UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TRAVAIL PRATIQUE 1

PRÉSENTÉ À

PHILIPPE GOULET COULOMBE

DANS LE CADRE DU COURS

APPLICATIONS DE MODÈLES ÉCONOMIQUES

ECO 8086, GROUPE 40

VAUDESCAL GUILLAUME VAUG30119904

PHILIPPE TOUSIGNANT TOUP11049704

REMIS LE

31 OCTOBRE 2021

Table des matières

INTRODUCTION :

Dans notre travail de prévision d'une variable macroéconomique, nous avons décidé de prévoir le chômage des États-Unis. En ce sens, nous avons utilisé les données de FRED[1] du chômage et plus précisément la première différence du taux de chômage. De plus, les données sont déjà transformées par l'organisme (désaisonnalisées et différentiées).

D'autre part, ces données sont mensuelles, et disponible de février 1948 à Août 2021. Bien que l'échantillon soit disponible depuis les années 40, nous avons décidé de garder seulement les observations à partir de 1961. En effet, nous avons pris cette décision en raison de l'absence de certaines variables exogène indispensable à d'autres modèle de 1948 à 1961. Il est impératif de faire cela afin de comparer les modèles sur un pied d'égalité.

Enfin, toujours dans une optique de description des données, on peut souligner que la première différence du chômage semble stationnaire d'ordre 2, à l'exception de la période de la Covid-19 (2020-02 à 2021-06).

Dans notre devoir, dans un premier temps on présentera les estimations et évaluations de nos différents modèles. Puis dans un deuxième temps, nous établirons des prévisions, dont notre analyse s'appuiera à l'aide de figures et tableaux, pour la période demandée c'est-à-dire de septembre 2021 à septembre 2022, et ce, en comparant nos résultats lorsqu'on inclut les données du covid ou non. Enfin dans une troisième partie nous discuterons et résumerons les résultats obtenus dans nos analyses et nous conclurons.

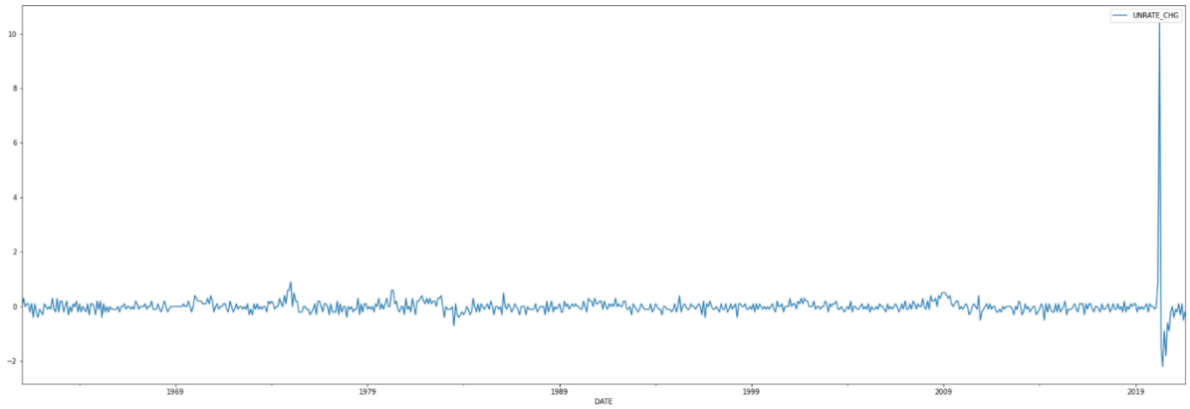
[1] U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, October 17, 2021.

Tableau 1 : Statistiques descriptives du chômage (*percent change*) :

	UNRATE_CHG
count	728.000000
mean	-0.001923
std	0.443875
min	-2.200000
25%	-0.100000
50%	0.000000
75%	0.100000
max	10.400000

Ce tableau présente une description sommaire des données comprenant la période de la pandémie (2020-02 à 2021-06). Dans une partie subséquente, nous séparerons la période du covid de l'échantillon afin de réaliser des prévisions comme indiqué ci-dessus.

Figure 1 : Représentation graphique des données (*percent change*) :



Ce graphique nous représente une série temporelle des données incluant la période de la Covid-19. Comme mentionné ultérieurement la série semble stationnaire d'ordre 2 à l'exception de la période de la pandémie (2020-02 à 2021-06).

PARTIE I - Estimation et évaluation des modèles :

Dans cette première partie nous allons estimer et évaluer 3 modèles obligatoires, soit la moyenne historique, l'autorégressif direct et le modèle à facteur, ainsi que nos 3 meilleurs modèles optionnels qui sont ARMA, ARDL et VAR.

Nos données d'entrainements sont de 1961-01 à 2014-12 inclusivement, et nos données pour le test (ou le *pseudo out of sample*) sont de 2015-01 à 2021-01 inclusivement. Il est important de noter que chaque modèle a été estimé 732 fois (61 périodes fois 12 horizons), de ce fait les données d'entrainements ont été rajoutés à chacune des 61 périodes. Les horizons sont modélisés simplement en modifiant les positions des inputs du modèle (sauf pour la moyenne historique). Enfin, dans une optique de faciliter la comparaison de la performance des modèles, toutes les MSE sont exprimés par leur rapport sur notre MSE *benchmark* qui est un AR(1) d'horizon 1.

A. Moyenne historique :

$$y_t = c + \epsilon_t$$

Une prévision avec la moyenne historique est exactement comme son nom l'indique, on utilise la moyenne historique des données au temps t comme prévision au temps t+h.

Tableau 2 : MSE du modèle de moyenne historique par rapport au MSE *benchmark* :

h	MSE moyenne historique /MSE Benchmark
1	0.918306
2	0.917571
3	0.917618
4	0.917411
5	0.918295
6	0.918070
7	0.918690
8	0.919093
9	0.918612
10	0.918548
11	0.919918
12	0.920439

Comme nous avons vu en cours, la moyenne historique rend quand même des résultats impressionnants bien qu'elle ne capture pas la variance de notre variable. En effet, elle minimise mieux les erreurs en moyenne que notre *benchmark*. On peut noter qu'en moyenne la moyenne historique à une MSE 10 point de pourcentage inférieur au *benchmark*.

B. Autorégressif direct :

$$y_{t+h} = c + \rho * y_t + \epsilon_{t+h}$$

Le modèle autorégressif est une modélisation d'une série temporelle en fonction d'une constante et de ses valeurs passées. Afin de choisir le bon ordre de notre modèle autorégressif, nous avons comparé les modèles d'ordre 1 à 12 et gardé celui avec le meilleur BIC. Ainsi le meilleur modèle est un AR(4).

Tableau 3 : MSE du modèle autorégressif direct par rapport au MSE *benchmark* :

h	MSE AR /MSE Benchmark
1	0.917270
2	0.914911
3	0.873599
4	0.839876
5	0.930982
6	0.902112
7	0.966242
8	0.951005
9	0.859011
10	0.846561
11	0.961640
12	0.964474

On remarque que notre AR(4) performe systématiquement mieux que notre *benchmark*. On peut noter sur le tableau que le AR(4) horizon 4 nous donnent les meilleurs résultats (0.839876). Ce qui de ce fait, équivaut à régresser le temps t avec ses 4,5,6 et 7 *lags*.

C. Modèle à facteurs :

$$y_{t+h} = c + \rho * y_t + \beta * F_t + \epsilon_{\{t+h)}$$

$$X_t = \lambda * F_t + u_t$$

Un modèle à facteur est un modèle qui modélise une série temporelle avec ses *lags* (ou délai) et des facteurs. Plus précisément, les modèles factoriels décomposent le comportement d'une variable économique (X_{ti}) en une composante déterminée par quelques facteurs inobservables

(F_t), communs à toutes les variables mais ayant des effets spécifiques sur celles-ci (λ_i), et une variable idiosyncratique spécifique.[2]

Dans notre cas, nos facteurs sont tirés de la base FREDMD[3], une grande base de données américaines mensuelles réalisé par la FRED. Les données sont également transformées par les codes de transformations prévu dans la base de données. Encore une fois, nous avons choisis le modèle optimal à l'aide du BIC. Dans notre modèle, le nombre de composante variant entre 1 à 5, les paramètres autorégressifs de 1 à 12 et les paramètres des composantes variant de 1 à 12. Le modèle optimal utilise 2 composantes, les deux *lags* de la première composante et les 5 lags de la deuxième composante respectivement, ainsi qu'un *lag* de la série elle-même.

