de (1-L)Taux_interet
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0210602
  statistique de test: tau_nc(1) = -1,74787
  p. critique asymptotique 0,07643
  Coeff. d'autocorrélation du 1er ordre pour e: 0,015
  différences retardées: F(2, 57) = 4,040 [0,0229]
```

```
Test de Dickey-Fuller augmenté pour d_Taux_interet
testing down from 20 lags, criterion AIC
taille de l'échantillon 61
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)d_Taux_interet
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): -0,740098
  statistique de test: tau_nc(1) = -5,80797
  p. critique 3,243e-08
  Coeff. d'autocorrélation du 1er ordre pour e: 0,073
```

Étant donné que les 2 séries sont intégrées du même ordre nous pouvons passer à l'analyse des résidus de long terme en commençant par estimer un modèle en régression linéaire simple composée seulement de ces 2 séries, comme pour la cointégration précédente. La sortie du modèle sous Gretl est disponible en annexe n°9, après sauvegarde des résidus on teste leur stationnarité à l'aide du test KPS. Également disponible en annexe n°9 la sortie de ce test indique que les résidus associés au modèle de long terme sont stationnaires puisque la valeur de la statistique "p" est supérieure au seuil de significativité fixé à 5%. Ici encore nous pouvons poursuivre les démarches de cointégration du taux de chômage avec le taux d'intérêt.

TABLEAU N°15 : Modèle de court terme (MCE) entre Y_t et X_2

```
Modèle 8: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                 coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const          -0,00680439   0,0325114     -0,2093        0,8349
  d_Taux_interet -0,112544     0,117268      -0,9597        0,3411
  uhat7_1        -0,0219267    0,0508102     -0,4315        0,6676

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    3,618237   Éc. type de régression  0,247641
R2                      0,017724   R2 ajusté              -0,015573
F(2, 59)                0,532308   p. critique (F)         0,590042
Log de vraisemblance    0,101383   Critère d'Akaike        5,797234
Critère de Schwarz      12,17864   Hannan-Quinn            8,302734
rho                     0,228340   Durbin-Watson           1,540318
```

La relation entre le taux de chômage et le taux d'intérêt est fallacieuse, elle n'est pas pertinente puisque le coefficient de correction associé au modèle de correction d'erreur est non significatif (confère tableau n°15). La cointégration de ces 2 séries n'est donc pas possible car la relation de court terme qui les lie est fallacieuse, il n'est ainsi pas possible d'établir une relation de long terme entre ces dernières.

→ Cointégration du taux de chômage avec la production industrielle (X_3)

Les résultats des tests pour la variable "production industrielle" sont les mêmes que ceux de la variable Google Trends ; X_3 est intégré d'ordre 1 comme le taux de chômage, donc il est possible de commencer les démarches de cointégration en passant par la régression linéaire simple. Comme pour les variables X_1 et X_2 les résidus sont stationnaires, c'est à dire intégrés d'ordre 0 (I[0]) car la p-value du test KPSS est supérieure à α=0,05. Enfin, lorsque l'on estime la relation de court terme par le MCE il apparaît que delta estimé est négatif et significatif, nous sommes dans le cas d'une relation non fallacieuse qu'il est donc possible d'interpréter. Le détail de cette analyse est disponible en annexe n°10. On voit ainsi que les séries du taux de chômage et de la production industrielle qui reflète l'activité économique convergent de 11% par trimestre ($\widehat{\delta}$=-0,1089), elles s'apparentent totalement au bout de 10 périodes ($\frac{100}{10,89}$ = 9,18) ce qui correspond à 2 ans et demi. Aussi, le coefficient estimé de X_3 à long terme s'élève à −0,106 : lorsque la production industrielle augmente d'une unité par rapport à la base 100 en 2015, le taux de chômage diminue de 0,11 points de pourcentage.

→ Cointégration du taux de chômage avec la population active (X_4)

Finalement, nous terminons cette partie en cointégrant la dernière variable X_4 avec le taux de chômage. D'après le tableau n°16 on voit que les 2 séries ont le même ordre d'intégration d=1 puisque X_4 est stationnaire lorsqu'il est différencié une fois. De plus, on voit que les résidus sauvegardés de la régression linéaire simple de long terme entre les 2 séries sont (enfin) stationnaires puisque la p-value du test ADF est de 0,022 ce qui nous permet de rejeter l'hypothèse nulle selon laquelle les résidus ne sont pas stationnaires.

TABLEAU N°16 : Développement démarche de cointégration de Y_t avec la population active

```
Test de Dickey-Fuller augmenté pour Population_active
testing down from 20 lags, criterion AIC
taille de l'échantillon 62
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)Population_active
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): 0,00141609
  statistique de test: tau_nc(1) = 2,71892
  p. critique 0,9982
  Coeff. d'autocorrélation du 1er ordre pour e: -0,212
```

```
Test de Dickey-Fuller augmenté pour d_Population_active
testing down from 20 lags, criterion AIC
taille de l'échantillon 61
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 0 retards de (1-L)d_Population_active
  modèle: (1-L)y = (a-1)*y(-1) + e
  valeur estimée de (a - 1): -1,06963
  statistique de test: tau_nc(1) = -8,25518
  p. critique 7,336e-21
  Coeff. d'autocorrélation du 1er ordre pour e: 0,007
```

```
Modèle 19: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                     coefficient   erreur std.  t de Student  p. critique
  ---------------------------------------------------------------------------
  const              -9,42918      2,84188        -3,318       0,0015     ***
  Population_active   0,000645274  9,95562e-05     6,482       1,79e-08   ***

Moy. var. dép.          8,981320   Éc. type var. dép.      0,916827
Somme carrés résidus    30,86155   Éc. type de régression  0,711285
R2                      0,407824   R2 ajusté               0,398116
F(1, 61)                42,00990   p. critique (F)         1,79e-08
Log de vraisemblance   -66,91398   Critère d'Akaike        137,8280
Critère de Schwarz      142,1142   Hannan-Quinn            139,5138
rho                     0,960119   Durbin-Watson           0,106989
```

```
Test de Dickey-Fuller augmenté pour uhat19
testing down from 20 lags, criterion AIC
taille de l'échantillon 60
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 2 retards de (1-L)uhat19
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,0948566
  statistique de test: tau_nc(1) = -2,2761
  p. critique asymptotique 0,02205
  Coeff. d'autocorrélation du 1er ordre pour e: -0,037
  différences retardées: F(2, 57) = 7,810 [0,0010]
```

Le fait que les résidus soient intégrés d'ordre 0 nous permet de valider la première condition pour que la relation entre Y_t et X_4 ne soit pas fallacieuse, on peut maintenant construire le modèle à correction d'erreur noté MCE qui modélise la relation de court terme qu'il existe entre le taux de chômage et la population active.

TABLEAU N°17 : MCE pour Y_t et X_4

