## MODELISATION  des séries pluviométriques et hydrométriques de Fiherenana. Première étape : étude de la stationnarité.

### Définitions fondamentales en séries temporelles

#### Stationnarité stricte

La connaissance de la stationnarité d’une série est primordiale si on veut trouver un modèle qui retranscrit au mieux nos données hydro-climatiques. Selon que la série est stationnaire ou non, les démarches de modélisation ne seront pas les mêmes. On entend par stationnarité le fait que la structure d’un processus stochastique évolue ou non avec le temps. Si la structure reste la même, le processus est alors dit stationnaire.

Par définition, « une série temporelle, est *stationnaire au sens strict* si la distribution conjointe de est identique à celle de quels que soient k le nombre d’instants considérés, (les instants choisis et t, le décalage. »  Si on choisit donc n’importe quel nombre de dates, si on décale ces dates d’une même quantité, et que la distribution ne change pas quels que soient les dates choisies, la série est stationnaire au sens strict. En d’autres termes, « la stationnarité stricte dit que la distribution conjointe de tout sous-vecteur de ,, quels que soient sa longueur et les instants choisis, est invariante quand on translate ces instants d’une même quantité ».

Bien évidemment, la stationnarité stricte est très difficile à satisfaire, et une définition plus faible la stationnarité est souvent utilisée.

#### Stationnarité faible

Une série temporelle, est *stationnaire au sens faible*  si :

* , constante indépendante de t
* , constante indépendant de t
* ne dépend que de l mais pas de t.

L’Espérance, et donc la moyenne, d’une série stationnaire est donc constante, ce qui suppose qu’elle ne doit pas présenter de tendance. Sa variance est aussi constante au cours du temps et n’est pas infinie. La troisième condition stipule que la covariance entre et (on parle aussi d’autocovariance, d’ordre ou de décalage l) ne dépend que de l’ampleur du décalage l et non pas de la position dans le temps t. On pourra remarquer que la troisième condition inclut la deuxième car pour l = 0, l’autocovariance est égale à la variance.

#### Bruit blanc

Afin de mieux comprendre l’analyse de nos données en utilisant la démarche préconisée pour les séries temporelles, certaine définitions sont importantes à savoir, comme celle d’un bruit blanc. Sa définition est la suivante :

«Un bruit blanc   est une suite de variables aléatoires non corrélées de moyenne nulle et de variance constante ». De par sa définition c’est donc une série stationnaire au sens faible. On peut adopter la notation suivante pour un bruit blanc

En statistique, les variables aléatoires non corrélées ne sont pas forcément indépendantes. On désigne par bruit blanc indépendant, une suite de variables aléatoires indépendants, de moyenne nulle et de variance constante.

Un bruit blanc est qualifié de gaussien quand c’est une suite de variables indépendantes, de moyenne nulle, de variance constante et qui suivent une loi normale. On note

Ce qui est important à retenir c’est qu’un bruit blanc, qu’il soit indépendant ou pas, gaussien ou pas, reste un processus stochastique stationnaire au sens faible.

### Appréciation de la stationnarité des séries par des méthodes graphiques :

On veut se faire une idée de la stationnarité de nos séries de pluies et de débits. La première chose à faire est l’observation du chronogramme. Les deux conditions nécessaires pour la stationnarité sont la constance de la moyenne et celle de la variance quand on prend n’importe quelle intervalle de la série, peu importe la position et la longueur de cette intervalle. Une moyenne constante implique que le graphe de la série montre toujours un niveau à peu près constant au fur et à mesure que le temps avance. Une variance constante suppose que les fluctuations autour de cette supposée moyenne restent toujours à peu près de même ampleur. Ces deux conditions se doivent d’être vérifiées quelle que soit la date autour de laquelle on examine la série.

##### Cas de la série hydrométrique

Une des premières propriétés d’une série stationnaire est la constance de sa moyenne. Examinons le chronogramme de la série hydrométrique qu’on a reconstituée (Figure X). Si l’on considère par exemple un intervalle de 3 mois que l’on fait glisser au fil du temps. On pourra constater que la moyenne des débits dans cet intervalle est de l’ordre de 20m3/s pour les premières années (1994 -1995) et que ça peut atteindre 35m3/s pour les dernières années (1998-1999). De plus, la moyenne sur 3 mois consécutifs est différente selon que ces mois soient centrés sur la période de crues ou la période d’étiage. Pour cette série, la moyenne dépend donc du temps, elle n’est donc pas stationnaire si l’on se réfère à une simple observation des outils graphiques.

##### Cas de la série pluviométrique

Le fait que la que la pluviométrie ait un caractère saisonnier implique plusieurs traits qui montre que notre série pluviométrique ne soit pas stationnaire. Tout comme pour la série des débits, la moyenne sur 3 mois consécutifs diffère énormément selon qu’on soit dans la saison des pluies ou pas. L’observation des monthplot et des lagplot nous permet de voir aussi ces indicateurs de non stationnarité (voir travaux précédents). Toutefois, la non-stationnarité n’est pas clairement affirmée par ces simples observations. On peut toujours constater une certaine régularité dans cette série pluviométrique, tout en sachant que régularité ne veut pas dire stationnarité.

### Evaluation de la stationnarité des séries par l’observation de leur correlogramme  :

L’examen du corrélogramme (empirique) d’une série permet de se faire une idée de la stationnarité de celle-ci. Il apporte un ajout non négligeable sur l’observation des autres types de graphiques (chronogramme, monthplot, lagplot etc. ). Le corrélogramme est le graphe qui permet de montrer la fonction d’autocorrélation.

Afin de définir ce qu’est la fonction d’autocorrélation, on doit d’abord connaître ce qu’est une fonction d’autocovariance. Pour une série donnée la covariance est tout simplement appelée *autocovariance d’ordre l*. On utilise aussi le terme décalage (ou retard), pour désigner l. Notons qu’ici, on suppose que soit stationnaire. **La fonction d’autocovariance** est la fonction qui à tout l, associe l’autocovariance. l étant un entier relatif.

On démontre que cette fonction est paire c'est-à-dire que . On ne la représente donc que pour l = 0,1,2,… Deux autres propriétés s’ajoutent à celle-ci.

 ;

.

*Le coefficient d’autocorelation d’ordre l* se définit comme suit :

Et **la fonction d’autocorrélation** (théorique) de la série et la fonction qui à tout l  entier naturel, associe

Cette fonction a les propriétés suivantes :

1

Ces deux fonctions sont les versions théoriques des fonctions d’autocovariance et d’autocorrelation. Il existe aussi les versions empiriques de ces fonctions.

L’autocovariance empirique d’ordre l se définit comme suit, en considérant une série observée, t = 1,…, T :

Avec

, le nombre d’observation.

Notons ici qu’on parle d’une série « observée » dont le nombre d’observation est égal à . La notation est différente de qui exprime le processus stochastique en général.

Le coefficient d’autocorrélation empirique d’ordre l est :

Par extension, la fonction d’autocovariance empirique est :

Et la fonction :

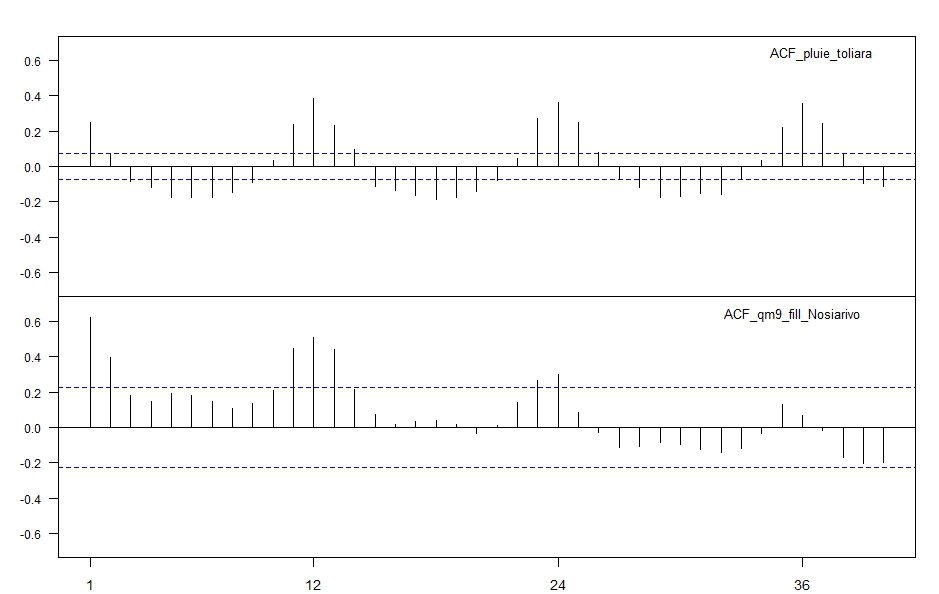
est **la fonction d’autocorrélation empirique**. Son graphique est donc le corrélogramme empirique. On peut calculer cette fonction pour toute série, qu’elle soit stationnaire ou non. Si la série n’est pas stationnaire, la fonction porte bien son nom car elle devient purement empirique. Cette fonction est largement utilisée pour l’analyse des séries temporelles. Une propriété qui a été démontrée et qui nous intéresse le plus dans cette partie est celle qui dit que **pour une série stationnaire, *le corrélogramme empirique*, graphe de**  **décroît exponentiellement vers 0,** avec éventuellement des oscillations. Ainsi, on peut dire qu’un graphe de fonction d’autocorrélation empirique sur lequel on ne remarque pas une décroissance rapide témoigne de la non-stationnarité de la série. Le Corrélogramme est donc un indicateur non négligeable.

Observons alors maintenant les graphes des fonctions d’autocorrélations (ACF) de nos séries hydro-climatiques (Figure X). Logiquement, l'ACF qui est presque nulle correspond à des données non corrélées, l'ACF tantôt positive, tantôt négative à des données périodiques, les autres à des données corrélées. Pour la pluie on constate qu’aux décalages 0 et 12, on a une forte corrélation et au décalage 6 on une corrélation inverse. Ça se passe de la même manière pour les décalages multiples de 6 et de 12 ( 18 , 24 et 30 , 36). La saisonnalité se voit alors très bien. On ne constate aucune décroissance pour l’ACF de la pluviométrie, ni rapide ni lente. Cela fait pencher l’interprétation vers la non-stationnarité de la série.

Pour le débit non plus, l’ACF ne décroît pas de manière rapide, ce qui indique une non-stationnarité. Par contre, on ne peut négliger le fait que les valeurs aux décalages supérieurs à 24 sont tous inférieurs aux valeurs correspondantes aux décalages précédents. De plus la valeur de l’ACF pour les décalages 6 et 18, ne sont pas négatives, contrairement à ce que l’on pourrait s’attendre pour une série ayant une composante saisonnière de composante 12. Et encore mieux, l’ACF pour les décalages 0, 12, 24, 36 sont positifs mais forment une série décroissante. Il y a donc une bien une décroissance de la fonction d’autocorrelaton mais elle est très lente. Cela pourrait être dû à la présence non négligeable d’une tendance. Cette décroissance n’est certes pas un indicateur de stationnarité. Elle permet juste d’extrapoler que la non-stationnarité de la série chronologique des débits n’est pas du même type que la non-stationnarité de la pluviométrie.

En effet, nous savons qu’une série peut être stationnaire ou non. Mais parmi les séries non-stationnaires, il y a différents types c'est-à-dire qu’il y a différentes formes de non-stationnarité. Par exemple, il y a des séries stationnaires en différence, ce qui veut dire que ces séries non-stationnaires deviennent stationnaires après différenciation. D’autres séries sont appelés stationnaires à une tendance près. Ces dernières sont des séries qui présentent une tendance et en enlevant cette tendance, on arrive à rendre ces séries stationnaires. Il est bon de remarquer qu’on peut très bien différencier une série non-stationnaire à une tendance près et aboutir quand même à une série stationnaire. L’inverse est aussi possible, même si ce n’est pas courant (soutirer une tendance à une série stationnaire en différence et aboutir à une série stationnaire).

Pour avoir une autre alternative en plus de l’observation des graphiques, il existe des tests statistiques qui permettent d’évaluer si la série est stationnaire ou pas. Et si elle ne l’est pas, des fois ces tests permettent d’avoir une idée sur la forme de la non-stationnarité.



