ENSTA

IP PARIS

STA202 – Séries Chronologiques

Etude du taux de natalité en France entre 1994 et 2021

Maxence JOSEPH
Tia ZOUEIN

Février 2022

Sommaire

I. Présentation du sujet

Nous avons choisi un jeu de données représentant le taux de natalité, c'est-à-dire le nombre de naissances pour 1000 habitants, en France entre le mois de janvier 1994 et le mois de décembre 2021. Ce taux est donné de manière mensuelle ; ainsi, le tableau initial de données comporte 336 lignes correspondant aux nombres de mois entre janvier 1994 et décembre 2021, et 3 colonnes. De plus, les chiffres prennent en compte le territoire de Mayotte à partir de l'année 2014.

Les données ont été trouvées sur le site de l'INSEE, qui collecte, produit, analyse et diffuse des informations sur l'économie et la société française.

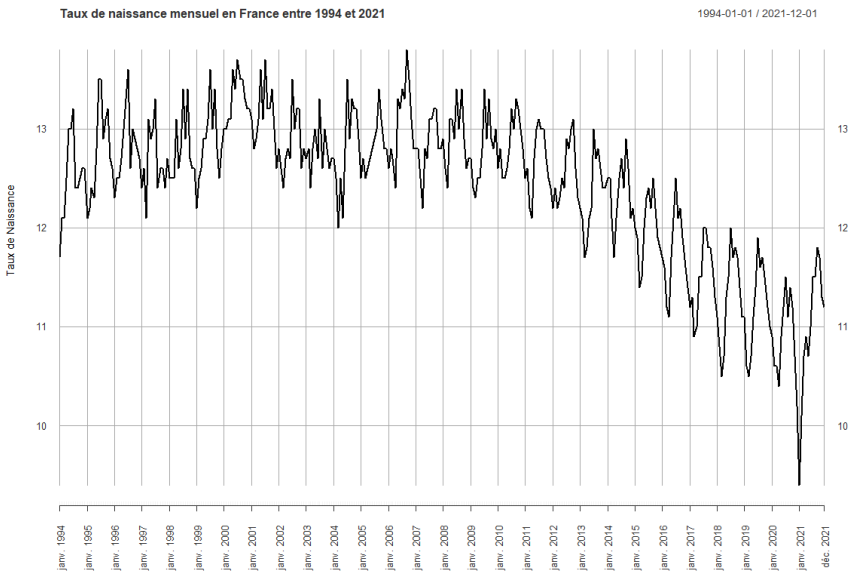
Après avoir téléchargé ce tableau au format .csv (les données peuvent être trouvées à l'adresse suivante : https://www.insee.fr/fr/statistiques/serie/001641602#Tableau), nous remarquons que les trois colonnes nous donnent différentes informations :

- la première colonne correspond à la date au format « année-mois »,
- la seconde colonne nous donne la valeur du taux de natalité pour chacun des mois,
- enfin, la dernière colonne fait correspondre à chaque ligne une lettre. La lettre P signifie que les données communiquées sont provisoires, autrement dit, une mise à jour sera nécessaire ultérieurement. On remarque d'ailleurs que cette lettre concerne toutes les données à partir de Janvier 2019. Avant cette date, toutes les autres lignes se voient attribuer la lettre A, qui a contrario, signifie que les données correspondantes ont été validées.

Nous décidons d'utiliser le logiciel R, et plus particulièrement la libraire *xts* pour l'étude des données. La dernière colonne ne donnant pas d'information pertinente dans un but d'analyse, de modélisation et de prédiction, nous choisissons de les retirer.

Par ailleurs, la création du dataframe nommé *Naissance* nécessite l'utilisation de la fonction *rev* pour inverser les valeurs de la colonne indiquant le Taux de natalité. En effet, la première valeur du tableau fournie par l'INSEE nous donne la valeur du taux associé à décembre 2021 ; or, nous voulons construire notre dataframe de telle sorte à ce que cette première valeur corresponde à Janvier 1994. Une fois cette inversion faite, nous pouvons convertir le dataframe en série temporelle *Naissance.xts*. Dans la suite du projet, nous utiliserons les deux objets en fonction de nos besoins.

Voici un premier aperçu de cette série temporelle :



II. Analyse descriptive des données

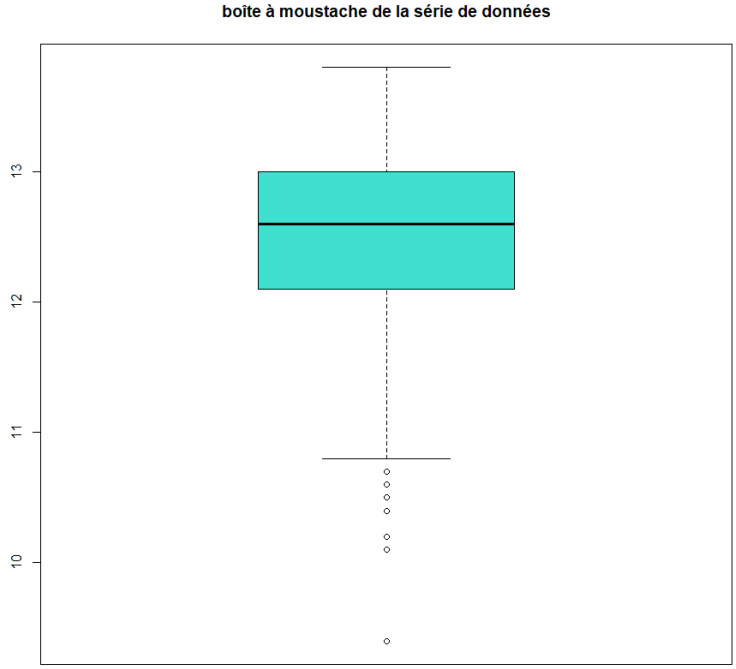
1) ETUDE MENSUELLE ET ANNUELLE

Une première analyse des données est nécessaire avant toute tentative de modélisation et de prédiction.

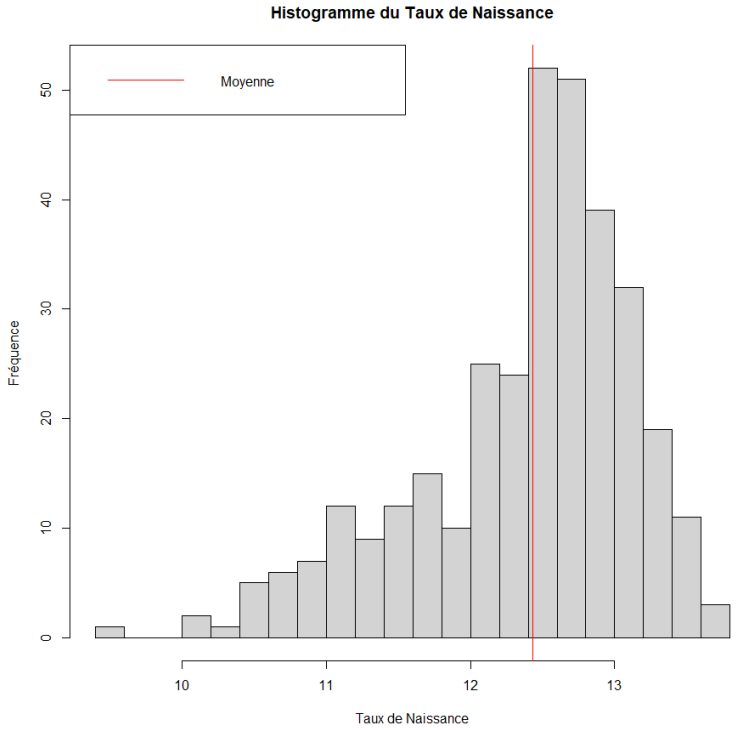
Les principaux indicateurs statistiques sont donnés par la fonction *summary()* et sont regroupés dans le tableau suivant :

	Taux
Minimum	9.40
1er quartile	12.10
Médiane	12.60
Moyenne	12.43
3ième quartile	13.00
Maximum	13.80
Variance (fonction std)	0.76

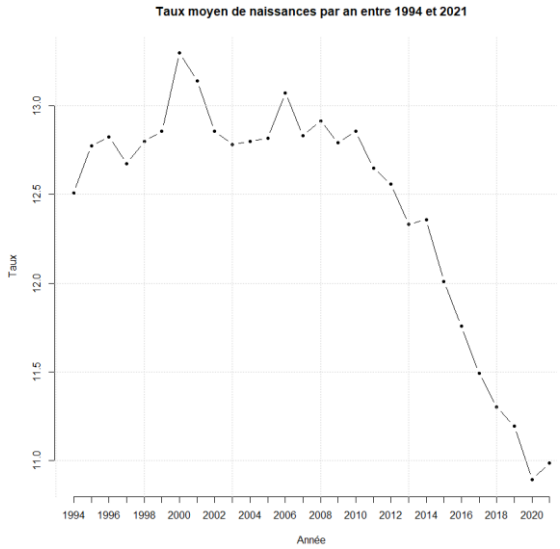
Ces indicateurs peuvent aussi être représentés par une boîte à moustache (boxplot) comme suit :

