COMPTE RENDU
Projet series temporelles

El Hadrami N'DOYE, Ismaïl RAMDÉ et MARAME DIAGNE

15 Fevrier 2022

1 Introduction

Dans le cadre du module séries temporelles, nous nous intéressons à l'analyse des données temporelles. Les données que nous allons étudier portent sur la température mensuelle (°C) de trois villes différentes du Brésil (São Paulo, Rio et Salvador). Notre objectif sera de comprendre le passé en analysant et en comparant les valeurs des températures observées dans les trois villes. En plus de cela, nous mettrons en place des méthodes d'apprentissage afin de faire des prédictions pour des valeurs non observées. Pour ce faire nous allons d'abord faire une statistique descriptive sur les différents jeux de données afin de voir le comportement des séries dans le temps. Ensuite nous réaliserons différentes prédictions, notamment les méthodes de lissages, ARMA, ARIMA et SARIMA puis à la fin nous sélectionnons le meilleur modèle en se basant sur quelques critères de qualité notamment l'AIC et BIC pour les modèles ARIMA et le RMSE pour les modèles de lissage.

2 Étude descriptive des données

Chaque jeu de données contient une série chronologique de températures pour des stations faisant référence à trois villes du Brésil (São Paulo, Rio et Salvador). Les séries chronologiques fournissent des enregistrements de températures (une mesure moyenne est calculée) par mois pour chaque année.
La série de Sao-Paulo contenait des mesures de température entre 1945 - 2019, celle de Rio entre 1973 - 2019 et celle de Salvador des mesures de températures entre 1961 - 2019. La figure ci-dessous nous montre un aperçu de l'une des séries :

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1973	24.51	25.18	22.22	23.85	18.73	18.97	18.42	17.28	18.28	19.44	19.82	22.63
1974	23.22	24.54	23.06	20.21	18.99	16.46	17.84	18.97	19.28	19.50	21.29	21.30
1975	22.74	24.16	23.81	20.31	18.16	17.59	16.10	21.05	20.27	20.70	21.15	23.46
1976	24.06	22.38	22.98	20.83	18.29	17.16	16.26	17.45	17.61	19.18	21.59	22.54
1977	23.57	25.54	24.14	20.51	18.94	18.53	20.25	20.44	20.29	21.99	22.16	21.41
1978	24.34	23.59	23.66	20.28	17.72	17.14	18.37	17.67	19.10	21.56	21.08	22.15
1979	20.78	23.42	999.90	20.22	19.33	999.90	999.90	19.63	18.37	20.68	21.11	22.80
1980	22.54	23.54	24.76	21.61	20.58	17.05	18.40	18.56	17.58	21.09	21.46	999.90
1981	999.90	24.97	23.23	999.90	20.05	17.01	15.55	18.35	20.78	19.25	22.47	22.24
1982	21.69	999.90	999.90	19.77	18.07	19.25	18.30	19.33	19.71	20.80	23.28	21.72
1983	23.02	23.62	21.72	21.22	19.62	17.32	18.02	17.52	16.42	19.32	21.82	22.42
1984	25.13	26.40	23.47	20.58	21.19	18.93	18.61	17.07	18.51	21.95	22.35	21.92
1985	21.90	24.37	23.74	22.34	18.85	16.45	16.47	19.51	19.22	21.50	22.54	23.00
1986	24.84	24.09	23.45	22.52	20.73	17.70	16.93	18.90	19.10	20.63	22.84	22.74
1987	24.28	24.01	22.73	22.71	18.35	16.75	19.61	17.54	18.06	21.25	22.58	23.52
1988	25.26	22.82	23.73	21.44	19.03	16.12	14.78	18.97	20.88	19.95	20.70	23.14

Figure 1: Représentation sous R des données

2.1 Pré-traitement

Le but de ce pré-traitement étant entre autres la comparaison du comportement des températures entre les trois villes, nous avons uniformisé les différentes séries en prenant les mesures de températures (janvier à décembre) entre les années 1973 et 2019. Nous avons ensuite supprimé certaines valeurs qui ne semblaient pas logique, par exemple les températures qui sont égales a 999.9(voir figure 14), puis nous les avons remplacées par la médiane des températures mensuelles au lieu de la moyenne pour être plus précis.

