Chap 1
Les Processus
SARIMA , ARIMAX et Dynamic Model

I- Modèle SARIMA

En analyse de séries chronologiques, Le modèle ARIMA **avec effet de saison**, souvent noté **SARIMA** (**Seasonal ARIMA**), est une extension du modèle ARIMA classique qui prend en compte les composantes saisonnières dans les données temporelles.

Il est utilisé lorsque les données présentent des tendances et des motifs saisonniers qui se répètent à intervalles réguliers.

Voici les composants d'un modèle SARIMA :

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}}\quad \underbrace{(P,D,Q)_m}_{\text{Seasonal part of the model}}$$

- Composantes **Non Saisonnières** ARIMA
- Composantes **Saisonnières SARIMA**

1. Composantes Non Saisonnières ARIMA :

- **AR(p)** (Auto-Regressive) : Termes qui capturent les relations linéaires entre les observations actuelles et les observations passées à différents retards (lags).
- **I(d)** (Integration) : Différenciation pour rendre les données stationnaires en supprimant les tendances.
- **MA(q): Moving average**

2.Composantes Saisonnières SARIMA :

SAR(P) (Seasonal Auto-Regressive) : Composante AR appliquée à la saison précédente.
SI(Q) (Seasonal Integration) : Différenciation saisonnière pour rendre les données stationnaires sur la saison.
SMA(Q) (Seasonal Moving Average) : Termes d'erreur saisonnière qui capturent les erreurs passées à différents retards saisonniers.

Ordre de Saisonnalité (s) :

Il représente le nombre de périodes par saison. Exemple: pour des données mensuelles, la saisonnalité est de 12.

Ordre d'Intégration Saisonnier (D) :

Le nombre de différenciations saisonnières nécessaires pour rendre les données stationnaires.

Ordre des Termes d'Erreur Saisonniers (Q) :

Le nombre de termes d'erreur saisonniers (SMA) à prendre en compte.

La notation générale d'un modèle SARIMA est SARIMA(p,d,q)(P,D,Q)s, où :

p : Ordre de l'AR non saisonnier.

d : Ordre de la différenciation non saisonnière.

q : Ordre du MA non saisonnier.

P : Ordre de l'AR saisonnier.

D : Ordre de la différenciation saisonnière.

Q : Ordre du MA saisonnier.

S : Période de saisonnalité.

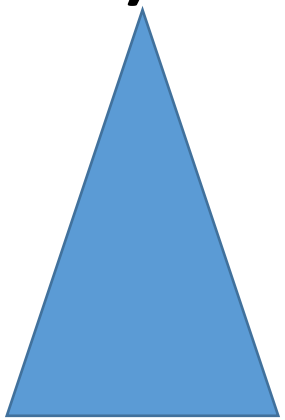
```
data("AirPassengers")
library(fpp2)
autoplot(AirPassengers)
a<-log(AirPassengers)
autoplot(a)
library(forecast)
auto.arima(a,seasonal = TRUE )
a<-log(AirPassengers)
```

auto.arima(a, seasonal = TRUE)

Series: a
ARIMA(2,1,1)(0,1,0)[12]
Coefficients:

ar1	ar2	ma1
0.5960	0.2143	-0.9819
s.e. 0.0888	0.0880	0.0292

sigma^2 = 132.3: log likelihood = -504.92
AIC=1017.85 AICc=1018.17
BIC=1029.35



> auto.arima(log(a), seasonal = TRUE)

Series: log(a)
ARIMA(0,1,1)(0,1,1)[12]
Coefficients:

ma1	sma1
-0.4018	-0.5569
s.e. 0.0896	0.0731

sigma^2 = 0.001371: log likelihood = 244.7
AIC=-483.4 AICc=-483.21
BIC=-474.77

Tests statistiques de validation du modèle

1. Test d'autocorrélation des résidus (Test de Ljung Box):

Il teste si les résidus sont indépendants ou non cad si les autocorrélations des résidus sont nulles ou non:

Hypothèse Nulle (H0) : Les autocorrélations jusqu'au décalage h h sont toutes zéro (les résidus sont indépendants)

Hypothèse Alternative (H1) : Au moins une des autocorrélations jusqu'au décalage h n'est pas nulle, (autocorrélation dans les résidus)

Le test génère une statistique Q qui suit approximativement une distribution du chi-carré sous l'hypothèse nulle. La significativité est déterminée en comparant la statistique Q à une valeur critique de la distribution du chi-carré.

