FAUJOUR Clara 19004656
GUIBERT Marie 20115711
PAILLARD Loevane

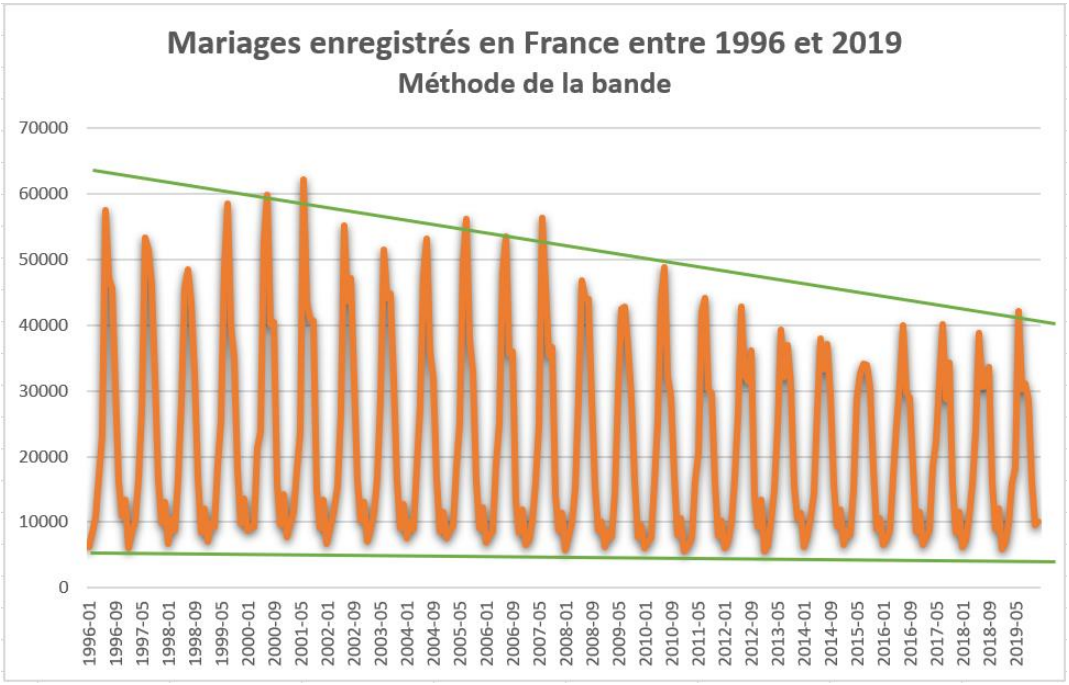
Rapport d'étude d'une série chronologique

Dans cette étude, nous avons choisi d'analyser le nombre de mariages enregistrés en France de 1996 à 2019. Afin de préparer les données à notre étude, nous avons éliminé les valeurs manquantes. En effet, l'année 1995 ne présentant aucune donnée, nous avons décidé de commencer notre étude à partir de 1996. Nous n'avons pas observé de valeur incohérente ni d'irrégularité temporelle, nous n'avons donc pas eu à faire de modification. Ces données sont issues d'un rapport de l'INSEE. Dans la base de données, nous retrouvons une colonne correspondant à la période, composée de l'année et du mois, une autre avec le nombre de mariages enregistrés et une dernière avec un code. Un document est fourni avec cette base de données, permettant d'expliquer ce code. La valeur la plus fréquente de cette colonne est la lettre "A", correspondant à la valeur normale alors que "O" signifie valeur manquante. Cette information n'est pas utile lors de l'analyse mais précise des caractéristiques sur les données observées. Concernant, la première colonne de cette base, 12 lignes forment une année et à chaque mois est associé un nombre de mariages. Ainsi, nous traiterons une série mensuelle. L'objectif de cette étude est de prédire le nombre d'unions pour l'année suivante en France. Dans un premier temps, nous analyserons la base de données et déterminerons quel modèle convient à la série temporelle. Dans un second temps, nous appliquerons une moyenne mobile qui annule la saisonnalité de celle-ci. Et enfin, nous étudierons les lissages applicables à cette étude, et nous en déduirons le plus adapté afin d'obtenir les meilleures prévisions possibles.

Pour commencer, nous procédons à une analyse descriptive de la série. En effet, la première approche est de mettre en lumière à quel genre de modèle nous avons à faire. La série temporelle que nous avons choisi d'analyser, présente des données plutôt régulières, même si nous observons une diminution du nombre de mariages en France au fil des années. La tendance à se marier de la population dépend entre autres de la proportion de personnes entre 25 et 59 ans. De plus, l'évolution du comportement de la population impacte grandement ces données. Aujourd'hui, les individus se marient moins et le font plus tardivement. Ainsi, le nombre de mariés tend à diminuer au fil du temps. Nous allons maintenant observer quel modèle correspond le plus à cette base de données.