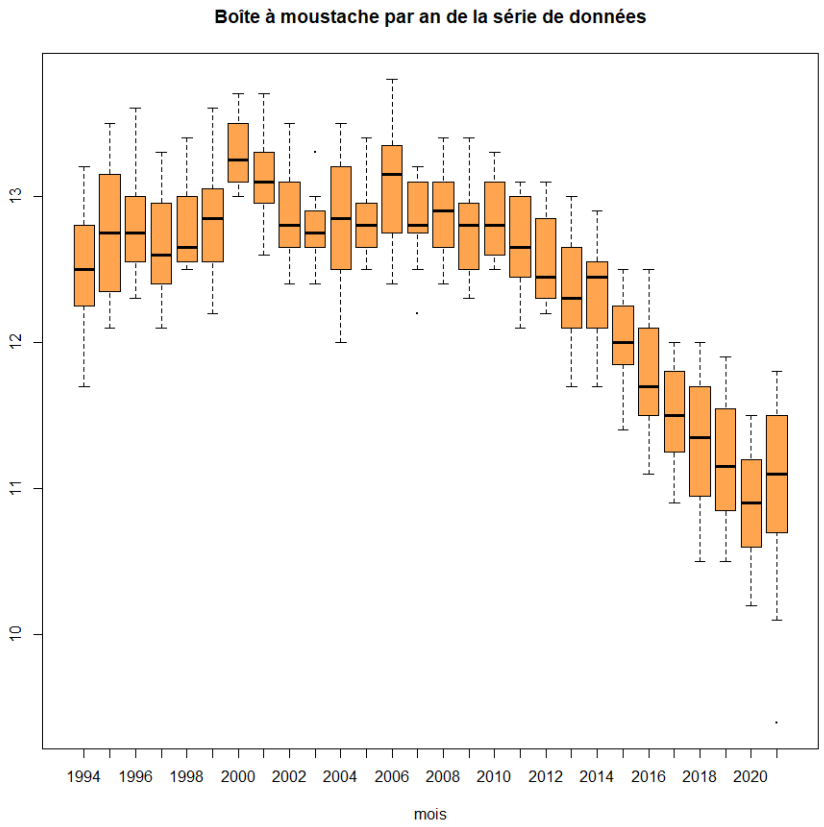


Le tracé de la série temporelle évoque une tendance du taux à la baisse. Les valeurs évoluent dans un intervalle important de valeur, entre 9.40 et 13.80. Ce minimum a été enregistré en Janvier 2021. En outre, cela laisse penser que le confinement de Mars et Avril 2020 mis en place pour lutter contre la crise sanitaire a eu un effet direct sur le nombre de naissance 9 mois plus tard, expliquant probablement la forte baisse enregistrée ce mois-là.



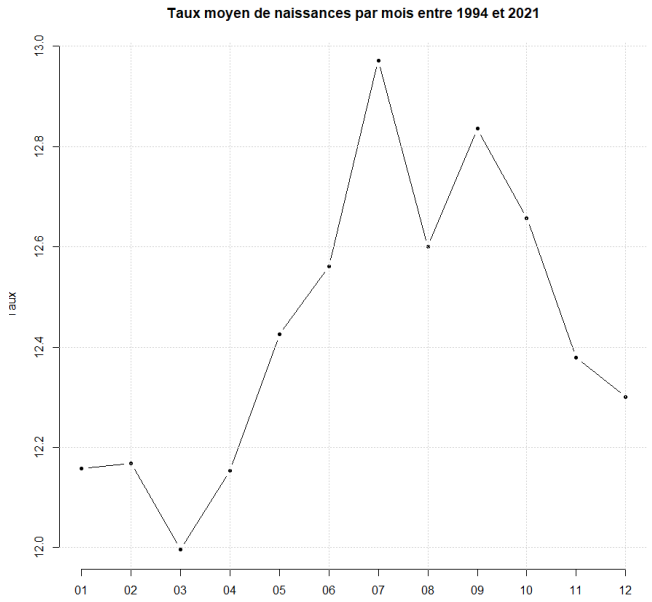
Par ailleurs, la moitié des valeurs enregistrée se trouvent dans l'intervalle [12.10, 13.00], qui demeure un intervalle beaucoup plus petit que celui donné par l'étendue. En réalité, cet intervalle est fortement biaisé par les valeurs enregistrées au début de l'étude, c'est-à-dire entre 1994 et 2010. Il ne montre pas qu'à partir de la moitié de l'année 2016, aucun mois n'enregistre un taux de natalité supérieur à 12.10. On voit d'ailleurs sur l'histogramme ci-dessus que l'étendue des valeurs est bien plus importante à gauche qu'à droite de la moyenne. Observons plus en détails l'évolution de la moyenne du taux de natalité selon les années, ainsi que la boîte à moustache associée :



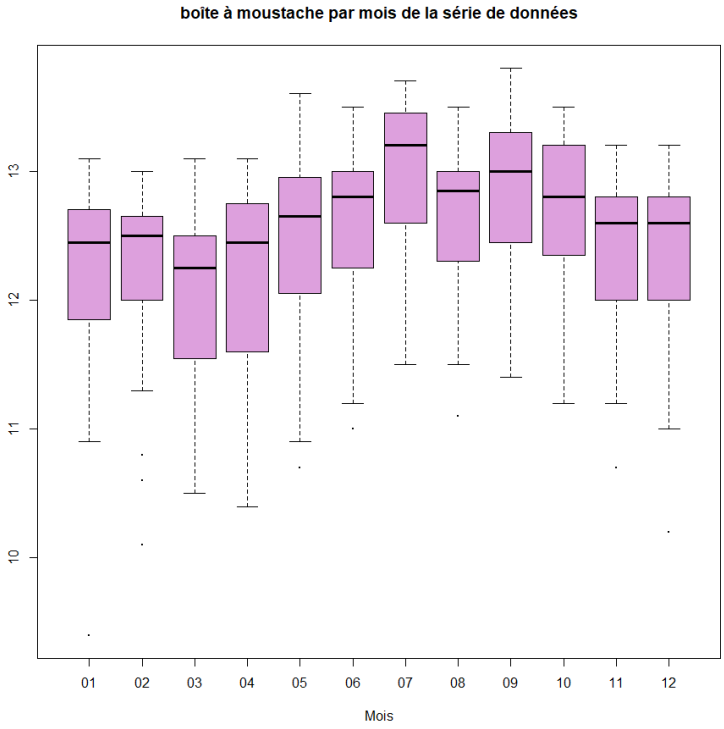


Ces données sont beaucoup plus significatives et donnent un premier aperçu de la tendance générale de l'évolution du taux de natalité. Nous pouvons dès à présent constater très clairement une diminution continue et importante du taux de natalité à partir de l'année 2010. Il semble alors intéressant d'essayer d'expliquer cette baisse et cette fracture du cycle qui semblait stable de 1994 à 2010. Une première hypothèse serait celle d'un lien avec la crise économique de 2008.