```
Modèle 20: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                      coefficient   erreur std.  t de Student  p. critique
  ---------------------------------------------------------------------------
  const               -0,0293921    0,0314125     -0,9357       0,3533
  d_Population_act~    0,000746067  0,000257412    2,898        0,0053     ***
  uhat19_1            -0,0409446    0,0431239     -0,9495       0,3463

Moy. var. dép.          0,000482   Éc. type var. dép.      0,245735
Somme carrés résidus    3,206684   Éc. type de régression  0,233132
R2                      0,129453   R2 ajusté               0,099942
F(2, 59)                4,386723   p. critique (F)         0,016745
Log de vraisemblance    3,844620   Critère d'Akaike       -1,689240
Critère de Schwarz      4,692163   Hannan-Quinn            0,816260
rho                     0,323455   Durbin-Watson           1,347502
```

Le tableau n°17 contient l'estimation du modèle à correction d'erreur de court terme englobant les résidus de long terme retardés d'une période. On voit que le coefficient de correction est bien négatif puisque $\widehat{\delta}$ =-0,041, en revanche il n'est pas significatif donc nous ne pouvons pas interpréter les résultats de la régression linéaire de long terme puisque la relation entre Y_t et X_4 est fallacieuse. Ainsi, les variables 'taux de chômage' et 'population

active' sont cointégrées (avec d=1 et $\widehat{\varepsilon}_t \sim I(0)$) mais leur relation est fallacieuse donc la cointégration de ces 2 séries n'est pas concluante.

Pour finir, le tableau n°18 résume les démarches de cointégration des 4 variables explicatives avec le taux de chômage, deux des 4 cointégrations ont été possibles. Parmis celles qui n'ont pas abouti cela était lié au non respect des 2 conditions à vérifier impérativement pour pouvoir parler de séries cointégrées comme énoncées antérieurement à savoir ; même ordre d'intégration des 2 séries et relation non fallacieuse.

TABLEAU N°18 : Récapitulatif des processus de cointégration des variables

Variables	**Ordre d'intégration**	**Stationnarité des résidus**	**Coefficient de correction CT**	**Relation fallacieuse**
Popularité du mot emploi et Y_t	d=1	p-value test ADF=0,029	$\widehat{\delta}$ =-0,041 et significatif	Non
Taux d'intérêt et Y_t	d=1	p-value test KPSS>0,05	$\widehat{\delta}$ =-0,022 mais non significatif	Oui
Production industrielle et Y_t	d=1	p-value test KPSS>0,05	$\widehat{\delta}$ =-0,109 et significatif	Non
Population active et Y_t	d=1	p-value test ADF=0,0221	$\widehat{\delta}$ =-0,041 mais non significatif	Oui

: test KPSS (H_0 : X_t non stationnaire)

: test ADF (H_0 : X_t stationnaire)

VII- Conclusion

De plus en plus présent dans notre société, le Big Data est devenu un facteur clé pour obtenir des informations pouvant aider l'analyse économique. Ainsi, dans cette étude nous avons essayé de prévoir le chômage et l'emploi à l'aide de l'outil Google Trends ainsi que d'autres variables économiques. La difficulté de l'utilisation de tels outils est de trouver quel mot représente le mieux la réalité de recherches internet des chercheurs d'emplois, nous avons donc regarder l'évolution de la popularité relative de 3 termes différents et avons choisi le mot 'emploi' comme variable prédictive du taux de chômage en France de 2004 à 2019.

Pour l'analyse de ce dernier nous avons utilisés la méthodologie ARIMA en suivant l'approche de Box-Jenkins pour déterminer le meilleur modèle qui puisse estimer les coefficients de manière valable. Aussi, après avoir différencié le taux de chômage qui était initialement non stationnaire nous avons trouvé un modèle ARIMA[1,1,0] bien spécifié et dont les résidus sont "bruit blanc" c'est à dire nuls. En utilisant ce modèle nous avons pu faire des prévisions du taux de chômage à horizon 1 pour un taux d'erreur de 105% (prévu ; 8,53%, réalisé : 8,1%). Enfin, nous avons tenté de cointégrer Y_t avec les 4 variables explicatives, les variables "popularité du mot 'emploi'" et "production industrielle" sont co-intégrégrées mais les 2 autres (taux d'intérêt et population active) ne le sont pas parce que la relation qui les liait au taux de chômage était fallacieuse.