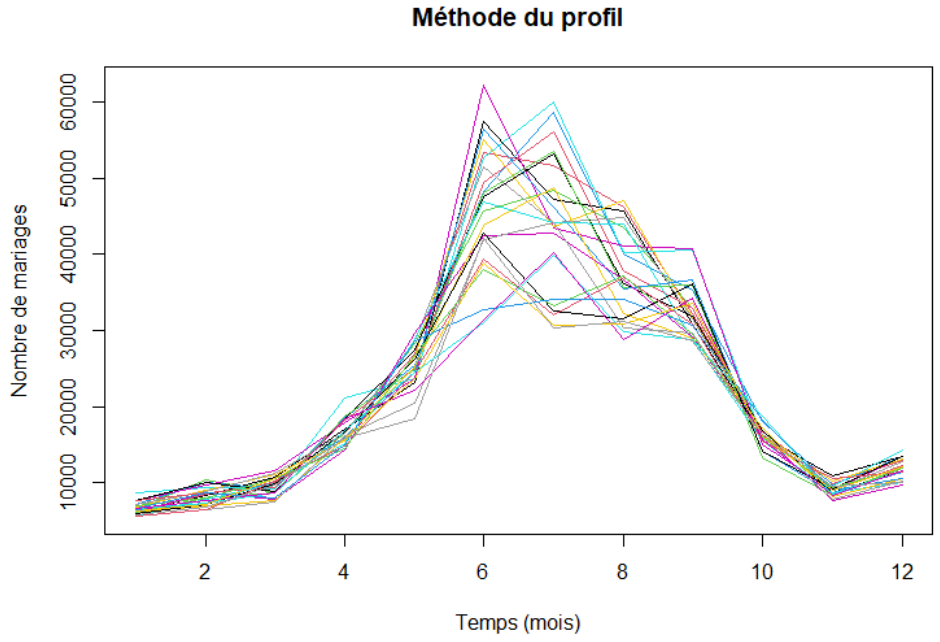
Il existe deux types de modèles : additif ou multiplicatif. Le modèle additif s'exprime ainsi : $xt = Tt + St + Ut$, alors que le modèle multiplicatif est défini par $xt = Tt * St * Ut$. Dans ces deux équations, on retrouve Tt décrivant la tendance de la série. Cette dernière reflète l'évolution de moyen-long terme. Dans notre situation, les données de la série dépendent du temps et présentent une tendance décroissante. St correspond aux coefficients saisonniers. Ils résultent les variations saisonnières mais n'influencent pas la tendance. Et enfin, l'aléas représenter par Ut dans l'équation, décrit les variations résiduelles qui peuvent avoir de nombreuses causes mais ne sont pas significatives pour la série temporelle. Pour mettre en lumière le modèle de notre série de données sur les mariages en France, nous disposons de deux méthodes graphiques et une méthode plus calculatoire : la méthode de la bande, la méthode du profil et le test de Buys-Ballot.

D'abord, la première méthode nécessite une représentation graphique des données, classées par ordre chronologique. La méthode de la bande consiste à tracer deux droites passant par les minimas et maximas de la série. Si ces droites semblent être parallèles alors le modèle s'approche d'un additif, en revanche, si elles s'apparentent à un entonnoir, alors le modèle est davantage multiplicatif. Pour appliquer cette méthode, le logiciel Excel est un très bon outil.



Dans notre situation, nous nous rapprochons davantage d'un modèle multiplicatif mais les droites reliant les minimas et les maximas ne s'accordent pas exactement à un entonnoir. L'hypothèse d'un modèle additif n'est donc pas totalement écartée. La suite de notre analyse va nous permettre d'éclairer ce point.

Pour continuer, la deuxième méthode possible est celle du profil. Cette dernière repose sur la représentation des saisons sur un même graphique. Il faut ensuite analyser le comportement des courbes de profils. Si ces dernières se superposent alors nous sommes dans le cas d'un modèle additif. En revanche, si elles s'entrecroisent parfaitement, le modèle est multiplicatif. Il existe des modèles particuliers, notamment lorsque sur le même graphique certaines courbes de profils se superposent et d'autres s'entrecroisent.