2.2 Visualisation des données

2.2.1 Séries (São Paulo, Rio et Salvador)

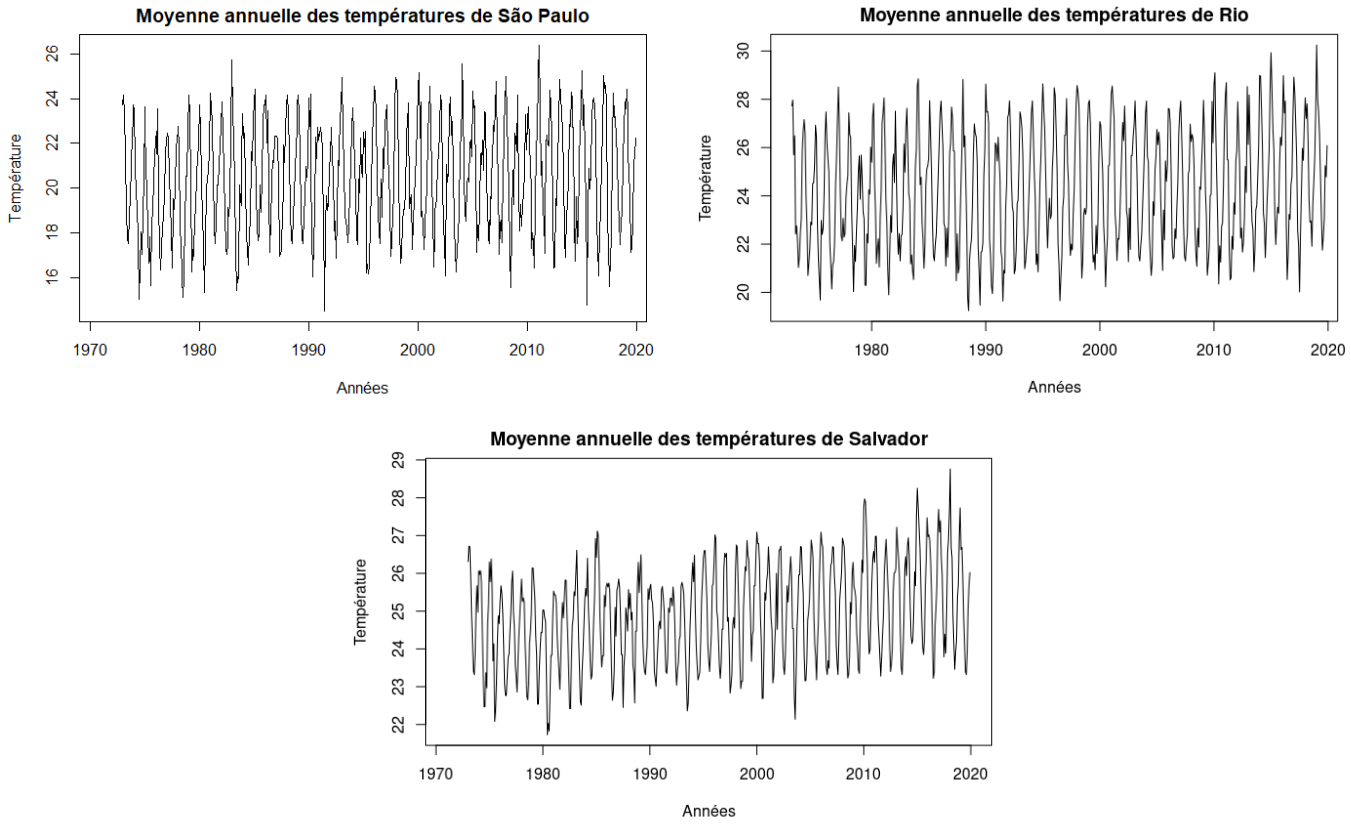

Figure 2: Représentation des trois séries

À partir des représentations graphiques, on remarque que la série de la ville de São Paulo et celle de Rio ont une tendance stationnaire. Quant à la série de Salvador, on voit une tendance plus ou moins constante de 1973 à 1995 et une croissance à partir des années 1995. Les trois séries présentent des variations entre les observations qui pourraient être associées à une saisonnalité. Nous vérifions cette hypothèse ainsi que celle de la tendance grâce aux graphes d'auto-corrélations ou en s'intéressant à une plus petite portion de temps et en faisant des ajustements des différentes tendances.

Moyennes des séries désaisonnalisées

Moyennes		
São Paulo	Rio	Salvador
20.47	23.28	23.92

La moyenne de la série désaisonnalisée la plus basse est celle de la ville de São Paulo (20.47), ensuite celle de Rio (23.28). La plus haute moyenne est celle de la ville de Salvador (23.92).