L'utilisation de 'Google Trends' sur notre échantillon est donc concluante, sa cointégration avec Y_t prouve qu'elle peut aider à prévoir le chômage et l'emploi et constitue ainsi un outil convenant car disponible gratuitement et en temps réel. Revenons dans une dernière partie sur les limites de cet outil malgré ses avantages incontournables et sur les raisons qui peuvent expliquer la non intégration des 2 autres séries temporelles.

VIII- Discussion

Le taux d'intérêt et la population active sont, dans la littérature, de bons indicateurs dans l'explication du taux de chômage. Quelles raisons économiques peuvent expliquer les conclusions de notre étude pour ces 2 variables ? Malgré ses apparences attractives, l'outil Google Trends par son jeune âge présente quelques limites, quelles sont-elles ?

Premièrement, la période étudiée s'étend de 2004 à nos jours, elle prend donc en compte la crise des Subprimes qui est venue bouleverser l'équilibre économique et donc créer des évolutions de grande variance qu'il est difficile d'analyser par la suite. Le principal problème que nous avons rencontrer dans cette étude était celui de la cointégration des 4 séries avec le taux de chômage. Celle-ci n'a pas été possible et ce peut être lié à l'atypicité de la période étudiée pour laquelle les fluctuations des indicateurs étaient inhabituelles ce qui explique que les séries ne convergent pas.

Dans un second temps on peut s'intéresser aux limites de l'indicateur lui-même, en effet l'outil du géant Google est disponible en temps réel et donc pertinent pour compléter les indicateurs économiques traditionnels. Mais les données qu'il fournit sont brutes, elles ne sont corrigées d'aucun effet saisonnier et ne sont ni stabilisées ni débarrassées d'éventuelles valeurs atypiques. La pérennité des résultats de cet outil est encore trop fragile, c'est pourquoi il faut savoir l'utiliser avec prudence.

De plus, bien que les recherches Google représentent 95% des recherches françaises et 90%[43] des recherches mondiales, il y a un manque à gagner lié au développement des applications. En effet, les internautes accèdent de plus en plus à leurs requêtes en ayant recours aux applications pour réaliser leurs achats par exemple (les marques commencent à développer leurs propres applications maintenant). Finalement, une hausse de la tendance de recherche n'est pas toujours la conséquence d'une hausse des recherches en volume mais à une migration de la population qui fait baisser le nombre d'utilisateurs et peut ainsi fausser les résultats de l'outil Google Trends.

Pour conclure cette étude on peut dire que Google Trends est un bon outil pour prédire le taux de chômage français de par ses nombreux avantages, mais qui doit être amélioré pour une utilisation officielle comme indicateur économique à part entière.

[43] https://www.planetoscope.com/Internet-/1474-recherches-sur-google.html (consluté le 17/02/2020)

IX- Bibliographie

Andrieu O., “Google Insights for Search disponible en français”, *Abondance*, 19/09/2019.
https://www.abondance.com/20090819-10009-google-insights-for-search-disponible-en-francais.html

Banque Mondiale Données, “Industrie, valeur ajoutée (% du PIB)”, consultable en ligne :
https://donnees.banquemondiale.org/indicateur/NV.IND.TOTL.ZS?view=chart

Bremme L., “Définition : Qu’est-ce que le Big Data ?”, *Le Big Data*, 07/2016.
https://www.lebigdata.fr/definition-big-data

Brief Eco, “Les déterminants du chômage”, 31/10/2018.
https://www.brief.eco/a/2018/10/31/on-fait-le-point/les-determinants-du-chomage/

Charpentier A., “Cours de séries temporelles”, *ENSAE Paris Dauphine*, pp.6-15.
https://www.math.u-bordeaux.fr/~hzhang/m2/st/TS1.pdf

Etner J. et Le Maitre P., “L'impact du taux d'intérêt sur l'évolution simultanée du chômage et de l'épargne”, *Persée*, 1999, pp. 917-935.
https://www.persee.fr/doc/reco_0035-2764_1999_num_50_5_410125

Garbay A, “Les Français cherchent un emploi sur Internet mais le trouvent grâce à leur réseau”, *Le Figaro*, 17/01/2017.
https://www.lefigaro.fr/emploi/2017/01/17/09005-20170117ARTFIG00290-les-francais-cherchent-un-emploi-sur-internet-mais-le-trouvent-grace-a-leur-reseau.php