Ici, nous sommes dans une situation particulière puisque les profils ne s'entrecroisent pas parfaitement. Aux extrémités de ce graphique, les courbes se superposent plus ou moins. Néanmoins, cette méthode nous guide plutôt vers un modèle multiplicatif qu'un modèle additif.

Enfin, le test de Buys-Ballot permet d'examiner l'existence d'une relation linéaire entre la moyenne et l'écart-type de la série de données. Pour détecter quel modèle correspond à la série, nous calculons le coefficient de détermination R^2. Si celui-ci est égal à 0, aucun lien n'existe et ce modèle est additif. A l'inverse, s'il est différent de 0, le modèle est multiplicatif, il y a donc une relation linéaire entre les variables. Ce test entraîne tout un calcul de moyennes et d'écart-type pour chaque année. Avec le logiciel R, le test de Buys-Ballot se réalise très simplement, comme indiqué ci-dessous :

```
#Test de buys-ballot
aganmean = aggregate (init$Mariages,list(an=init$an),mean)
agansd = aggregate(init$Mariages,list(an=init$an),FUN="sd")
tbb=lm(agansd$x~aganmean$x)
summary(tbb)
#P-value est de 7.732e-14
```

Pour mettre en lumière le modèle qui concorde avec la série, nous nous intéressons à la "P-value". C'est cette valeur qui précise l'existence d'un lien linéaire entre les variables. Calculée par le logiciel, si elle est inférieure à 0.05 alors le test conduit vers un modèle multiplicatif, en revanche si cette valeur est supérieure à 0.05, le modèle est additif. En effet, cette valeur permet de rejeter ou non l'hypothèse d'additivité du modèle.

```
Call:
lm(formula = agansd$x ~ aganmean$x)

Residuals:
   Min     1Q Median     3Q    Max
-860.9 -486.6 -243.9  503.8 1418.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.878e+03  1.509e+03  -6.545 1.39e-06 ***
aganmean$x   1.125e+00  6.845e-02  16.433 7.73e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 680.8 on 22 degrees of freedom
Multiple R-squared:  0.9247,    Adjusted R-squared:  0.9212
F-statistic:   270 on 1 and 22 DF,  p-value: 7.732e-14
```

Dans notre cas, la "P-value" est de $7.732e^{-14}$, cette valeur est largement inférieure à 0.05. L'hypothèse est donc rejetée et le test de Buys-Ballot nous conduit vers un modèle multiplicatif. Toutefois ce test peut être biaisé, cette méthode rejette souvent le modèle additif par manque d'observations.

En conclusion, après avoir analysé cette série temporelle, nous supposons que le modèle est multiplicatif. En revanche, cette approche n'est pas exhaustive. La détection du modèle est une étape importante dans l'étude d'une série car la suite de celle-ci en dépend. En effet, selon le résultat obtenu, les données doivent, ou non, connaître une transformation pour faciliter leur analyse.

Dans notre cas, les maximas de notre série semblent diminuer. Cette baisse se confirme par une moyenne annuelle plus faible au fil du temps. Nous constatons un nombre de mariages moyen de 23 928 en 1996 et de 18 728 lors de l'année 2019. Nous avons tout intérêt à transformer la série pour pouvoir prédire de manière précise les données. Trois possibilités s'offrent à nous : l'estimation de la série initiale ainsi que les transformations de type Box-Cox et logarithmique. Ces séries doivent être désaisonnalisées pour poursuivre leur analyse. Il existe plusieurs méthodes de désaisonnalisation, comme la régression linéaire ou

encore l'application de moyenne mobile. Nous avons choisi d'utiliser cette dernière pour notre étude.

Les moyennes mobiles sont l'une des premières méthodes utilisées pour analyser les séries chronologiques. Le principe est de créer un filtre qui isole la tendance lorsqu'il est appliqué à la série. Les moyennes mobiles permettent de lisser et de gommer les irrégularités de la série pour extraire sa structure globale. Le concept de base est de calculer une moyenne de quelques données autour de la date à laquelle on s'intéresse. Ainsi une moyenne mobile est une somme pondérée de valeurs de *x* correspondant à des dates proches de *t*. Elle s'écrit de la manière suivante :

$$M_{m_1+m_2+1}X_t = \sum_{i=-m_1}^{m_2} \theta_i X_{t+i} = \theta_{-m_1}X_{t-m_1} + \ldots + \theta_{-1}X_{t-1} + \theta_0 X_t + \theta_1 X_{t+1} + \ldots + \theta_{m_2}X_{t+m_2}$$

$\theta_{-m_1}, \ldots, \theta_{m_2}$ sont des réels et *m1*, *m2* appartiennent aux entiers naturels.