### Test de stationnarité :

Il existe des tests statistiques qui permettent d’évaluer si une série donnée est stationnaire ou non. En général, il y a deux types de test, les tests qui prennent la stationnarité comme hypothèse nulle et celles qui la considèrent comme étant l’hypothèse alternative. Pour le premier type de test, on a par exemple le test KPSS (Kwiatkowski et al., 1992)ou le test de Leybourne et McCabe (Leybourne and McCabe, 1994). Les tests groupés dans la deuxième catégorie sont aussi appelés, test de racine unité. On observe par exemple le test de Dickey Fuller (Dickey et Fuller, 1979) ou sa version augmentée, ADF (Said et Dickey, 1984). Il y a aussi le test de Phillips-Perron (Phillips et Perron, 1988) et le test ERS d’Elliot (Elliott et al., 1996).

#### KPSS test

La fonction ur.kpss() de R permet d’effectuer le test KPSS de Kwiatkowski et al., (1992). Il nécessite le package **urca.** Ce test fait partie des premiers types de test de stationnarité, à savoir, celles qui prennent la stationnarité comme hypothèse nulle. Plus précisément, dans ce test l’hypothèse nulle est « la série est stationnaire soit en tendance, soit à une moyenne non nulle près « et l’hypothèse alternative est « la série est non stationnaire en un certain sens ».

Dans le test, il est supposé que la série résulte de l’addition de trois composantes : Une marche aléatoire, une tendance déterministe et une erreur stationnaire.

On désigne par marche aléatoire, une série de la forme :

La tendance déterministe peu avoir la forme suivante : . (Expliquer tendance déterministe et stochastique). Si on désigne par Ut, l’erreur stationnaire, la série est supposée avoir la forme suivante :

Afin de tester que la série est stationnaire à une tendance près, l’hypothèse nulle suppose qu’il n’y pas de composante « marche aléatoire » c’est à dire = 0.

La statistique du test de KPSS est la suivante (Kwiatkowski et al.,1992) :

Avec , est le résidu de la régression de sur la composante déterministe supposée

un estimateur convergent de la variance de , basé sur

est ici le nombre d’observation. C'est-à-dire que c’est une série observée, avec t = 1,…,T.

Sous l’hypothèse nulle, la loi de KPSS converge vers une loi qui dépend seulement de la forme de la tendance qui peut être un niveau β1 = 0 ou une tendance linéaire β1 +β2t

***On rejette l’hypothèse nulle de stationnarité pour de grandes valeurs de la statistique de test.***

##### Cas des pluies

Intéressons-nous d’abord aux pluies. Comme on a dit, il y a deux types de test KPSS. Le premier qui va suivre permet de savoir si notre série pluviométrique p.ts est stationnaire de moyenne constante. On utilise la fonction ur.kpss() avec l’option type = "mu". La fonction retourne la valeur de la statistique de test ainsi que les valeurs critiques pour les seuils de significativité.

summary(ur.kpss(p.ts,type="mu"))

#######################

# KPSS Unit Root Test #

#######################

Test is of type: mu with 6 lags.

2

Value of test-statistic is: 0.195

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.347 0.463 0.574 0.739

La statistique de test 0.195 est inferieur au seuil significatif de 10 % (ici 0.347) sans être largement inferieur. On ne rejette pas l’hypothèse de stationnarité.

Dans un deuxième temps, on cherche à savoir si la série pluviométrique p.ts est stationnaire à une tendance près. On utilise alors l’option type = "tau" dans ur.kpss()

> summary(ur.kpss(p.ts,type="tau"))

#######################

# KPSS Unit Root Test #

#######################

Test is of type: tau with 6 lags.

Value of test-statistic is: 0.0417

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.119 0.146 0.176 0.216

La statistique de test est très inferieure aux seuils significatifs. La stationnarité n’est pas rejetée. On comparant les valeurs des statistiques des deux tests, on pourrait privilégier une stationnarité à tendance linéaire plutôt qu’à une moyenne constante. Etrangement, le caractère saisonnier de la série n’est pas assez conséquent pour enlever sa stationnarité. Cela peut être dû à la longueur de la série. (Plus la série est longue elle est « diluée » dans la moyenne.)

##### Cas des débits

Procédons de la même manière pour notre série de débit reconstituée qm9\_fill. Testons pour savoir si qm9\_fill est stationnaire de moyenne constante

> summary(ur.kpss(qm9\_fill.ts,type="mu"))

#######################

# KPSS Unit Root Test #

#######################

Test is of type: mu with 3 lags.

Value of test-statistic is: 1.1838

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.347 0.463 0.574 0.739

La statistique à une très grande valeur (1.1838) par rapport aux niveaux de signification. On conclut très clairement au rejet de la stationnarité à une moyenne non nulle.

Refaisons le teste encore une fois mais cette fois-ci afin de savoir si qm9\_fill est stationnaire à une tendance linéaire près

> summary(ur.kpss(qm9\_fill.ts,type="tau"))

#######################

# KPSS Unit Root Test #

#######################

Test is of type: tau with 3 lags.

Value of test-statistic is: 0.0809

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.119 0.146 0.176 0.216

La statistique a une très petite valeur (0.0809). Si on considère ce résultat, on aboutit au non rejet de l’hypothèse de stationnarité à tendance près. Il peut arriver que les résultats des tests contredisent l’observation des graphiques, dans ce cas, il faut prêter attention avant tout aux chronogrammes. Mais compte tenu du caractère particulier de notre série ( cf, reconstitution d’un série présentant des manquants aux paragraphes précédents ), on aurait très bien pu s’attendre au résultat car on a plus ou moins « imposé » cette tendance linéaire lors dans notre méthode de reconstitution de la série. Ce résultat peut donc s’expliquer par la très forte influence de la tendance linéaire dans la série. Rappelons qu’une série stationnaire à une tendance près est une série non stationnaire mais qui le devient dès qu’on enlève sa tendance.

### Conclusions sur l’étude de la stationnarité des séries hydro-climatiques

Pour les deux séries (pluies et débits), l’observation des graphiques ramène à la non-stationnarité. L’Analyse des ACF nous suggère que les deux séries n’ont pas le même comportement même s’ils sont tous les deux non-stationnaires. Le test de stationnarité, en particulier celui du KPSS permet de trancher sur la forme de la non-stationnarité de la série hydrométrique. En effet le test rejette de façon claire l’hypothèse que la série de débits soit stationnaire à une moyenne non nulle près. Par contre, l’hypothèse de la stationnarité à une tendance près est loin d’être rejeté. En observant encore une fois le chronogramme des débits (Figure X), on arrive à comprendre qu’effectivement il se peut que la série soit non-stationnaire mais après mise en évidence de la tendance, elle serait stationnaire. Toute porte à croire que cette hypothèse est envisageable.

En ce qui concerne la série pluviométrique, l’analyse devient moins évidente. L’observation des graphiques suggère la non-stationnarité. Le test par contre nous dit que la série peut prendre les deux formes de non-stationnarité (c'est-à-dire être à la fois stationnaire à une tendance près et stationnaire à une moyenne non nulle près). Même si en comparant la valeur des statistiques des deux tests KPSS, on peut extrapoler que l’hypothèse de stationnarité à une tendance près puisse prendre le dessus, on ne peut pas trancher car la différence dans les résultats n’est pas aussi flagrante que pour le cas des débits.

Notons que dans la pratique, la plupart des séries ne sont généralement pas stationnaires. Les modèles statistiques sont toutefois plus aisés à mettre en œuvre pour les séries stationnaires. De ce fait, il est courant de rendre d’abord stationnaires les séries à étudier en utilisant la méthode adéquate (différenciation, retrait de la tendance etc.) avant d’entamer la modélisation proprement dite.

## MODELISATION. Deuxième étape : construction d’un modèle Arma.

Dans la suite, nous allons entamer la partie modélisation c'est-à-dire que nous allons rechercher des modèles statistiques qui représentent au mieux nos séries hydro-climatiques. Il faudrait que ces modèles puissent prendre en compte toute la dynamique de nos chroniques de pluies et de débits afin de pouvoir envisager une quelconque prédiction.

Nous avons tourné notre choix vers des types de modèles suffisamment complètes pour expliquer toute la complexité de nos chroniques. Les modèles ARMA en font parties. En statistique, les modèles ARMA (modèles autorégressifs et moyenne mobile), ou aussi modèle de Box-Jenkins (Références), sont parmi les modèles classiques les plus utilisés en séries temporelles. Le modèle ARMA est un outil pour comprendre et prédire, éventuellement, les valeurs futures d’une série temporelle donnée. Un processus ARMA est composé de deux parties : une part autorégressive. Dans les prochaines paragraphes nous allons expliquer un par un ces deux aspects du processus. Notons qu’un série se doit d’être stationnaire pour qu’on puisse y ajuster un ARMA. Nous avons vu pourtant qu’il était difficile de conclure à la stationnarité de nos chroniques hydro-climatiques. Nous verrons plus avant qu’il y a des méthodes pour rendre des séries stationnaires afin de les rendre adéquat pour l’ajustement d’un modèle ARMA.

### L’operateur retard

L’évolution d’une série est s’exprime souvent en fonction de son passé. Pour faciliter les manipulations des séries temporelles et avoir plus d’aisance dans la notation, on utilise souvent l’opérateur retard, note **B** ou **L**, selon les auteurs. C’est l’opérateur qui à tout élément d’une série temporelle, associe l’observation précédente.

Et par extension, on a

Grâce à cet opérateur, on peut écrire la plupart de séries temporelles usuelles comme un polynôme de l’operateur retard.

### Le processus autorégressif (AR)

Un processus { est un processus autoregressif d’ordre p AR(p), s’il obéit à

Avec

Avec l’opérateur retard, on peut l’écrire comme ceci :

, et la partie

est appelé opérateur d’autoregression

#### Conditions pour la stationnarité d’un processus autorégressif

Contrairement au processus moyenne mobile MA(q), un autorégressif AR(p) n’est pas forcément stationnaire. Pour qu’il le devienne, il faut vérifier certaines conditions.

### Le processus moyenne mobile (MA)

Un processus { est un processus autoregressif d’ordre p MA(q), s’il a la forme :

Où

En utilisant l’opérateur retard, on a :

On peut poser , qui est appelé opérateur moyenne mobile.

L’expression nous montre qu’un processus MA(q) revient à la somme d’un ou plusieurs bruits blancs, il est donc toujours stationnaire quels que soient les valeurs de , en veillant juste que la valeur de soit différente de zéro. Sa moyenne est égal à

### Le processus autorégressif et moyenne mobile (ARMA (p,q))

Comment son nom l’indique, un processus ARMA (p,q), il revient donc à la combinaison des deux processus précédents. De ce fait, il aura donc la forme suivante :

Avec c, constante et

Ou, avec l’opérateur retard,

Elle peut aussi s’écrire de la manière suivante, (vérifier la démonstration pour aboutir à cette expression).

### Le modèle ARMA saisonnier (SARMA)

Etant donné qu’on travaille sur des données hydro-climatiques, on s’attend énormément la présence d’une certaine saisonnalité dans nos chroniques. Le modèle sur lequel on se propose d’ajuster nos données se doit donc de posséder ce caractère saisonnier. C’est le cas du SARMA ou ARMA saisonnier.

Soit un ARMA (p,q) :

Dans lequel n’est plus forcément un bruit blanc dû à la présence de la saisonnalité (c’est d’ailleurs pour cela qu’on l’a nommé au lieu de). Afin de gérer cette non-blancheur, on pourrait modéliser par un ARMA dont l’unité de temps est la période de la saisonnalité. Dans le cas de nos chroniques hydro-climatiques, 12 serait le plus approprié. On peut donc supposer une modélisation de tel que :

Avec s, la période , on a alors :

Avec (à vérifier)

, polynômes de degré p et q en B

polynômes de degré P et Q en .

S’il est stationnaire et qu’il vérifie cet équation (X.X), on dit que est un SARMA(p,q)(P,Q)s

#### Conditions de stationnarité d’un SARMA

### Intégration des séries, ARIMA et SARIMA

On a vu que dans la pratique, la plupart des séries sont non stationnaires. On a vu aussi que cette non stationnarité peut prendre différentes formes. Un de ces aspects est la stationnarité en différence. Cela concerne les séries non stationnaires à partir desquels on obtient des séries stationnaires après différenciation. Les processus ARIMA et SARIMA (pour la version saisonnière) sont des processus non stationnaires qui deviennent des processus ARMA (ou SARMA) après une différenciation simple ou saisonnière.