Google Trends consultable en ligne :
https://trends.google.fr/trends/?geo=FR

Hellier J., “Économie du travail”, *Research Gate*, 01/2018.
https://www.researchgate.net/publication/322540093_Economie_du_Travail_Master_1_Doss

ier

INSEE, *Enquête Emploi*, 13/02/2020, n°36.
https://www.insee.fr/fr/statistiques/4309346

INSEE, définition des enquêtes de conjoncture, consultable en ligne :
https://www.insee.fr/fr/metadonnees/definition/c1422

INSEE, définition des intérêts, consultable en ligne :
https://www.insee.fr/fr/metadonnees/definition/c1287

INSEE, définition du taux de chômage, consultable en ligne :
https://www.insee.fr/fr/metadonnees/definition/c1687

Inwin - Digital Expert, "Un outil puissant et gratuit mais méconnu Google Trends", 11/03/2019.
https://www.inwin.fr/blog/un-outil-puissant-et-gratuit-mais-meconnu-google-trends/

Journal du net, "Google Trends (ex Google Insight) : définition", mis à jour le 09/01/2019.
https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203505-google-trends-ex-google-insight-definition/

Journal du Net, "Nombre d'internautes en France", mis à jour le 14/02/2014.
https://www.journaldunet.com/ebusiness/le-net/1071394-nombre-d-internautes-en-france/

Lambert M., "Comment passer de l'idée à la stratégie grâce à Google Trends", *La Tranchée*, 11/10/2018.
https://www.latranchee.com/comment-passer-de-lidee-a-la-strategie-grace-a-google-trends/

Le Figaro, "La BCE maintient ses taux directeurs au plus bas", 12/12/2019.
https://www.lefigaro.fr/conjoncture/la-bce-maintient-ses-taux-directeurs-au-plus-bas-1-20191212

Levasseur S., “Vieillissement de la population”, *Revues de l’OCFE*, 2015, pp. 339-370.
https://www.cairn.info/revue-de-l-ofce-2015-6-page-339.htm#

L’Express, “Taux de chômage à 8% en 2007”, 06/03/2008.
https://lexpansion.lexpress.fr/actualite-economique/taux-de-chomage-a-8-en-2007_470864.html

Moyou E., “Part des ménages ayant un accès internet en France de 2006 à 2018”, Statista, 08/01/2020.
https://fr.statista.com/statistiques/509227/menage-francais-acces-internet/

OCDE consultable en ligne :
https://data.oecd.org/fr/

Ouest France, “Le taux de chômage au plus bas depuis 10 ans”, 27/12/2019.
https://www.ouest-france.fr/economie/emploi/chomage/le-taux-de-chomage-au-plus-bas-depuis-dix-ans-6671745

Tavernier J-L., “Note de conjoncture”, *INSEE*, 03/2015, pp.43-56.
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&ved=2ahUKEwi2nLu42ojnAhWJzYUKHUuRBlEQFjAIegQICRAB&url=https%3A%2F%2Fwww.insee.fr%2Ffr%2Fstatistiques%2Ffichier%2F1408926%2Fmars2015_d2.pdf&usg=AOvVaw0Kee7rCU0qt0_SnakRGtid

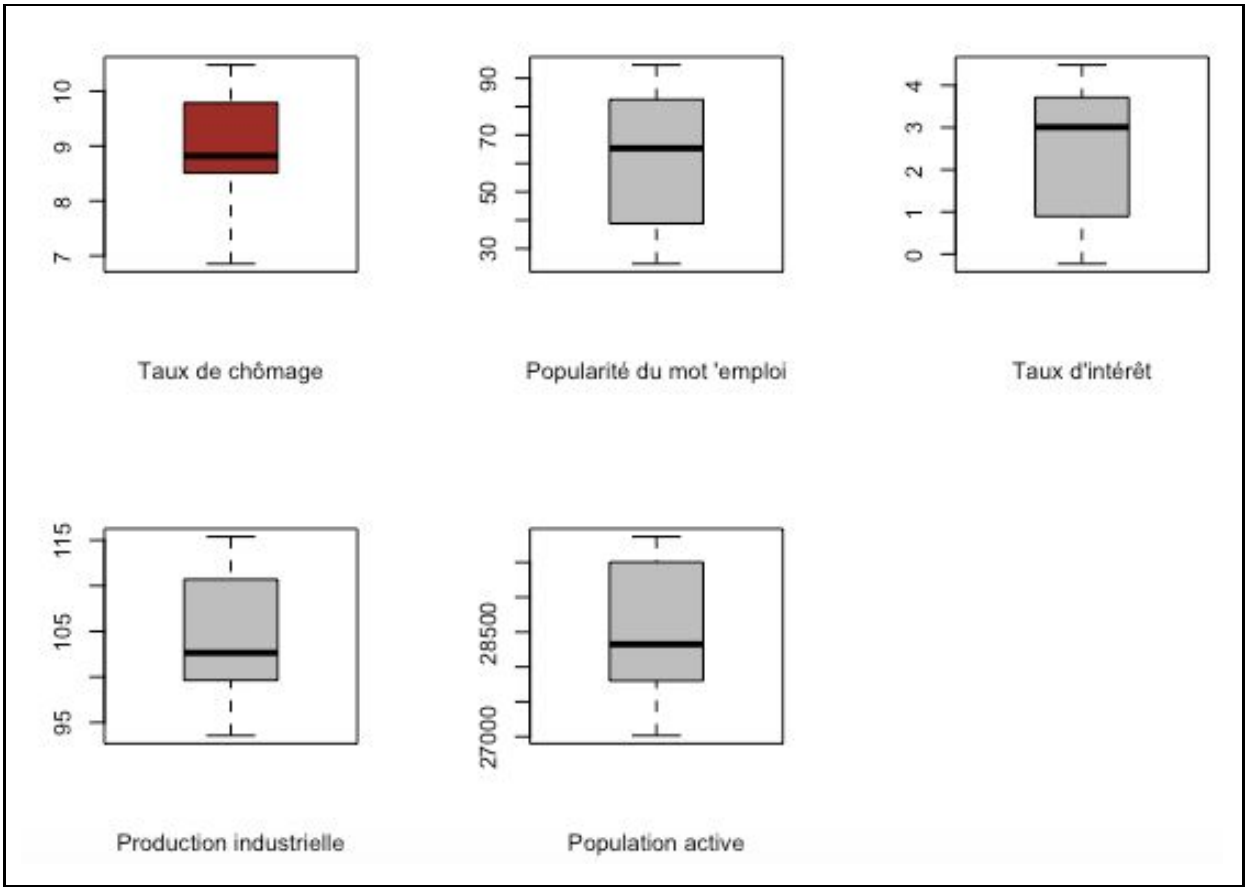
Tsay R. S., “Analysis of financial times series”, *University of Chicago*, 2002, pp.28-53.

Vaté M., “Statistiques chronologiques et prévisions”, *Economica*, 1993, pp.217-225.

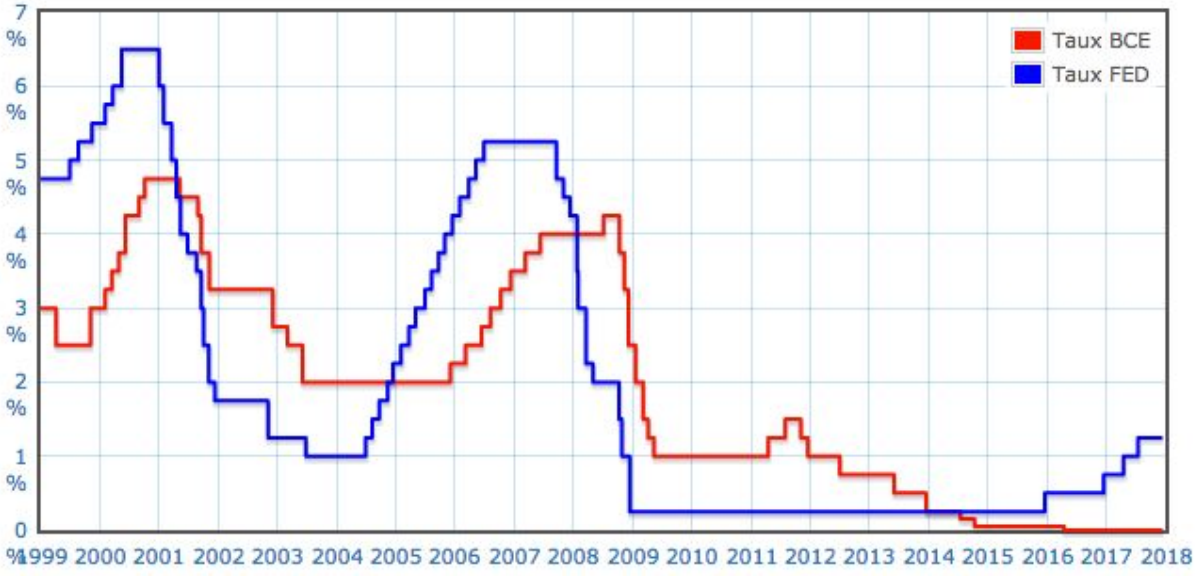
Yassine A., “Google Trends : popularité ou volume de recherche d’un terme, de quoi parle-t-on ?”, *Ya-Graphic*, 04/07/2016.
https://www.ya-graphic.com/google-trends-popularite-volume-de-recherche/

X- Annexes

Annexe n°1 : Boxplots des 5 variables réalisés sous R studio

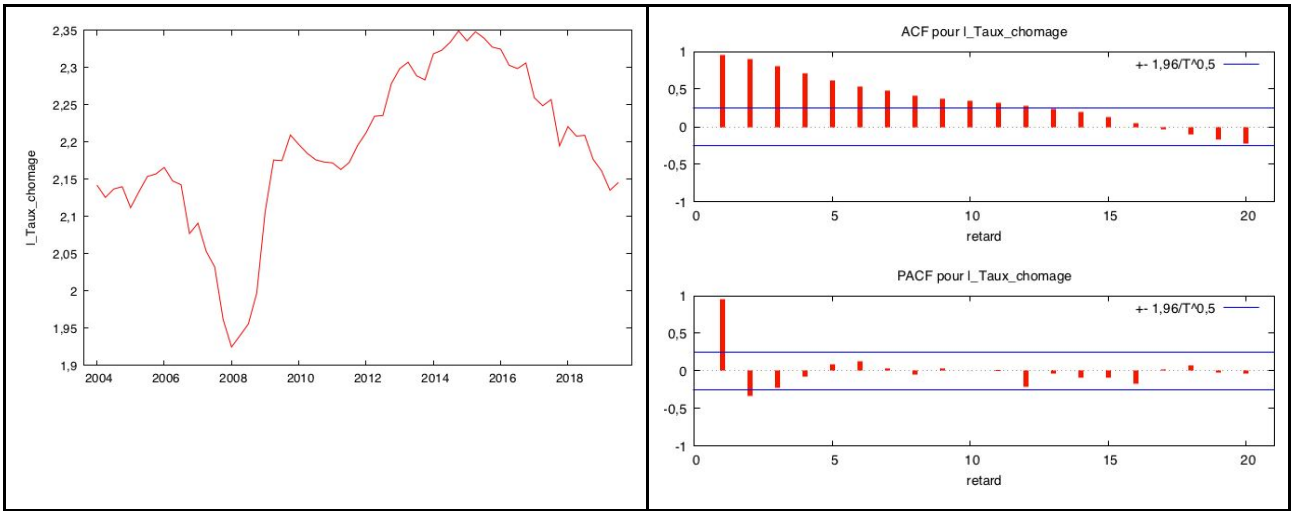


Annexe n°2 : Évolution des taux directeurs de la FED et de la BCE depuis janvier 1999



Source : https://france-inflation.com/taux-directeurs-bce-fed.php (consulté le 07/02/2020

Annexe n°3 : Évolution temporelle et corrélogramme du logarithme du taux de chômage

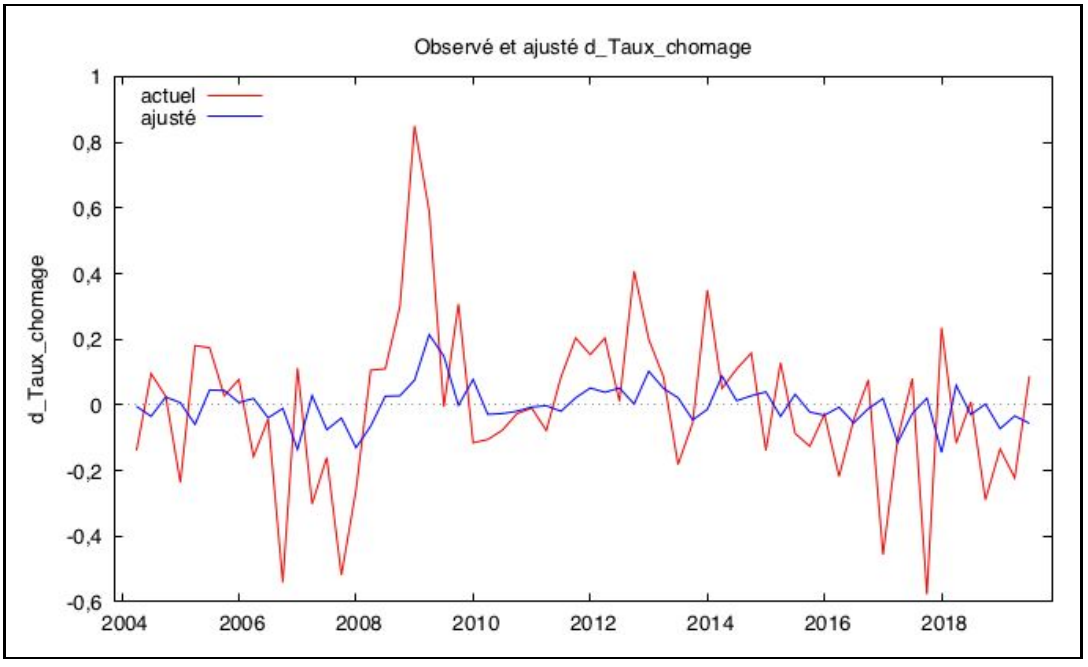


Annexe n°4 : Fonctions d'auto-corrélation du taux de chômage différencié une fois

Fonction d'auto-corrélation pour d_Taux_chomage
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5

RETARD	ACF	PACF	Q	[p. crit.]
1	**0,2537** **	**0,2537** **	4,1855	[0,041]
2	**0,2916** **	**0,2428** *	9,8070	[0,007]
3	0,1770	0,0674	11,9131	[0,008]

Annexe n°5 : Graphique des valeurs prédites et des valeurs observées du modèle ARIMA[1,1,0]



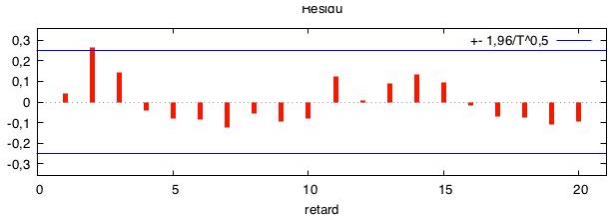
Annexe n°6 : Détails du modèle B : ARIMA(0,1,1)

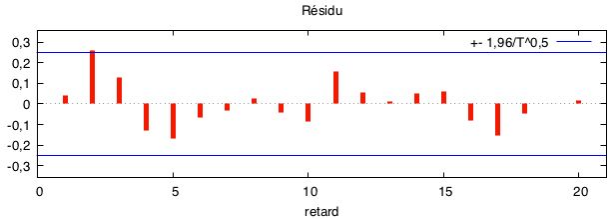
```
Évaluations de la fonction : 28
Évaluations du gradient : 10