Le tracé de la série temporelle suggère également une certaine périodicité des données. En outre, cette période semble égale à T = 12 mois, ce qui paraît plausible. Il semble en effet possible que certaine période de l'année soient propices à la natalité, contrairement à d'autres. Analysons plus précisément l'exactitude de cette information en traçant le graphe correspondant à la moyenne du taux de natalité pour chaque mois de l'année :

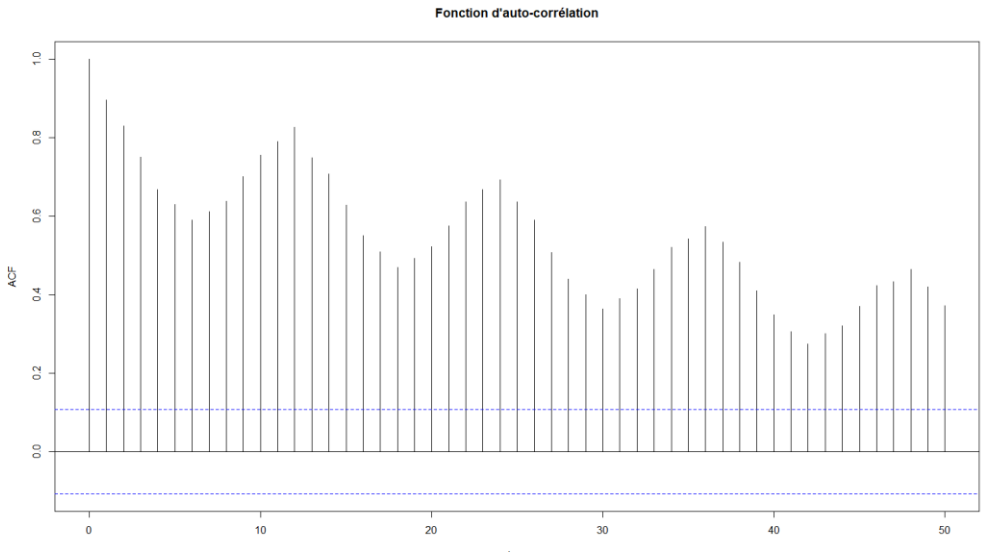


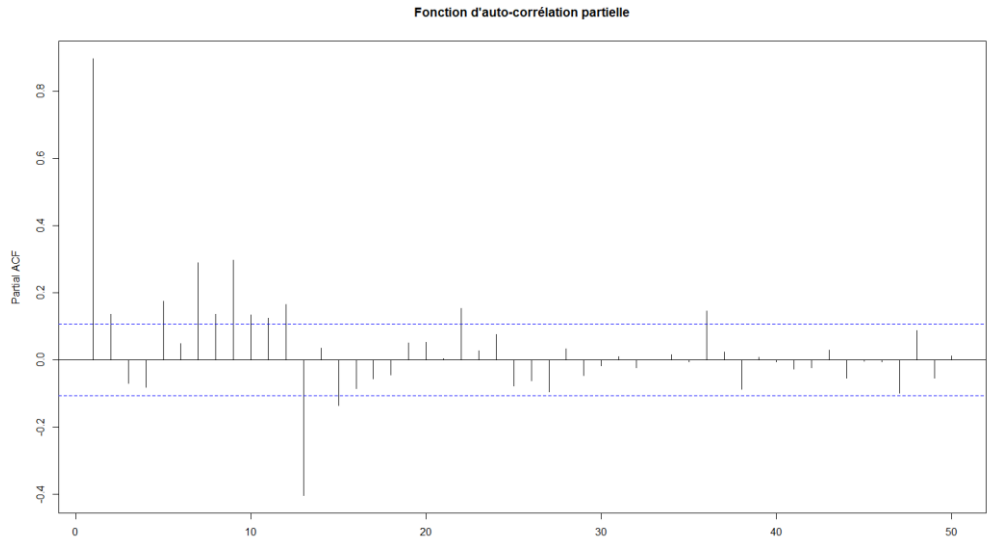
Ainsi, sur la période étudiée (de 1994 à 2021), on constate en effet que ce taux varie selon les mois. La période estivale semble d'avantage propice aux naissances, comme le laissait suggérer le tracé de la série temporelle, qui augmentait systématiquement au milieu de chaque année. On retrouve sur les différentes boîtes à moustache le fait que l'étendue entre le minimum et le 1er quartile est plus conséquent que l'écart entre le maximum et le 3^{ième} quartile. Cela s'explique par la baisse générale du taux de natalité à partir de 2010.



2) AUTO-CORRELATION ET AUTO-CORRELATION PARTIELLE

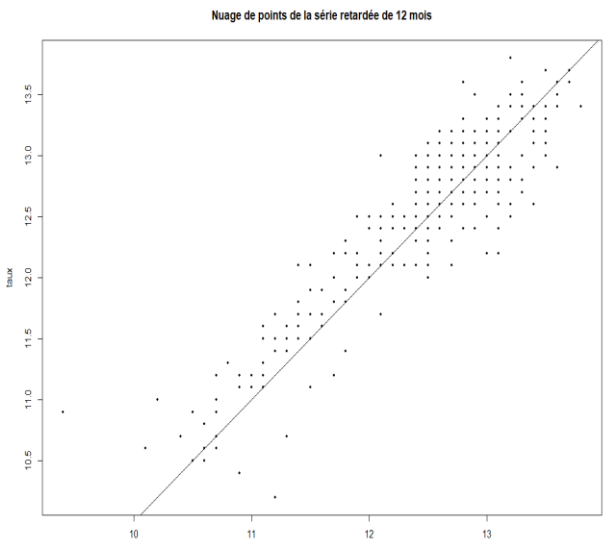
Enfin, afin de déterminer plus précisément le lien entre les données et de préciser notre première intuition concernant la période (estimée à T=12), déterminons l'auto corrélation et l'auto corrélation partielle de la série.

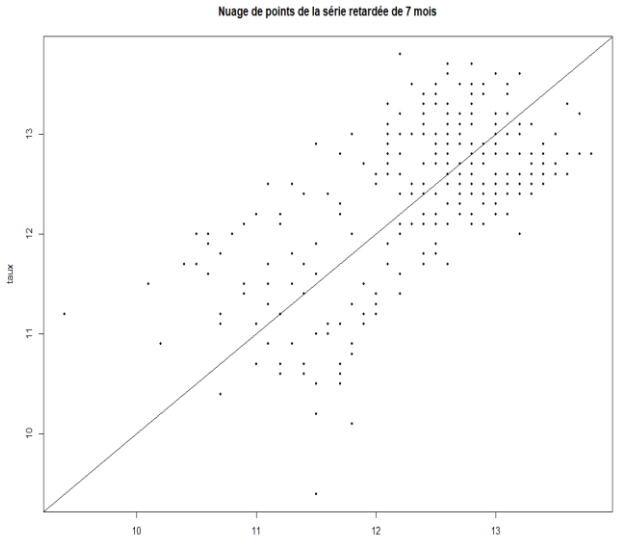




Ces deux graphes nous montrent instantanément que le taux du mois t dépends fortement du taux relevé 12 mois auparavant. De plus, ce taux actuel dépend également de ce qui s'est passé tous les 12*i mois avant, avec i un entier naturel, soit tout ce qui passé les années précédentes au même moment. La décroissance des pics indique par ailleurs que ce qu'il s'est passé au mois t de l'année n-1 influe d'une manière plus significative que ce qui s'est passé au mois t de l'année n-2, etc.… Ceci semble cohérent et confirme l'intuition que nous avions.

Le fait que les données évoluent selon une période de 12 mois est confirmé par le nuage de points de la série retardée de 12 mois, qui admet une représentation presque linéaire, et dont la variance (dispersion des points par rapport à la droite linéaire moyenne) est moindre que tout autre nuage de points de la série retardée. Nous avons ici le graphe du nuage de points de la série retardée de 12 mois avec celui de la série retardée de 7 mois à titre d'exemple.





Cela nous conforte dans l'idée qu'une période de T=12 constitue une bonne approximation de la période qui régit la série temporelle.

III. Modélisation des données

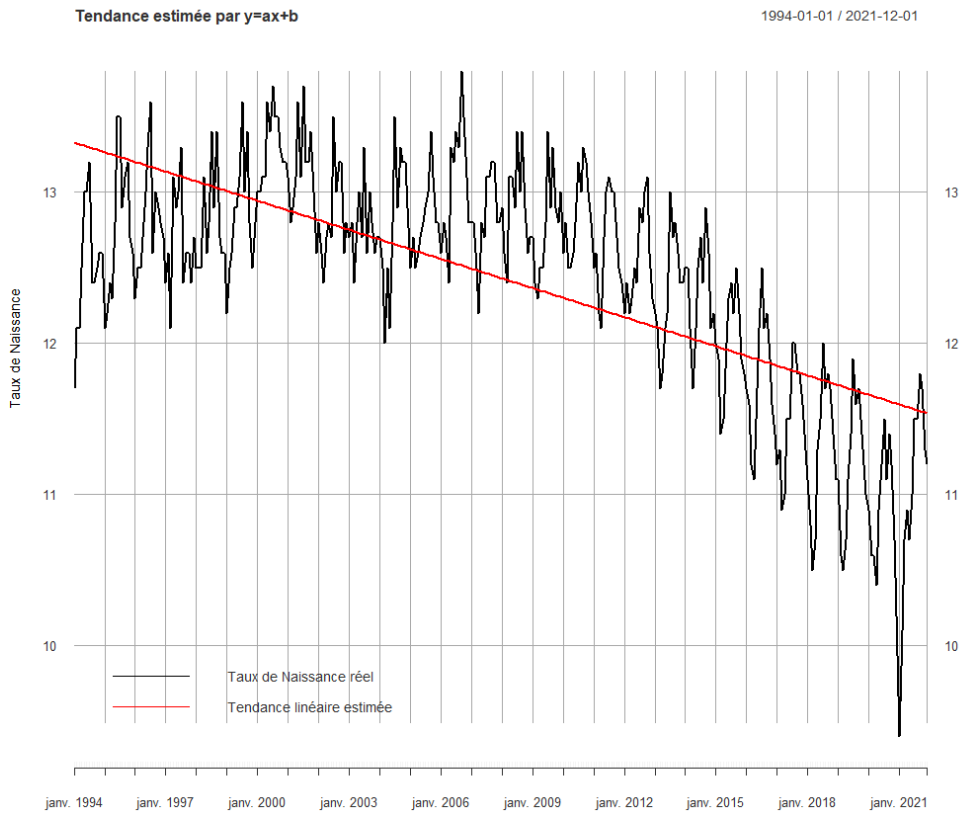
1) Estimation de la tendance

Nous allons maintenant nous pencher sur la question de la modélisation des données. Leur analyse a permis l'émergence de plusieurs conclusions, et nous a donné la possibilité de mieux comprendre leur évolution.