Test de Ljung-Box sous R:

```
# Chargement de la bibliothèque
library(stats)

# Génération de résidus (à titre d'exemple)
residuals <- rnorm(100)

# Test de Ljung-Box
Box.test(residuals, lag = 10, type = "Ljung-Box")
```

Test de Normalité des Résidus

Le test de normalité des résidus évalue si les résidus du modèle suivent une distribution normale.

Hypothèse Nulle (H0) : Les résidus suivent une distribution normale.

Hypothèse Alternative (H1) : Les résidus ne suivent pas une distribution normale.

Plusieurs tests peuvent être utilisés: Test de Shapiro-Wilk, Test de Jarque Bera JB

```
# Chargement de la bibliothèque
library(stats)

# Génération de résidus (à titre d'exemple)
residuals <- rnorm(100)

# Test de Shapiro-Wilk
shapiro.test(residuals)
```

Test d'Autocorrélation Partielle des Résidus :

Le test d'autocorrélation partielle des résidus examine s'il y a une autocorrélation dans les résidus après avoir pris en compte les corrélations à des décalages inférieurs. Cela permet de détecter s'il y a des modèles manquants dans la spécification du modèle.

Hypothèse Nulle (H0) : Les résidus sont indépendants, et il n'y a pas d'autocorrélation partielle.

Hypothèse Alternative (H1) : Il y a de l'autocorrélation partielle dans les résidus.

Le test génère une statistique t qui suit approximativement une distribution du Student.

```
# Chargement de la bibliothèque
library(stats)

# Génération de résidus (n=100)
residuals <- rnorm(100)

# Calcul de l'autocorrélation partielle
pacf_resid <- pacf(residuals)

# Affichage de l'autocorrélation partielle
print(pacf_resid)
```

Tests de Racines Unitaires (pour I(d)) :

Ces tests, tels que le test Augmented Dickey-Fuller (ADF) ou le test de Phillips-Perron, évaluent si une ou plusieurs différenciations sont nécessaires pour rendre les données stationnaires.

Hypothèse Nulle (H0) : Les données ne nécessitent pas de différenciations supplémentaires et sont stationnaires.

Hypothèse Alternative (H1) : Les données nécessitent des différenciations supplémentaires pour devenir stationnaires.

Le test génère une statistique de test qui suit approximativement une distribution du t sous l'hypothèse nulle.

```
# Chargement de la bibliothèque
library(urca)

# Génération de données non stationnaires (à titre d'exemple)
non_stationary_data <- cumsum(rnorm(100))

# Test Augmented Dickey-Fuller
adf_test <- ur.df(non_stationary_data, type = "drift", lags = 1)
summary(adf_test)
```

```
> summary(adf_test)
Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
Residuals:
    Min      1Q  Median      3Q     Max
-3.0068 -0.7999  0.1609  0.7780  2.9455
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14161    0.12386  -1.143
0.2558
z.lag.1     -0.07447    0.04069  -1.830
0.0704 .
z.diff.lag  -0.04094    0.10360  -0.395
0.6936
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
Residual standard error: 1.084 on 95
degrees of freedom
Multiple R-squared:  0.03967,
        Adjusted R-squared:  0.01945
F-statistic: 1.962 on 2 and 95 DF,  p-
value: 0.1462
Value of test-statistic is: -1.83 1.7273
Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.51 -2.89 -2.58
phi1  6.70  4.71  3.86
```