Modèle 8: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z      p. critique
  ---------------------------------------------------------------
  const      0,000397620   0,0354865    0,01120    0,9911
  theta_1    0,175201      0,105315     1,664      0,0962     *

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000393
  Ec. type des innovations   0,238335
  R2                         0,045743
  R2 ajusté                  0,045743
  Log de vraisemblance       0,922932
  Critère d'Akaike           4,154135
  Critère de Schwarz         10,53554
  Hannan-Quinn               6,659635

                        Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  MA
  Racine  1          -5,7077     0,0000     5,7077     0,5000
  ---------------------------------------------------------------
```

```
Évaluations de la fonction : 17
Évaluations du gradient : 7

Modèle 9: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.     z     p. critique
  ---------------------------------------------------------------
  theta_1    0,175202      0,105314     1,664     0,0962     *

  Moy. var. dép.             0,000482
  Éc. type var. dép.         0,245735
  Moyenne des innovations    0,000732
  Ec. type des innovations   0,238336
  R2                         0,045749
  R2 ajusté                  0,045749
  Log de vraisemblance       0,922870
  Critère d'Akaike           2,154261
  Critère de Schwarz         6,408529
  Hannan-Quinn               3,824594

                        Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  MA
  Racine  1          -5,7077     0,0000     5,7077     0,5000
  ---------------------------------------------------------------
```

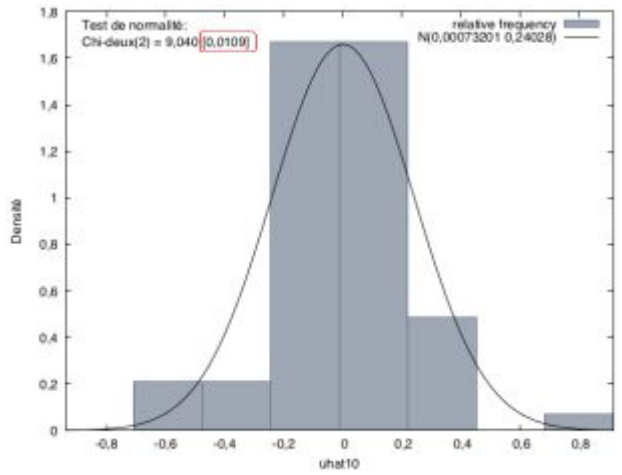
```
2018:1      0,235416    -0,103681     0,339097
2018:2     -0,116902     0,059411    -0,176313
2018:3      0,009221    -0,030890     0,040111
2018:4     -0,289367     0,007028    -0,296395
2019:1     -0,134507    -0,051929    -0,082578
2019:2     -0,222715    -0,014468    -0,208247
2019:3      0,087860    -0,036485     0,124345

Note : * indique un résidu supérieur à 2,5 fois l'écart type
  Statistiques d'évaluation des prédictions using 62 observations