Tableau 4 : Extrait de FREDMD transformées de 1961-01 à 2021-08 :

DATE	RPI	W875RX1	DPCERA3M086SBEA	CMRMTSPLx	RETAILx	INDPRO	IPFPNSS	IPFINAL	IPCONGD
1961-01-01	0.008982	0.008686	0.002483	-0.036080	0.000780	0.001215	-0.001189	-0.002385	-0.008258
1961-02-01	0.003817	0.000083	0.003686	0.012134	-0.003571	-0.001215	0.001189	0.000000	0.005905
1961-03-01	0.004064	0.004912	0.012333	0.023524	0.010510	0.006065	0.004742	0.001193	0.000000
1961-04-01	0.002170	0.003453	0.000488	-0.029094	-0.017860	0.020348	0.015268	0.016547	0.023285
1961-05-01	0.007748	0.005932	0.006053	0.028427	0.014924	0.015284	0.006969	0.008174	0.011441
...	...	...	...	...	...	...	...	...	...
2021-04-01	-0.152394	0.000453	0.004273	-0.003919	0.008998	0.000689	-0.006716	-0.008300	-0.005042
2021-05-01	-0.026821	0.001040	-0.004792	-0.020754	-0.013824	0.006232	0.006009	0.008166	0.006653
2021-06-01	-0.003488	0.001618	0.005309	0.004279	0.008488	0.004889	-0.001673	-0.000942	-0.003311
2021-07-01	0.006882	0.002319	-0.001446	NaN	-0.017895	0.008412	0.014105	0.015796	0.007083
2021-08-01	NaN	NaN	NaN	NaN	0.007077	0.004019	0.006693	0.007003	0.008141

[2] Massimiliano Marcellino « An Introduction to Factor Modelling » Bocconi University, 2017

[3] McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. Journal of Business and Economic Statistics, 34(4):574–589

Ce tableau représente une partie de la base de données utilisés pour créer les composantes principales. En réalité cette base de données contient 728 variables de 1961-01 à 2021-08. Bien entendu, pour notre entrainement nous avons utilisé les mêmes périodes que mentionnées antérieurement.

Tableau 5 : MSE du modèle à facteur par rapport au MSE *benchmark* :

MSE Factor model /MSE Benchmark	
h	
1	0.938159
2	0.983749
3	0.985827
4	0.957578
5	0.989934
6	0.966122
7	0.989230
8	0.998396
9	0.984777
10	1.005463
11	0.971915
12	0.975762

On peut noter qu'en moyenne la performance du modèle à facteur est meilleur que le AR(1) *benchmark*. Toutefois, il est moins performant que les deux précédents modèles, à savoir la moyenne historique et l'autorégressif direct. Enfin, le meilleur horizon est h1 (0.938159).

D. Modèle ARMA :

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{\mathrm{q}} \theta_i \varepsilon_{t-i}$$

Le modèle ARMA est la combinaison d'une part d'un modèle autorégressif (AR) et d'autre part d'une moyenne mobile (MA). Le modèle est généralement noté ARMA (p, q), où p est l'ordre de la partie AR et q l'ordre de la partie MA. Nous avons également choisi l'ordre de notre modèle ARMA par le BIC. Il s'avère que notre modèle optimal est d'ordre (1,2).

Tableau 6 : MSE du modèle ARMA par rapport au MSE *benchmark* :

h	MSE ARMA /MSE Benchmark
1	0.925398
2	0.941758
3	1.020180
4	0.901306
5	1.016643
6	0.941547
7	0.966342
8	1.054757
9	0.998546
10	0.945392
11	0.963681
12	0.957872

On remarque que le modèle ARMA (1,2) performe relativement moins bien que les précédents modèles présentés. Cependant il reste sensiblement meilleur que le AR(1) *benchmark*. Enfin le meilleur horizon est h4 (0.901306).

E. Modèle ARDL :

$$Y_t = \varphi + \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=0}^{q} b_j X_{t-j} + \epsilon_t$$

Les modèles « *Autoregressive Distributed Lag/ARDL* », ou « modèles autorégressifs à retards échelonnés ou distribués/ARRE » en français, sont des modèles dynamiques. Ces derniers ont la particularité de prendre en compte la dynamique temporelle (délai d'ajustement, anticipations, etc.) dans l'explication d'une variable (série chronologique), améliorant ainsi les prévisions et efficacité des politiques (décisions, actions, etc.), contrairement au modèle simple (non dynamique) dont l'explication instantanée (effet immédiat ou non étalé dans le temps) ne restitue qu'une partie de la variation de la variable à expliquer.[4]

Nous avons essayé quelques combinaisons de variables disponibles dans FREDMD comme régresseur dans le modèle ARDL selon la pertinence théorique de ceux-ci. Nous avons choisi **CLAIMSx** et **PAYEMS** comme variables exogènes. Par la suite, afin de déterminer l'ordre optimal du modèle, nous avons procédés de la même façon que les modèles antérieurs en utilisant le BIC. En ce sens, l'ordre optimal est 1 *lag* de la variable endogène (*unrate percent change*) et 3 et 2 *lags* respectivement de **CLAIMSx** et **PAYEMS**.

Tableau 7 : MSE du modèle ARDL par rapport au MSE *benchmark* :

	MSE ARDL /MSE Benchmark
h	
1	0.867007
2	0.903617
3	0.909578
4	0.869122
5	0.910491
6	0.877789
7	0.899670
8	0.929164
9	0.910818
10	0.915436
11	0.875917
12	0.897662

[4] Jonas Kibala Kuma. Modélisation ARDL, Test de cointégration aux bornes et Approche de TodaYamamoto : éléments de théorie et pratiques sur logiciels. Licence. Congo-Kinshasa. 2018. ffcel01766214f

On peut noter que le modèle ARDL performe relativement bien comparés aux précédents modèles présentés. En effet, en moyenne chacun de ses horizons est plus performant que le *benchmark*. Enfin le meilleur horizon est h1 (0.867007).

F. Modèle VAR :

$$y_t = v + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t$$

Le Vecteur Autorégressif (VAR) est un modèle statistique qui permet de capturer les interdépendances entre plusieurs séries temporelles. Dans un modèle VAR, les variables sont traitées symétriquement de manière que chacune d'entre elles soit expliquée par ses propres valeurs passées et par les valeurs passées des autres variables.

Dans notre modèle, les variables sont **UNRATE** puis **CLAIMSx**. L'ordre du VAR encore une fois calculé par le BIC est 6 lags de chaque.

Tableau 8 : MSE du modèle VAR par rapport au MSE *benchmark* :

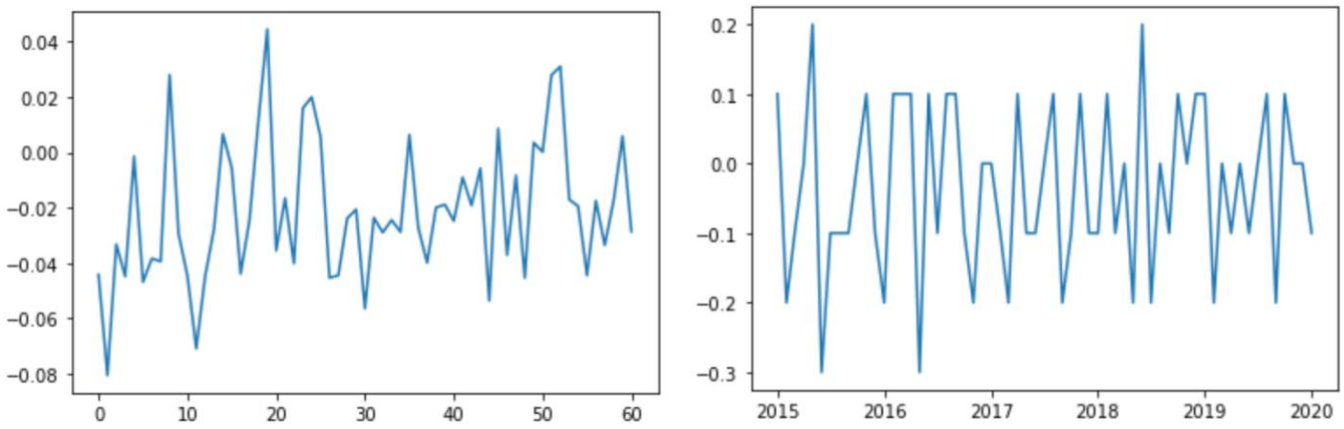
h	MSE VAR /MSE Benchmark
1	0.821966
2	1.016794
3	0.981744
4	1.039604
5	0.941307
6	0.950520
7	0.893554
8	0.954298
9	0.934350
10	0.920388
11	0.908603
12	0.919592

On peut observer que le modèle VAR performe relativement bien comparés aux précédents modèles présentés. En effet, en moyenne chacun de ses horizons est plus performant que le *benchmark*. Également on peut préciser l'excellente performance de l'horizon h1 (0.821966) comparativement aux autres modèles.

Ainsi, après avoir passés en revus plusieurs modèles afin de prédire la première différence du taux de chômage, on peut noter que les 3 meilleurs modèles comparativement au *benchmark*, sont : Le VAR h1, le AR(4) h4 et le ARDL h1. De surcroit, tous les modèles semblent bien minimiser la somme des erreurs mais, ne capture pas adéquatement notre variance de notre variable d'intérêt. Afin d'illustrer cela, on peut montrer l'exemple du modèle autorégressif.