Un processus est un SARIMA(p,d,q)(P,D,Q) s’il obéit à :

Les termes et expriment respectivement la différenciation simple et la différenciation saisonnière. Une série est dite intégrée d’ordre d s’il faut la différencier d fois pour obtenir une série stationnaire. De même, une série est dite intégrée saisonnièrement d’ordre D s’il faut la différencier saisonnièrement D fois pour obtenir une série stationnaire. L’ordre est donc l’ordre de différenciation simple, et D, l’ordre de différenciation saisonnière. La série différenciée est appelée la série des accroissements de .

### Démarche adoptée pour la construction d’un modèle ARMA Saisonnier

Afin d’aboutir à nos modèles de type SARMA (SARMA, saisonnier dans le cas de nos chroniques hydroclimatiques). Nous allons suivre la démarche classique de construction de ce genre de modèle (Aragon, 2011). Nous verrons que selon les propriétés de nos séries, nous allons adopter telle ou telle décision dans la mise en œuvre de nos modèles.

Tout d’abord, il faut vérifier la stationnarité des séries. Selon le cas, il faut faire les manipulations nécessaires pour aboutir à une série stationnaire modélisable par une démarche de type ARMA. Cela peut être une différenciation, le retrait d’une tendance ou toutes autres transformations susceptibles de donner une série stationnaire. Nous avons déjà fait une analyse de la stationnarité de nos chroniques hydro-climatiques dans les paragraphes précédents.

Une fois la série supposée stationnaire, on peut passer au choix des ordres p, q et P, Q. L’examen de la fonction d’auto corrélation ACF peut aider à faire le choix de ces valeurs. C’est l’étape d’*identification*. L’examen d’une autre fonction est aussi très utile dans cette étape, la fonction d’auto corrélation partielle PACF, nous allons en parler plus en avant dans les paragraphes qui suivent. Notons qu’il peut s’ajouter un autre type d’ordre appelé ordre de différenciation, d. C’est le cas si on est en présence d’une série stationnaire en différenciation. L’ordre d’identification d est alors le nombre de différenciation dont la série a besoin pour être stationnaire, donc modélisable. Donc, au lieu d’avoir les ordres p, q et P, Q pour un SARMA simple, on aura les ordres, p, d, q et P, D ,Q. d, est l’ordre de différenciation simple et D, l’ordre de différenciation saisonnier. On parle alors de modèle SARIMA (I , pour Integrated).

Une fois les ordres choisi, il faut estimer les paramètres et du modèle correspondant. On vérifie alors la qualité du modèle. Si le modèle n’est pas satisfaisant, on choisit d’autres ordres et on refait l’étape d’*estimation*. Dès qu’on abouti à un modèle convenable, on peut entamer la phase de *prédiction*. Les critères pour évaluer la qualité d’un modèle sont diverses mais la plus importante est la blancheur du résidu. Des tests permettent de conclure ou non à cette blancheur, comme le test du portemanteau (Box et Pierce, 1970). Si le résidu n’est pas blanc, c’est que « le modèle n’as pas capté toute le dynamique du phénomène » (Aragon, 2011), et il faut en chercher un autre. Quand cette blancheur est confirmée, il faut aussi tester la significativité des paramètres. On essaie alors d’aboutir à un modèle le plus simple possible.

#### Identification automatique des ordres

Il existe des outils qui permettent d’automatiser certaines étapes dans la construction d’un modèle SARMA. C’est le cas de la fonction *auto.arima()* disponible avec le package **forecast** de R (Hyndman et Khandakar, 2008). L’idée générale de l’automatisation c’est de trouver les ordres p,d,q et P,D,Q adéquats parmi un panel de modèle proposé. Les ordres de différenciations, d et D sont recherchés en utilisant une succession de différenciation et de test de stationnarité (KPSS test pour la différenciation simple et un test adapté appelé test de Canova\_Hansen pour la différenciation saisonnière). C'est-à-dire qu’on fait d’abord une différenciation d’ordre 1, on fait le test de stationnarité. Si la stationnarité n’est pas rejetée, l’ordre de différenciation est donc égal à 1. Sinon, passe à une différenciation d’ordre 2 et ainsi de suite, jusqu’à trouver le bon ordre. La même chose est faite pour l’ordre de différenciation saisonnière. Une fois les ordres d et D, trouvées, les ordres p, q , P et Q sont sélectionnés en utilisant un critère d’information, celui d’Akaike ou AIC( Akaike information criterion) (Akaike, 1974).

Si l’on considère un processus ARIMA(p,d,q)( P,D,Q)s , tel que :

L’AIC aura la forme :

Avec si et autrement.

(Likelihood) est le maximum de vraisemblance du modèle ajustés aux données différenciées.

La procédure automatique fait donc varier les ordres p, q, P et Q et **le modèle qui minimise la valeur de l’AIC sera la modèle sélectionnée**. Par contre cette variation des ordres ne se fait pas de manière aléatoire. La procédure n’essaie pas non plus de balayer toutes les valeurs possibles des ordres dans le but d’avoir toutes les combinaisons possibles, car cela prendrai trop de temps de calcul.

Hyndman et Khandakar (2008) ont proposé une démarche méthodique qui permet de sélectionner le meilleur modèle sans avoir recours à un temps de calcul top exagéré.

Pour commencer, la procédure essaie les quatre modèles suivants :

* ARIMA(2,d,2)(1,D,1)s
* ARIMA(0,d,0)(0,D,0)s
* ARIMA(1,d,0)(1,D,0)s
* ARIMA(0,d,1)(0,D,1)s

Parmi ces modèles, on sélectionne celle qui a la plus petite valeur d’AIC. Si l’une des valeurs d ou D est nulle, le modèle est ajusté avec c ≠ 0, sinon on a c = 0. Le modèle qui a le plus petit AIC devient alors « le modèle courant ». On génère alors d’autres modèles en faisant varier les ordres à partir de ce modèle courant. Les règles de variations sont les suivantes :

* On fait varier l’un des ordres p, q, D et Q de ± 1 par rapport au modèle courant.
* Puis, on fait varier p et q, en même temps, de ± 1 par rapport au modèle courant.
* Puis, on fait varier P et Q, en même temps, de ± 1 par rapport au modèle courant.
* Puis, on inclut la constant c1, si le modèle courant avait c1 = 0. Sinon on met c1 = 0 si le modèle courant avait c1 ≠ 0.

Ces variations nous permettent d’obtenir une trentaine de modèles. On compare alors ces modèles entre eux, le modèle courant inclus, et on sélectionne celle qui a la petite valeur d’AIC. Ce modèle sélectionné devient alors le nouveau modèle courant. On génère d’autres modèles, en suivant les règles de variations expliquées précédemment, à partir de ce modèle courant, puis on compare encore une fois leurs AIC. Celle qui a la plus petite valeur devient le nouveau modèle courant, et on recommence la procédure jusqu’à trouver le modèle idéal. La procédure s’arrête quand plus aucun modèle ne possède un AIC inférieur à celui de modèle courant. On peut spécifier des limites aux ordres qui seront pris en compte. Par défaut, cette limite est égale à 5 pour les ordres (p,q) et 2 pour les ordres (P,Q).

Notons que cette procédure automatique proposée par Hyndman et Khandakar (2008) consiste essentiellement en une sélection de modèles en se basant sur le critère d’information d’Akaike (AIC). Elle ne teste pas la qualité du modèle obtenu en ce qui concerne la blancheur de son résidu et la significativité des coefficients.

Il existe aussi d’autres moyens pour sélectionner les ordres en utilisant par exemple la méthode des moindres carrés (minimisation de la somme des carrés des résidus), au lieu du critère AIC. La fonction montre aussi d’autres critères tels que le BIC (Critère d’ Information Bayesien) ou le AICc (Correction de l’AIC pour le cas des échantillons de petite taille). Mais dans la suite de notre travail on utilisera exclusivement l’AIC.

Lors de la recherche du modèle, on peut très bien utiliser à la fois une méthode manuelle et une méthode automatique. Ce qui veut dire que si par exemple, à partir de nos observations, on constate qu’un ordre se doit d’avoir une valeur donnée, on peut imposer que le modèle doit obligatoirement avoir celle valeur là pour l’ordre donné et lancer la procédure automatique pour trouver les autres ordres. On parle alors de méthode semi-automatique. C’est cette manière de faire qu’on essaiera d’utiliser le plus car elle permet d’avoir une démarche réfléchie.

#### Identification manuelle des ordres par observation des graphiques d’auto corrélation (à rédiger)

#### Test de blancheur.

La condition nécessaire pour accepter la qualité d’un modèle ARIMA est la qualité de son résidu. Nous avons vu la définition d’un bruit blanc dans le paragraphe X. On rappelle que c’est une suite de variables aléatoires non corrélées de moyenne nulle et de variance constante. Pour vérifier cette non corrélation, Il faut que la fonction d’autocorrélation soit nulle à tous les décalages. La manière la plus courante de vérifier cette condition de blancheur est de faire un test de portemanteau. Comme son l’indique, à la manière d’un portemanteau qui rassemble plusieurs vêtements, ce test traite de manière simultanée plusieurs coefficients d’autocorrélation. Ce test utilise la statistique de Box Pierce test (Box and Pierce, 1970), ou la version de Ljung Box (Ljung and Box, 1978) adapté aux petits échantillons.

La *statistique de Box-Pierce* est la suivante, en considérant la série observée

Pour les petits échantillons, un version plus adaptée est utilisée, la *statistique de Ljung Box*

Avec , le décalage choisi, le nombre d’observation, et , le coefficient d’autocorrélation (empirique dans ce cas, car on parle d’une série observée). Le principe du test est simple. peut être considéré comme étant la distance du vecteur au vecteur. Si la but du test est de vérifier la nullité des coefficients d’autocorrélation, donc ces coefficients sont nuls pour des valeurs proche de zéro. En d’autres termes, si l’hypothèse nulle H0 pour ce test est «   » contre l’hypothèse alternative H1 « au moins un des est non nul, alors on rejette l’hypothèse nulle H0  pour les grandes valeurs de .

Dans le contexte d’une démarche de modélisation, comme notre cas, le test ne sera pas forcément appliqué sur une série de variables aléatoires mais sur les résidus de l’ajustement d’un modèle

Il existe la fonction Box.test2() du package **cashchrono** de R qui permet d’effectuer ce test de portemanteau. La statistique de Ljung Box semble adapté à nos données hydrométriques étant données que nous de disposant des de chroniques assez courtes. Par contre, pour la longue série pluviométrique dons nous disposons, on peut envisager l’utilisation de celle de Box Pierce.

Si p-value = 0, blancheur rejeté.

#### Significativité des coefficients

Une fois que la qualité principale du modèle est confirmée, à savoir, la blancheur du résidu, on peut alors vérifier si tous les coefficients sont significatifs. Pour ce faire, il faut calculer la statistique du test. Elle est égale à la valeur du coefficient divisé par son écart-type. Il s’agit de l’erreur-type du coefficient. La p-value associé à ce test est approximativement la probabilité qu’une variable aléatoire dépasse ce quotient en valeur absolu. **On rejette l’hypothèse que le coefficient soit non significatives pour des valeurs très faibles de la p-value**.