Prédiction

Pour faire des prédictions avec un modèle SARIMA sous R, on utiliser la fonction ***predict*** mais après avoir trouvé le meilleur modèle d'estimation:

```
library(forecast)
forecast_values <-
 predict(mod, n.ahead = 12)
forecast_values
```

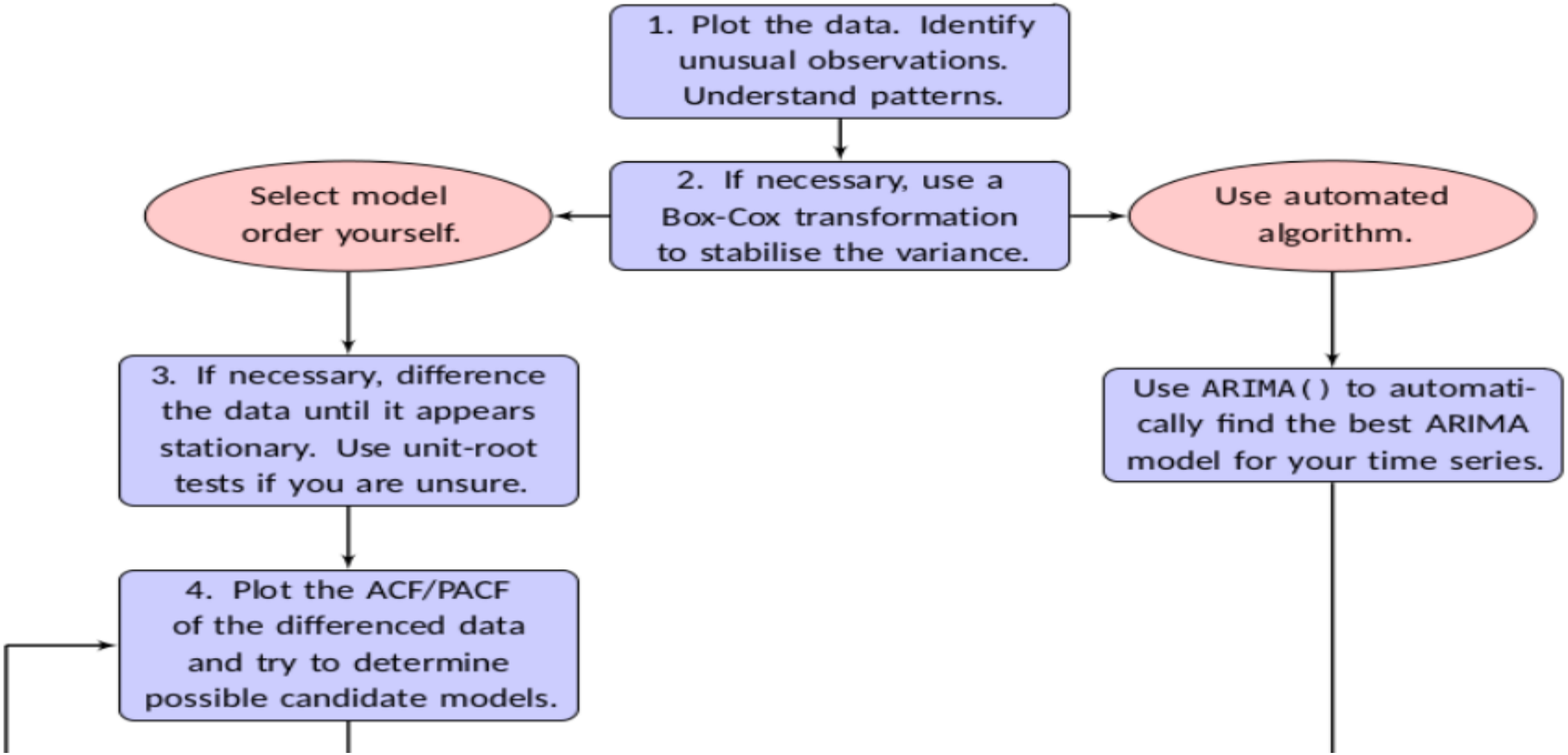
```
$pred
       Jan      Feb      Mar      Apr      May
1961 6.110186 6.053775 6.171715 6.199300 6.232556
       Jun      Jul      Aug      Sep      Oct
1961 6.368779 6.507294 6.502906 6.324698 6.209008
       Nov      Dec
1961 6.063487 6.168025
$se
         Jan       Feb       Mar       Apr
1961 0.03703056 0.04314989 0.04850323 0.05332179
         May       Jun       Jul       Aug
1961 0.05773962 0.06184266 0.06568991 0.06932399
         Sep       Oct       Nov       Dec
1961 0.07277682 0.07607310 0.07923236 0.08227039
```