Figure 2 : Graphique *Pseudo out of sample* avec AR(4) h4 2015-01 à 2020-01 (à gauche) :

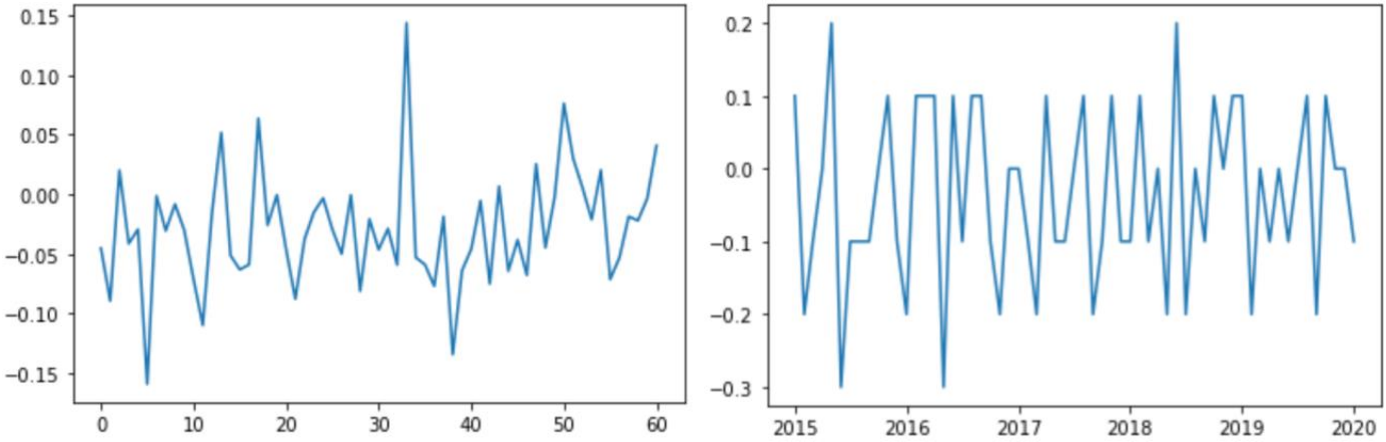
Figure 3 : Graphique *Unrate (Percent change) testset* 2015-01 à 2020-01 (à droite) :



On voit rapidement que les prévisions ont une variance beaucoup plus petite que les valeurs réelles. Également, celui qui capture le mieux la variance est le modèle VAR, comme on peut le voir avec les deux graphiques ci-dessous.

Figure 4 : *Pseudo out of sample* avec CLAIMSx de 2015-01 à 2020-01 (à gauche) :

Figure 5 : Unrate *testset* de 2015-01 à 2020-01 (à droite) :



En guise de limite et d'extension de nos analyses on soulève les points suivants : D'une part, on peut noter que les MSE des différents modèles sont relativement proche. En ce sens, on pourrait alors réaliser des tests statistiques en guise d'extension, afin de voir s'ils sont

statistiquement différents. D'autre part, il y a une possibilité *d'overfitting* dans les modèles plus complexes (par exemple modèles à facteur ou ARDL) a l'inverse de capturer la généralité de la série. Dernièrement, il est intéressant de noter que dans la plupart des modèles, l'horizon 1 et 4 sont très performant pour réduire la MSE.

PARTIE II - Prévisions graphiques :

Dans cette partie, notre objectif est de faire du *out-of-sample forecast* de la première différence du taux de chômage américain pour la période de 2021-09 à 2022-09. Tous les meilleurs modèles présentés dans la section précédente ont été réestimés deux fois, la première avec les données de 1961-01 à 2021-08 et la deuxième avec cette même période mais en retirant a période d'extrême volatilité dû à la pandémie de covid-19. (2020-02 à 2021-06). En ce sens, nous aurons 12 prédictions de 13 périodes.

Il est intéressant de comparer nos deux séries temporelles, c'est-à-dire avec et sans covid.

Figure 6 : Représentation graphique des données avec la période de la Covid (*percent change*) :

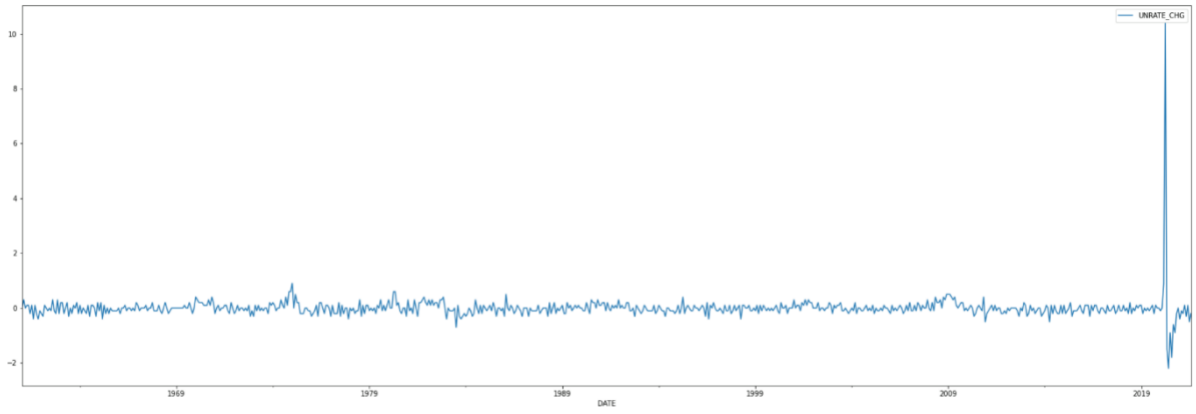
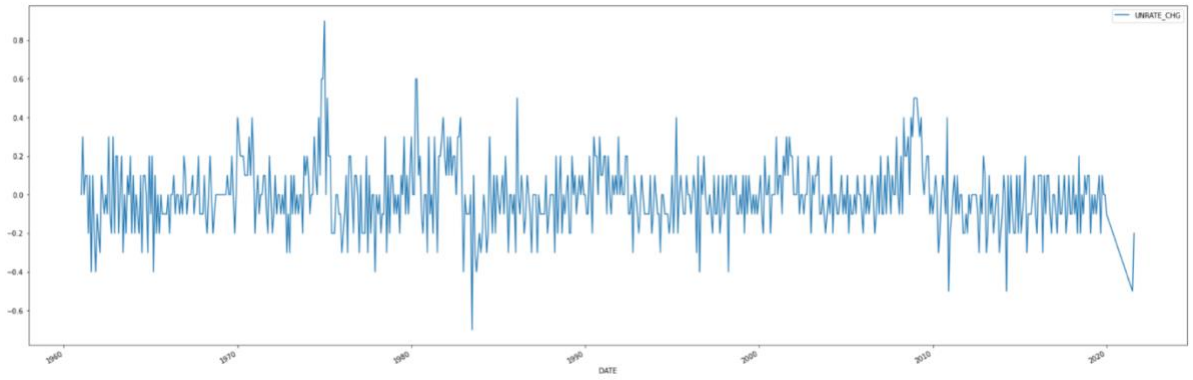


Figure 7 : Représentation graphique des données sans la période de la Covid (*percent change*) :



On voit clairement que la période du covid nous amène des valeurs extrêmes, et qu'il est fort probable que nos estimations soient différentes et faussés dans les deux cas.

Tableau 9 : Statistiques descriptives avec données covid (à gauche) :

Tableau 10 : Statistiques descriptives sans données non covid (à droite) :

	UNRATE_CHG
count	728.000000
mean	-0.001923
std	0.443875
min	-2.200000
25%	-0.100000
50%	0.000000
75%	0.100000
max	10.400000

	UNRATE_CHG
count	711.000000
mean	-0.005345
std	0.174339
min	-0.700000
25%	-0.100000
50%	0.000000
75%	0.100000
max	0.900000

On peut voir à travers toutes ces variables de distribution que la période du covid apporte des valeurs extrêmes, apporte une grande variance et augmente significativement la moyenne.