2.2.2 Graphique des auto-corrélations

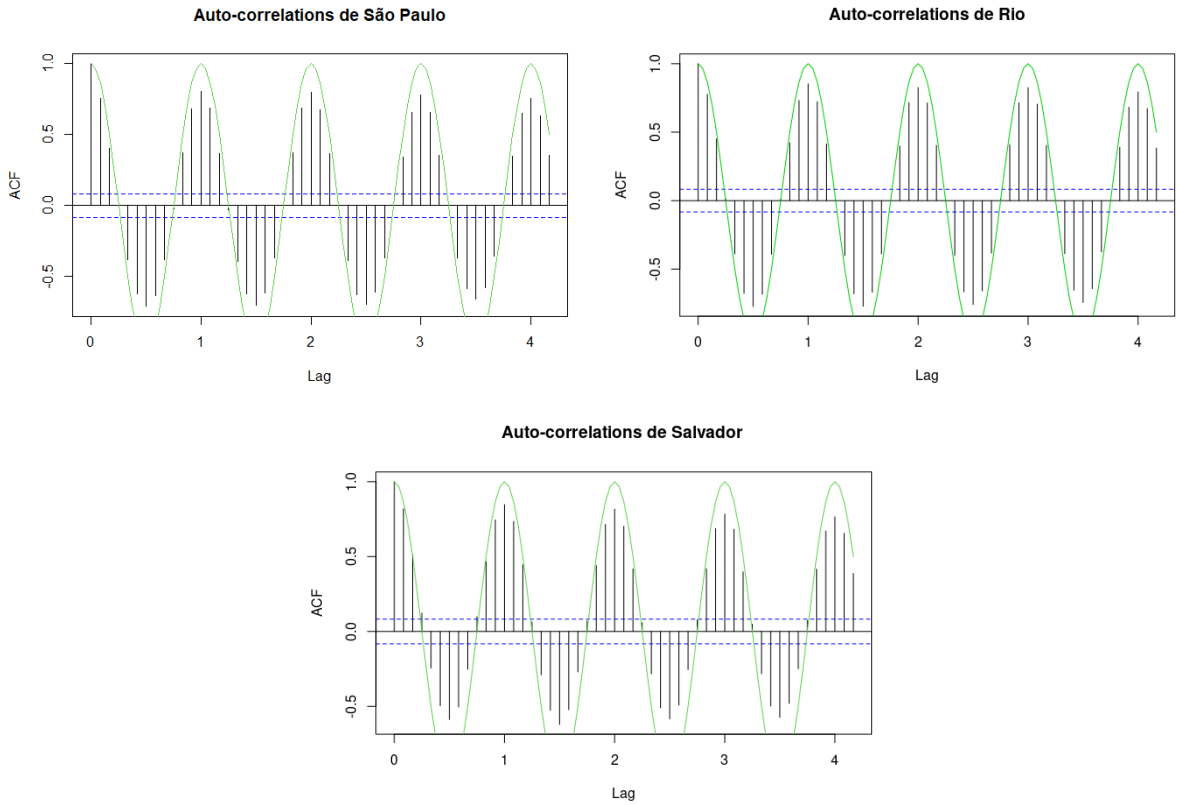

Figure 3: Représentation des auto-corrélations

On fait les mêmes observations sur les séries des trois villes. En effet, pour une suite de fonctions d'auto-corrélation à 50, on observe que la fonction périodique (période de 12) représentée en vert est bien similaire aux fonctions d'auto-corrélations. On peut donc valider notre première hypothèse selon laquelle les trois séries sont toutes saisonnières.

2.2.3 Ajustement de la tendance

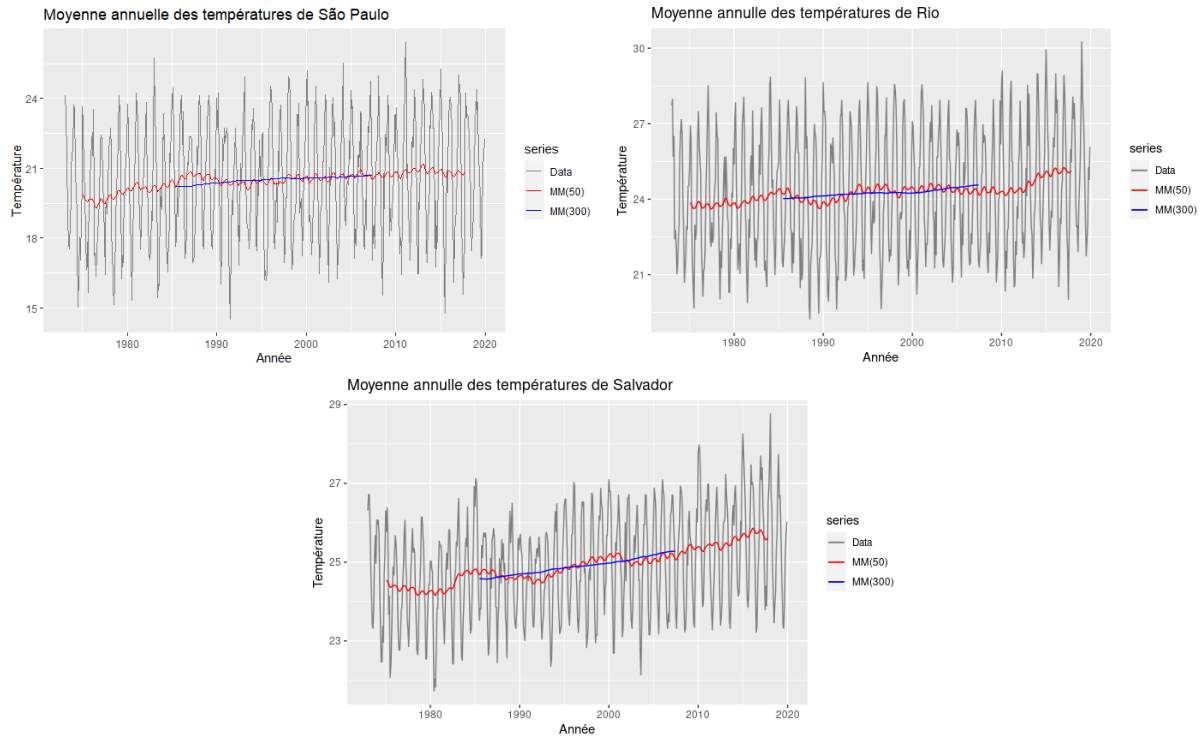

Figure 4: Représentation de l'ajustement de la tendance

Dans cette section nous effectuons un ajustement de tendance à l'aide des moyennes mobiles (MM) afin de faire ressortir les tendances respectives.

On constate que les séries de São Paulo et de Rio ont une tendance qu'on peut qualifier de stationnaire (constantes) tandis que celle de la ville de Salvador est plutôt légèrement croissante ce qui nous amènera plus tard de l'estimer par un polynôme de degré compris entre 1 et 3.