  Erreur Moyenne                              0,00073201
  Racine de la moyenne des erreurs au carré   0,23834
  Erreur absolue moyenne                      0,17545
  Mean Percentage Error                       111,31
  Mean Absolute Percentage Error              118,34
  U de Theil                                  1,018
```

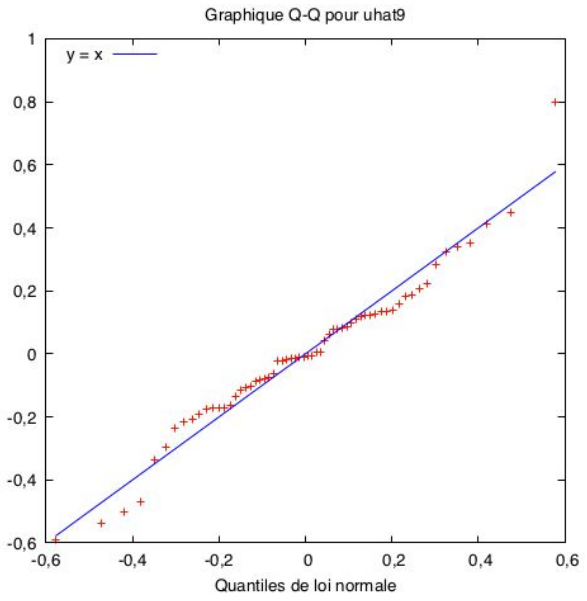
On estime le modèle ARIMA(0,1,1) une première fois, puis un seconde sans la constante qui n'est pas significative. On regarde ensuite l'erreur de prévision, puis on vérifie ce modèle par les 2 approches d'analyse des résidus. Enfin, on regarde la normalité de ces derniers par l'histogramme et le diagramme de distribution.






```
Fonction d'auto-corrélation résiduelle
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5

 RETARD   ACF       PACF          Q   [p. crit.]

   1    0,0433      0,0433
   2    0,2632  **  0,2618 **     4,7031  [0,030]
   3    0,1412      0,1306         6,0432  [0,049]
   4   -0,0431     -0,1272         6,1703  [0,104]
   5   -0,0826     -0,1688         6,6452  [0,156]
   6   -0,0855     -0,0662         7,1636  [0,209]
   7   -0,1224     -0,0328         8,2445  [0,221]
   8   -0,0540      0,0251         8,4590  [0,294]
   9   -0,0929     -0,0436         9,1046  [0,334]
  10   -0,0822     -0,0830         9,6195  [0,382]
  11    0,1245      0,1559        10,8249  [0,371]
  12    0,0053      0,0558        10,8271  [0,458]
  13    0,0882      0,0115        11,4565  [0,490]
  14    0,1353      0,0489        12,9701  [0,450]
  15    0,0960      0,0595        13,7486  [0,469]
  16   -0,0149     -0,0822        13,7678  [0,543]
  17   -0,0692     -0,1534        14,1898  [0,585]
  18   -0,0741     -0,0485        14,6844  [0,618]
  19   -0,1081     -0,0001        15,7633  [0,609]
  20   -0,0969      0,0179        16,6508  [0,614]
```





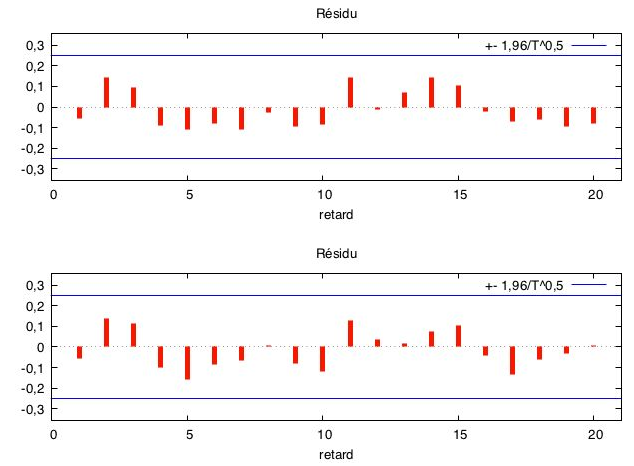
Annexe n°7 : Détails du modèle C : ARIMA(1,1,1)

```
Évaluations de la fonction : 34
Évaluations du gradient : 13

Modèle 16: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z       p. critique
  ---------------------------------------------------------------
  const      -0,00120802   0,0492284    -0,02454    0,9804
  phi_1       0,654963     0,189138      3,463      0,0005     ***
  theta_1    -0,411393     0,211810     -1,942      0,0521     *

  Moy. var. dép.              0,000482
  Éc. type var. dép.          0,245735
  Moyenne des innovations     0,000898
  Ec. type des innovations    0,231424
  R2                          0,098595
  R2 ajusté                   0,083571
  Log de vraisemblance        2,704093
  Critère d'Akaike            2,591814
  Critère de Schwarz          11,10035
  Hannan-Quinn                5,932481

                        Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  AR
  Racine  1           1,5268     0,0000     1,5268     0,0000
  MA
  Racine  1           2,4308     0,0000     2,4308     0,0000
  ---------------------------------------------------------------
```

```
Évaluations de la fonction : 29
Évaluations du gradient : 11

Modèle 15: ARMA, utilisant les observations 2004:2-2019:3 (T = 62)
Estimated using AS 197 (MV exacte)
Variable dépendante: d_Taux_chomage
Écarts type basés sur la matrice hessienne

             coefficient   erreur std.      z       p. critique
  ---------------------------------------------------------------
  phi_1       0,654820     0,189073      3,463      0,0005     ***
  theta_1    -0,411256     0,211741     -1,942      0,0521     *

  Moy. var. dép.              0,000482
  Éc. type var. dép.          0,245735
  Moyenne des innovations     0,000179
  Ec. type des innovations    0,231425
  R2                          0,098572
  R2 ajusté                   0,083548
  Log de vraisemblance        2,703792
  Critère d'Akaike            0,592416
  Critère de Schwarz          6,973820
  Hannan-Quinn                3,097916

                        Réel  Imaginaire   Modulo   Fréquence
  ---------------------------------------------------------------
  AR
  Racine  1           1,5271     0,0000     1,5271     0,0000
  MA
  Racine  1           2,4316     0,0000     2,4316     0,0000
  ---------------------------------------------------------------
```

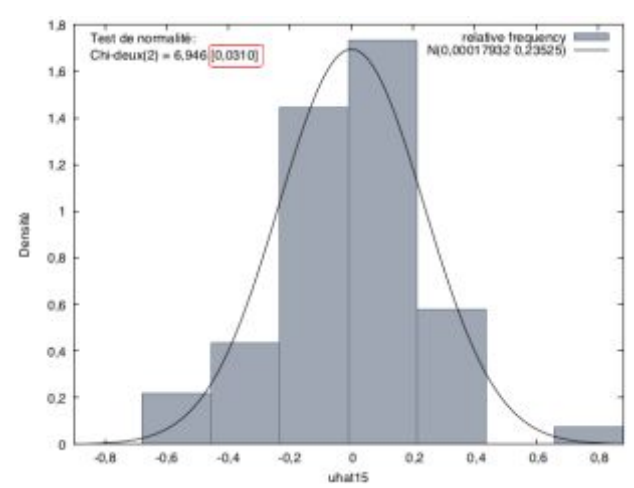
```
2018:1       0,235416    -0,144448     0,379864
2018:2      -0,116902    -0,002066    -0,114836
2018:3       0,009221    -0,029323     0,038544
2018:4      -0,289367    -0,009813    -0,279554
2019:1      -0,134507    -0,074515    -0,059992
2019:2      -0,222715    -0,063406    -0,159309
2019:3       0,087860    -0,080322     0,168182