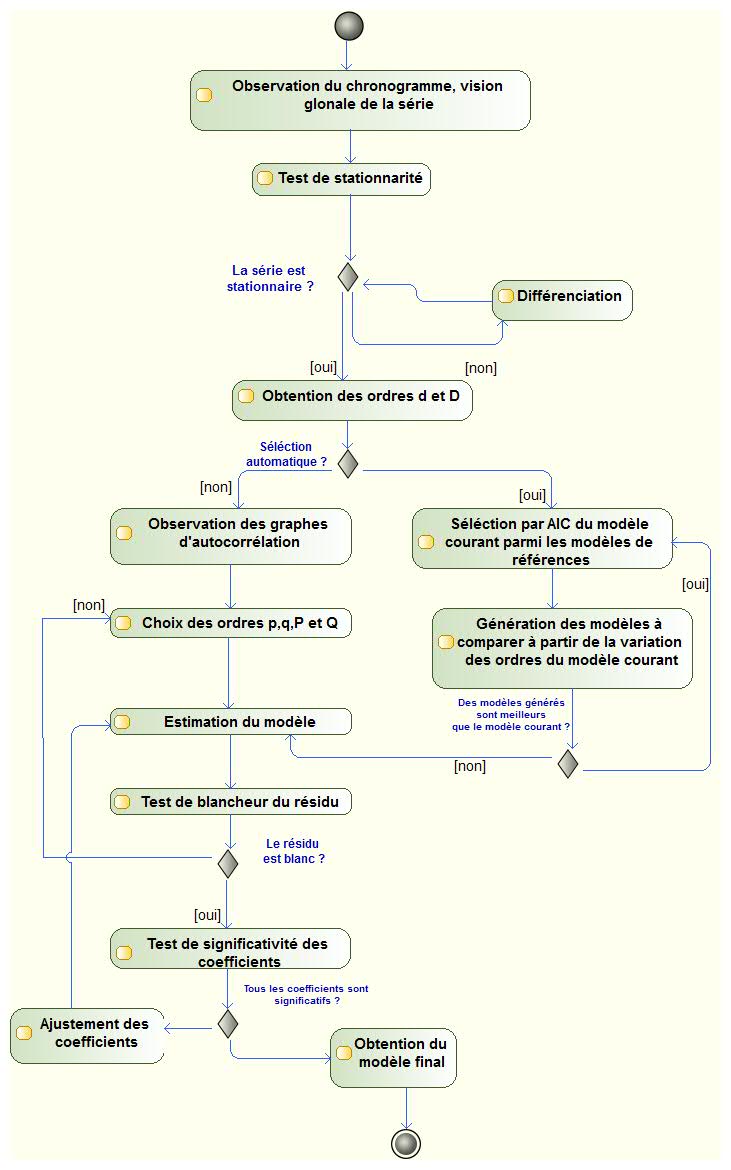
La fonction t\_stat() de **cashchrono** donne directement la valeur de la statistique de test et de sa p-value.

#### Résumé de la procédure de construction d’un modèle ARIMA,

En résumé, la construction du modèle peut être divisée en plusieurs étapes :

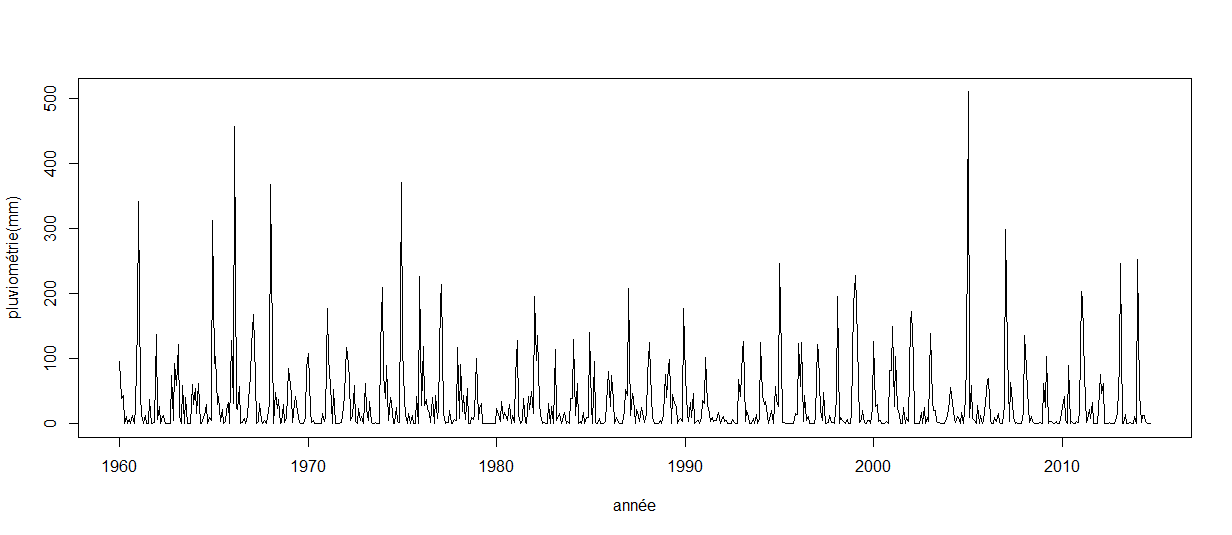
* Observation du chronogramme et d’autres outils graphiques si nécessaires (monthplot, lagplot) pour avoir une vision globale de la chronique à étudier. Cette étape peut s’avérer banale mais elle est très importante. Si jamais pendant les autres phases, on constate que les résultats des tests aboutissent à une conclusion différente de ce que le chronogramme montre d’une façon évidente. C’est le chronogramme qu’il faut prendre en compte.
* Tester la stationnarité de la série et dans le cas d’une non stationnarité, faire les différenciations pour la rendre stationnaire. C’est ici donc qu’on obtient les valeurs des ordres d et D.
* Choisir les ordres p, q, P et Q du modèle. Ce choix peut se faire par l’observation des graphes d’auto corrélation ou d’auto corrélation partielle. Une fois les ordres choisi, on estime le modèle. Notons que le choix peut également se faire en prenant des valeurs et en comparer les critères AIC de modèles sélectionnés.
* Tester la qualité du modèle obtenu, en vérifiant la blancheur du résidu (par l’observation de l’ACF ou par un test KPSS).
* Vérifier la significativités des coefficients. Si des coefficients ne sont pas significatifs, on peut les supprimer en les contraignant à zéro, tout en essayant de garder les ordres du modèle. Dans ce cas, on ré estime le modèle dans ce sens et on refait les tests de qualité (blancheur du résidu et significativité des paramètres).

Les trois premières étapes sont les étapes qui peuvent être automatisées. Il faut donc toujours tester la qualité du modèle, tant avec une méthode manuelle qu’avec une méthode automatique. Ces étapes sont résumées sur la figure X.



### Modélisation de la série pluviométrique

Considérons nos données pluviométriques avec une vue d’ensemble par l’intermédiaire du chronogramme des pluies. Dans toute la suite, la chronique pluviométrique considérée en tant que série temporelles sera nommée *p.ts*



#### Approche par automatisation

Essayons de voir tout de suite ce que donne une modélisation automatique. La fonction auto.arima() nous suggère un SARIMA (1,0,1)(1,0,1)12. Nommons le *modp0*.

modp0=auto.arima(p.ts)

Series: p.ts

ARIMA(1,0,1)(1,0,1)[12] with non-zero mean

Coefficients:

ar1 ma1 sar1 sma1 intercept

0.2266 -0.2134 0.9431 -0.7757 25.1366

s.e. 0.1122 0.1058 0.0164 0.0343 8.4551

sigma^2 estimated as 2447: log likelihood=-3495.45

AIC=7002.97 AICc=7003.1 BIC=7029.89

Testons la qualité du modèle à travers un test de blancheur du résidu. La fonction Box.test2() du package **cashchrono** nous permet de faire un test de portemanteau dessus. Les arguments demandés par la fonction sont la série à tester, les retards sur lesquels on veut faire le test, le type de statistique à utiliser, les décimales à prendre et le nombre de degré de liberté qu’il faut soustraire dans le cas d’un test appliqué sur les résidus d’un ajustement (ce qui est notre cas). On avait dit dans le paragraphe X de description du test qu’il est préférable vu la longueur de nos chronique de pluies d’utiliser la statistique de Box-Pierce. Pour le dernier argument, vu qu’on applique le test sur les résidus provenant de l’ajustement d’un ARMA (p,q)(P,Q)s, les auteurs suggèrent que la meilleure valeur de cet argument est égal à p+q+P+Q (Box and Pierce, 1970; Ljung and Box, 1978). Faisons le test pour les retards multiples de 6 c'est-à-dire 6, 12, 24 et 30.

> Box.test.2(residuals(modp0),nlag=ret,type="Box-pierce",decim=5,fitdf=4)

Retard p-value

[1,] 6 0.94430

[2,] 12 0.99514

[3,] 18 0.97072

[4,] 24 0.97375

[5,] 30 0.99513

Dans le cas pratique, la blancheur est rejetée si les p-value sont nulles pour tous les retards. Ce qui n’est pas notre cas ici, on ne rejette donc pas la blancheur du résidu.

Testons maintenant la significativité des coefficients. Rappelons qu’on rejette la significativité des coefficients pour des valeurs de la p-value qui ne sont pas très faible.

> t\_stat(modp0)

ar1 ma1 sar1 sma1 intercept

t.stat 2.018895 -2.018115 57.54925 -22.62183 2.972939

p.val 0.043498 0.043579 0.00000 0.00000 0.002950

Dans le cas de ce premier modèle, on ne rejette pas la significativité des coefficients. On peut aussi essayer une autre approche en prenant les ordres p,q,P et Q suggéré par auto.arima() et faire l’ajustement d’un autre modèle en utilisant Arima() du package **forecast**. Cette fonction permet donc d’ajuster un ARIMA sur une série donnée en permettant à l’utilisateur de spécifier les ordres. Soit *modp1* ce nouveau modèle.

modp1=Arima(p.ts,order=c(1,0,1),seasonal=list(order=c(1,0,1),period=12))

Series: p.ts

ARIMA(1,0,1)(1,0,1)[12] with non-zero mean

Coefficients:

ar1 ma1 sar1 sma1 intercept

0.6223 -0.6054 0.9962 -0.9294 32.5298

s.e. 0.5482 0.5507 0.0024 0.0196 11.2423

sigma^2 estimated as 2197: log likelihood=-3470.07

AIC=6952.13 AICc=6952.26 BIC=6979.06

Testons la blancheur du résidu pour *modp1.*

Box.test.2(residuals(modp1),nlag=ret,type="Box-pierce",decim=5,fitdf=4)

Retard p-value

[1,] 6 0.98654

[2,] 12 0.99763

[3,] 18 0.98873

[4,] 24 0.99386

[5,] 30 0.99915

Les p-value sont largement supérieur à zéro, on conclut à la blancheur du résidu. Testons la significativité des coefficients (si la p\_value est très faible, le coefficient est significatif)

> t\_stat(modp1)

ar1 ma1 sar1 sma1 intercept

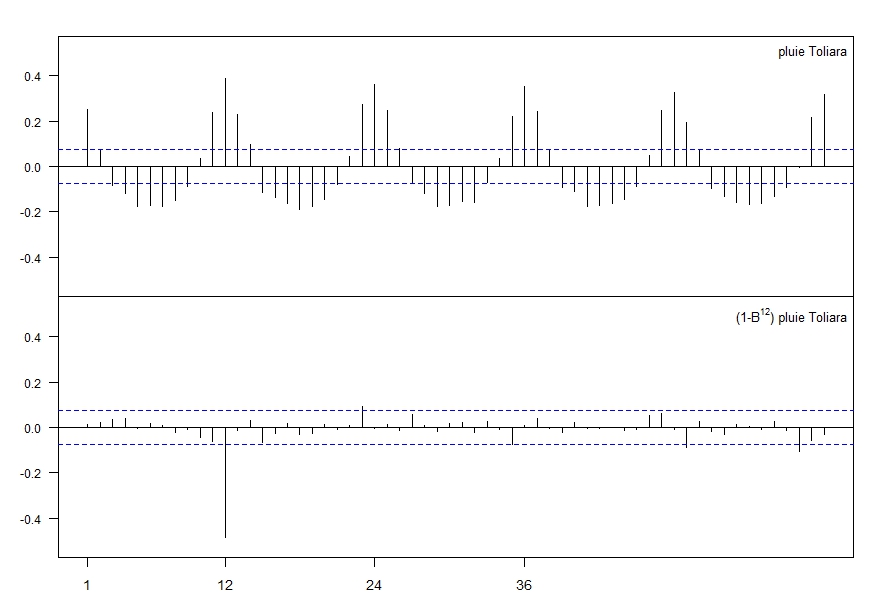
t.stat 1.135114 -1.099367 412.5805 -47.32158 2.893514

p.val 0.256327 0.271608 0.0000 0.00000 0.003810

On constate que les coefficients saisonniers sont très significatifs, par contre les paramètres autoregressifs et moyennes mobiles simples ne le sont pas trop. La significativité des coefficients est meilleure pour le modèle *modp0.* Par contre en comparant les AIC, le modèle *modp1* a une valeur de critère plus petite (6952.13) que celle de *modp0* (7002.97). Avant de trancher sur le modèle à prendre, passons à une démarche manuelle, plus réfléchie.