2.2.4 Évolutions mensuelles (saisonnières) par année entre 1973 et 2019

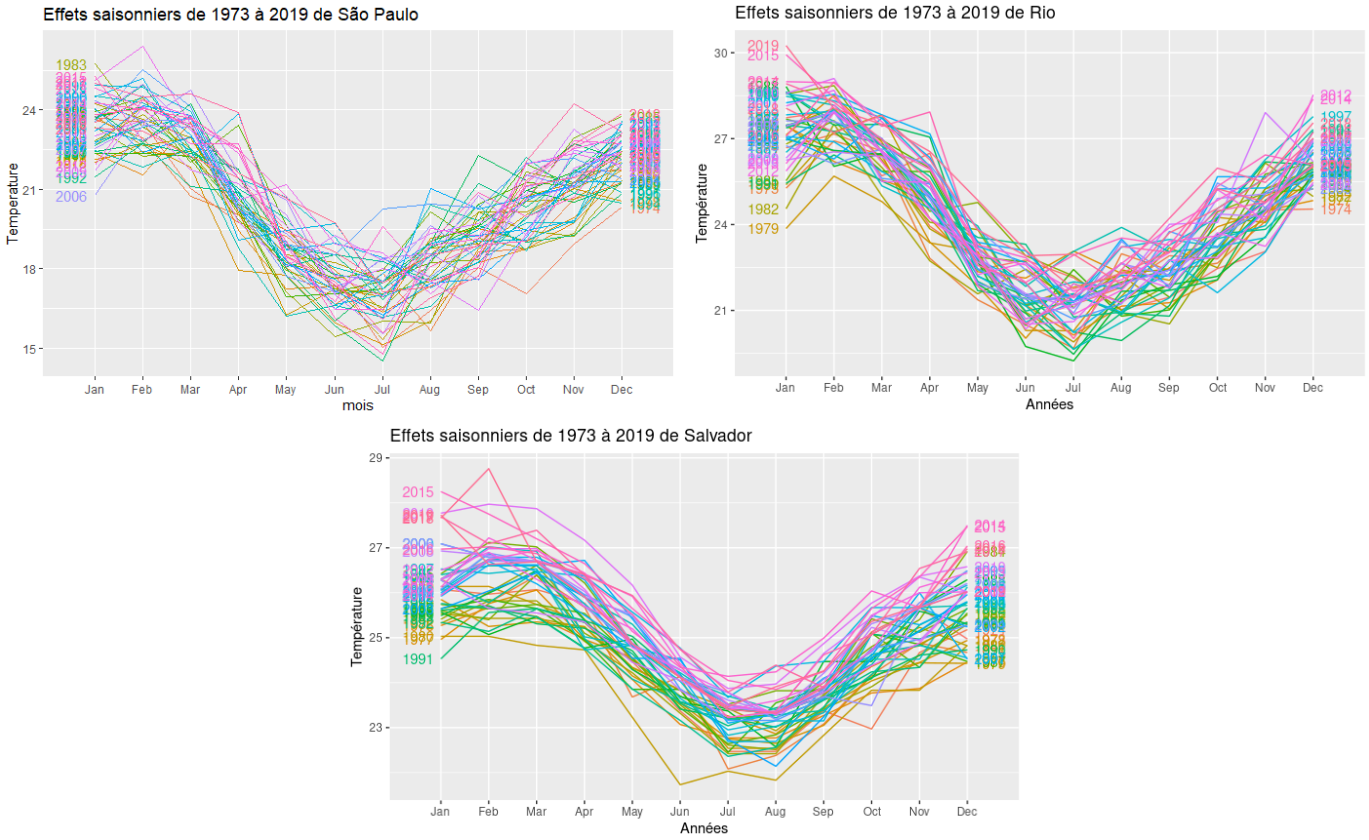

Figure 5: Représentation des évolutions mensuelles

Pour les trois villes, on observe de façon globale, les mêmes évolutions mensuelles entre les années 1973 et 2019. Les températures les plus élevées se situent aux mois de décembre, janvier, février et mars. Et les plus basses températures sont enregistrées autour des mois de juin, juillet et août.

3 Estimation et filtrage de la partie déterministe

3.1 Les méthodes de lissages

Choix du meilleur modèle de lissage

RMSE pour São Paulo				
LES	LED	HWNS	HWSA	HWSM
1.49	1.82	2.86	1.16	1.2
RMSE pour Rio				
LES	LED	HWNS	HWSA	HWSM
2.04	1.61	4.11	0.72	0.74
RMSE pour Salvador				
LES	LED	HWNS	HWSA	HWSM
0.25	3.60	1.23	0.02	0.06

Nous avons répertorié dans le tableau ci-dessus l'erreur quadratique moyenne (RMSE) de 5 méthodes de lissage que nous avons implémentées. Pour ce faire nous avons divisé chaque jeu de données en données d'apprentissage sur lesquels nous avons appris les modèles et de tests sur lesquels nous avons testé les modèles. Par la suite, nous avons calculé le RMSE de chaque méthode afin d'en choisir le meilleur modèle (RMSE petit). Les méthodes de lissages utilisées sont :

- lissage exponentiel simple (LES)
- lissage exponentiel double (LED)
- lissage de HoltWinters non saisonnier (HWNS)
- lissage de HoltWinters saisonnier additif (HWSA)
- lissage de HoltWinters saisonnier multiplicatif (HWSM)

