


**UBS/Faculté des Sciences et Sciences pour L'Ingénierie**



## **Programmation et traitement des données**



Rédigés par : DOUMBALY CAMARA SY & Montfort Niyonzima

**Professeur** : Larjane Salim

2018-2019

# Avant-propos

L'objectif de ce projet est de fournir un environnement interactif d'analyse de données, doté d'outils graphiques performants et permettant une adaptation aisée aux besoins des utilisateurs, depuis l'exécution de tâches routinières jusqu'au développement d'applications entières. Toute l'analyse s'est faite à l'aide des logiciels statistique R, SAS et PYTHON. À cet égard, nous remercions ici toute la communauté des chercheurs qui travaillent continuellement à améliorer ces outils de travail de qualité, accessible gratuitement et d'une grande puissance. R est un langage de programmation pour l'analyse et la modélisation des données. R peut être utilisé comme un langage orienté objet tout comme un environnement statistique dans lequel des listes d'instructions peuvent être exécutées en séquence sans l'intervention de l'utilisateur.

SAS est un système intégré pour la manipulation, l'analyse et la présentation des données. C'est un système modulaire, de nombreux modules pouvant être ajoutés au système de base : SAS Base.

Le langage Python est très compact mais dispose d'une grande quantité d'extensions et de bibliothèques qui permettent d'effectuer un grand nombre de tâches • Les plus utilisées pour la Statistique, le Data Mining et le Machine Learning sont NumPy, SciPy et Matplotlib.

Étant conscient que la science est basée sur la critique, nous sommes toujours intéressés à recevoir toutes les remarques ou suggestions de correction ou d'amélioration de ce qui a été présenté dans ce mémoire.

## INTRODUCTION

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. Parmi ces méthodes on peut citer les séries temporelles et la théorie des sondages.

**La théorie des sondages** qui est une méthode statistique visant à évaluer les proportions de différentes caractéristiques d'une population à partir de l'étude d'une partie seulement de cette population, appelée échantillon. Les proportions sont déterminées avec des marges d'erreur, dans lesquelles se situent les proportions recherchées avec telle ou telle probabilité.

Une **série temporelle**, ou série chronologique, est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. De telles suites de variables aléatoires peuvent être exprimées mathématiquement afin d'en analyser le comportement, généralement pour comprendre son évolution passée et pour en prévoir le comportement futur.

Nous allons dans un premier temps importer les jeux de données dans l'environnement de R, SAS et PYTHON ensuite l'analyser et enfin réaliser un sondage et des prévisions là-dessus.

Dans ce présent rapport nous disposons de deux jeux de données le premier note IPI contenant 1260 individus et 31 variables (la première variable num sera très utile dans la partie concernant les sondages). Nous tenons aussi à préciser que le jeu de données ne dispose pas de données manquantes et que l'ensemble du traitement de données se fera avec les logiciels précités précédemment.

## **PREMIERE SECTION : ANALYSES AVEC LE LOGICIEL R**

# Partie I:

# Exploitation des données &

# Théorie des Sondages

## **I. Importation du jeu de données**

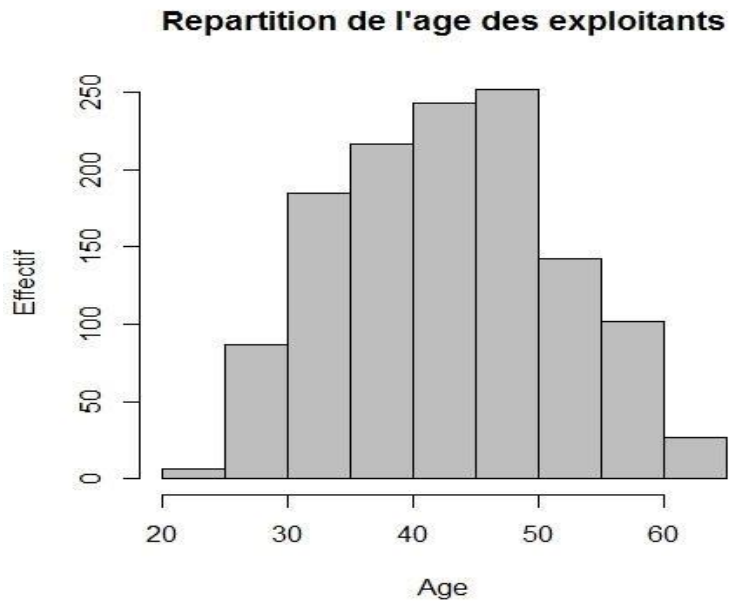
1. Cette partie sera beaucoup plus explicitée au niveau du scripte des codes en annexes avec une explication détaillée de la démarche suivie.
2. Dans ce jeu de données les exploitants sont les individus et les variables sont au nombre de 31 dont 6 qualitatives (5 qualitatives cardinales et 1 qualitative ordinale à savoir la variable num) et 25 quantitatives.