A. Moyenne historique :

$$y_t = c + \epsilon_t$$

Figure 8 : Graphique moyenne historique avec données Covid (2021-09 à 2022-09) :

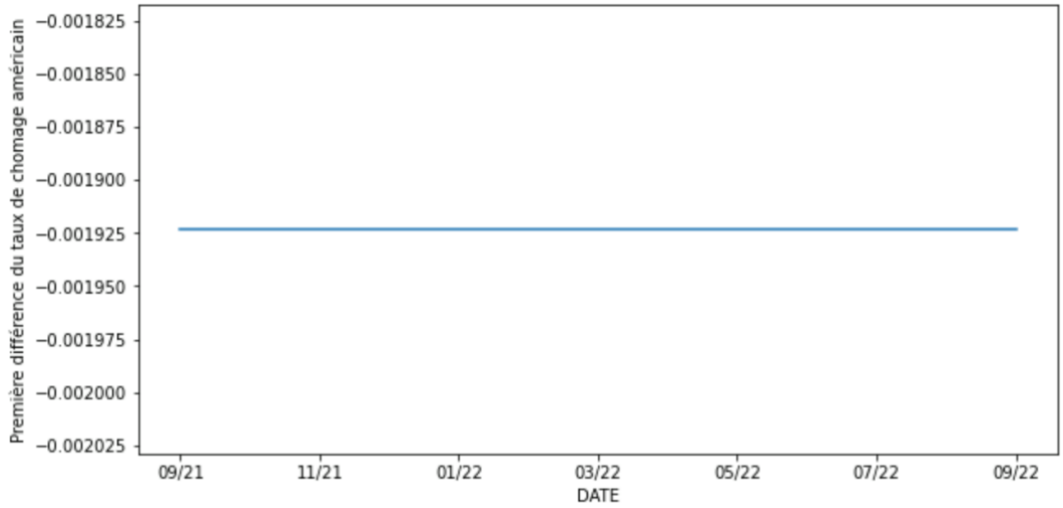
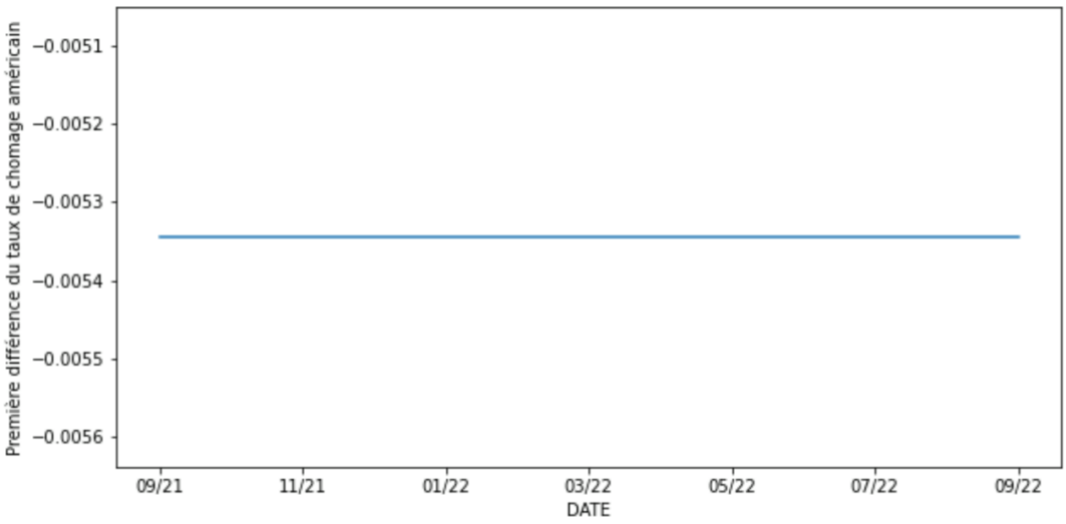


Figure 9 : Graphique moyenne historique sans données Covid (2021-09 à 2022-09) :



On peut noter que la moyenne historique avec la Covid est plus grande comparativement que sans le covid (-0.001925 > -0.0053). Ce qui est dans un sens logique car la première différence du chômage a connu des variables extrêmes positives lors du confinement.

B. Autorégressif direct :

$$y_{t+h} = c + \rho * y_t + \epsilon_{t+h}$$

Tableau 11 : Résultats d'estimation AR(4) h4 avec données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	728
Model:	Restr. AutoReg(7)	**Log Likelihood**	-437.596
Method:	Conditional MLE	**S.D. of innovations**	0.444
Date:	Sun, 24 Oct 2021	**AIC**	887.193
Time:	11:50:34	**BIC**	914.676
Sample:	08-01-1961	**HQIC**	897.802
	- 08-01-2021		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0026	0.017	-0.158	0.875	-0.035	0.030
UNRATE_CHG.L4	-0.0768	0.037	-2.060	0.039	-0.150	-0.004
UNRATE_CHG.L5	-0.0037	0.037	-0.099	0.921	-0.077	0.069
UNRATE_CHG.L6	-0.0337	0.037	-0.902	0.367	-0.107	0.039
UNRATE_CHG.L7	-0.0080	0.037	-0.214	0.831	-0.081	0.065

Figure 10 : Graphique AR(4) h4 avec données Covid (2021-09 à 2022-09):

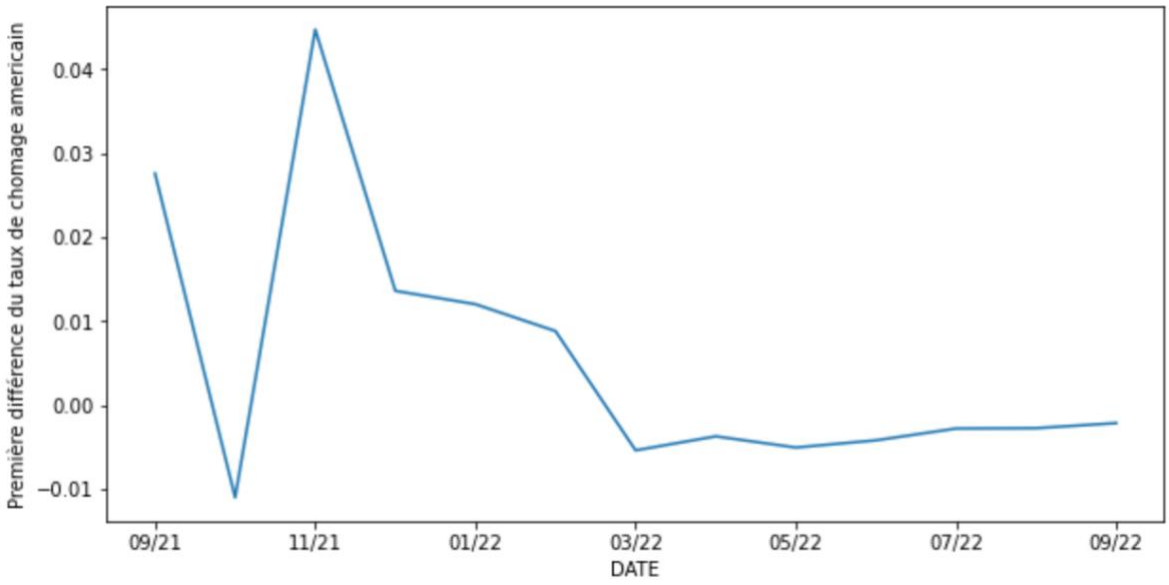
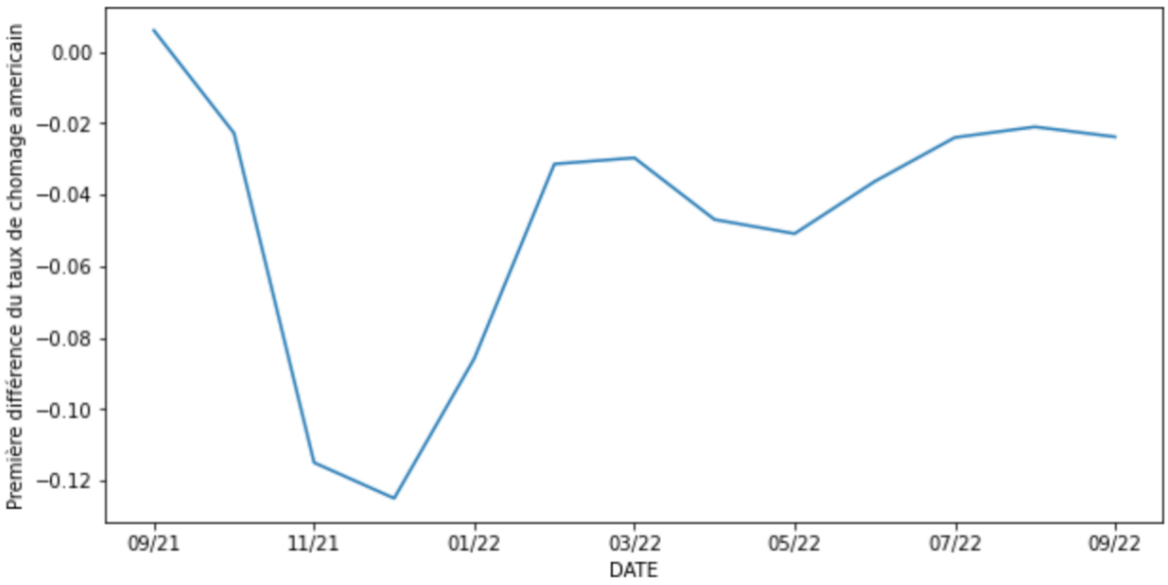