Graphique du modèle choisi pour les 3 séries

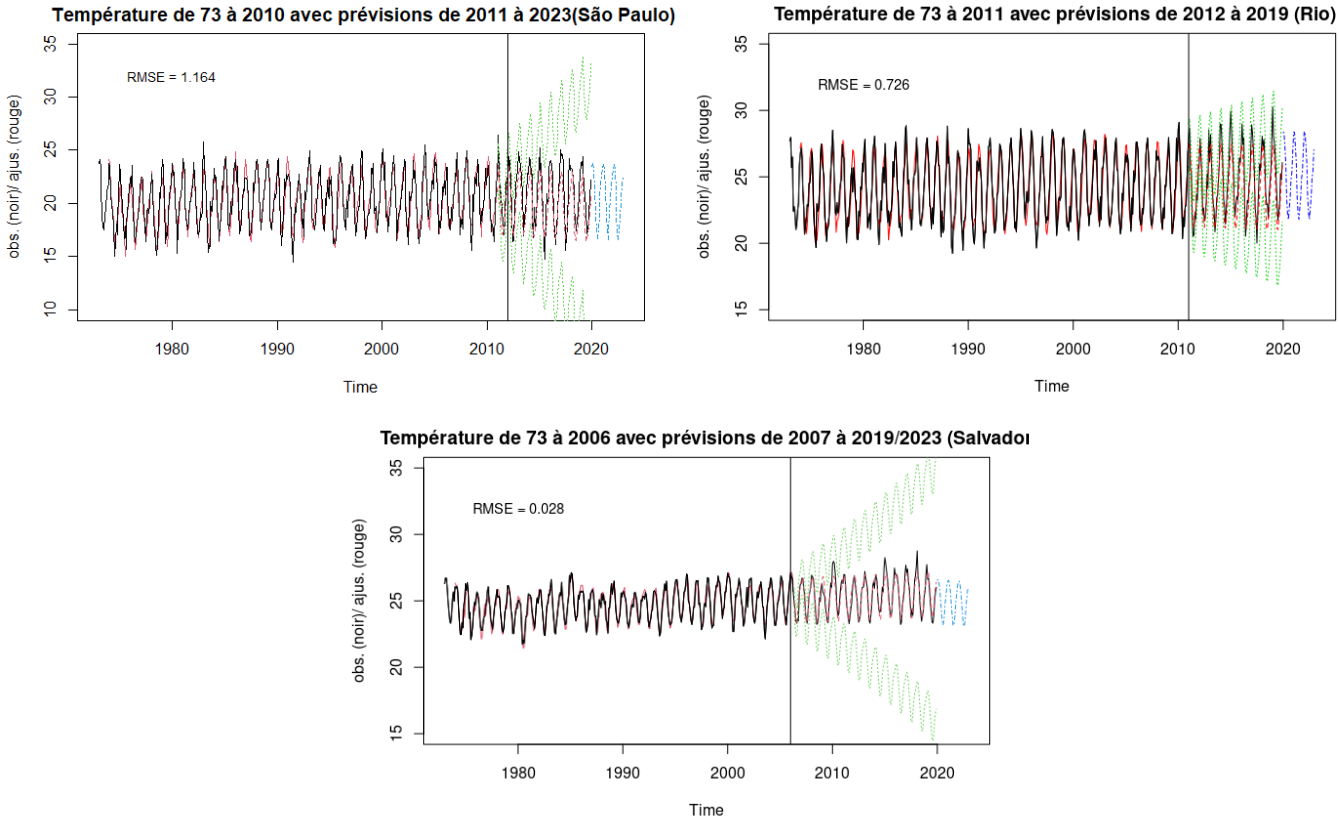

Figure 6: lissage de HoltWinters saisonnier additif

À l'issue de de ces 5 méthodes, la meilleure prédiction est obtenue avec le lissage de HoltWinters saisonnier et additif (HWSA) pour les villes de São Paulo, Rio et Salvador avec respectivement un RMSE égal à 1.16, 0.72 et 0.02. Sur les trois graphiques, la serie d'origine a été représentée en noir et les prévisions du lissage de HoltWinters saisonnier additif (HWSA) en rouge. On peut voir aussi l'intervalle de confiance en vert. Nous avons représenté des prévisions pour h = 36 (3 ans) en bleu pour voir le comportement des températures sur jusqu'en 2023. On constate que pour les prévisions sur ces 3 années, les températures ne présentent pas de hausses ou de baisses significatives pour les trois villes.

3.2 ARMA, ARIMA et SARIMA

A l'aide de l'acf et pacf nous avons estimés les coefficients p et q afin de réaliser les méthodes ARMA, ARIMA et SARIMA.

3.2.1 Sélection du meilleur modèle de São Paulo

Table 1: Les différents modèles effectués

Modèle	BIC	AIC
SARIMA(1,0,2)(1,0,2)[12]	1694.81	1660.13
AUTO.ARIMA(3,0,0)(2,1,2)[12]	1645.02	1610.51

A l'issue de ces deux modèles, le meilleur modèle est obtenue par la méthode AUTO.ARIMA car elle minimise le BIC et le AIC. A la suite nous allons vérifier si les résidus du modèle auto.arima sont satisfaisants.

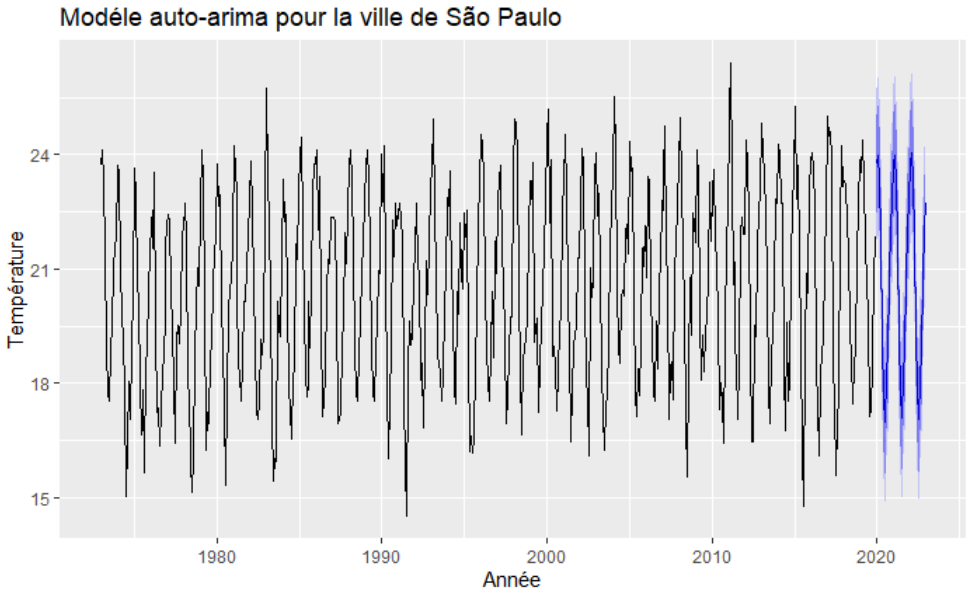

Figure 7: Modèle AUTO ARIMA et prévision (2020 - 2023)

Sur la figure 7 on voit la prédiction de la température dans la ville de São Paulo sur 3 années consécutives suivant la dernière observation.