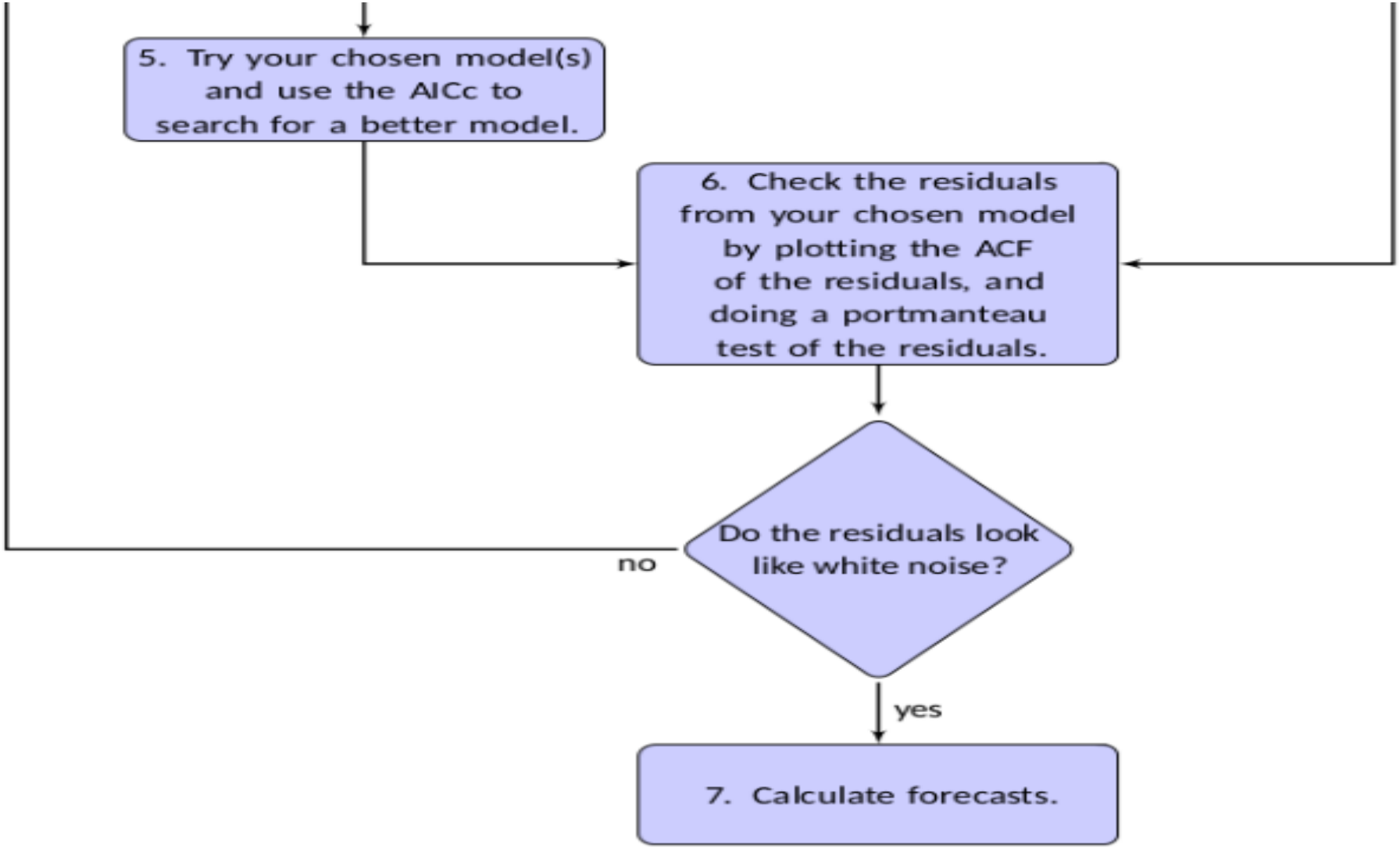
Charger **fpp3**

Exercice 1: Traitez la série ausAirpassengers