Dans un but de prévision du taux de natalité, nous allons mettre en place plusieurs modèles de modélisation. Celui que nous estimerons le plus performant, c'est-à-dire le plus proche de la réalité, sera sélectionné. Ce critère de performance sera explicité formellement plus loin.

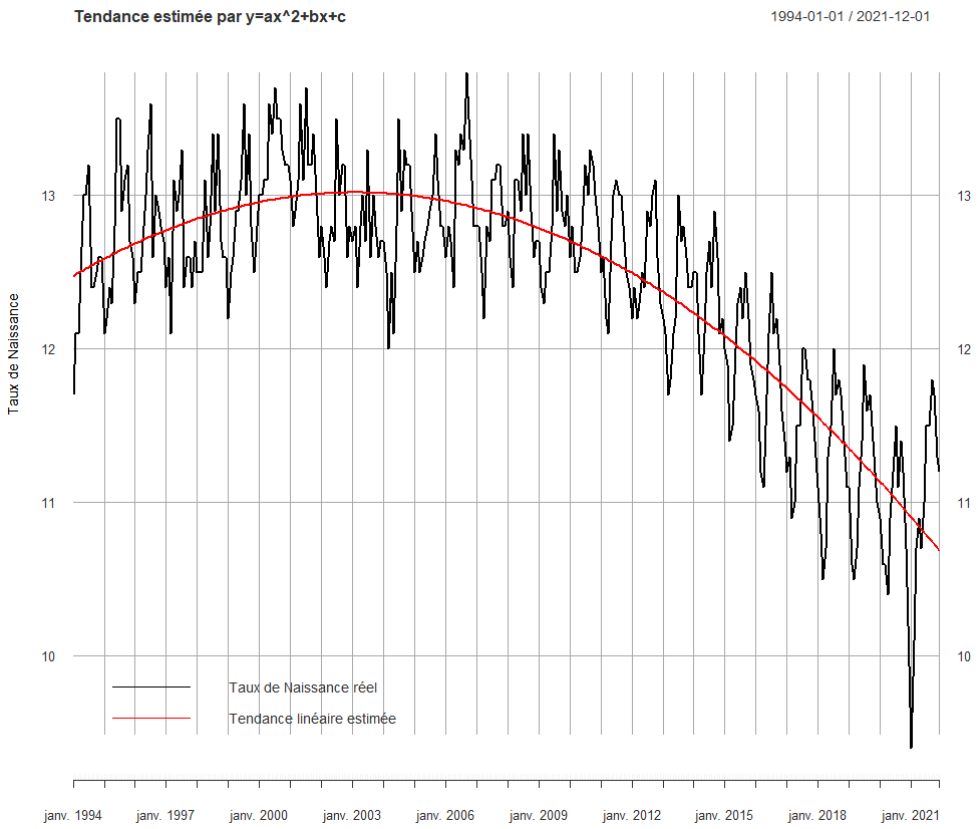
a) Regression Lineaire

Le premier modèle que nous décidons d'implémenter est la régression linéaire. Commençons par approcher le taux de natalité par une fonction affine du temps.



La fonction summary() permet d'obtenir l'équation de droite : Taux = at + b avec a = -0.0053 et b=13.3348.

On réalise de même une régression approchant la tendance par un polynôme de degré 2, et l'on trouve une droite d'équation Taux = $at^2 + bt + c$ avec a=-4.572, b=1.006 et c=12.47.



Intuitivement, on constate que l'approximation par un polynôme du second degré semble plus satisfaisante. Cela est confirmé numériquement par les données fournies grâce au logiciel R :

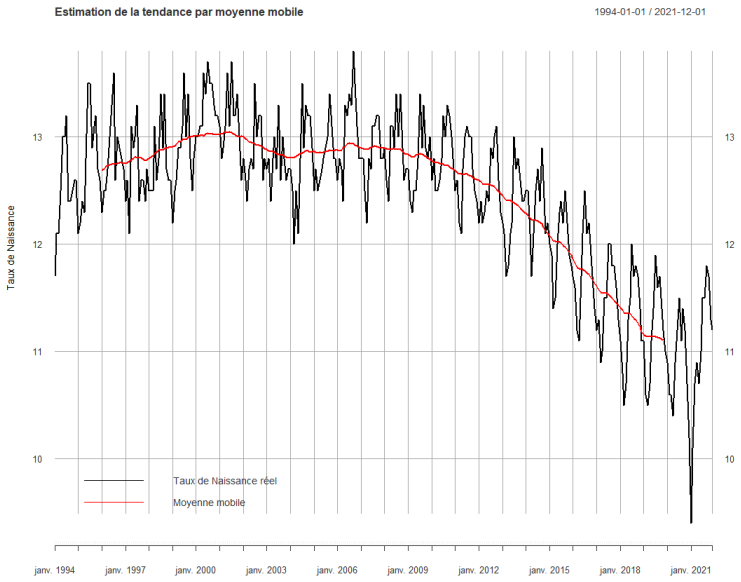
	Approximation linéaire	Approximation quadratique
Coefficient de Pearson R^2 ajusté	0.4634	0.7191
Erreur des résidus	0.5580	0.4038

Le coefficient de Pearson mesure la qualité de prédiction de la régression. Au plus ce nombre est proche de 1, au mieux le modèle approxime correctement les données. Dans notre cas, on constate que le modèle quadratique approche mieux nos données que le modèle linéaire. Ceci est confirmé par l'erreur des résidus, correspondant à l'écart entre la réalité et l'approximation réalisée, qui est supérieure dans le cas linéaire.

Nous retenons donc pour l'instant le modèle de régression par un polynôme de degré 2, comme nous pouvions le deviner.

b) Moyenne mobile

Dans un second temps, nous développons la méthode de la moyenne mobile. Nous choisissons une fenêtre h=50, de telle sorte à prendre suffisamment de données autour d'un instant t (nous voulons surtout prendre en compte plusieurs périodes, que nous avons estimée à T=12 mois). De plus, ce choix de fenêtre élimine le calcul de 50 moyennes (les 25 premiers mois et les 25 derniers), ce que l'on considère comme raisonnable par rapport aux 336 mois composants la série (cela correspond à environ 15% des valeurs).



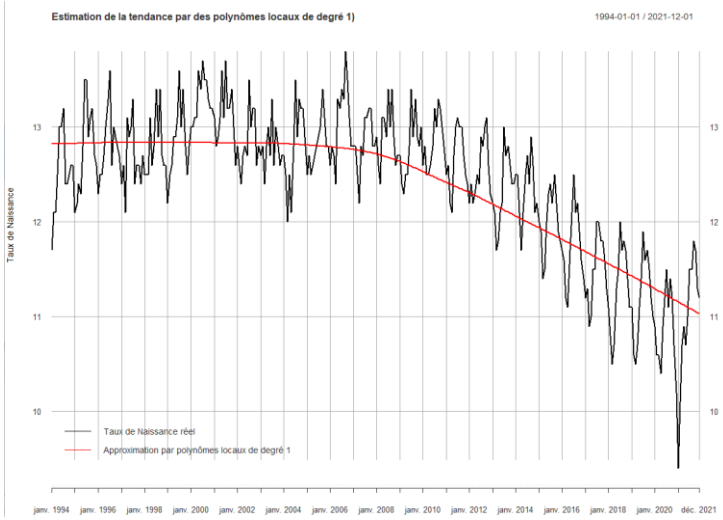
	Taux réel	Moyenne mobile
Moyenne	12.43	12.53
Variance	0.76	0.55