#### Approche manuelle

Ce qui nous a le plus étonné dans cette démarche automatique c’est que auto.arima() n’a suggéré aucune différenciation que ça soit simple ou saisonnière. Pourtant, en regardant l’ACF de la pluviométrie de Toliara on constate la non-stationnarité de cette série pluviométrique. Cette stationnarité est par contre obtenue après une différenciation saisonnière de période 12. Ci-dessous l’ACF de la pluviométrie et de la pluviométrie différenciée à l’ordre 12 (Figure X). On voit clairement que la série pluviométrique ne décroit pas vers zéro alors que c’est le cas après différenciation.



On peut alors extrapoler en disant que la non stationnarité peut être due en grande partie à la saisonnalité. Il est donc judicieux de faire une différenciation saisonnière. On peut toujours faire appel a auto.arima() mais cette fois-ci en imposant une différenciation saisonnière c'est-à-dire en spécifiant D = 1. On obtient alors un SARIMA (0,0,0)(0,1,1)12. Nommons le modèle obtenu *modp2*.

modp2=auto.arima(p.ts, D=1)

Series: p.ts

ARIMA(0,0,0)(0,1,1)[12]

Coefficients:

sma1

-0.9268

s.e. 0.0186

sigma^2 estimated as 2204: log likelihood=-3409.62

AIC=6823.25 AICc=6823.27 BIC=6832.19

Testons de la blancheur du résidu.

> Box.test.2(residuals(modp2),nlag=ret,type="Box-Pierce",decim=5,fitdf=1)

Retard p-value

[1,] 6 0.95356

[2,] 12 0.99004

[3,] 18 0.98398

[4,] 24 0.99354

[5,] 30 0.99928

On ne rejette pas la blancheur du résidu.

On a ici qu’un seul coefficient, un coefficient moyenne mobile saisonnier. Testons la significativité de ce coefficient.

t\_stat(modp2)

sma1

t.stat -49.95521

p.val 0.00000

La p-value est très faible, ce qui confirme la significativité du coefficient.

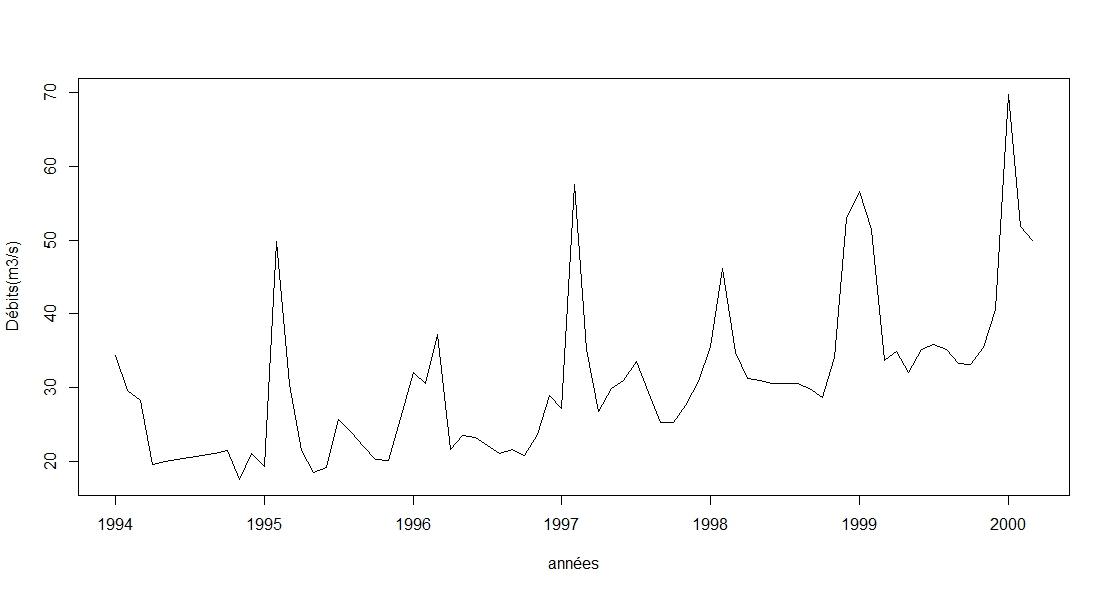
Essayons de trancher parmi les trois modèles présentés précédemment. Si on regarde les critères AIC, le choix se penche vers *modp2.* L’AIC de de *modp0* est de 7002 , celle de *modp1* 6952, et celle de *modp2* 6823 .

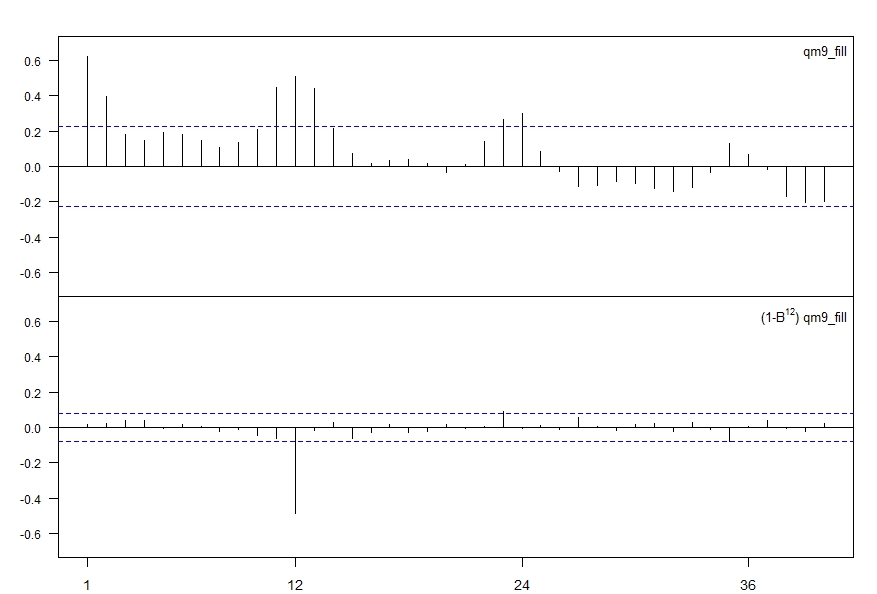
Le plus important c’est que modp2 suggère une différenciation, ce qui n’est pas le cas des deux autres. Or, en observant l’ACF de la série de références, on constate que cette différenciation est nécessaire. La blancheur du résidu provenant de l’ajustement de *modp2* et la significativité de son coefficient nous amène à sélectionner ce modèle SARIMA (0,0,0)(0,1,1)12 .

### Modélisation de la série hydrométrique

En ce qui concerne nos chroniques hydrométriques, prenons *qm9\_fill* la série « reconstituée » obtenue selon la méthode de reconstitution proposée dans les paragraphes précédents (voir travaux précédents).

Observons le chronogramme de cette série (Figure X), ainsi que ses corrélogrammes (Figure X).





L’observation de la fonction d’auto-corrélation ACF de la chronique de débits reconstitué et l’ACF de sa série différencié à l’ordre 12 montre qu’une fois la différenciation faite la série est stationnaire. En effet, le graphe d’autocorrelation (correlogramme) décroît exponentiellement vers zéro. L’ACF de la série non différenciée présente aussi une décroissance mais elle n’est pas assez rapide. Par contre, cette décroissance est clairement affichée pour la série différenciée. Les pic aux ordres 12, 24 et 36 témoignent de l’existence de la composante saisonnière.

#### Modélisation par automatisation

La fonction auto.arima() suggère un SARIMA(0,0,0)(1,1,0). C’est fois-ci elle a suggéré un ordre de différenciation saisonnière D = 1 c'est-à-dire qu’il faut différencier une fois la série à l’ordre 12 (différenciation saisonnière) avant d’aboutir à une série stationnaire, condition requise pour une estimer un ARMA dessus. Cette suggestion automatique appuie alors les observations des graphiques.

Le modèle suggéré par auto.arima est un SARIMA(0,0,0)(1,1,0)12 avec dérive (drift) nommons le *modq0*.

modq0=auto.arima(qm9\_fill.ts)

Series: qm9\_fill.ts

ARIMA(0,0,0)(1,1,0)[12] with drift

Coefficients:

sar1 drift

-0.5799 0.2886

s.e. 0.1122 0.0497

sigma^2 estimated as 48.07: log likelihood=-167.72

AIC=341.45 AICc=341.85 BIC=347.88

Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set -0.09905707 5.717544 3.346651 -3.218458 10.10402 0.6434194 0.01473467

Testons la blancheur du résidu. Cette fois-ci, nous allons utiliser la statistique de Ljung-Box au lieu de celle de Box-Pierce car on est face à une chronique assez courte.

> Box.test.2(residuals(modq0),nlag=ret,type="Ljung-Box",decim=5,fitdf=1)

Retard p-value

[1,] 6 0.99081

[2,] 12 0.99339

[3,] 18 0.98743

[4,] 24 0.99881

[5,] 30 0.99941

La valeur des p-value nous amène à rejeter la non blancheur du résidu.

Passons au test de significativité des coefficients.

> t\_stat(modq0)

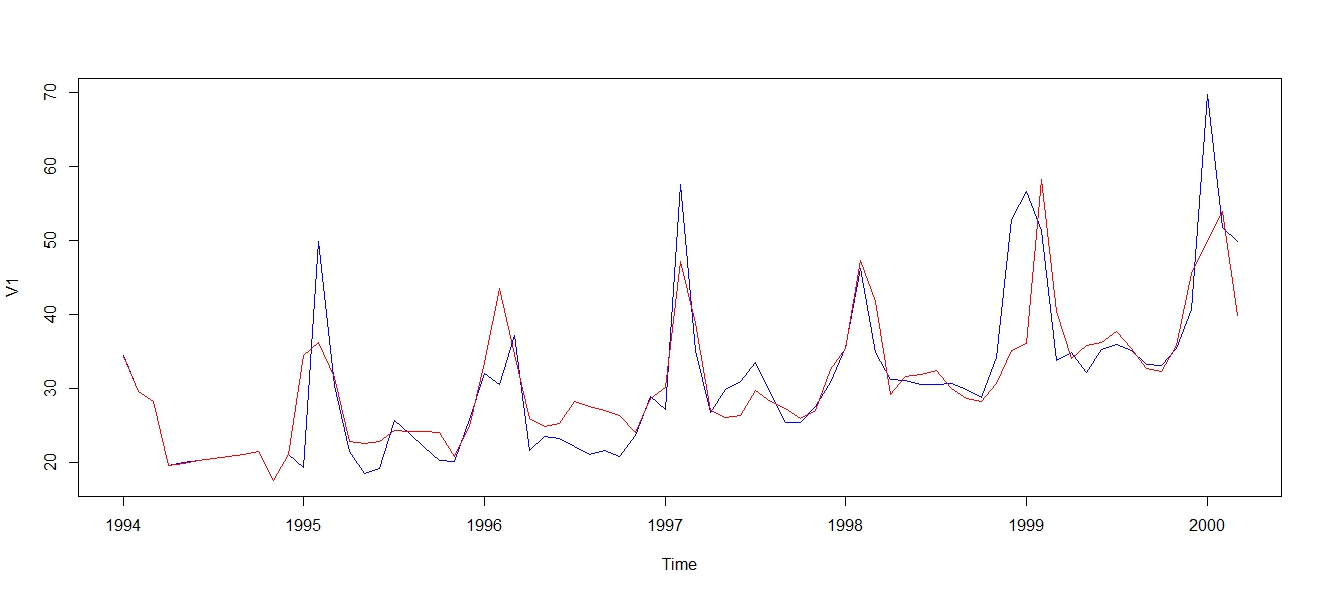
sar1 drift

t.stat -5.165986 5.808765

p.val 0.000000 0.000000

Les p-value sont très faibles, les coefficients sont donc significatifs.

Ci-dessous un aperçu du modèle comparé avec la série de référence.