Il existe plusieurs types de moyennes mobiles comme les moyennes mobiles centrées, symétriques et arithmétiques. Chacune d'entre elles présentent des particularités. Par exemple, les moyennes mobiles arithmétiques permettent de conserver la tendance d'une série mais elle supprime la saisonnalité. Dans notre situation, pour effacer notre saisonnalité mensuelle, il suffit d'appliquer la composition d'une MM2 et d'une MM12. Ce processus revient donc à utiliser une MM13 pour que la série soit « corrigée des variations saisonnières ». Cette moyenne mobile est centrée et symétrique. Elle s'écrit de la manière suivante :

$$MMA_2 * MMA_{12} = \left\{[13], \left[\frac{1}{24}; 6 * \frac{1}{12}\right]\right\}$$

Cette technique de désaisonnalisation est beaucoup utilisée et se réalise très efficacement avec le logiciel R. Nous avons donc le code suivant :

```
# MM arithmétique d'ordre 13----------------------------------------------------
init0 = init[1:276,]
filter12=c(1/24,rep(1/12,11),1/24)
init0$mariages13=filter(init0$Mariagest, filter12, method = "convolution",sides = 2, circular = FALSE)
```

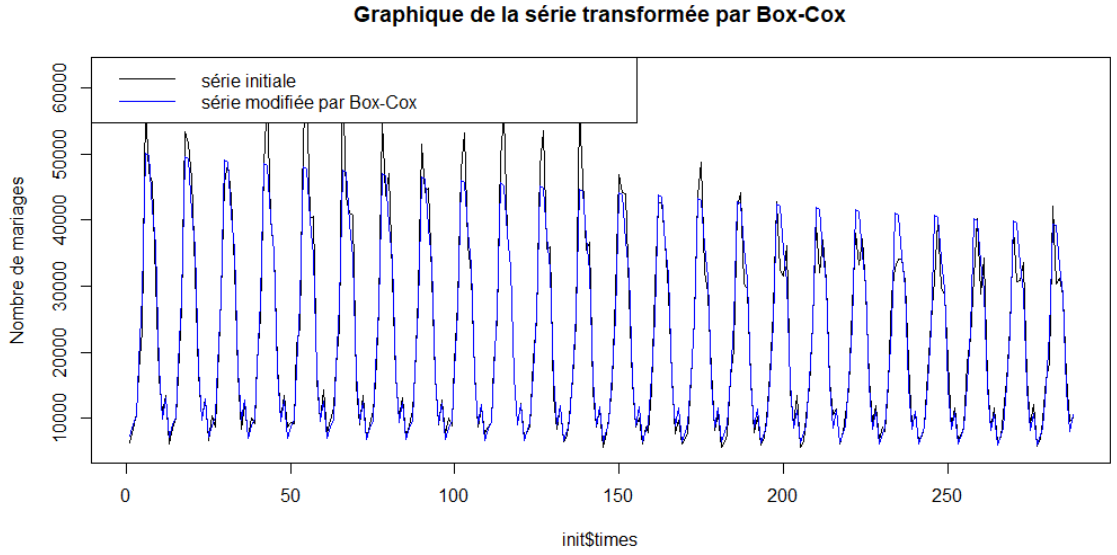
A présent, nous transformons la série temporelle. Pour chacune de ces transformations nous calculons les coefficients saisonniers, extrapolons la tendance puis estimons les nouvelles séries. L'ensemble de ces démarches nous mènent au calcul des SCR. Nous obtenons 3549685045 pour l'estimation de la série initiale, 2761258647 concernant la transformation Box-Cox et 2832251861 après le passage en logarithme. La méthode la plus adaptée est la transformation Box-Cox du fait de sa SCR inférieure. La suite de notre analyse va donc porter sur cette dernière.