3.2.2 Sélection du meilleur modèle de Rio

Table 2: Les différents modèles effectués

Modèle	BIC	AIC
SARIMA(1,0,3)(1,0,3)[12]	1536.32	1492.97
AUTO.ARIMA(2,0,2)(2,1,0)[12]	1561.89	1527.38

On choisit le modèle obtenu par SARIMA car il minimise le AIC.

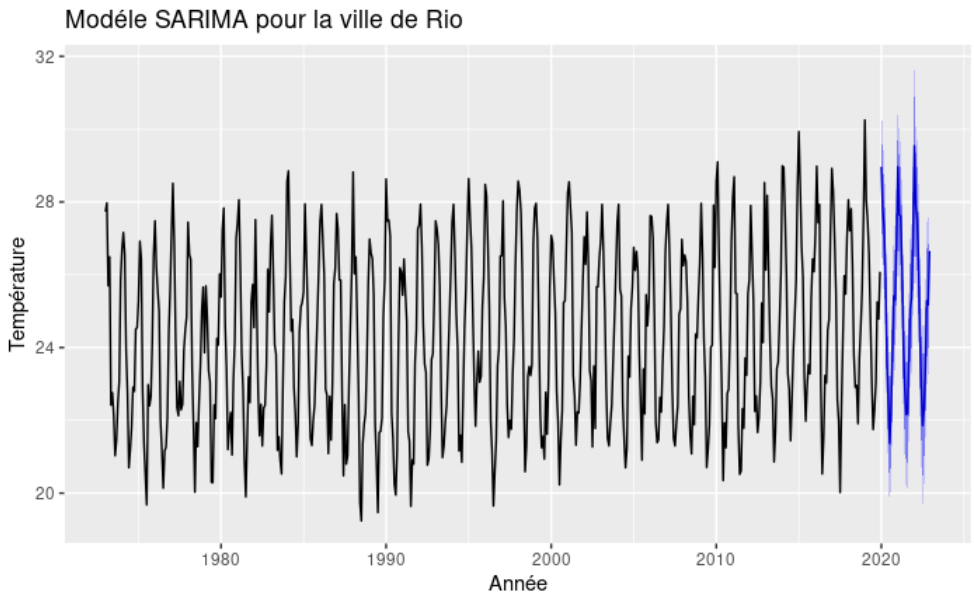

Figure 8: Modèle SARIMA et prévision (2020 - 2023)

3.2.3 Sélection du meilleur modèle de Salvador

Table 3: Les différents modèles effectués

Modèle	BIC	AIC
AR(3)	995.56	973.99
MA(7)	1947.198	928.9
ARMA(3,7)	927.96	876.19
SARIMA(1,1,3)(1,1,3)[12]	692.54	653.73
AUTO.ARIMA(2,0,2)(1,1,0)[12]	823.73	793.53

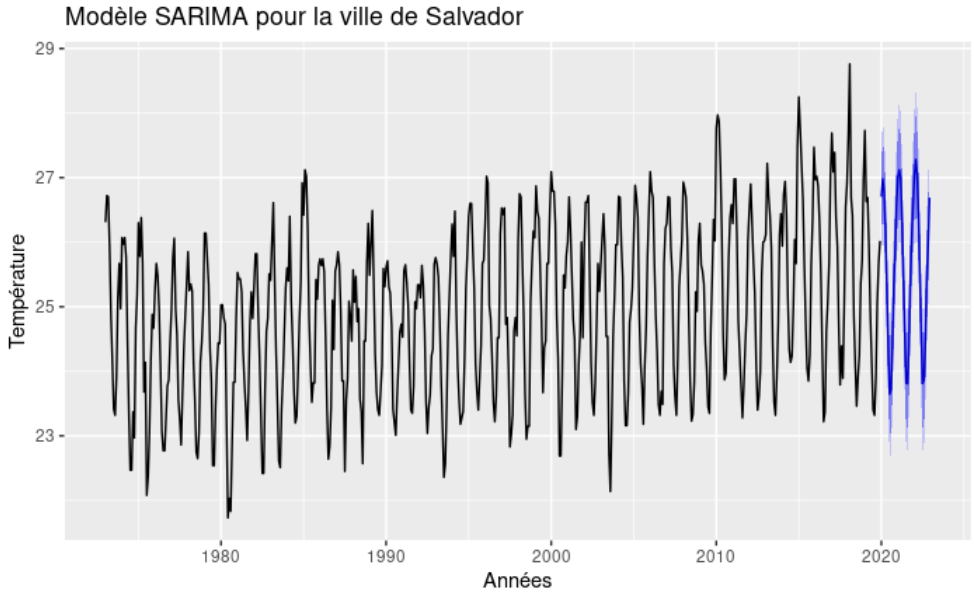

Figure 9: Modèle SARIMA et prévision (2020 - 2023)

Au vu des AICs, le meilleur modèle pour la série de Salvador est le modèle SARIMA représenté ci-dessus avec les prévisions sur 3 ans (2020 - 2023) en bleu.

4 Analyse descriptive de la partie résiduelle

Pour le traitement de la partie résiduelle, nous avons utilisé un Test de blancheur (Box-Pierce ou Box-Ljung unilatéral) pour tester si le résidu est un bruit blanc.

Soient les deux hypothèses suivantes :
H_0 : La série résiduelle est un bruit blanc.
H_1 : La série résiduelle n'est pas un bruit blanc.