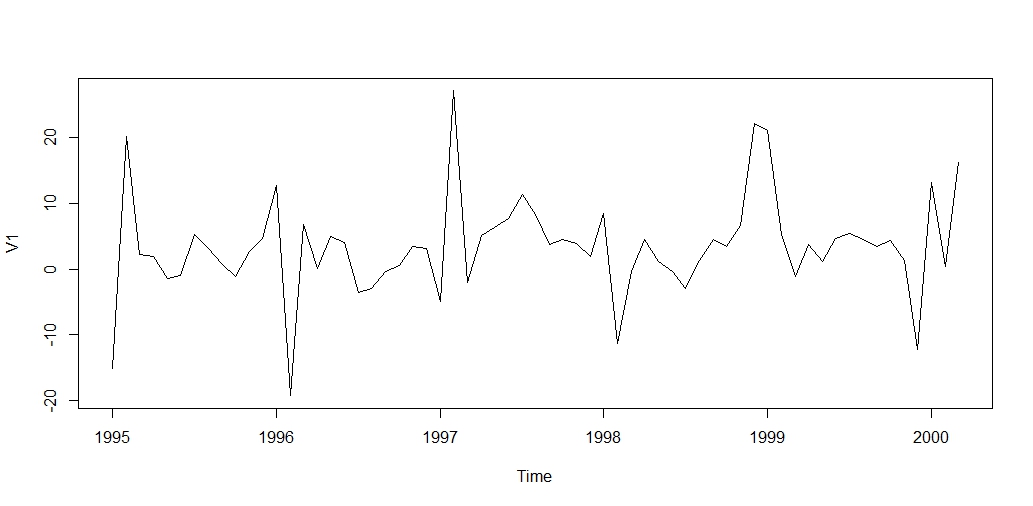
Le drift , égal à 0.2886 ici est la quantité qui s’ajoute à yt à chaque période.

#### Modélisation manuelle

Précédemment, en examinant l’ACF de notre série hydrométrique ainsi que l’ACF de la série différenciée saisonnièrement, on a conclu que pour avoir une stationnarité, il faut passer par cette différenciation saisonnière.

Examinons donc cette série différenciée.

qm9\_diff.ts=diff(qm9\_fill.ts,12)



Série des débits différenciée saisonnièrement.

En observant cette série différenciée, on pourrait penser qu’elle est de moyenne nulle. Calculons alors cette moyenne. On constate que ce n’est pas le cas. Il faut donc introduire une dérive dans le modèle de qm9\_fill.

mean(qm9\_diff.ts)

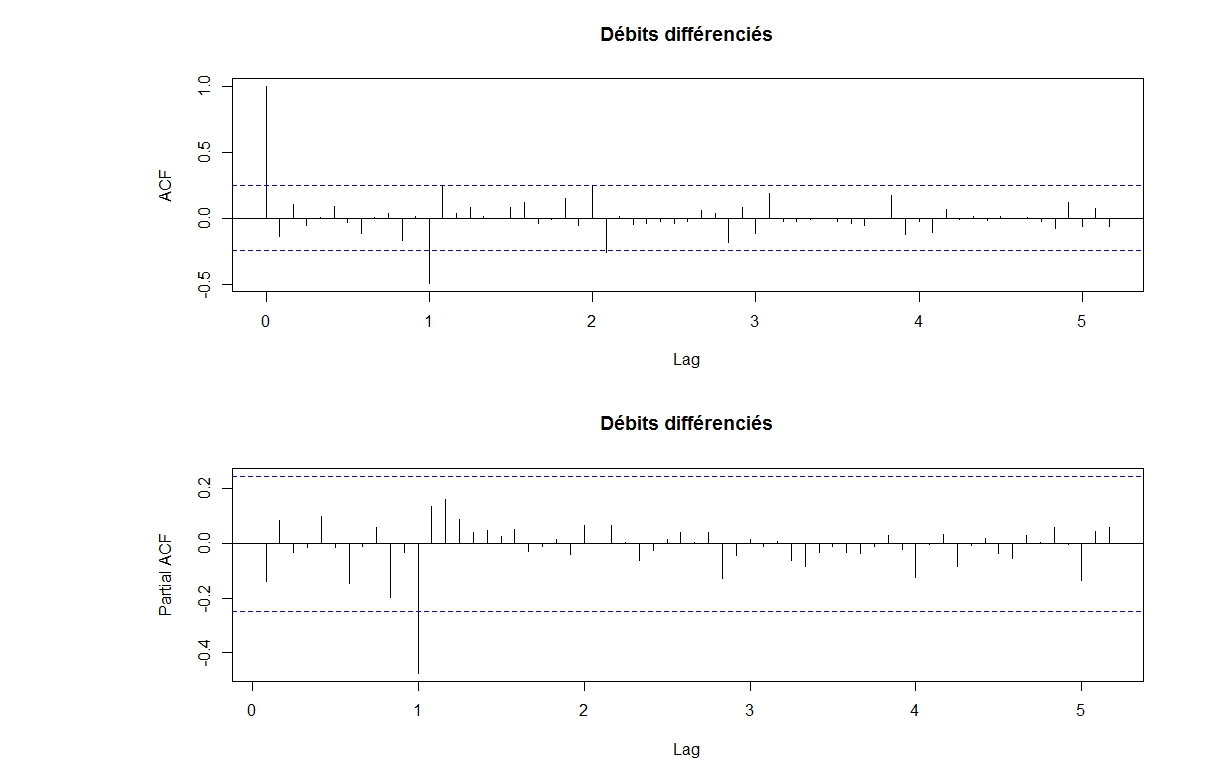
[1] 3.37619

Examinons alors les ACF et PACF de cette série différenciée

par(mfrow = c(2,1))

> acf(qm9\_diff.ts,main="Débits différenciés",lag.max=4000)

> pacf(qm9\_diff.ts,main="Débits différenciés",lag.max=4000)



Après examens des ACF et PACF, on constate que le PACF semble décroître plus vite que l’ACF. Essayons alors de privilégier un modèle autorégressif saisonnier d’ordre 1. (Justification dans les propriétés des ACF et des PACF pour les AR et MA, paragraphe à rédiger) en incluant une dérive.

|  |
| --- |
| > modq1=Arima(qm9\_fill.ts,seasonal=list(order=c(1,1,0),period=12),include.drift=TRUE) |
|  |
| |  | | --- | |  | |

Series: qm9\_fill.ts

ARIMA(0,0,0)(1,1,0)[12] with drift

Coefficients:

sar1 drift

-0.5799 0.2886

s.e. 0.1122 0.0497

sigma^2 estimated as 48.07: log likelihood=-167.72

AIC=341.45 AICc=341.85 BIC=347.88

Testons la blancheur du résidu

> Box.test.2(residuals(modq1),nlag=ret,type="Ljung-Box",decim=5,fitdf=1)

Retard p-value

[1,] 6 0.99081

[2,] 12 0.99339

[3,] 18 0.98743

[4,] 24 0.99881

[5,] 30 0.99941

Testons la significativité des coefficients.

> t\_stat(modq1)

sar1 drift

t.stat -5.165986 5.808765

p.val 0.000000 0.000000

Ils sont significatifs

## PREDICTIONS

### Prévisions issu des modèles SARIMA estimés (à faire)

### Modèle simple de prévision à court terme : Le lissage exponentiel (basé sur les travaux précédents).

Le lissage exponentiel est un modèle de prévision qui utilise les valeurs passées d’une série temporelle. Il en existe plusieurs types selon les caractéristiques de la série à explorer/modéliser/prévoir.

* Le lissage exponentiel simple : pour les chroniques dépourvues d’une tendance ou d’une saisonnalité.
* Le lissage exponentiel double : pour les séries présentant une tendance.
* Le lissage de hot Winters : pour les séries possédant une tendance et une composante saisonnière juxtaposées (de manière additive ou multiplicative).
* Enfin on peut aussi faire des modèles ayant une composante saisonnière mais sans aucune tendance distincte (comme le cas de notre chronique de pluie à Toliara par exemple).

#### Le lissage simple

Le principe est simple. Par exemple, pour le lissage exponentiel simple, on cherche à obtenir une valeur lissée en t pour le reporter en t + 1.

On a donc comme paramètres **une prévision initiale** (c’est souvent une moyenne des dernières observations) et **une constante de lissage** (idéalement entre 0.1 et 0.3). On peut alors chercher les valeurs optimales des paramètres en comparant les valeurs observées et simulées et en se basant sur les erreurs suivantes : Erreur moyenne (ME), Erreur quadratique moyenne (MSE), et Erreur absolue moyenne (MAE).

L’ajustement peut se faire avec le package forecast de R. ets() (Hyndman et Khandakar , 2008).

#### Le lissage double

Pour les séries sans saisonnalité à tendance localement linéaire, le modèle de prévision possède un **niveau** et une **pente.**

Niveau :

Pente :

La prévision à la date t pour l’horizon h (c'est-à-dire pour t + h) est :

Les paramètres sont donc **le niveau initial, la pente initiale, .**

#### Le lissage de Hot Winters

Ce modèle est adapté pour les séries ayant à la fois une tendance et une saisonnalité. Les composants du modèle sont : **le niveau**, la **pente** auxquels on ajoute **la saisonnalité**.

Dans le cas d’un schéma additif, on a ceci.

Niveau :

Pente :

Saisonnalité :

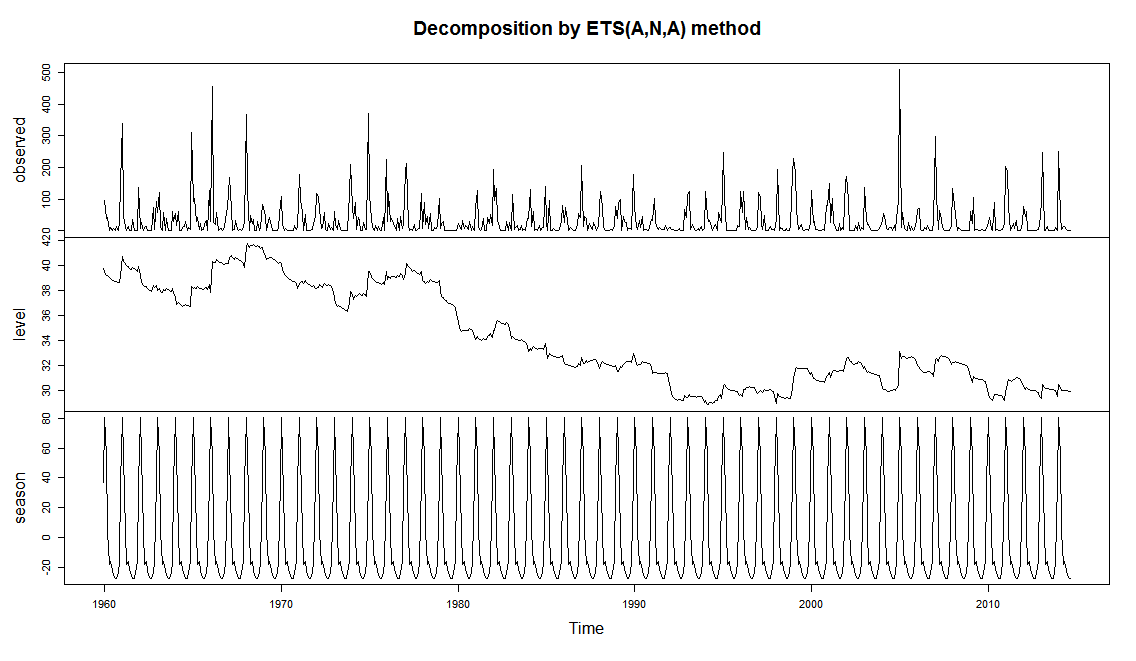
La prévision à la date t pour l’horizon h (c'est-à-dire pour t + h) est :

Les paramètres sont donc,  **le niveau initial, la pente initiale et les coefficients saisonniers initiaux.**

Enfin, pour les modèles ayant une composante saisonnière mais sans aucune tendance distincte (comme le cas de notre chronique de pluie à Toliara par exemple), on aura donc un niveau, une saisonnalité mais pas de pente.

#### Prévision des pluies :

Série avec saisonnalité et tendance constante.