Premièrement, la transformation de type Box-Cox se réalise à l'aide du logiciel R de cette façon :

```
# Transformation de la série----------------------------------------------------
powerTransform(init$Mariages)
init$mod=((init$Mariages^(-0.1156859))-1)/(-0.1156859)

plot(init$mod, type="l", col="red", main="Transformation BOX-COX de la série", xlab="init$times")
```

Cette démarche permet de stabiliser la série en réduisant la dispersion des données et donc sa variance. La représentation graphique des séries initiales et transformées par Box-Cox est la suivante :



Deuxièmement, nous appliquons à la série transformée la MM13, pour isoler la tendance et procéder au calcul des coefficients saisonniers. En effet, nous retranchons la tendance à la série modifiée afin d'extraire sa saisonnalité. Nous extrapolons la tendance pour trouver les coefficients nécessaires à l'estimation de la série modifiée. Le détail de cette étude est présenté ci-dessous.

```
#etude de la serie modifie par BOX-COX ----------------------------------------

#MM13 sur serie modifie
init0$mod13 = filter(init0$mod, filter12, method="convolution", sides=2, circular=FALSE)

#calcul des coefs saisonnier
init0$modsais1 = init0$mod-init0$mod13
sais1mod = tapply(init0$modsais1, init0$mois, mean, na.rm=TRUE)
sais1modmoy = mean(sais1mod)
sais1modbis = sais1mod - sais1modmoy
init0$sais1modbis = rep(sais1modbis, 23)
init$sais1modbis = rep(sais1modbis, 24)

#extrapolation de la tendance
init0$desais1 = init0$mod - init0$sais1modbis
reg5=lm(init0$desais1 ~ init0$times)
summary(reg5)

#estimation de la serie mod
init0$tchap1 = 5.881+(-0.0002517)*init0$times
init0$modprev1 = init0$tchap1 + init0$sais1modbis
init$tchap1 = 5.881+(-0.0002517)*init$times
init$modprev1 = init$tchap1 + init$sais1modbis


init0$prev1 = ((-0.1156859)*init0$modprev1+1)^(1/(-0.1156859))
init$prev1 = ((-0.1156859)*init$modprev1+1)^(1/(-0.1156859))

#calcul de la SCR
res5car = (init0$Mariages-init0$prev1)^2
SCR5 = sum(res5car)

#graphique
plot(init$Mariages~init$times, type="l", col="black", lwd=1, main="Graphique de la série transformée par Box-Cox", ylab="Nombre de mariages")
lines(init$prev1~init$times, type="l", col="red", lwd=1)
legend("topleft",legend = c("série initiale","série modifiée par Box-Cox"),col=c("black", "red"),lty=1)
```

Nous avons ainsi les éléments nécessaires pour nos prévisions. Nous les réalisons de cette manière :

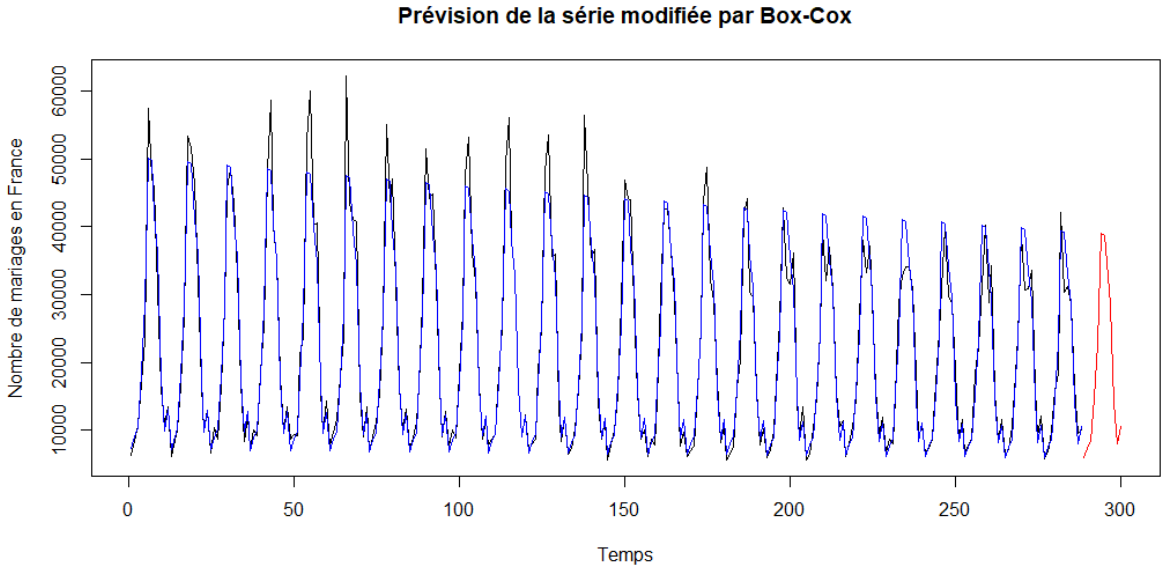
```
#prevision de la serie mod
ptimes = seq(289,300)
modtchap1 = 5.881+(-0.0002517)*ptimes
psais = c(sais1modbis)
modprevision = modtchap1 + psais
mariagesprev1 = ((-0.1156859)*modprevision + 1)^(1/(-0.1156859))

plot(init$Mariages~init$times, xlim=c(1,300), col="black", type='l')
lines(init$prev1~init$times, col="blue", type="l")
lines(mariagesprev1~ptimes, col="red", type="l")
```