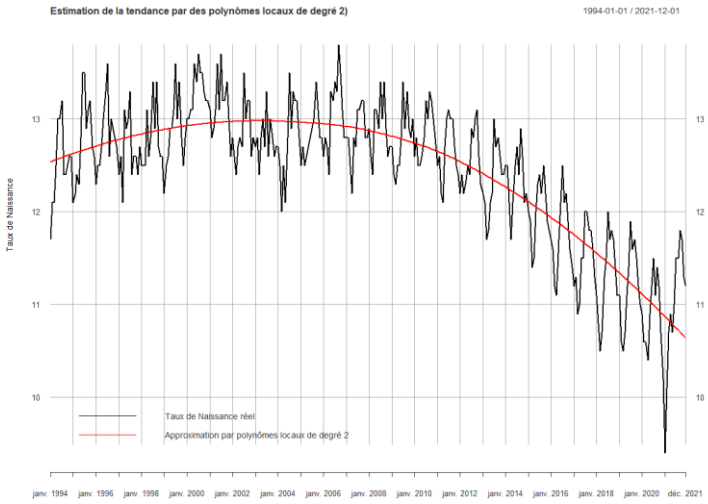
La moyenne mobile possède une moyenne proche de celle du taux réel. Sa courbe donne une bonne approximation de la tendance de la série temporelle. La variance de la moyenne mobile est moindre que celle de la série temporelle, mais ceci est expliqué par le fait que cette moyenne élimine toutes oscillations et perturbations dues au bruit (et dans une moindre mesure à la saisonnalité).

c) POLYNOMES LOCAUX

Nous nous intéressons maintenant à la méthode des polynômes locaux. Nous utilisons la fonction *loess* premièrement pour approcher la tendance par un polynôme de degré 1, puis ensuite par un polynôme de degré 2.

Afin d'éliminer la saisonnalité, nous choisissons de prendre un paramètre de fenêtre h=0.9 ; ainsi la taille du voisinage inclut une proportion h=0.9 des points du jeu de données.





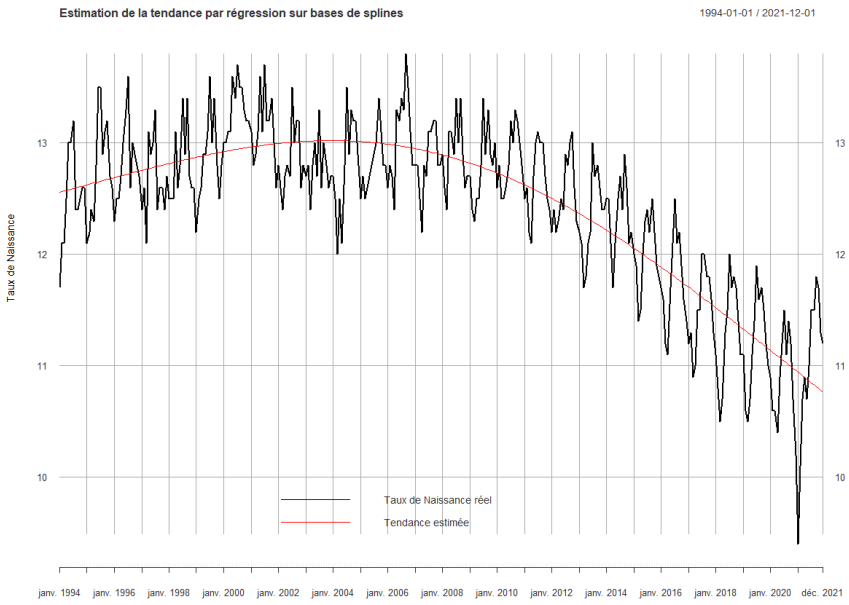
Les erreurs résiduelles, obtenues grâce à la fonction *summary()*, sont récapitulées dans le tableau suivant :

	Polynôme de degré 1	Polynôme de degré 2
Erreur résiduelle	0.4344	0.4027

De même, nous estimons ici une meilleure approximation des données par les polynômes locaux de degré 2. Inévitablement, réduire le paramètre de fenêtre h réduirait l'erreur résiduelle. Nous ne faisons cependant pas ce choix, car il impliquerait de prendre en compte une partie de la saisonnalité.

d) REGRESSION SUR BASE DE SPLINES

La méthode de régression sur bases de splines nous donne la courbe suivante :



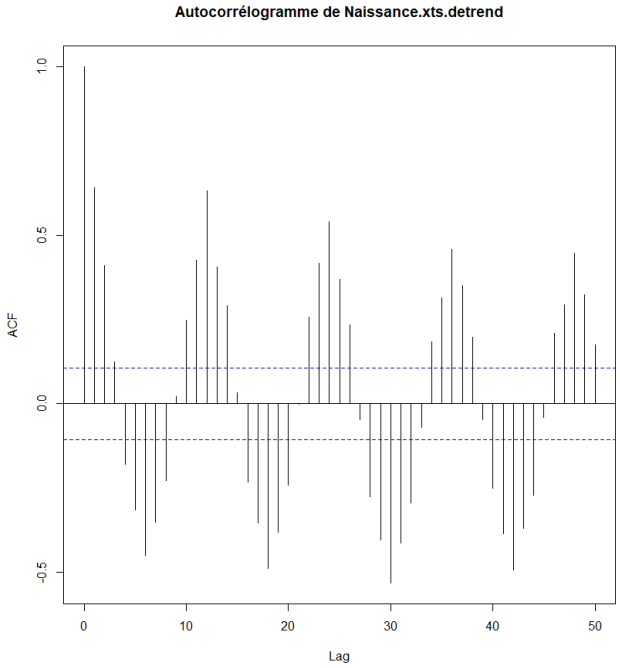
Le coefficient de Pearson associé est de $R^2 = 0.721$. Cela est supérieur à la valeur obtenue par régression linéaire de degré 2.

2) ESTIMATION DE LA SAISONNALITE

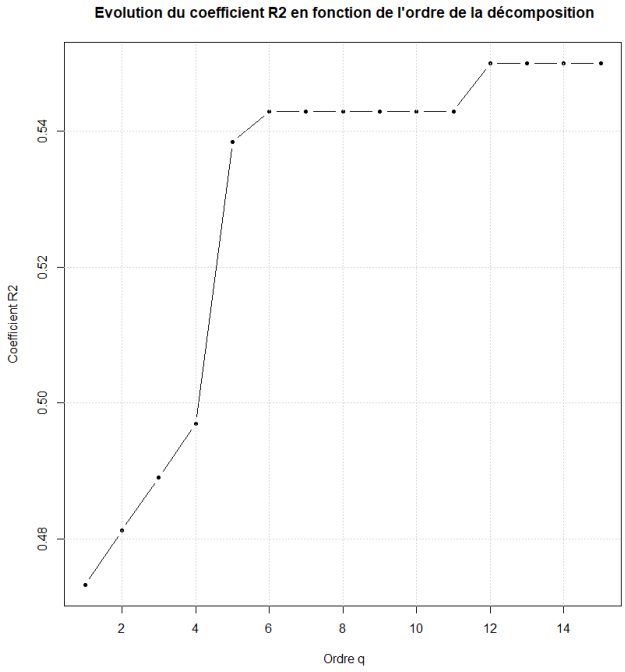
La partie précédente a permis de mettre en évidence que certains modèles étaient plus adaptés que d'autres pour modéliser la tendance de la série temporelle. On choisit ici le modèle de régression linéaire pour estimer cette tendance, ce modèle ayant fournit des critères de performances que l'on a jugé satisfaisants. On cherche maintenant a capter la saisonnalité de la série temporelle.

a) REGRESSION LINEAIRE

Pour déterminer la saisonnalité, nous décidons premièrement de poursuivre avec le modèle de régression linéaire. Cette régression se fait ici en décomposant la série sur une série de Fourier, c'est-à-dire que l'on cherche à exprimer la saisonnalité comme une somme de cosinus et de sinus de période estimée T=12, ce que nous confirme ci-dessous l'autocorrélogramme de la série privée de sa tendance.