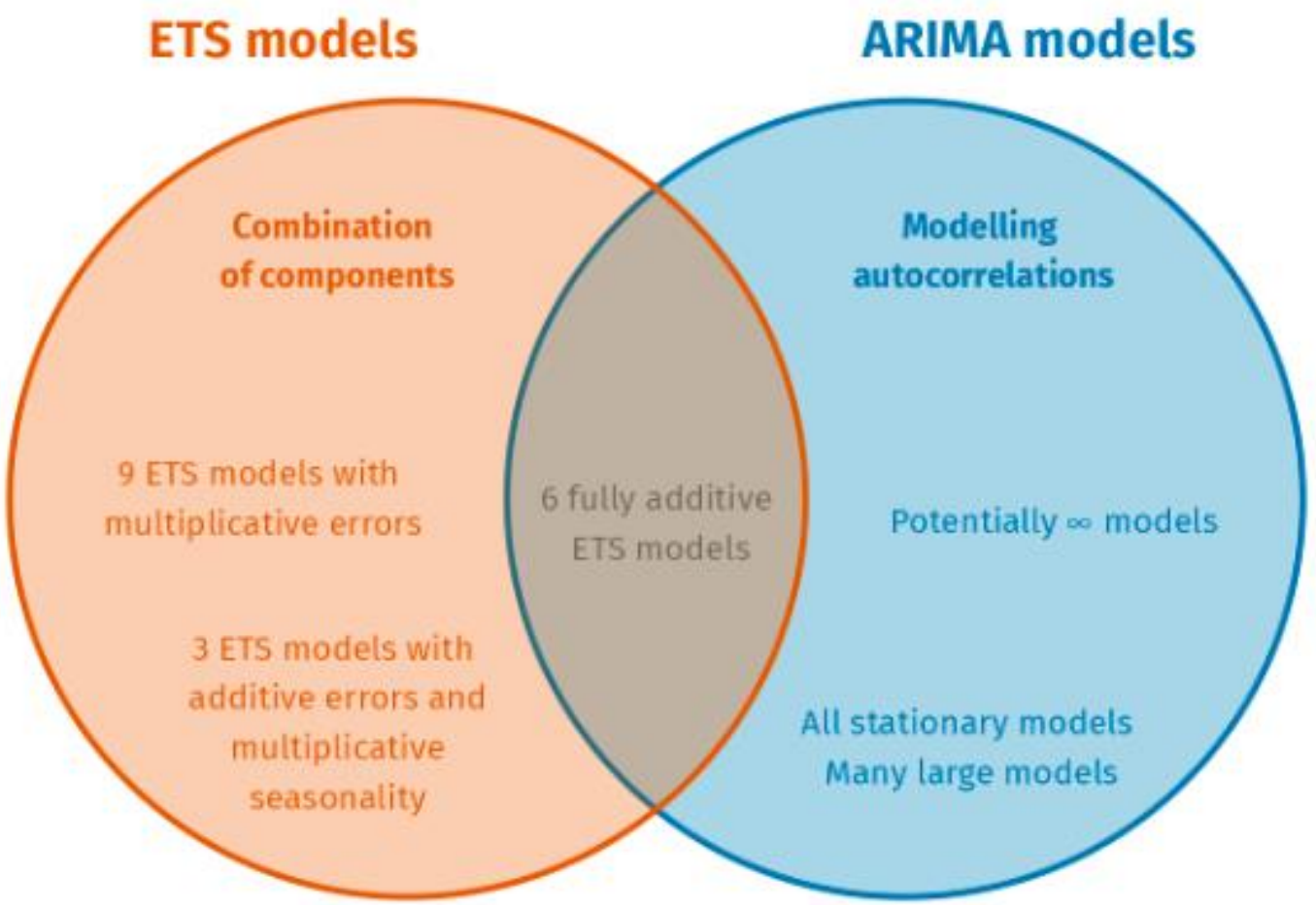
Exercice 2: Même question pour austourists