Tableau 12 : Résultats d'estimation AR(4) h4 sans données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	711
Model:	Restr. AutoReg(7)	**Log Likelihood**	264.075
Method:	Conditional MLE	**S.D. of innovations**	0.166
Date:	Sun, 24 Oct 2021	**AIC**	-516.150
Time:	11:50:34	**BIC**	-488.810
Sample:	7	**HQIC**	-505.584
	711		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0039	0.006	-0.621	0.535	-0.016	0.008
UNRATE_CHG.L4	0.1935	0.038	5.059	0.000	0.119	0.269
UNRATE_CHG.L5	0.1431	0.038	3.804	0.000	0.069	0.217
UNRATE_CHG.L6	0.1080	0.038	2.877	0.004	0.034	0.182
UNRATE_CHG.L7	0.0047	0.038	0.124	0.901	-0.070	0.080

Figure 11 : Graphique AR(4) h4 sans données Covid (2021-09 à 2022-09) :



Plusieurs points intéressants à notre analyse ressortent des tableaux et graphiques du modèle AR(4) h4. Premièrement, les paramètres des tableaux 11 et 12 sont différents car on a estimé ces derniers avec des données différentes (avec et sans Covid). En ce sens, les *forecasts* sont différents car nous n'avons pas les mêmes valeurs de paramètres et les mêmes valeurs de *lags* dû à l'inclusion ou l'exclusion de la période de la Covid. Deuxièmement, la variance des prédictions de notre modèle semble déraisonnable en comparaison avec l'historique. En effet, les variances prédites sont trop faibles, ne capturant que relativement la variance des données passés. Troisièmement, comme on *forecast* beaucoup de période, les prévisions convergent vers leur constante respective plus on avance dans le temps. Enfin, on peut noter que le point de départ de la figure 10 avec les données du covid (0.03) est plus élevé que celui de la figure 11 sans les données de la Covid (0.005), et que ce dernier possède une plus faible variance. Un autre aspect intéressant sont les valeurs de *p-value* des paramètres des deux tableaux. On voit clairement que les estimations contenant des données Covid augmente l'incertitude quant aux valeurs des paramètres. Cela sera constant au travers de plusieurs modèles.

C. Modèle à facteurs :

$$y_{t+h} = c + \rho * y_t + \beta * F_t + \epsilon_{\{t+h)}$$

$$X_t = \lambda * F_t + u_t$$

Afin de réaliser du vrai *out of sample forecast* avec le modèle à facteur, il nous faut des variables exogènes *out of sample*. De ce fait pour les générer les deux composantes été générées comme des processus ARMA. L'ordre de ces processus ont été calculé par BIC, à savoir pour la première composante (0,1) et la deuxième composante (1,1).

Tableau 13 : Modèle à facteur avec les données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	728
Model:	ARDL(1, 2, 5)	**Log Likelihood**	-360.955
Method:	Conditional MLE	**S.D. of innovations**	0.398
Date:	Sun, 24 Oct 2021	**AIC**	741.909
Time:	12:26:37	**BIC**	787.798
Sample:	06-01-1961	**HQIC**	759.617
	- 08-01-2021		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0042	0.015	-0.279	0.781	-0.033	0.025
UNRATE_CHG.L1	-0.6726	0.065	-10.279	0.000	-0.801	-0.544
comp_0.L1	-9.9390	0.795	-12.507	0.000	-11.499	-8.379
comp_0.L2	0.4550	0.460	0.989	0.323	-0.449	1.359
comp_1.L1	-1.4856	0.486	-3.059	0.002	-2.439	-0.532
comp_1.L2	-0.3948	0.455	-0.868	0.386	-1.288	0.499
comp_1.L3	-0.6069	0.452	-1.342	0.180	-1.495	0.281
comp_1.L4	-0.4732	0.451	-1.050	0.294	-1.358	0.412
comp_1.L5	0.0506	0.447	0.113	0.910	-0.827	0.928

Figure 12 : Graphique modèle à facteur avec les données Covid (2021-09 à 2022-09) :

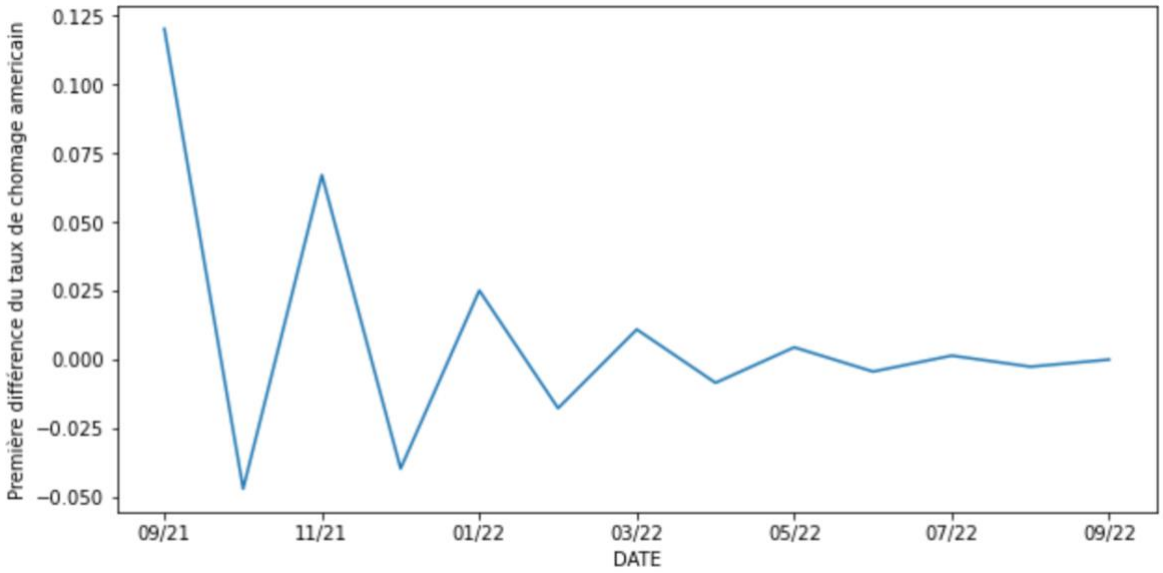
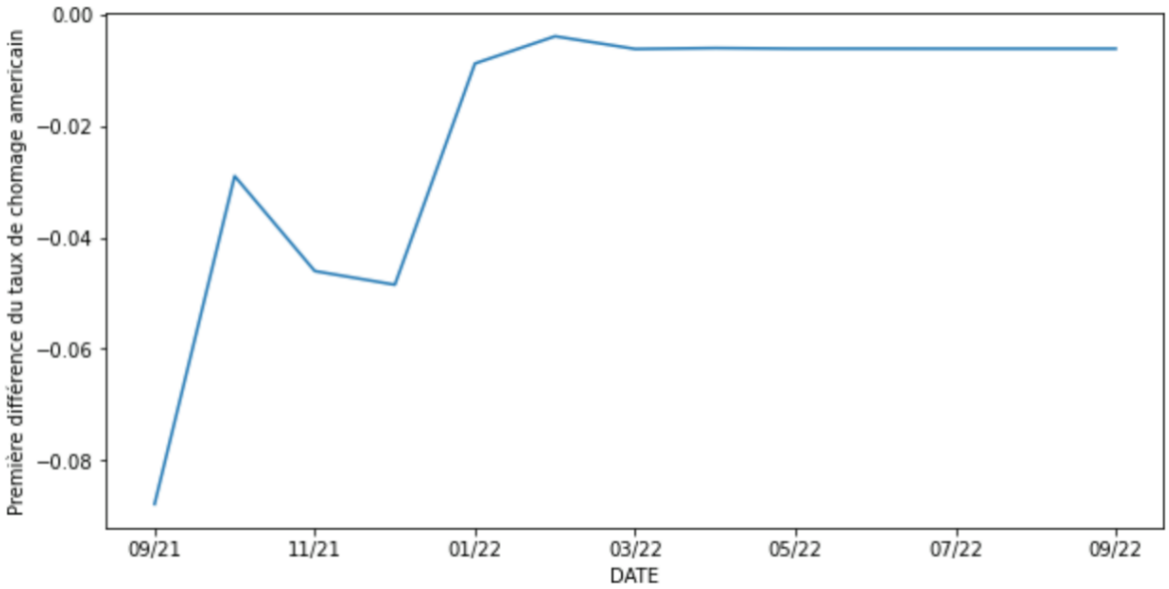