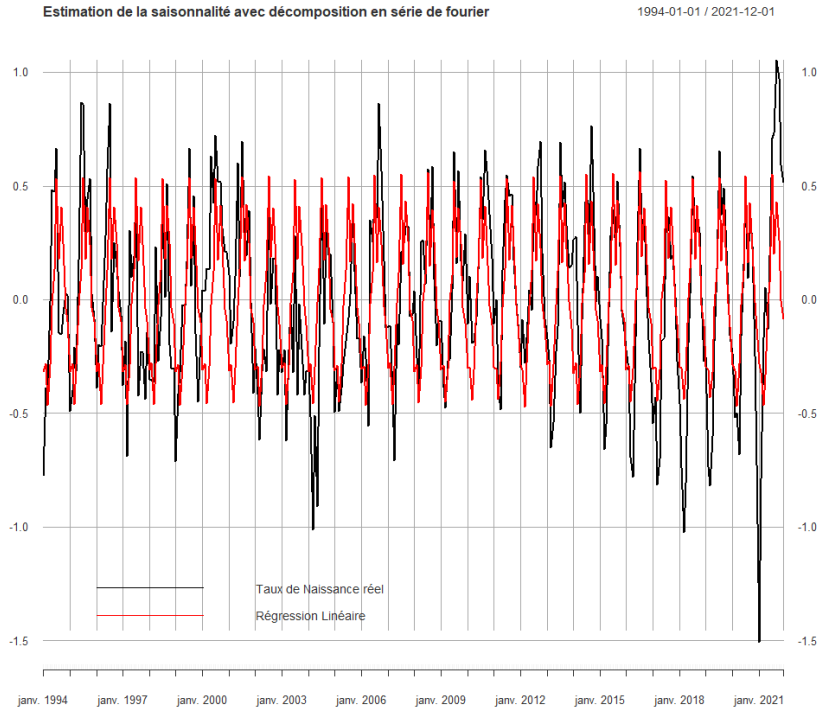


On cherche d'abord l'ordre de la décomposition en série de Fourier. L'implémentation d'un petit programme en R nous permet d'obtenir la courbe suivante :



On décide donc de choisir un ordre de décomposition q=6, ce qui paraît être un bon compromis entre précision de la régression linéaire et complexité algorithmique (augmenter l'ordre signifie augmenter le temps de calcul).

La régression linéaire permet d'obtenir une saisonnalité telle que présentée ici :

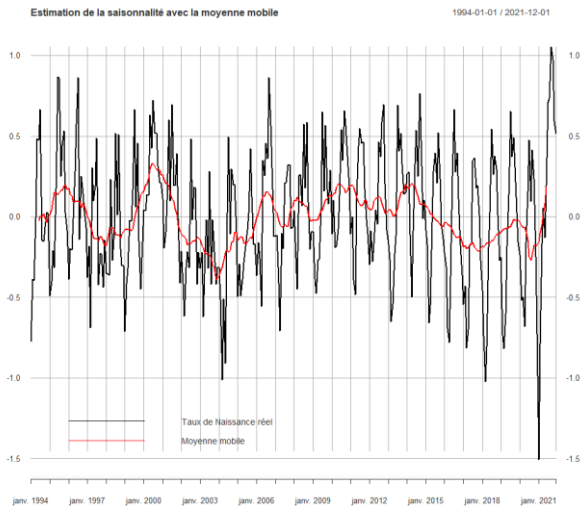


Les coefficients associés sont exposés dans le tableau ci-dessous.

	Régression Linéaire sur une série de Fourier
Coefficient de Pearson R^2 ajusté	0.526
Erreur des résidus	0.2768

b) MOYENNE MOBILE

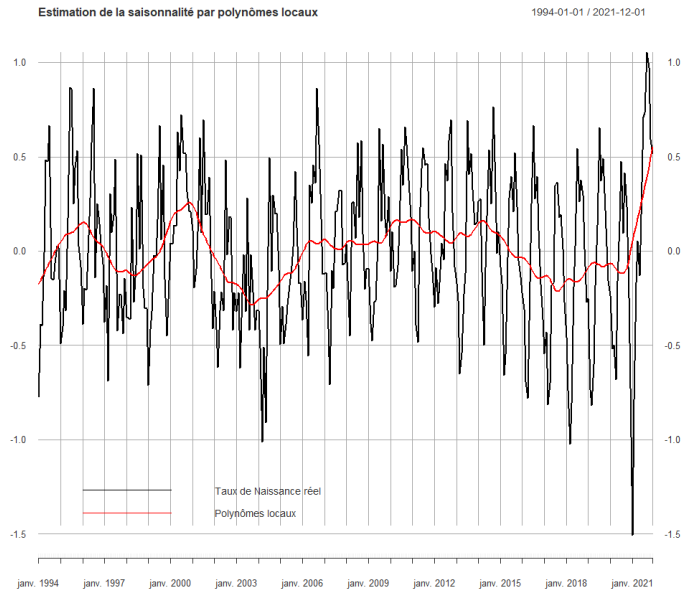
Nous essayons d'implémenter la méthode la moyenne mobile avec une fenêtre égale à la période estimée, c'est-à-dire h=12.



Nous considérons l'approximation réalisée comme moins bonne que celle obtenue par la décomposition en série de Fourier. Il semble difficile ici d'apercevoir un phénomène périodique de période de T=12.

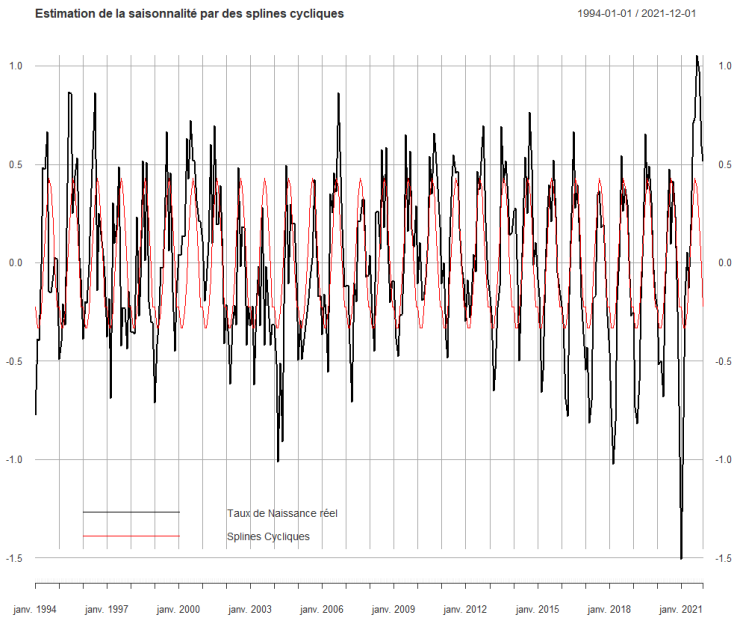
c) POLYNOMES LOCAUX

De même, la méthode des polynômes locaux réalisés avec une fenêtre de h=0.1 et des polynômes de degrés 2 ne fournit pas un modèle satisfaisant, comme en témoigne la figure obtenue ci-après :



d) SPLINES CYCLIQUES

Enfin, la régression sur des bases de splines cycliques, avec un cycle de période T=12, fournit quant à lui une approximation qui semble pertinente :

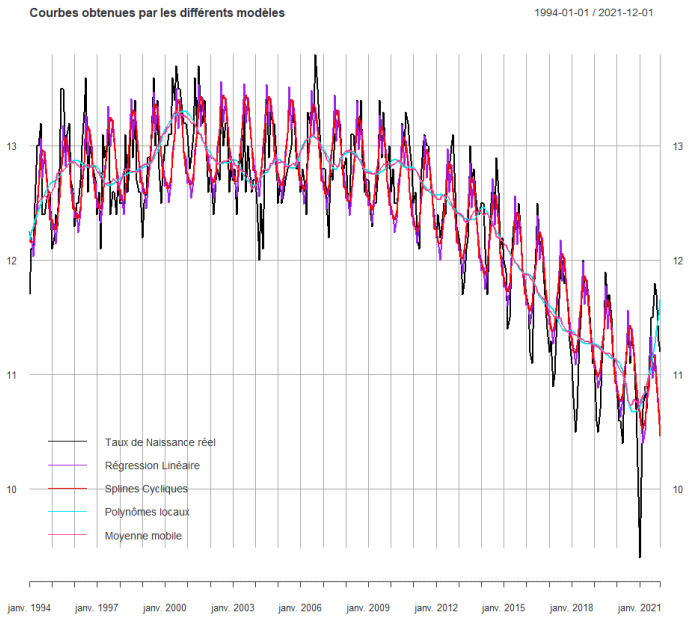


Le coefficient de Pearson associé est R^2 =0.457.

3) COMPARAISON DES METHODES

Nous avons jusqu'à présent établi plusieurs modèles pour estimer la tendance et la saisonnalité de notre série temporelle.

La tendance a été estimée à partir d'un modèle de régression linéaire par un polynôme de degré 2. Cette estimation a été ensuite retranchée à la série temporelle pour former une nouvelle série, *Naissance.detrend*, que l'on a étudié par plusieurs modèles pour déterminer la saisonnalité restante.



Nous avons, pour chacun de ces modèles, calculer deux coefficients que sont le RMSE et le MAE.