> summary(p.ets)

ETS(A,N,A)

Call:

ets(y = p.ts, model = "ANA")

Smoothing parameters:

alpha = 0.0068

gamma = 1e-04

Initial states:

l = 39.7168

s=37.2062 -16.3968 -23.5387 -27.5563 -27.5865 -25.6063

-21.7816 -16.2614 -17.9401 3.8544 55.0851 80.5221

sigma: 46.1177

AIC AICc BIC

9324.602 9325.256 9387.430

Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set -2.176327 46.11774 26.16178 -Inf Inf 0.6946571 -0.02475023

> p.predict=predict(p.ets,12)

> p.predict

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Oct 2014 6.412307 -52.689960 65.51457 -83.976810 96.80142

Nov 2014 13.552008 -45.551635 72.65565 -76.839213 103.94323

Dec 2014 67.150956 8.045938 126.25597 -23.242369 157.54428

Jan 2015 110.480664 51.374271 169.58706 20.085236 200.87609

Feb 2015 85.030124 25.922354 144.13789 -5.367408 175.42766

Mar 2015 33.802234 -25.306911 92.91138 -56.597402 124.20187

Apr 2015 12.006512 -47.104009 71.11703 -78.395228 102.40825

May 2015 13.683249 -45.428646 72.79514 -76.720593 104.08709

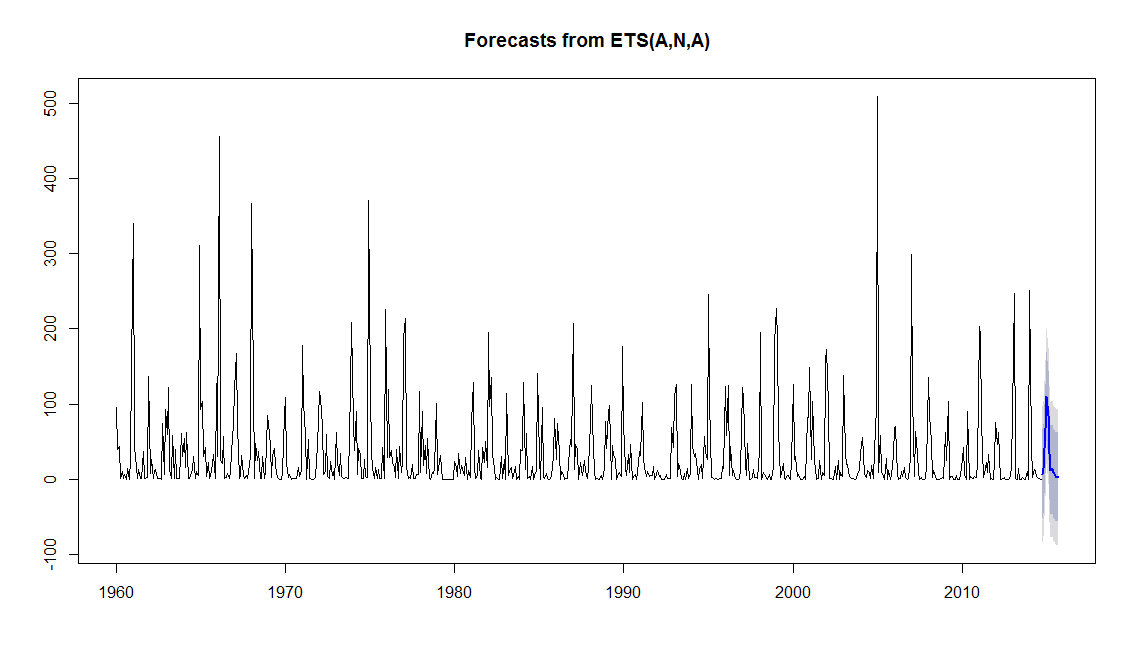
Jun 2015 8.167815 -50.945456 67.28109 -82.238131 98.57376

Jul 2015 4.336208 -54.778438 63.45085 -86.071841 94.74426

Aug 2015 2.367916 -56.748146 61.48398 -88.042299 92.77813

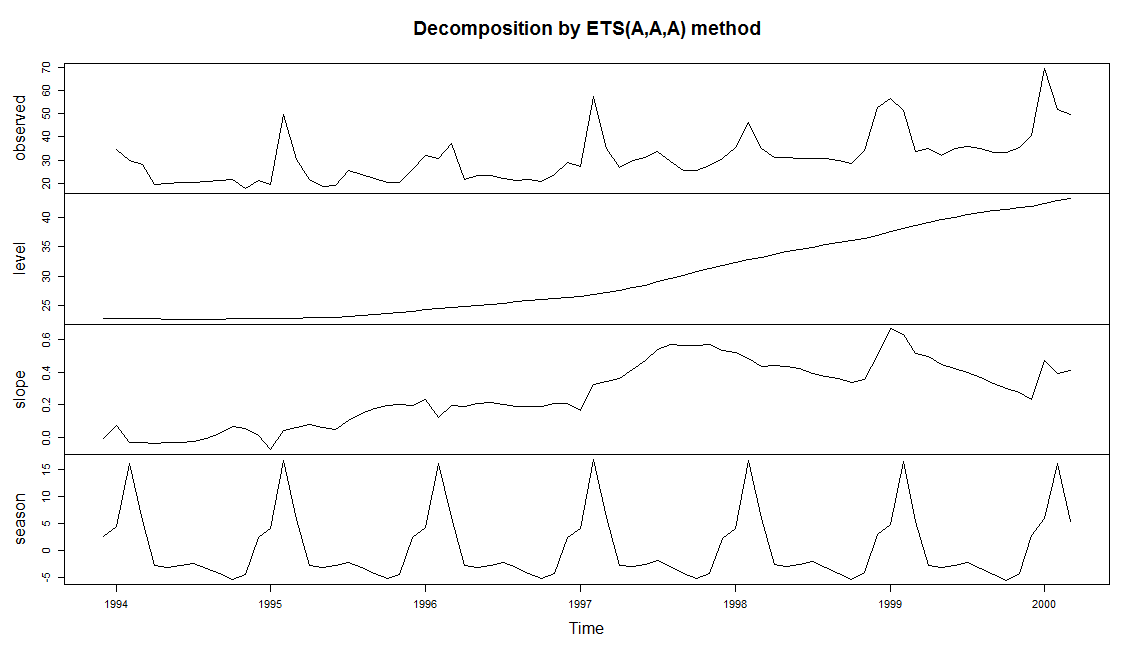
Sep 2015 2.402273 -56.715165 61.51971 -88.010046 92.81459

> plot(p.predict)



#### Le modèle du debit:

Série avec tendance et saisonnalité



> summary(qm9\_fill.ets)

ETS(A,A,A)

Call:

ets(y = qm9\_fill.ts, model = "AAA")

Smoothing parameters:

alpha = 0.0104

beta = 0.0104

gamma = 0.0476

Initial states:

l = 22.7572

b = -0.0028

s=2.6497 -4.3324 -5.5198 -4.4642 -3.3641 -2.4488

-2.7982 -3.1532 -2.8113 5.5902 16.5488 4.1032

sigma: 5.5134

AIC AICc BIC

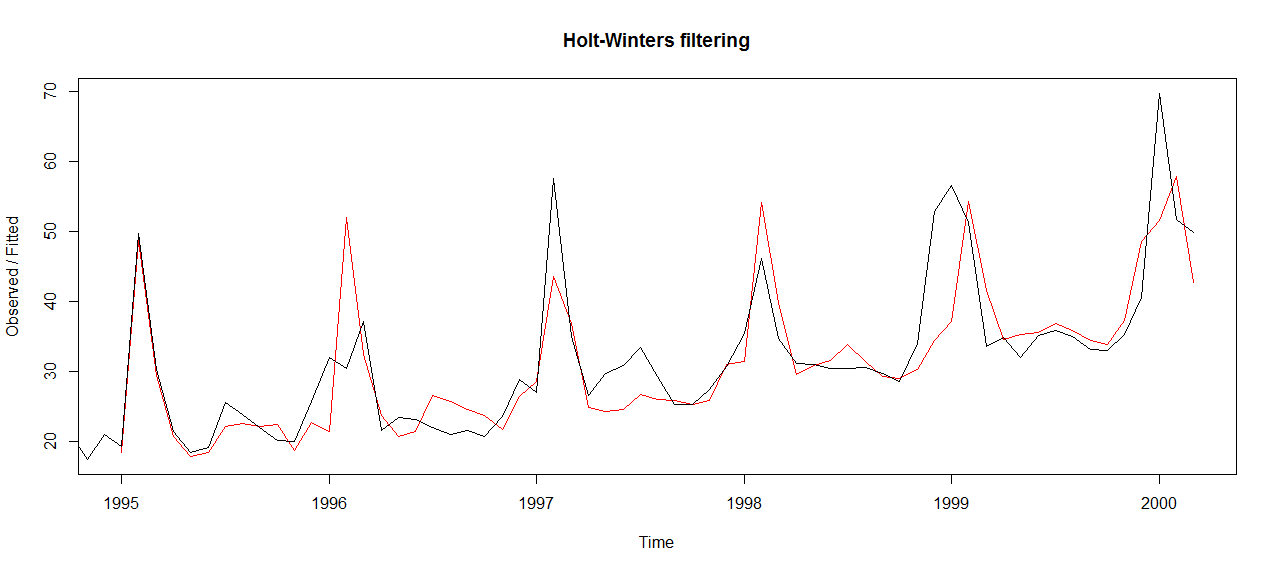
611.8887 621.2680 648.9685

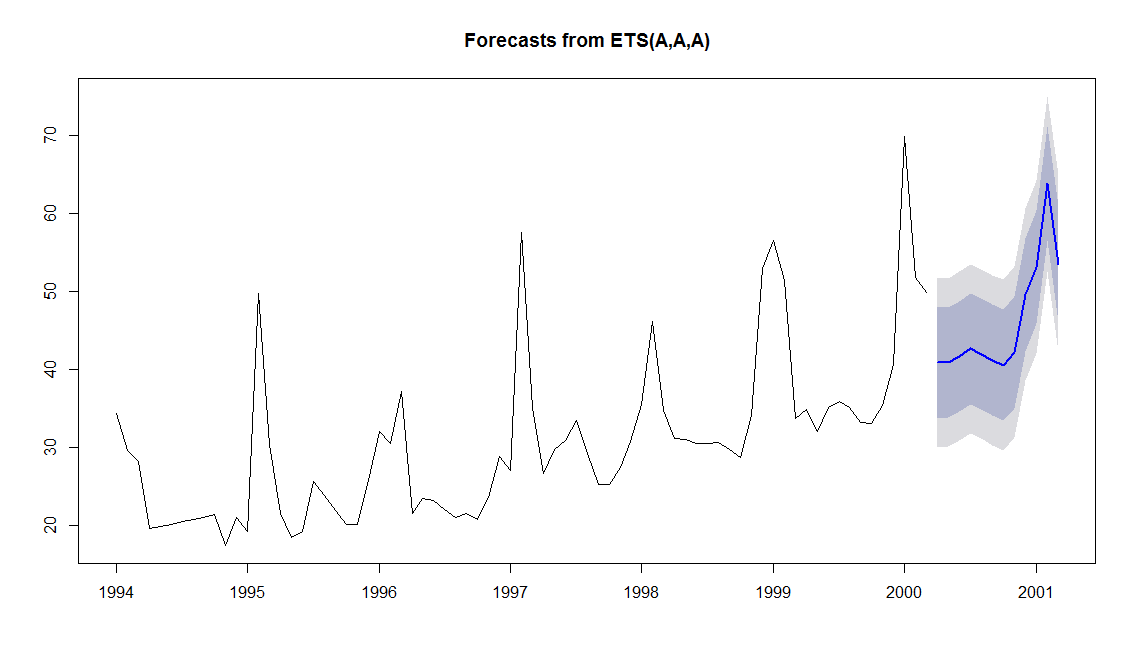
Training set error measures:

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.5226603 5.513393 3.620531 0.3714323 10.60376 0.6960751 -0.02808592