Ce procédé nous permet d'obtenir des prévisions précises sur le nombre de mariages en France pour l'année suivante. Ces données sont cohérentes avec la concentration du nombre de mariages durant la période estivale. Aussi, nous remarquons un creux en début d'année pour diverses raisons, qui nécessiterait plutôt une analyse sociologique.

```
> mariagesprev1
       1        2        3        4        5        6        7        8        9       10       11       12
5950.321 7388.785 8391.521 14781.870 22629.419 39043.381 38726.679 32661.772 28385.765 14191.451 8011.622 10547.537
```

Graphiquement, nous retrouvons ces valeurs, en rouge ci-dessous.



Finalement, nous sommes parvenus à estimer les valeurs pour l'année 2020 grâce à la méthode des moyennes mobiles. Depuis les années 60, une nouvelle technique de prévision a vu le jour, les lissages. Il s'agit d'une méthode plus récente, plus efficace, que nous allons utiliser.

L'objectif d'un lissage est d'émettre des prévisions. Il en existe plusieurs types.
D'abord, on distingue le lissage exponentiel simple. Cette méthode est souvent appliquée lorsque la série est constante, c'est-à-dire quand elle ne présente pas de tendance ni de saisonnalité. Nous devons déterminer lambda, la constante de lissage, pour faire ces prévisions. Cette constante dépend des préférences de l'individu. Selon sa valeur, la prévision va dépendre plus ou moins du passé. Par exemple, la prévision est "naïve" lorsqu'elle consiste à affecter la dernière observation à la prochaine prévision (lambda proche de 1). Le phénomène de prévision est peu réactif aux dernières observations quand lambda est proche de 0. Dans notre situation, la série initiale présente une tendance et une saisonnalité, ce type de lissage n'est donc pas adapté.
Ensuite, le lissage de Holt fonctionne de la même façon que le lissage exponentiel. En effet, l'utilisation d'un lambda est nécessaire. En revanche ce lissage s'applique aux séries qui sont localement linéaires. Quand la série présente seulement une tendance, le lissage de Holt est la meilleure option pour continuer l'étude. Nos données s'accompagnent d'une saisonnalité, nous ne pouvons donc pas utiliser ce lissage pour nos prévisions.

Enfin, la dernière méthode de lissage est celle de Holt-Winters. Ce lissage est une amélioration du lissage de Holt, il est plus précis car il prend en compte la saisonnalité. Dans notre étude, nous avons décidé d'opter pour ce dernier car notre série révèle une tendance à la baisse et une saisonnalité. Cette méthode nous renvoie les coefficients saisonniers et la tendance. Le lissage de Holt-Winters s'emploie de deux manières différentes selon le type de modèle analysé avec deux fonctions de prévision distinctes.

Notre analyse descriptive nous a conduit vers un modèle multiplicatif. Cependant, l'hypothèse du modèle additif n'est pas complètement écartée. C'est pourquoi nous avons décidé d'appliquer le lissage de Holt-Winter dans le cas d'un modèle additif et d'un modèle multiplicatif. Le logiciel R est une nouvelle fois utilisé pour réaliser cette démarche.

```
#lissage de Holt-Winters SANS declaration des coefficients de lissage, MODELE ADDITIF----------------------
HW1=HoltWinters(init$Mariagest)
plot(HW1)

HW1fit=fitted(HW1)

p1=predict(HW1,12,prediction.interval=TRUE)
plot(HW1,p1)
SCR1=sum((HW1fit-init$Mariagest)^2)

#lissage de Holt-Winters SANS declaration des coefficients de lissage, MODELE MULTIPLICATIF----------------
HW2=HoltWinters(init$Mariagest, seasonal="multiplicative")
plot(HW2)

HW2fit=fitted(HW2)

p2=predict(HW2,12,prediction.interval=TRUE)
plot(HW2,p2)

SCR2=sum((HW2fit-init$Mariagest)^2)
```