4.0.1 Série de São Paulo

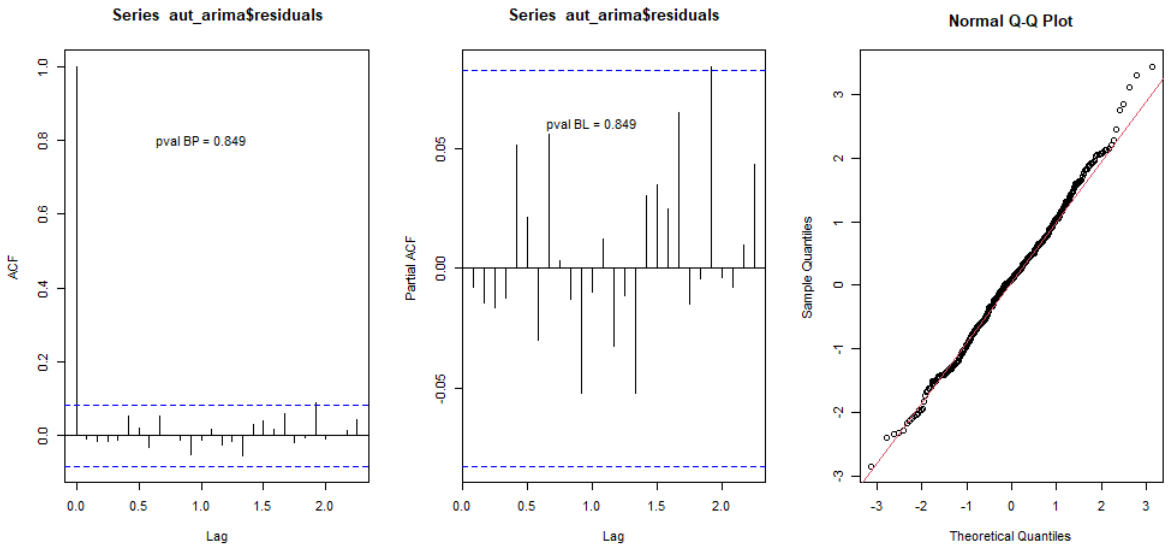

Figure 10: ACF, PACF, P-valeur des residus et qqnorm

La p-valeur obtenue est très grande donc on peut affirmer que la série résiduelle est un bruit blanc.

4.0.2 Série de Rio

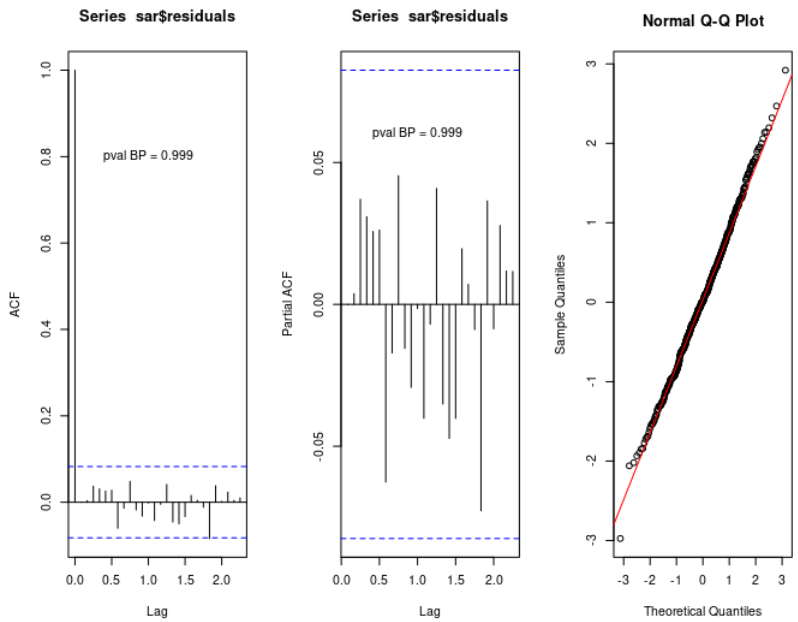

Figure 11: ACF, PACF, P-valeur des residus et qqnorm

La p-valeur indique que la partie résiduelle est un bruit blanc car étant très grande.

4.0.3 Série de Salvador

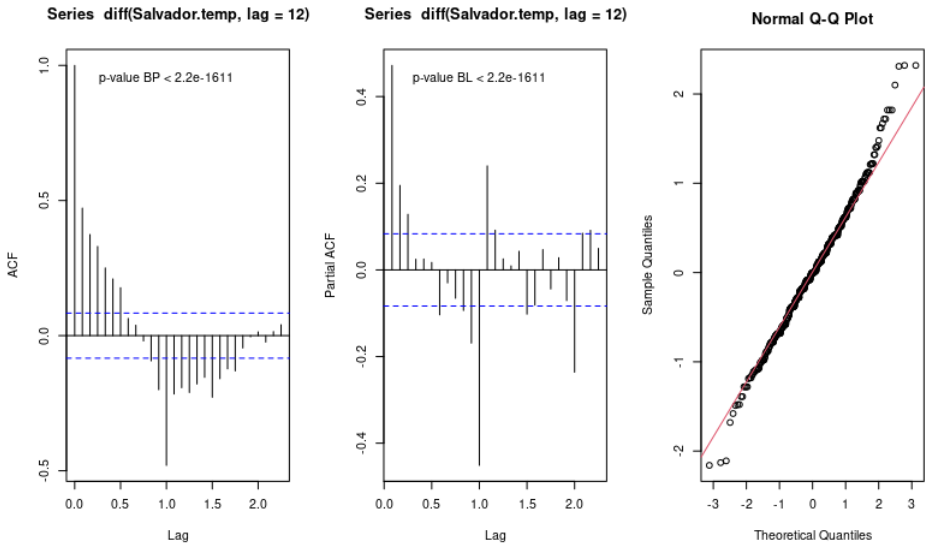

Figure 12: ACF, PACF, P-valeur des residus et qqnorm

On remarque que la p-valeur de Salvador est très petite ($< 2.2e-16$), ce qui signifie que le bruit n'est pas blanc. On peut donc en extraire de l'information.