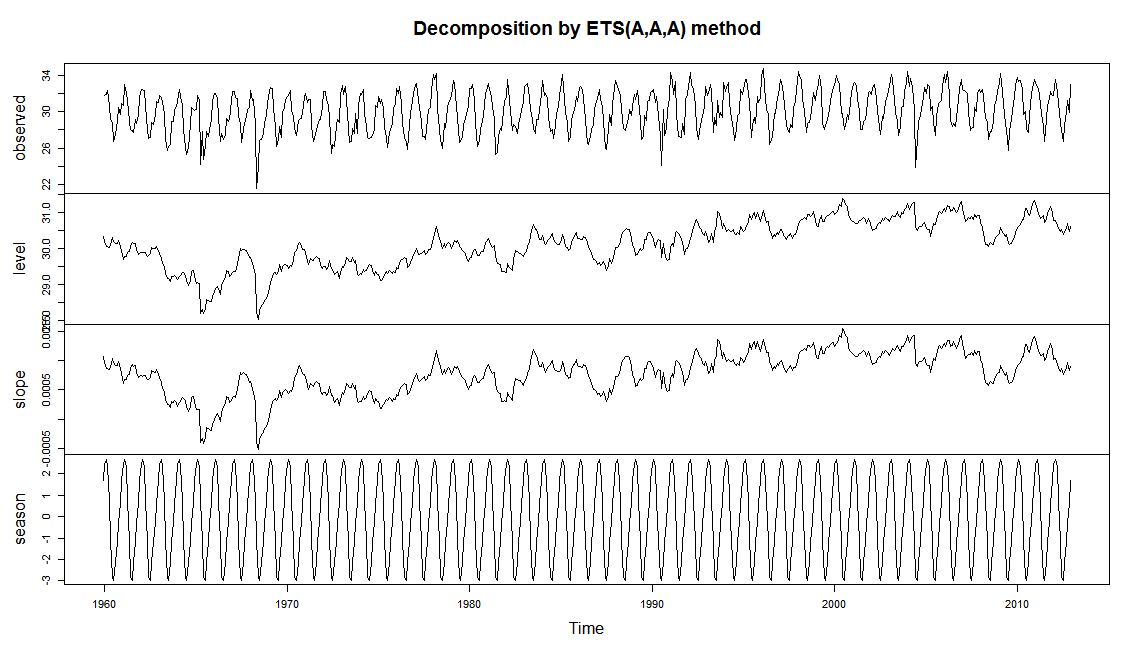
Prevision à l’horizon 12





#### Modélisation de la Température :

Série avec tendance et saisonnalité.



> Tmax.ets

ETS(A,A,A)

Call:

ets(y = Tmax.ts, model = "AAA")

Smoothing parameters:

alpha = 0.1487

beta = 1e-04

gamma = 1e-04

Initial states:

l = 30.3415

b = 0.0011

s=1.6921 0.6231 -0.2815 -1.1876 -2.1542 -2.9616

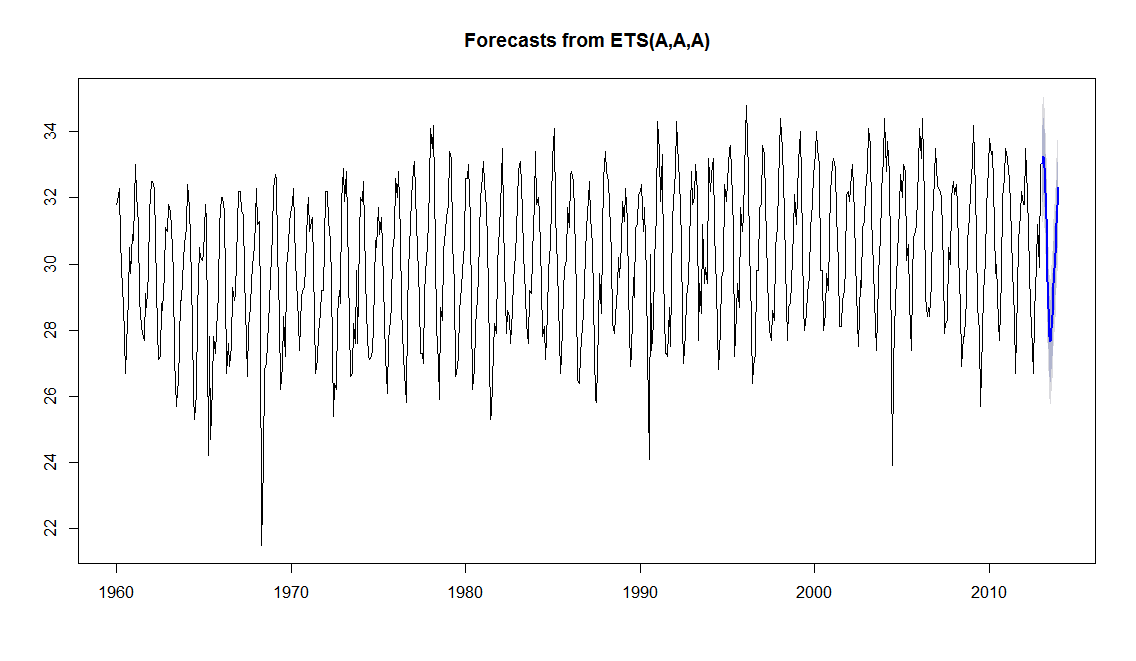
-2.7899 -1.1268 0.7691 2.3211 2.6626 2.4336

sigma: 0.9065

AIC AICc BIC

4012.653 4013.532 4083.936

Prévision à l’horizon 12.



### Comparaison des prévisions

## REFERENCES

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723. doi:10.1109/TAC.1974.1100705

Aragon, Y., 2011. Séries temporelles avec R Méthodes et cas. New York : Springer, Paris.

Box, G.E.P., Pierce, D.A., 1970. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. Journal of the American Statistical Association 65, 1509–1526. doi:10.1080/01621459.1970.10481180

Dickey, D.A., Fuller, W.A., 1979. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association 74, 427. doi:10.2307/2286348

Elliott, G., Rothenberg, T.J., Stock, J.H., 1996. Efficient Tests for an Autoregressive Unit Root. Econometrica 64, 813. doi:10.2307/2171846

Hyndman, R.J., Khandakar, Y., 2008. Automatic Time Series Forecasting: The **forecast** Package for *R*. Journal of Statistical Software 27. doi:10.18637/jss.v027.i03

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics 54, 159–178. doi:10.1016/0304-4076(92)90104-Y

Leybourne, S.J., McCabe, B.P.M., 1994. A Consistent Test for a Unit Root. Journal of Business & Economic Statistics 12, 157. doi:10.2307/1391480

Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303. doi:10.1093/biomet/65.2.297

Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346. doi:10.1093/biomet/75.2.335

Said, S.E., Dickey, D.A., 1984. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. Biometrika 71, 599. doi:10.2307/2336570

[A - Vue générale sur le bassin versant de Fiherenana et ses caractéristiques hydro-climatiques. 1](#_Toc455139005)

[1. Objectifs. 1](#_Toc455139006)

[2. Le bassin de Fiherenana et son suivi , les données existants. 1](#_Toc455139007)

[2.1. Les données pluviométriques 1](#_Toc455139008)

[2.2. Les données hydrométriques 2](#_Toc455139009)

[3. Caractérisation de l’Aléa. Constats sur la variabilité hydro-climatique au niveau du bassin versant de Fiherenana. 2](#_Toc455139010)

[3.1. Les tendances de la variabilité hydrologique et des débits extrêmes 3](#_Toc455139011)

[3.2. Variabilité climatique. 8](#_Toc455139012)

[B - Considérations des chroniques hydro-climatiques en tant que séries temporelles. 10](#_Toc455139013)

[1.1. Définitions, séries temporelles 10](#_Toc455139014)

[1.2. Les démarches classiques concernant l’étude des séries temporelles 10](#_Toc455139015)

[1.3. Cas particulier des chroniques de Fiherenana. Nécessité de l’ajout d’une démarche compensatrice 10](#_Toc455139016)

[2. DESCRIPTION : Etude Descriptive : 11](#_Toc455139017)

[3. RECONSTITUTION : Technique proposée pour la reconstitution de la série de hydrométrique de référence par décomposition en vue de la modélisation 19](#_Toc455139018)

[3.1. Décomposition d’une série temporelle (Aragon, 2011) 19](#_Toc455139019)

[La tendance : 19](#_Toc455139020)

[La composante saisonnière : 19](#_Toc455139021)

[La composante irrégulière : 19](#_Toc455139022)

[Cycle : 19](#_Toc455139023)

[3.1.1. Décomposition par moyenne mobile (sur R) 20](#_Toc455139024)

[3.2. Décomposition de la série des températures 20](#_Toc455139025)

[3.3. Décomposition de la pluviométrie 22](#_Toc455139026)

[3.4. Décomposition de la série de débits et technique de reconstitution 22](#_Toc455139027)

[3.4.1. Reconstitution et Décomposition (Méthode 1) : 23](#_Toc455139028)

[3.4.2. Reconstitution et Décomposition (Méthode 2) : 25](#_Toc455139029)

[4. MODELISATION  des séries pluviométriques et hydrométriques de Fiherenana. Première étape : étude de la stationnarité. 28](#_Toc455139030)

[4.1. Définitions fondamentales en séries temporelles 28](#_Toc455139031)

[4.1.1. Stationnarité stricte 28](#_Toc455139032)

[4.1.2. Stationnarité faible 28](#_Toc455139033)

[4.1.3. Bruit blanc 28](#_Toc455139034)

[4.2. Appréciation de la stationnarité des séries par des méthodes graphiques : 29](#_Toc455139035)

[a) Cas de la série hydrométrique 29](#_Toc455139036)

[b) Cas de la série pluviométrique 29](#_Toc455139037)

[4.3. Evaluation de la stationnarité des séries par l’observation de leur correlogramme  : 29](#_Toc455139038)

[4.4. Test de stationnarité : 32](#_Toc455139039)

[4.4.1. KPSS test 32](#_Toc455139040)

[a) Cas des pluies 33](#_Toc455139041)

[b) Cas des débits 34](#_Toc455139042)

[4.5. Conclusions sur l’étude de la stationnarité des séries hydro-climatiques 35](#_Toc455139043)

[5. MODELISATION. Deuxième étape : construction d’un modèle Arma. 36](#_Toc455139044)

[5.1. L’operateur retard 36](#_Toc455139045)

[5.2. Le processus autorégressif (AR) 36](#_Toc455139046)

[5.2.1. Conditions pour la stationnarité d’un processus autorégressif 37](#_Toc455139047)

[5.3. Le processus moyenne mobile (MA) 37](#_Toc455139048)

[5.4. Le processus autorégressif et moyenne mobile (ARMA (p,q)) 37](#_Toc455139049)

[5.5. Le modèle ARMA saisonnier (SARMA) 37](#_Toc455139050)

[5.5.1. Conditions de stationnarité d’un SARMA 38](#_Toc455139051)

[5.6. Intégration des séries, ARIMA et SARIMA 38](#_Toc455139052)

[5.7. Démarche adoptée pour la construction d’un modèle ARMA Saisonnier 38](#_Toc455139053)

[5.7.1. Identification automatique des ordres 39](#_Toc455139054)

[5.7.2. Identification manuelle des ordres par observation des graphiques d’auto corrélation 40](#_Toc455139055)

[5.7.3. Test de blancheur. 41](#_Toc455139056)

[5.7.4. Significativité des coefficients 41](#_Toc455139057)

[5.7.5. Résumé de la procédure de construction d’un modèle ARIMA, 41](#_Toc455139058)

[5.8. Modélisation de la série pluviométrique 44](#_Toc455139059)

[5.8.1. Approche par automatisation 44](#_Toc455139060)

[5.8.2. Approche manuelle 46](#_Toc455139061)

[5.9. Modélisation de la série hydrométrique 48](#_Toc455139062)

[5.9.1. Modélisation par automatisation 49](#_Toc455139063)

[5.9.2. Modélisation manuelle 50](#_Toc455139064)

[6. PREDICTIONS 53](#_Toc455139065)

[6.1. Modèle de prévision issu des modèles SARIMA estimés 53](#_Toc455139066)

[6.2. Modèle simple de prévision à court terme : Le lissage exponentiel. 53](#_Toc455139067)

[6.2.1. Le lissage simple 53](#_Toc455139068)

[6.2.2. Le lissage double 53](#_Toc455139069)

[6.2.3. Le lissage de Hot Winters 54](#_Toc455139070)

[6.2.4. Prévision des pluies : 54](#_Toc455139071)

[6.2.5. Le modèle du debit: 56](#_Toc455139072)

[6.2.6. Modélisation de la Température : 58](#_Toc455139073)

[6.3. Comparaison des prévisions 60](#_Toc455139074)

[7. REFERENCES 61](#_Toc455139075)