- Le MAE, 'Mean Absolute Error', est un coefficient qui mesure l'écart entre les valeurs prédites et les valeurs réelles.
- Le RMSE, littéralement 'Root Mean Square Erreur', est également une mesure de performance du modèle qui détermine à quel point ce modèle réalise des erreurs lors de ses prévisions. Il donne cependant un poids plus important que le MAE aux grandes erreurs.

Ces coefficients sont récapitulés dans le tableau suivant :

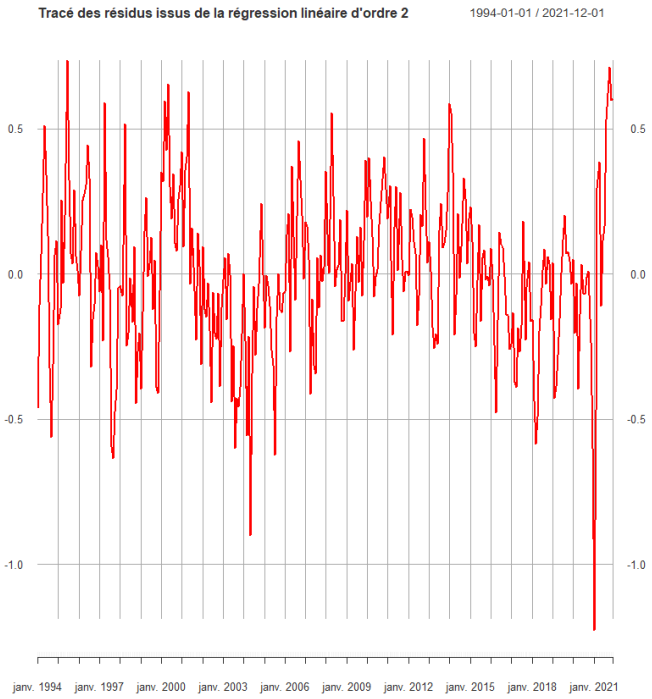
	RMSE	MAE
Fourier	0.271	0.205
Splines	0.295	0.226
Polynômes Locaux	0.370	0.300
Moyenne mobile	0.360	0.288

Ainsi, en choisissant comme modèle une régression linéaire de degré 2 pour estimer la tendance, on constate que le meilleur modèle en termes de minimisation d'erreur est également celui de la régression linéaire utilisant la décomposition en série de Fourier.

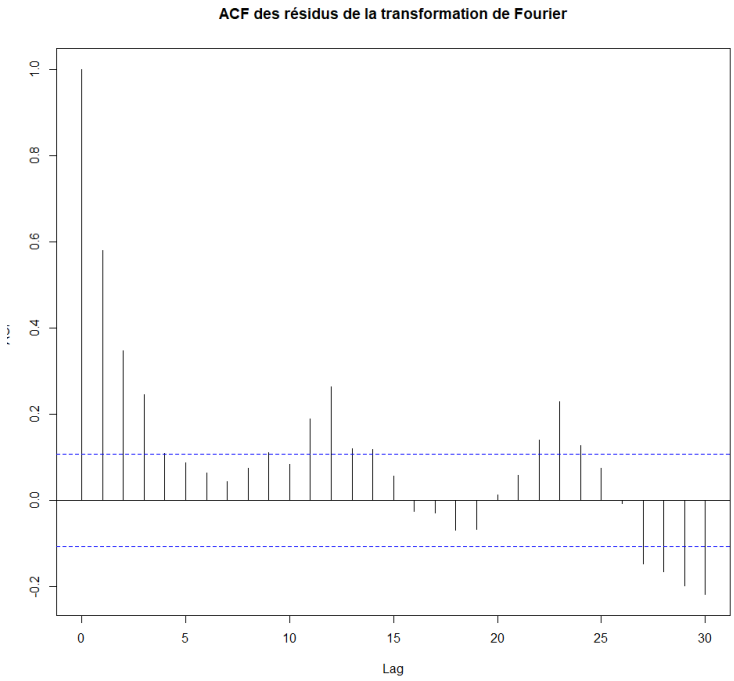
On constate par ailleurs que, comme nous l'avions supposé par les courbes tracées, les modèles de polynômes locaux et de moyenne mobile sont moins performants que les autres.

4) ETUDE DES RESIDUS

Nous nous intéressons donc à la modélisation qui utilise les deux régressions linéaires (polynomiale de degré 2 pour la tendance, et la décomposition de Fourier pour la saisonnalité). Il reste à présent un bruit, que l'on cherche à étudier. Voici un aperçu de ces résidus :



Commençons par regarder son autocorrélogramme :



Ce dernier n'est pas satisfaisant. En effet, l'on remarque qu'il reste une dépendance des données entre les instants t et t+T, les pics à lag=12 et lag=24 dépassant le seuil de non-significativité.

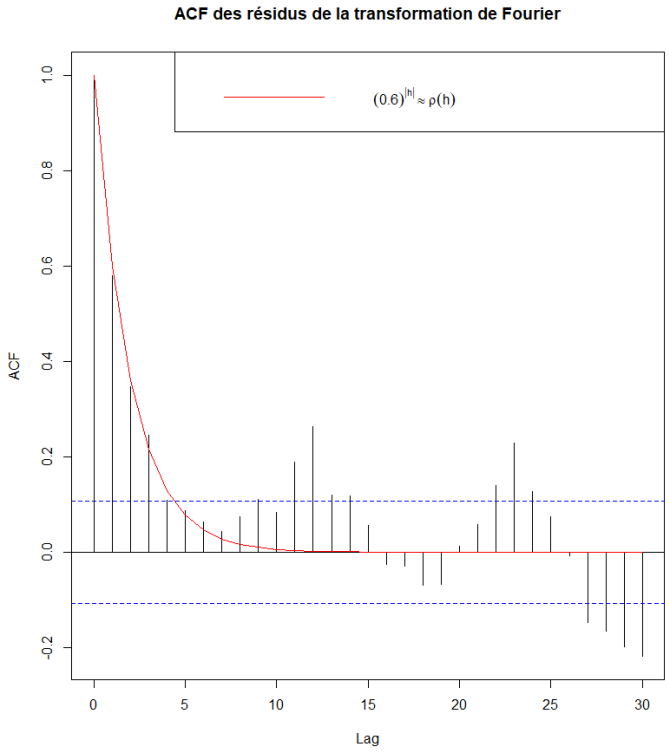
On essaye pour cela de capter une nouvelle fois la saisonnalité restante dans le bruit. On utilise de même un modèle de régression linéaire. Cependant, cela ne change rien et l'autocorrélogramme obtenue est strictement identique à celui du bruit initial.

Il semble donc qu'il reste dans le bruit une dépendance aux valeurs passées. Cela n'est toutefois pas incompatible avec la stationnarité du bruit.

Cependant, nous souhaitons tout de même être en mesure de modéliser et corriger notre erreur. Suite au tracé des résidus et de leur autocorrélogramme, nous pensons modéliser cette erreur par un processus autorégressif d'ordre 1. Ainsi, nous faisons le choix d'écrire que le bruit Y_t vérifie :

$$Y_t = aY_{t-1} + \varepsilon_t$$

Avec ε_t tel que $E[\varepsilon_t]=0$. Après avoir testé avec plusieurs valeurs de a, nous choisissons a-0.6 qui semble la valeur la mieux adaptée pour modéliser la décroissance exponentielle de la fonction d'autocorrélation. Toutefois, il est normal que les pics obtenus aux lag qui sont des multiples de 12 ne soient pas pris en compte dans cette modélisation. C'est pourquoi une étude plus poussée serait peut-être nécessaire.



IV. Simulation de prévision

1) PREVISION DE L'ANNEE 2021

Nous avons étudié différentes approches permettant de modéliser la tendance et la saisonnalité de notre série temporelle.