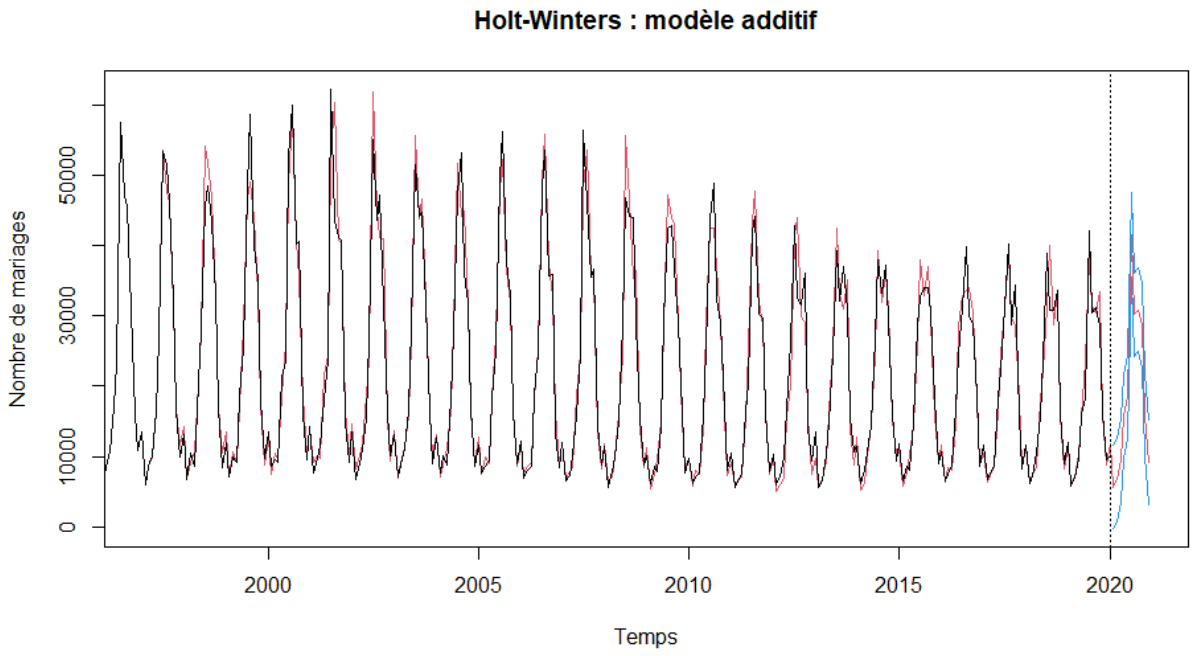
Nous obtenons deux prévisions différentes à l'aide du code ci-dessus. Nous pouvons observer les estimations chiffrées ainsi que leurs représentations graphiques.

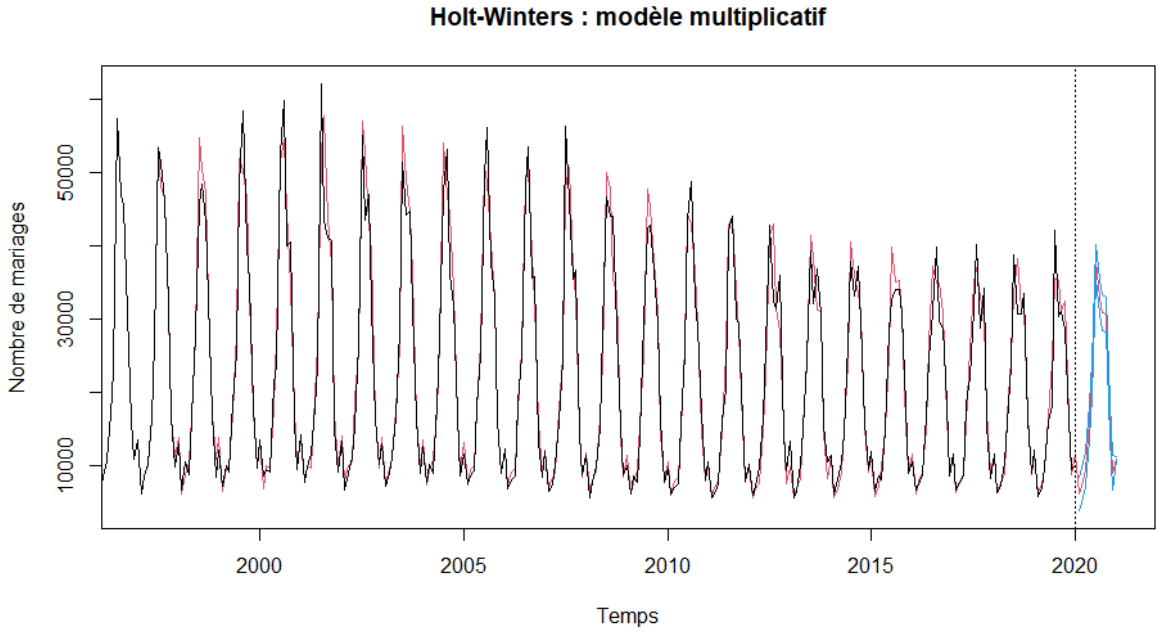
Prévisions selon Holt-Winters : modèle additif