Tableau 14 : Modèle à facteur sans les données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	711
Model:	ARDL(1, 2, 5)	**Log Likelihood**	336.613
Method:	Conditional MLE	**S.D. of innovations**	0.151
Date:	Sun, 24 Oct 2021	**AIC**	-653.227
Time:	12:26:37	**BIC**	-607.574
Sample:	5	**HQIC**	-635.591
	711		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0075	0.006	-1.314	0.189	-0.019	0.004
UNRATE_CHG.L1	-0.2069	0.040	-5.144	0.000	-0.286	-0.128
comp_0.L1	1.9292	0.240	8.047	0.000	1.459	2.400
comp_0.L2	0.9144	0.220	4.151	0.000	0.482	1.347
comp_1.L1	-0.2797	0.156	-1.798	0.073	-0.585	0.026
comp_1.L2	-0.2807	0.156	-1.802	0.072	-0.587	0.025
comp_1.L3	-0.4902	0.155	-3.161	0.002	-0.795	-0.186
comp_1.L4	-0.6505	0.155	-4.205	0.000	-0.954	-0.347
comp_1.L5	-0.5313	0.156	-3.407	0.001	-0.838	-0.225

Figure 13 : Graphique modèle à facteur sans les données Covid (2021-09 à 2022-09) :



On remarque que les tableaux 13 et 14 sont très différents. Cela montre l'effet des données de la pandémie sur l'estimation des paramètres. La différence entre les deux *forecast* peut être expliquer par les différentes valeurs de paramètres et de *lags*. Ces différences semblent si prononcées que nos deux graphiques sont le miroir de l'autre. De plus, on peut noter que le point de départ de la figure 12 avec les données du Covid (0.125) est plus élevé que la figure 13 sans les données du Covid (-0.009). L'effet des données Covid sur les *p-value* est toujours présent.

D. Modèle ARMA :

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{\mathrm{q}} \theta_i \varepsilon_{t-i}$$

Tableau 15 : ARMA avec données covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	728
Model:	ARIMA(1, 0, 2)	**Log Likelihood**	-437.331
Date:	Sun, 24 Oct 2021	**AIC**	884.661
Time:	12:32:26	**BIC**	907.613
Sample:	01-01-1961	**HQIC**	893.517
	- 08-01-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0017	0.033	-0.053	0.958	-0.066	0.063
ar.L1	0.6218	0.126	4.936	0.000	0.375	0.869
ma.L1	-0.5872	0.128	-4.585	0.000	-0.838	-0.336
ma.L2	-0.0997	0.015	-6.781	0.000	-0.128	-0.071
sigma2	0.1947	0.002	121.479	0.000	0.192	0.198

Figure 14 : Graphique ARMA avec données covid (2021-09 à 2022-09) :

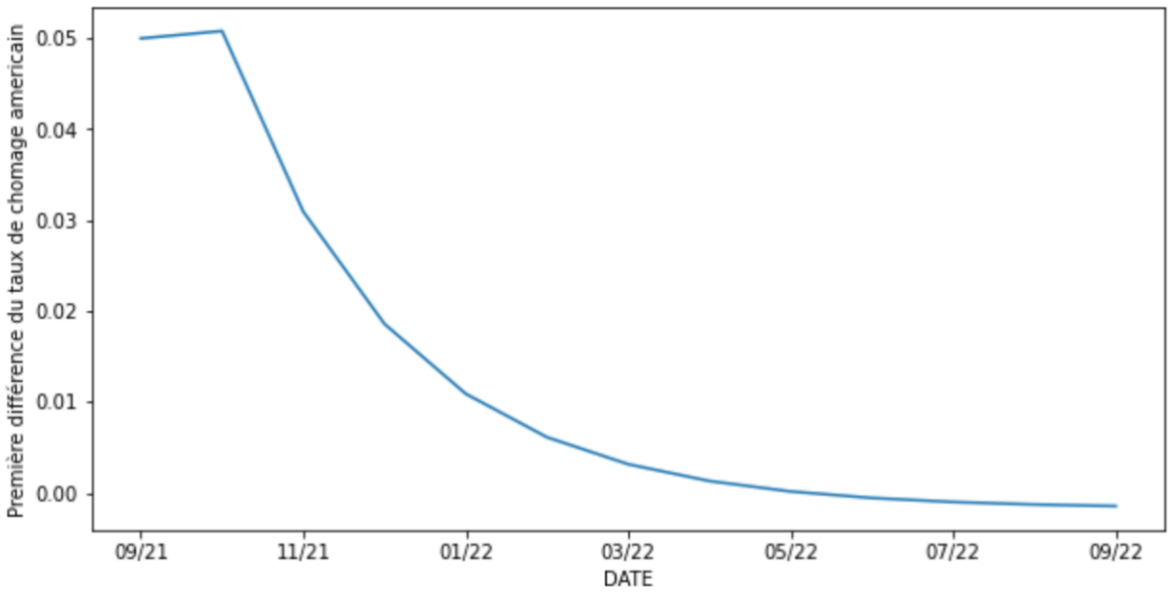
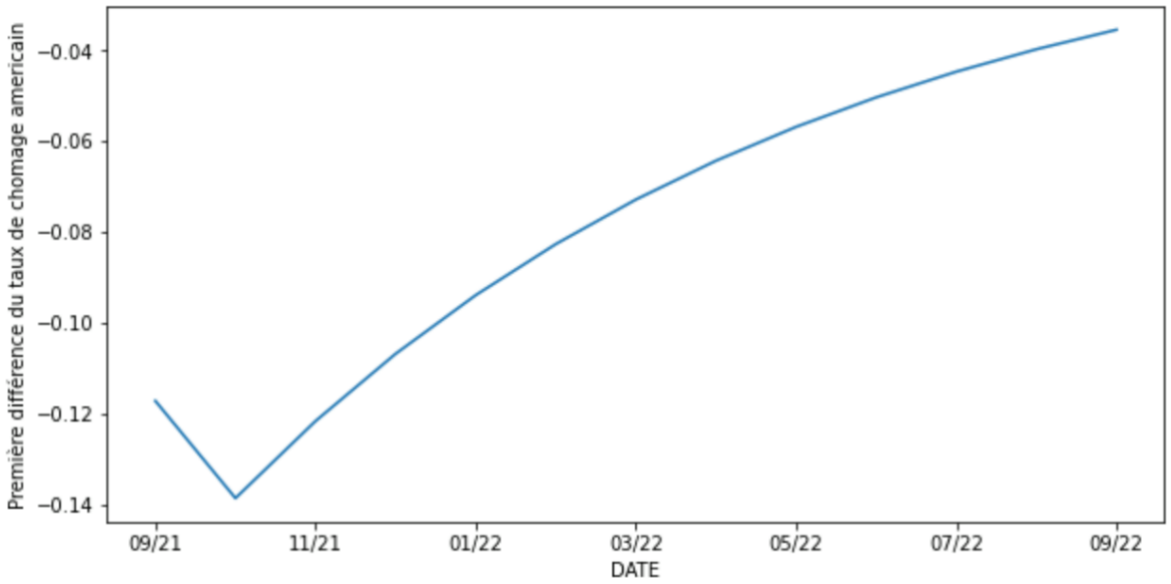


Tableau 16 : ARMA données sans Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	711
Model:	ARIMA(1, 0, 2)	**Log Likelihood**	284.560
Date:	Sun, 24 Oct 2021	**AIC**	-559.119
Time:	12:32:26	**BIC**	-536.286
Sample:	0	**HQIC**	-550.299
	- 711		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0062	0.016	-0.390	0.696	-0.037	0.025
ar.L1	0.8716	0.034	26.012	0.000	0.806	0.937
ma.L1	-0.8735	0.043	-20.541	0.000	-0.957	-0.790
ma.L2	0.1987	0.039	5.086	0.000	0.122	0.275
sigma2	0.0263	0.001	23.140	0.000	0.024	0.029

Figure 15 : ARMA données sans Covid (2021-09 à 2022-09) :



Toujours dans la même optique, les *lags* et les paramètres sont différents ce qui donnent deux *forecast* très distincts. L'effet miroir mentionné antérieurement pour le modèle à facteur est également présent pour ce modèle. Les deux *forecasts* convergent vers leur constante respective, ce qui est constant avec l'approche employé. Aussi, on peut noter que le point de départ de la figure 14 avec les données du Covid (0.05) est plus élevé que la figure 15 sans les données du Covid (-0.12).