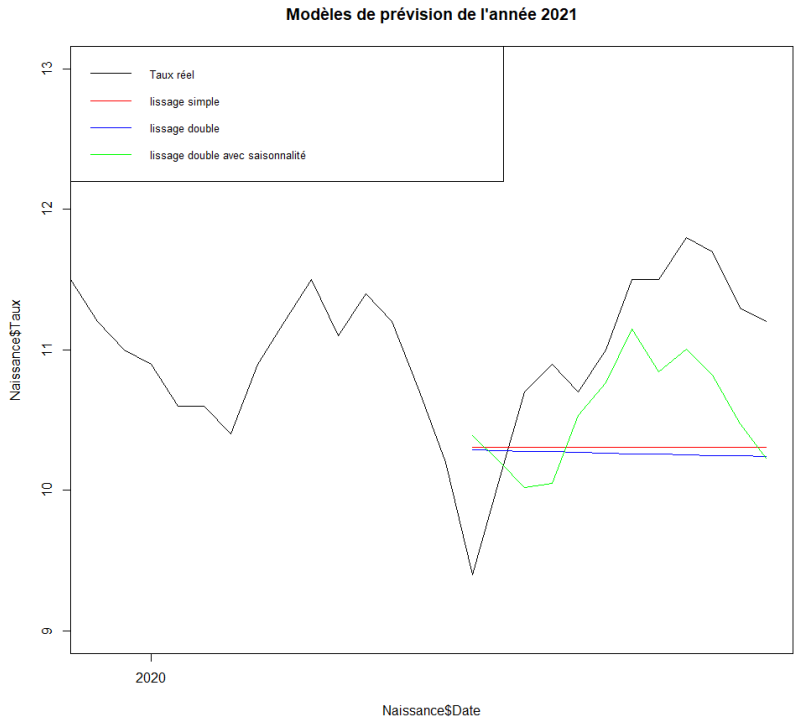
Dans cette partie, nous cherchons à modéliser notre série par des méthodes de lissage exponentiel. Pour cela, nous éparons notre jeu de données en deux :

- La première partie est une partie d'apprentissage : elle constitue toutes les données comprises entre 1994 et 2020
- La seconde partie est une partie de test constituée des données réelles : ces données vont être comparées avec les prévisions obtenus avec la partir de la partie d'apprentissage. Elles sont constituées des taux de natalité de l'année 2021.

Nous utilisons ici trois approches :

- Le lissage exponentiel simple, qui consiste à prévoir les données par une constante. Nos données n'étant pas un bruit (il possède une tendance et une saisonnalité), nous ne nous attendons pas à ce que ce modèle soit très performant.
- Le lissage exponentiel double, qui approche les données à prévoir par une fonction affine.
- Et enfin le lissage double en prenant en compte la saisonnalité, et qui, a priori, semble être le plus adapté ici.

Vérifions ces hypothèses à l'aide de la libraire *forecast* du logiciel R, et plus particulièrement de la fonction ets qui permet de réaliser ces lissages exponentiels.



Il est curieux de noter que le lissage double nous donne une fonction affine décroissante, alors que les données réelles sont en réalité à la hausse. De même, le lissage double avec saisonnalité commencent par prédire une baisse du taux, avant de remonter. Il sous-estime aussi les taux des mois suivants.

Les coefficients d'erreur pour chacun des modèles sont récapitulés ici :

	RMSE	MAE
Lissage simple	0.327	0.265
Lissage double	0.327	0.265
Lissage double avec saisonnalité	0.213	0.169

Comme attendu, le lissage double avec saisonnalité est le modèle permettant une prévision des données avec une erreur minimale. Il est naturellement le modèle retenu ici.

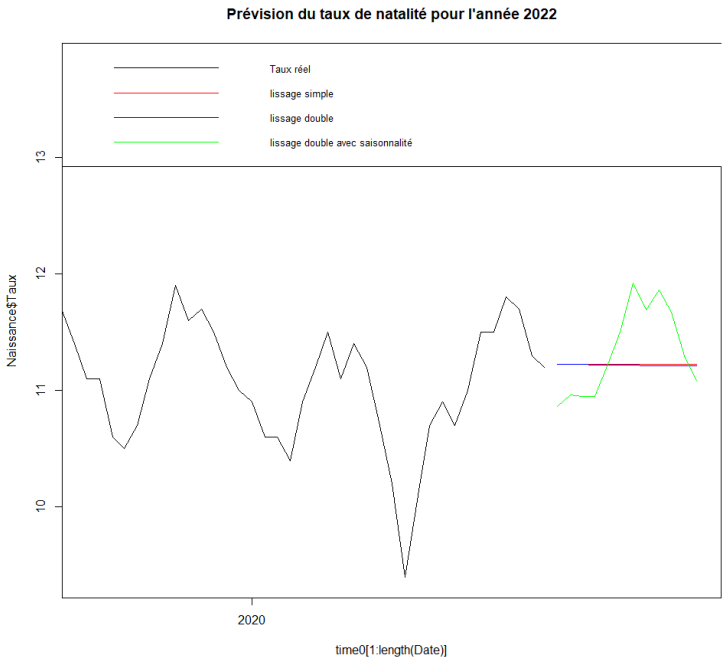
Il est par ailleurs intéressant de constater que les lissages simple et double nous donnent des coefficients d'erreur presque égaux (ces coefficients diffèrent par leurs décimales suivantes, non affichées ici). Graphiquement, la fonction affine obtenue par lissage double reste très proche de la constante évaluée par lissage simple, ce qui peut expliquer la similarité du RMSE et MAE.

2) PREVISION DE L'ANNEE 2022

Enfin, il peut être intéressant d'essayer de prévoir les taux de natalités pour l'année 2022.

a) LISSAGE EXPONENTIEL

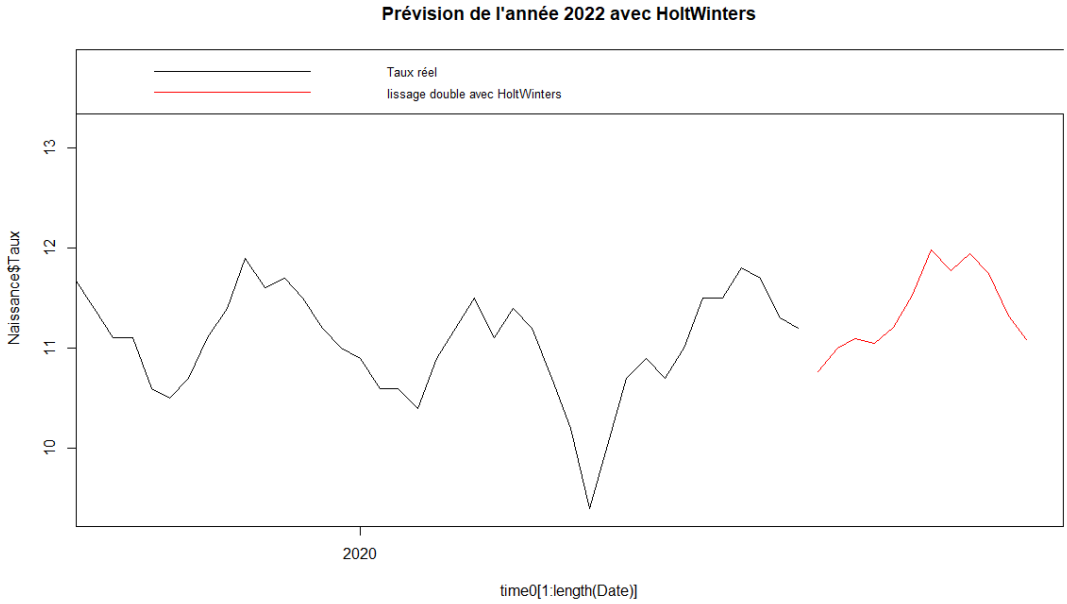
En utilisant les mêmes méthodes, voici les résultats obtenus :



Comme annoncé précédemment, nous nous fions aux prévisions données par le lissage exponentiel double avec saisonnalité. Ces données peuvent être utilisées à cours termes, comme première approximation du taux de natalité de Janvier et Février 2022 par exemple.

b) FONCTION HOLTWINTERS

Nous effectuons une deuxième prévision en utilisant la fonction HoltWinters de R. Nous choisissons de laisser estimer les paramètres de lissage par la méthode.



Cette méthode fournit une prévision similaire que celle obtenue avec la fonction ets précédemment. Nous nous contenterons donc de cette prévision comme approximation du taux de natalité dans un cours termes.

Conclusion

Lors de ce projet nous avons pu étudier et analyser nos données concernant le taux de natalité en France.

Plusieurs méthodes de représentation de la série ont été menées, mais c'est la régression linéaire d'ordre 2 qui a retenu notre attention. Elle nous a permis de modéliser une bonne estimation de la tendance, puis d'extraire une saisonnalité annuelle. L'étude du bruit restant nous laisse à penser que toute la saisonnalité n'a pu être captée, mais cela ne signifie pas que ce bruit n'est pas stationnaire.

De plus, nous avons obtenu une prédiction avec une erreur RMSE très raisonnable en utilisant le lissage exponentiel double avec saisonnalité. Les résultats obtenus sur l'année 2021 étaient satisfaisant dans le cas d'un lissage double avec saisonnalité. Nous avons pu construire à partir de ce modèle une prévision de l'évolution du taux de natalité en France pour l'année 2022, même s'il faut garder à l'esprit que cette approximation ne peut être utilisée efficacement qu'à court terme.