Graphique, saisonnalité, modèle, tests, prédiction


1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilise the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use ARIMA() to automatically find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.


5. Try your chosen model(s) and use the AICc to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Différences entre ARIMA ET ETS



II- Modèle ARIMAX

Un modèle ARIMAX (AutoRegressive Integrated Moving Average with eXogenous inputs) est une extension du modèle ARIMA classique qui permet d'intégrer **des variables exogènes** dans la prédiction d'une série temporelle. Ces variables exogènes sont des variables explicatives externes à la série temporelle elle-même, mais qui peuvent influencer son évolution.

II- Modèle ARIMAX

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_{t-k} + \epsilon_t$$

Where:

- Yt is the time series value at time tt.
- c is a constant.
- ϕi are the coefficients for the autoregressive terms.
- $Yt-i$are the lagged values of the time series.
- θj are the coefficients for the moving average terms.
- ϵt is the error term at time tt.
- βk are the coefficients for the exogenous variables.
- $Xt-k$ are the lagged values of the exogenous variables.

Pourquoi utiliser un modèle ARIMAX:

- Amélioration de la précision des prévisions: En incorporant des variables exogènes pertinentes, on peut mieux capturer les variations de la série temporelle et obtenir des prévisions plus précises.
- Compréhension des relations causales: ils permettent d'identifier les relations de causalité entre la série temporelle et les variables exogènes.
- Flexibilité: Ils sont très flexibles et peuvent s'adapter à une grande variété de situations.

Étapes pour construire un modèle ARIMAX

- Identification des variables exogènes: Il est crucial de sélectionner des variables exogènes qui sont susceptibles d'avoir un impact significatif sur la série temporelle.
- Stationnarité: Vérifier si la série temporelle et les variables exogènes sont stationnaires. Si ce n'est pas le cas, appliquer des transformations (différences, log, etc.) pour les rendre stationnaires.

- Sélection des ordres AR et MA: Utiliser des outils comme l'ACF et le PACF pour déterminer les ordres p et q du modèle ARMA.
- Estimation des paramètres: Estimer les coefficients du modèle à l'aide d'une méthode d'estimation appropriée (moindres carrés ordinaires, maximum de vraisemblance).
- Validation du modèle: Évaluer la qualité du modèle à l'aide de critères comme le RMSE, le MAE, ou des tests statistiques.

Activité pratique sous R

III- Modèles de régression dynamique

Les modèles de régression dynamique (ou modèles à retard) sont une classe plus large de modèles ARIMAX qui incluent à la fois des décalages des variables dépendantes et des variables explicatives. Ces modèles se concentrent sur la relation entre la variable dépendante et les variables explicatives en incluant leurs valeurs passées.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=0}^{q} \beta_j x_{t-j} + \epsilon_t$$

Le diagnostic inclut l'analyse de la significativité des retards, des relations dynamiques, et la vérification des hypothèses classiques de la régression linéaire (comme l'indépendance des erreurs, l'absence de multicolinéarité, etc.).

Activité pratique


© Everett Collection / Rex Feature