E. Modèle ARDL :

$$Y_t = \varphi + \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=0}^{q} b_j X_{t-j} + \epsilon_t$$

Tableau 17 : ARDL avec données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	728
Model:	ARDL(1, 3, 2)	**Log Likelihood**	-2.419
Method:	Conditional MLE	**S.D. of innovations**	0.243
Date:	Sun, 24 Oct 2021	**AIC**	20.838
Time:	12:47:58	**BIC**	57.550
Sample:	04-01-1961	**HQIC**	35.005
	- 08-01-2021		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0047	0.011	0.410	0.682	-0.018	0.027
UNRATE_CHG.L1	-0.2954	0.055	-5.418	0.000	-0.402	-0.188
CLAIMSx.L1	3.6921	0.092	40.059	0.000	3.511	3.873
CLAIMSx.L2	-0.0183	0.170	-0.108	0.914	-0.352	0.316
CLAIMSx.L3	0.0530	0.141	0.375	0.708	-0.225	0.331
PAYEMS.L1	1.7781	4.164	0.427	0.670	-6.398	9.954
PAYEMS.L2	-6.9812	2.829	-2.468	0.014	-12.534	-1.428

Figure 16 : Graphique ARDL avec données Covid (2021-09 à 2022-09) :

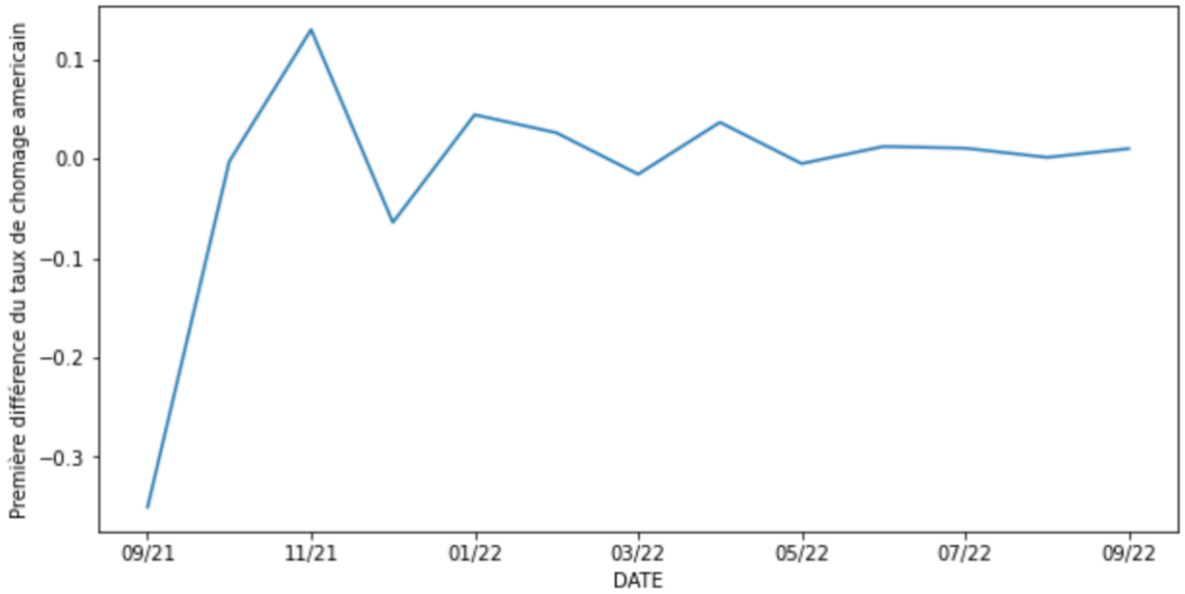
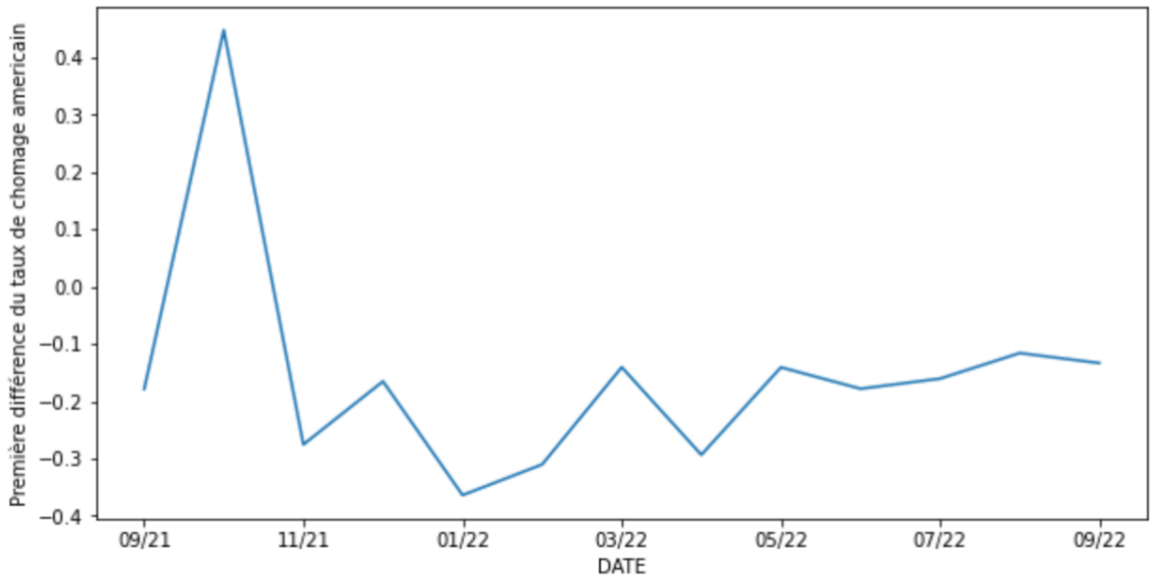


Tableau 18 : ARDL sans données Covid :

Dep. Variable:	UNRATE_CHG	**No. Observations:**	711
Model:	ARDL(1, 3, 2)	**Log Likelihood**	347.366
Method:	Conditional MLE	**S.D. of innovations**	0.148
Date:	Sun, 24 Oct 2021	**AIC**	-678.733
Time:	12:47:58	**BIC**	-642.211
Sample:	3	**HQIC**	-664.624
	711		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0500	0.008	6.478	0.000	0.035	0.065
UNRATE_CHG.L1	-0.1946	0.038	-5.091	0.000	-0.270	-0.120
CLAIMSx.L1	0.8284	0.128	6.497	0.000	0.578	1.079
CLAIMSx.L2	0.5942	0.125	4.749	0.000	0.349	0.840
CLAIMSx.L3	0.5306	0.123	4.305	0.000	0.289	0.772
PAYEMS.L1	-20.9768	3.703	-5.665	0.000	-28.246	-13.707
PAYEMS.L2	-16.2893	3.573	-4.559	0.000	-23.305	-9.274

Figure 17 : Graphique ARDL sans données Covid (2021-09 à 2022-09) :



Encore une fois, les prévisions sont très distinctes dans les deux approches pour les mêmes raisons. Comme on *forecast* beaucoup de période, les prévisions convergent vers la constante plus on est loin (0.00) avec les données du covid (figure 16) et (-0.15) sans les données Covid (figure 17). Enfin, on peut noter que le point de départ de la figure 16 avec les données du Covid (-0.3) est plus faible que la figure 17 sans les données de la Covid (-0.2). Les valeurs des paramètres sont encore plus incertaines avec les données Covid.

F. Modèle VAR :

$$y_t = v + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t$$

Tableau 19 : VAR avec données Covid (2021-09 à 2022-09) :