Note : * indique un résidu supérieur à 2,5 fois l'écart type
  Statistiques d'évaluation des prédictions using 62 observations

  Erreur Moyenne                              0,00017932
  Racine de la moyenne des erreurs au carré   0,23143
  Erreur absolue moyenne                      0,17408
  Mean Percentage Error                       142,41
  Mean Absolute Percentage Error              170,32
  U de Theil                                  0,82646
```

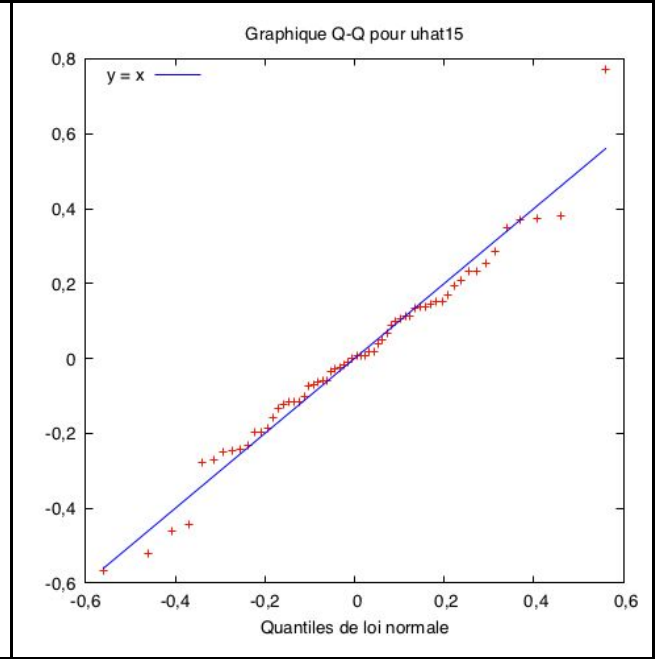
Même chose pour ce modèle ARIMA(1,1,1).




```
Fonction d'auto-corrélation résiduelle
***, **, * indicate significance at the 1%, 5%, 10% levels
using standard error 1/T^0,5
```

RETARD	ACF	PACF	Q	[p. crit.]
1	-0,0576	-0,0576		
2	0,1411	0,1383		
3	0,0952	0,1127	2,1425	[0,143]
4	-0,0909	-0,1020	2,7073	[0,258]
5	-0,1080	-0,1558	3,5192	[0,318]
6	-0,0814	-0,0846	3,9886	[0,408]
7	-0,1099	-0,0636	4,8603	[0,433]
8	-0,0250	0,0088	4,9063	[0,556]
9	-0,0950	-0,0802	5,5818	[0,589]
10	-0,0843	-0,1169	6,1242	[0,633]
11	0,1427	0,1294	7,7077	[0,564]
12	-0,0102	0,0348	7,7160	[0,657]
13	0,0689	0,0186	8,1000	[0,704]
14	0,1446	0,0776	9,8288	[0,631]
15	0,1026	0,1026	10,7178	[0,634]
16	-0,0211	-0,0433	10,7562	[0,705]
17	-0,0707	-0,1346	11,1967	[0,739]
18	-0,0631	-0,0603	11,5559	[0,774]
19	-0,0939	-0,0302	12,3703	[0,777]
20	-0,0780	0,0063	12,9457	[0,795]





<u>Annexe n°8</u> : Valeurs de $\widehat{Z_T}$ à gauche et de Y_T à droite

2018:1	0,235416	-0,145123	0,380539
2018:2	-0,116902	0,059200	-0,176102
2018:3	0,009221	-0,029397	0,038618
2018:4	-0,289367	0,002319	-0,291686
2019:1	-0,134507	-0,072767	-0,061740
2019:2	-0,222715	-0,033824	-0,188891
2019:3	0,087860	-0,056006	0,143866

2018:1	9,21447
2018:2	9,09757
2018:3	9,10679
2018:4	8,81742
2019:1	8,68292
2019:2	8,46020
2019:3	8,54806

Annexe n°9 : Sortie du modèle OLS de long terme entre le taux de chômage et le taux d'intérêt, et test KPSS des résidus

```
Modèle 7: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const          10,0980      0,168045       60,09         5,42e-56    ***
  Taux_interet   -0,446613    0,0583922      -7,648        1,77e-10    ***

Moy. var. dép.          8,981320   Éc. type var. dép.       0,916827
Somme carrés résidus    26,60301   Éc. type de régression   0,660390
R2                      0,489537   R2 ajusté                0,481169
F(1, 61)                58,49945   p. critique (F)          1,77e-10
Log de vraisemblance   -62,23665   Critère d'Akaike         128,4733
Critère de Schwarz      132,7596   Hannan-Quinn             130,1591
rho                     0,967973   Durbin-Watson            0,157724
```

```
Test KPSS pour uhat7

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,232453

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique > .10
```

Annexe n°10 : Développement démarche de cointégration de Y_t avec la production industrielle

```
Test de Dickey-Fuller augmenté pour Production_indus_manu
testing down from 20 lags, criterion AIC
taille de l'échantillon 54
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 8 retards de (1-L)Production_indus_manu
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -0,00162168
  statistique de test: tau_nc(1) = -0,650521
  p. critique asymptotique 0,4357
  Coeff. d'autocorrélation du 1er ordre pour e: 0,034
  différences retardées: F(8, 45) = 2,310 [0,0361]
```

```
Test de Dickey-Fuller augmenté pour d_Production_indus_
testing down from 20 lags, criterion AIC
taille de l'échantillon 54
hypothèse nulle de racine unitaire : a = 1

  test sans constante
  avec 7 retards de (1-L)d_Production_indus_manu
  modèle: (1-L)y = (a-1)*y(-1) + ... + e
  valeur estimée de (a - 1): -1,06309
  statistique de test: tau_nc(1) = -3,62015
  p. critique asymptotique 0,0002915
  Coeff. d'autocorrélation du 1er ordre pour e: 0,036
  différences retardées: F(7, 46) = 1,370 [0,2407]
```

```
Modèle 9: MCO, utilisant les observations 2004:1-2019:3 (T = 63)
Variable dépendante: Taux_chomage

                     coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const              20,0043       1,48879       13,44          6,91e-20    ***
  Production_indus~  -0,105857     0,0142742     -7,416         4,46e-10    ***

Moy. var. dép.          8,981320   Éc. type var. dép.       0,916827
Somme carrés résidus    27,40643   Éc. type de régression   0,670288
R2                      0,474121   R2 ajusté                0,465500
F(1, 61)                54,99632   p. critique (F)          4,46e-10
Log de vraisemblance   -63,17388   Critère d'Akaike         130,3478
Critère de Schwarz      134,6340   Hannan-Quinn             132,0336
rho                     0,944581   Durbin-Watson            0,116061
```

```
Test KPSS pour uhat9

T = 63
Paramètre du délai de troncation = 3
Statistique de test = 0,383945

                          10%      5%      1%
Valeurs critiques: 0,351   0,462   0,728
P. critique interpolée 0,085
```