Analyse de la partie résiduelle de Salvador

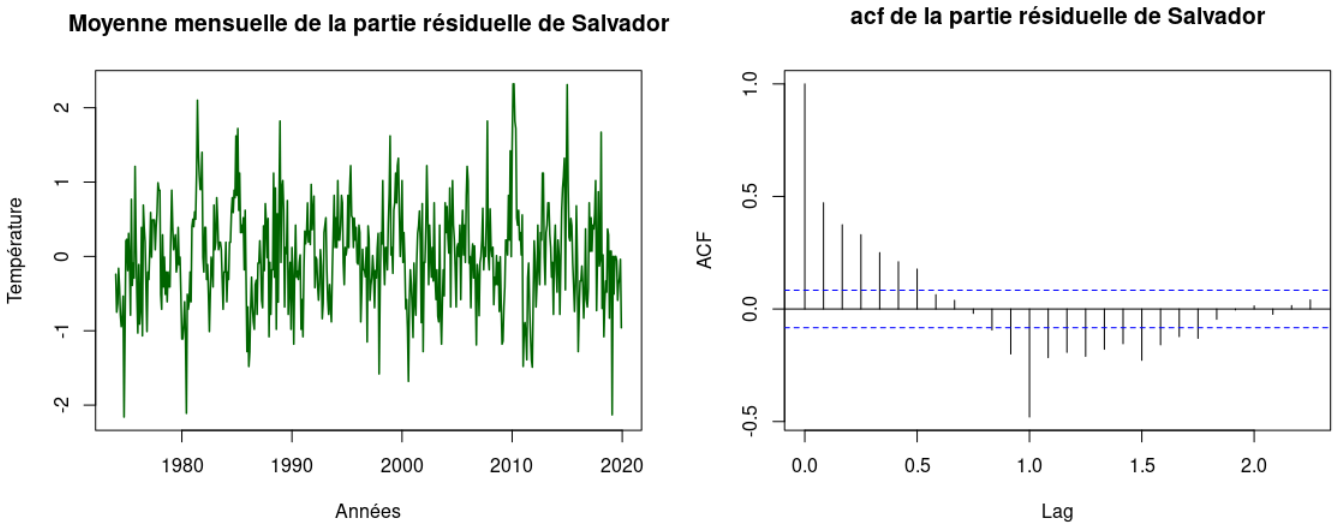

Figure 13: Série résiduelle de Salvador

La représentation de la série résiduelle de la ville de Salvador nous montre qu'elle est stationnaire et désaisonnalisée. Nous pouvons réaliser une estimation ARMA et auto.arima. Nous choisissons $p = 3$ et $q = 7$ à l'aide de l'acf et du pacf pour l'estimation ARMA.

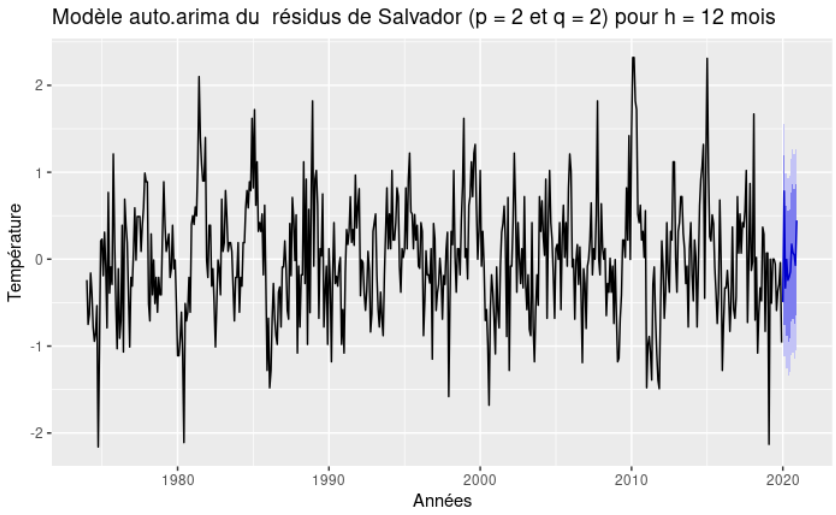

Figure 14: Auto-arima de la partie résiduelle de Salvador

Le meilleur modèle est celui de l'estimation faite avec "auto.arima" qui donne un AIC de 793.53 qui est plus bas que ceux des modèles ARMA(3,7)/ARMA(7,3) égal à 876.19/913.17, AR(3) égal à 973.99 et MA(7) égal à 928.9.

5 Conclusion

En somme, nous pouvons dire que les températures (climat) dans la ville de São Paulo et la ville de Rio sont assez similaires du point de vue de la tendance, la saisonnalité et de l'évolution mensuelle de la température. En outre, les prévisions (2020 - 2023) faites avec le modèle "auto.arima" pour les deux ne présentent pas de changement significatif. Par contre la ville de Salvador avec des températures saisonnières et légèrement croissantes depuis 1995, continuera d'avoir de légères croissances dans les 3 prochaines années selon les prévisions du modèle "SARIMA". Cependant pour aller plus loin et au-delà de l'aspect statistique, il serait intéressant de chercher à comprendre les causes des différences de comportement des températures dans les trois villes.