***Les paramètres de l'équation **CLAIMSx** ne sont pas présentés.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Sun, 24, Oct, 2021
Time:                     12:55:04
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -7.27681
Nobs:                     722.000    HQIC:                  -7.37812
Log likelihood:           663.548    FPE:                0.000586225
AIC:                     -7.44181    Det(Omega_mle):     0.000565671
--------------------------------------------------------------------
Results for equation UNRATE_CHG
================================================================================
                    coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const                -0.002612         0.008922           -0.293           0.770
L1.UNRATE_CHG        -0.315277         0.039697           -7.942           0.000
L1.CLAIMSx            3.708234         0.090336           41.049           0.000
L2.UNRATE_CHG         0.019589         0.041003            0.478           0.633
L2.CLAIMSx            0.010168         0.167737            0.061           0.952
L3.UNRATE_CHG         0.080577         0.041602            1.937           0.053
L3.CLAIMSx            0.192557         0.166399            1.157           0.247
L4.UNRATE_CHG         0.017761         0.042137            0.422           0.673
L4.CLAIMSx           -0.274620         0.170781           -1.608           0.108
L5.UNRATE_CHG         0.150678         0.041538            3.627           0.000
L5.CLAIMSx           -0.400077         0.168789           -2.370           0.018
L6.UNRATE_CHG        -0.008743         0.025184           -0.347           0.728
L6.CLAIMSx           -0.199079         0.164726           -1.209           0.227
================================================================================
```

Figure 18 : Graphique VAR avec données Covid (2021-09 à 2022-09) :

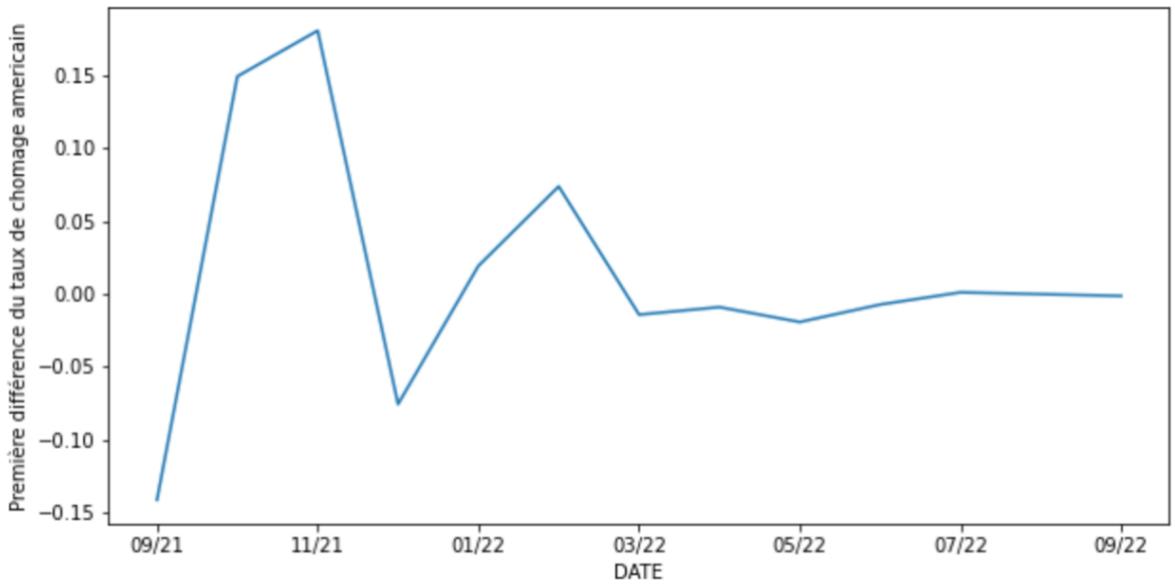
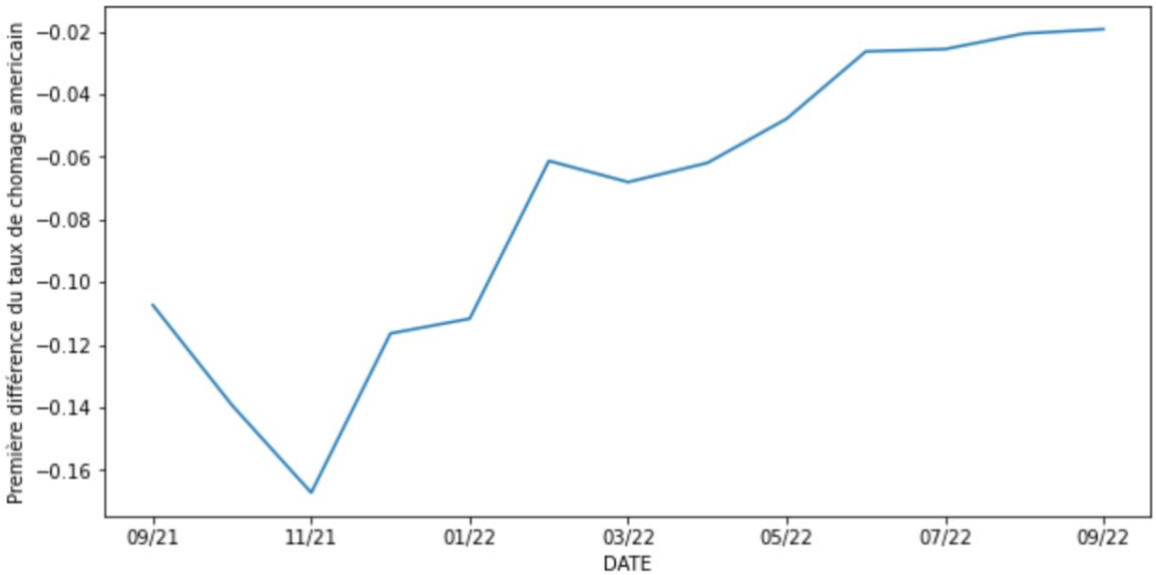


Tableau 20 : VAR sans données Covid :

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Sun, 24, Oct, 2021
Time:                     12:55:04
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                   -9.79573
Nobs:                     705.000    HQIC:                  -9.89887
Log likelihood:           1537.55    FPE:                4.70723e-05
AIC:                     -9.96384    Det(Omega_mle):     4.53831e-05
--------------------------------------------------------------------
Results for equation UNRATE_CHG
================================================================================
                    coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const                -0.002173         0.005685           -0.382           0.702
L1.UNRATE_CHG        -0.170869         0.039137           -4.366           0.000
L1.CLAIMSx            1.073411         0.127034            8.450           0.000
L2.UNRATE_CHG         0.009487         0.039917            0.238           0.812
L2.CLAIMSx            0.907066         0.131521            6.897           0.000
L3.UNRATE_CHG         0.058713         0.038778            1.514           0.130
L3.CLAIMSx            0.704187         0.133887            5.260           0.000
L4.UNRATE_CHG         0.128469         0.037429            3.432           0.001
L4.CLAIMSx            0.168576         0.134262            1.256           0.209
L5.UNRATE_CHG         0.124662         0.036663            3.400           0.001
L5.CLAIMSx            0.195611         0.132125            1.480           0.139
L6.UNRATE_CHG         0.140589         0.036780            3.822           0.000
L6.CLAIMSx            0.171404         0.130314            1.315           0.188
================================================================================
```

Figure 19 : Graphique VAR sans données Covid (2021-09 à 2022-09) :



Comme les autre modèles, les prévisions sont très distinctes dans les deux approches pour les mêmes raisons. Cependant, comme dans l'entrainement, le modèle VAR semble mieux reproduire la variance historique de notre variable. Également, on peut noter que le point de départ de la figure 16 avec les données du Covid (-0.15) est plus faible que la figure 17 sans les données du Covid (-0.10).

PARTIE III - Discussion/Conclusion :

La prévision de variables macroéconomiques est très complexe. Nous tentons d'approximer un processus générateur de données à l'aide de modèle statistiques, mais la taille des échantillons est faible et le bruit est souvent élevé. De plus, des évènements de type *black swan*, comme la Covid, sont quasi-impossible à prévoir. En ce sens, nos prévisions pour les 12 prochains mois seront probablement peu précises. Il est tout de même pertinent d'examiner l'exercice ci-dessus dans son ensemble.

L'objectif était de prévoir le taux de chômage américain. Nous avons opté pour la première différence de cette variable vu sa stationnarité et la dessaisonalisation effectué préalablement.

Dans un premier temps, nous avons entrainé et présenté six modèles soient : Moyenne historique, AR, À facteur, ARMA ARDL et VAR. Tous ces modèles ont relativement bien performé pour prévoir notre variable en *pseudo out of sample*. Le VAR et le AR(4) ont particulièrement bien minimiser la moyenne des erreurs au carré. Cependant, toutes les prévisions avaient une variance nettement inférieure aux réelles valeurs. La source de cette erreur systématique est probablement due à nos modélisations.

Dans un deuxième temps, nous avons prévu notre variable dans le futur, de septembre 2021 à septembre 2022. Nous avons effectué cet exercice 12 fois, avec et sans la période de la Covid, et ce, pour le meilleur modèle de chaque catégorie. Les prévisions sont très différentes entre les différents modèles ayant les mêmes données. Cela est normal vu que l'hypothèse de processus générateur de données est différente entre les modèles. Les prévisions sont très différentes entre le même modèle, même en incluant ou en excluant les données de pandémie. Comme expliqué dans la section précédente, cela est due aux valeurs des paramètres et de *lags* différents. Ce constat amène une question importante pour la prévision. Que devrions-nous faire des données résultats de *black swan* pour la prévision ? Il semble logique d'exclure ces

valeurs aberrantes pour approximer le processus générateur de données. Pour confirmer ou infirmer cette hypothèse, il faudrait comparer ces deux cas à travers plusieurs modèles dans un exercice *pseudo out of sample.* Pour nos prévisions, il faudra attendre quelques mois pour voir la meilleure approche.

Notre approche pour la création des prévisions comporte quelques limites. La quantité de données disponibles en est une comme mentionné antérieurement. Également, rien ne garantit que nos modèles soient les meilleurs modèles disponibles pour prévoir notre variable d'intérêt. Il est possible que des modèles d'apprentissage automatique non supervisé soient plus avantageux que les modèles utilisés dans le cadre de ce travail pratique. Il serait donc pertinent de refaire cet exercice en utilisant des modèles différents de ce type.

BIBLIOGRAPHIE :

- Jonas Kibala Kuma. Modélisation ARDL, Test de cointégration aux bornes et Approche de TodaYamamoto : éléments de théorie et pratiques sur logiciels. Licence. Congo-Kinshasa. 2018. ffcel01766214f
- McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. Journal of Business and Economic Statistics, 34(4):574–589
- Massimiliano Marcellino « An Introduction to Factor Modelling » Bocconi University, 2017
- U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, October 17, 2021.