```
Modèle 11: MCO, utilisant les observations 2004:2-2019:3 (T = 62)
Variable dépendante: d_Taux_chomage

                     coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------------
  const              -0,00445474   0,0261143     -0,1706        0,8651
  d_Production_ind~  -0,0508457    0,0144891     -3,509         0,0009      ***
  uhat9_1            -0,108863     0,0418107     -2,604         0,0116      **

Moy. var. dép.          0,000482   Éc. type var. dép.       0,245735
Somme carrés résidus    2,483029   Éc. type de régression   0,205147
R2                      0,325910   R2 ajusté                0,303059
F(2, 59)                14,26268   p. critique (F)          8,85e-06
Log de vraisemblance    11,77312   Critère d'Akaike        -17,54624
Critère de Schwarz     -11,16484   Hannan-Quinn            -15,04074
rho                     0,134515   Durbin-Watson            1,727062
```

Une différenciation a été nécessaire pour rendre X_3 stationnaire. Après estimation de la relation de long terme qui lie les 2 variables via une régression linéaire simple, on trouve que les résidus ceux eux aussi stationnaires. Enfin, lors de l'estimation de court terme on voit que la relation est pertinente car $\hat{\delta}$ est significatif et négatif.

Annexe n°11 : Base de données utilisée pour l'analyse

	Taux de chômage	Popularité du mot 'emploi'	Taux d'intérêt	Production industrielle	Population active
T1 2004	8,52	41,00	4,11	110,77	27015,61
T2 2004	8,38	39,33	4,31	110,97	27033,46
T3 2004	8,47	45,00	4,16	110,63	27184,71

T4 2004	8,50	38,00	3,83	111,26	27178,72
T1 2005	8,26	38,67	3,64	110,97	27199,43
T2 2005	8,44	36,00	3,37	110,88	27324,99
T3 2005	8,62	39,00	3,23	110,62	27362,27
T4 2005	8,65	32,33	3,39	111,63	27326,01
T1 2006	8,72	33,67	3,51	111,59	27396,38
T2 2006	8,57	29,67	3,99	113,42	27396,62
T3 2006	8,52	33,00	3,90	112,44	27575,51
T4 2006	7,98	28,67	3,79	113,06	27514,68
T1 2007	8,09	29,00	4,05	114,05	27605,50
T2 2007	7,79	25,67	4,39	114,69	27656,26
T3 2007	7,63	28,67	4,44	114,44	27782,50
T4 2007	7,11	25,00	4,33	114,23	27788,73
T1 2008	6,85	26,00	4,08	115,38	27812,27
T2 2008	6,96	24,67	4,47	112,67	27855,83
T3 2008	7,07	28,33	4,48	110,18	27898,07
T4 2008	7,37	28,00	3,90	102,45	28021,53
T1 2009	8,22	40,67	3,64	93,57	28081,58
T2 2009	8,81	47,00	3,79	93,64	28196,39
T3 2009	8,80	61,33	3,64	96,02	28103,08
T4 2009	9,11	64,00	3,53	96,84	28147,15
T1 2010	9,00	68,67	3,48	97,50	28235,62
T2 2010	8,89	65,33	3,18	99,00	28232,12
T3 2010	8,81	73,67	2,78	99,43	28257,01
T4 2010	8,79	71,00	3,02	100,77	28218,24
T1 2011	8,78	83,00	3,55	104,23	28186,24
T2 2011	8,70	79,00	3,54	103,10	28229,05
T3 2011	8,78	83,67	3,01	102,08	28253,21
T4 2011	8,99	88,00	3,19	102,64	28324,70
T1 2012	9,14	93,00	3,05	101,03	28341,48
T2 2012	9,34	82,00	2,77	99,71	28454,37
T3 2012	9,35	90,00	2,21	100,68	28487,46
T4 2012	9,76	90,00	2,11	98,09	28644,10

T1 2013	9,96	94,67	2,16	98,45	28530,25
T2 2013	10,05	86,33	1,96	100,10	28644,76
T3 2013	9,86	91,00	2,36	98,50	28680,20
T4 2013	9,81	85,33	2,33	99,56	28644,89
T1 2014	10,16	89,67	2,26	99,21	29378,03
T2 2014	10,21	80,67	1,86	98,39	29385,46
T3 2014	10,32	88,00	1,44	99,17	29367,67
T4 2014	10,48	87,33	1,11	98,07	29479,44
T1 2015	10,34	89,67	0,59	99,39	29380,31
T2 2015	10,47	81,67	0,84	100,09	29467,86
T3 2015	10,38	87,67	1,04	100,04	29533,92
T4 2015	10,25	83,00	0,89	100,48	29527,50
T1 2016	10,22	75,33	0,65	100,92	29552,88
T2 2016	10,01	68,00	0,47	99,86	29531,59
T3 2016	9,96	70,00	0,17	100,60	29558,66
T4 2016	10,04	69,00	0,58	101,03	29581,77
T1 2017	9,58	72,67	0,97	101,32	29474,86
T2 2017	9,48	63,33	0,79	102,83	29726,55
T3 2017	9,56	69,33	0,75	103,83	29708,93
T4 2017	8,98	70,67	0,73	105,68	29759,76
T1 2018	9,21	72,67	0,89	103,43	29801,86
T2 2018	9,10	61,33	0,77	103,49	29792,80
T3 2018	9,11	62,33	0,71	103,99	29833,30
T4 2018	8,82	60,00	0,76	103,81	29867,44
T1 2019	8,68	62,00	0,55	105,02	29716,09
T2 2019	8,46	54,67	0,25	104,64	29665,85
T3 2019	8,55	61,00	-0,23	103,74	29558,35

XI- Table des matières

I- Introduction **3**

II- Analyse économique du sujet **7**

III- Analyse des 4 variables **10**

1- Présentation des données 10

2- Google Trends : X_1 12

3- Taux d'intérêt : X_2 14

4- Production industrielle : X_3 16

5- Population active : X_4 17

6- Graphiques d'évolution des variables explicatives 19

IV- Présentation de la méthodologie ARIMA **21**

→ Stationnarisation de la série 23

→ Identification du modèle 24

→ Estimation du modèle 25

→ Vérification des résidus notés at 25

→ Prévisions 26

V- Choix du modèle ARIMA pour Yt **27**

→ Stationnarisation 27

→ Identification et estimation du modèle 29

→ Vérification et choix du modèle 32

→ Prévisions à une période 36

VI- Test de cointégration selon Engle-Granger **38**

→ Cointégration du taux de chômage avec la popularité du mot 'emploi' (X_1) 40

→ Cointégration du taux de chômage avec le taux d'intérêt (X_2) 43

→ Cointégration du taux de chômage avec la production industrielle (X_3) 44

→ Cointégration du taux de chômage avec la population active (X_4) 44

VII- Conclusion **47**

VIII- Discussion **48**

IX- Bibliographie **49**

X- Annexes **52**