	fit	upr	lwr
Feb 2020	5687.466	11645.13	-270.1951
Mar 2020	6661.501	12619.23	703.7727
Apr 2020	9471.746	15429.60	3513.8962
May 2020	15669.723	21627.76	9711.6821
Jun 2020	18508.145	24466.46	12549.8291
Jul 2020	41517.840	47476.53	35559.1474
Aug 2020	30142.531	36101.71	24183.3475
Sep 2020	30837.868	36797.67	24878.0615
Oct 2020	28725.602	34686.18	22765.0264
Nov 2020	16001.001	21962.51	10039.4939
Dec 2020	9180.305	15142.92	3217.6883
Jan 2021	9895.279	15859.20	3931.3596

Prévisions selon Holt-Winters : modèle multiplicatif

	fit	upr	lwr
Feb 2020	6169.928	8515.975	3823.882
Mar 2020	7237.566	9583.878	4891.255
Apr 2020	9535.567	11882.672	7188.461
May 2020	16144.841	18495.954	13793.729
Jun 2020	21758.387	24117.773	19399.002
Jul 2020	37901.423	40302.248	35500.598
Aug 2020	33073.937	35474.477	30673.398
Sep 2020	30891.329	33297.602	28485.055
Oct 2020	30697.586	33117.111	28278.062
Nov 2020	16312.500	18684.015	13940.985
Dec 2020	9016.924	11372.349	6661.499
Jan 2021	10932.857	11224.562	10641.152





Pour déterminer la meilleure prévision nous devons comparer les SCR des deux situations. De ce fait, le lissage de Holt-Winter pour le modèle additif nous renvoi une valeur inférieur à celle du modèle multiplicatif :

```
> SCR1
[1] 417090500264
> SCR2
[1] 438584736434
```

Ainsi, cette méthode nous précise une prédominance à l'additivité (SCR1). Les valeurs prévisionnelles admises par le lissage de Holt-Winter dans le cas additif présentent une suite logique aux données observées auparavant. Pour conclure, le lissage de Holt-Winters appliqué au modèle additif, nous permet d'émettre des prévisions sur le nombre de mariages qui seront enregistrés en France en 2020.

En conclusion, nous sommes parvenues à émettre une série prévisionnelle pour l'année 2020. Pour en arriver à ce résultat nous avons dû définir le modèle. Trois méthodes nous ont dirigé vers un modèle multiplicatif. En se basant sur ce résultat, nous avons transformé notre série pour faciliter son analyse. Par la méthode des moyennes mobiles, nous avons pu déterminer les coefficients saisonniers ainsi que la tendance. A l'aide de cette démarche nous avons émis une première prévision. Ensuite, nous avons utilisé une seconde méthode, les lissages. Contrairement à l'analyse descriptive qui nous a mené vers un modèle multiplicatif, le lissage de Holt-Winter nous a tourné vers un modèle additif. En comparant les SCR, nous nous sommes rendu compte que ces dernières étaient très proches, et que la confusion dans notre analyse était possible. Par suite, nous avons donc appliqué la méthode de Holt-Winter sur un modèle additif et nous sommes parvenus à des résultats. Nos deux estimations présentent des similitudes, plusieurs valeurs obtenues sont comparables. Cependant, pour certaines périodes, des écarts apparaissent. La méthode des moyennes mobiles demande plus de manipulations. Des erreurs de calculs ou d'approximation peuvent altérer les résultats. Tandis que les lissages fournissent des prévisions rapides et efficaces. Les écarts entre ces deux méthodes peuvent être dus à des erreurs humaines, tout comme à un choix de modèle erroné.

Dans un cadre plus concret, les multiples études de séries temporelles sont profitables pour les entreprises. Elles permettent de prédire les données, d'appréhender au mieux les besoins des consommateurs ou du pays par exemple. Ces observations servent à optimiser au mieux le budget d'une entreprise, ou des pouvoirs publics, mais aussi de maximiser leur profit en adoptant les meilleures stratégies.