On peut tout de même préciser que les variables qualitatives seront labélisées par la suite afin que R puisse bien interpréter ces variables.

## **II. Analyse des exploitations étudiées**

3. Les départements français étudiés dans ce jeu de données sont les départements :
  - ❖ 27 : Eure dans la région de Normandie
  - ❖ 59 : Nord dans la région de Haut de France
  - ❖ 61 : Orne dans la région de Normandie
  - ❖ 76 : Seine-Maritime dans la région de Normandie
4. Le nombre d'exploitants correspond au nombre d'individus qui est de 1260.
  - 48,1746% sont des exploitants en défaut de paiement.
  - 39,36508% sont des exploitants propriétaires.
5. Représentation de la répartition de l'âge des exploitants.

**Graphique 1** : Répartition de l'âge des exploitants



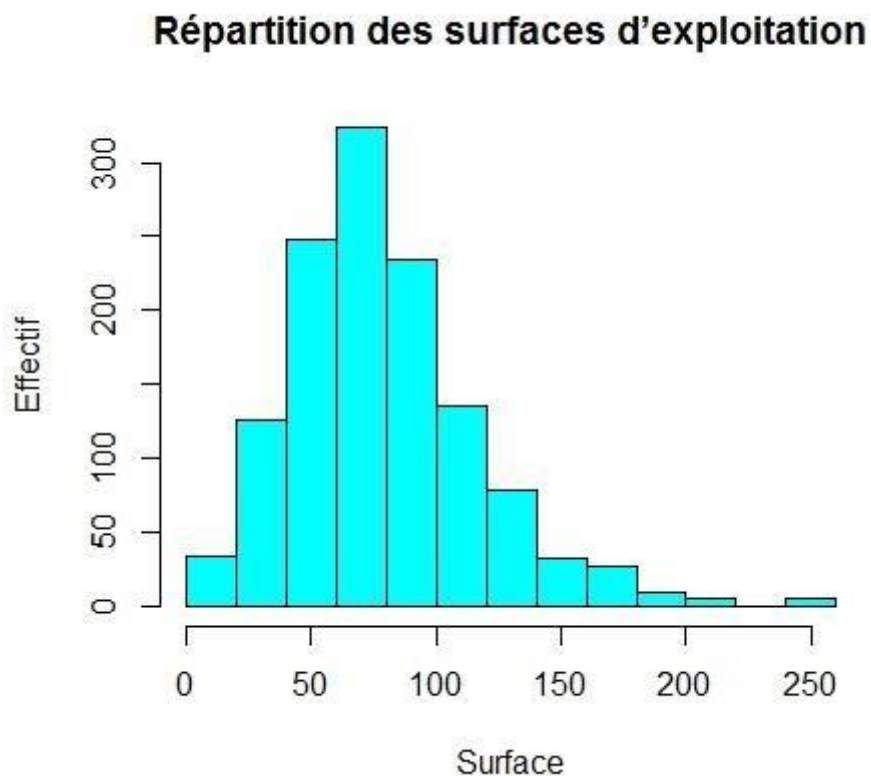
A travers l'histogramme on peut supposer que la variable âge suit une distribution normale de moyenne 43,54 et variance 76,77.  $N \sim (43,54 ; 76,77)$ .

L'âge moyen, médian, minimum, maximum correspond respectivement à 43,54 ; 44 ; 24 et 65.

L'intervalle interquartile correspond à [Q1, Q3] [et il contient 50% des observations, ici pour la variable âge l'intervalle interquartile est [37 ; 50] la proportion approximative est de 50%.

6. Représentons la répartition des surfaces d'exploitation.

**Graphique 2:** Répartition des surfaces d'exploitation



On peut supposer ici que la variable HECTARE suit une distribution normale étalée vers la droite.

La surface moyenne, maximale et le quantile a 90% (décile d'ordre 9) sont respectivement : 79,22 ; 247,50 et 125,3.

7. Dans cet échantillon le type d'exploitation le plus présent est le type mix av ec 468 sur 1260 exploitations.

**Tableau 1 :** Répartition du type d'exploitation

Elevage	Cereales	Mix	Autres
339	306	468	147



## 8. Représentons la répartition des surfaces par département

Pour ce fait nous devons en premier procéder au recodage en classe de la variable HECTARE afin d'avoir un tableau facile à interpréter. Nous utiliserons la méthode des quartiles pour le découpage des classes. Nous aurons 4 classes.

La première sera considérée comme les surfaces très petites [0-54), la seconde [54-74) les surfaces petites, la troisième [74-97) les surfaces moyennes et la 4ème [97-248) les grandes surfaces.

On obtiendra le tableau et le graphe suivants :

**Tableau 2** : Répartition des surfaces par département

	[0, 54)	[54, 74)	[74, 97)	[97, 248]
Eure (27)	30	75	78	165
Nord (59)	130	47	39	66
Orne (61)	78	84	96	75
Seine-Maritime (76)	69	114	99	15

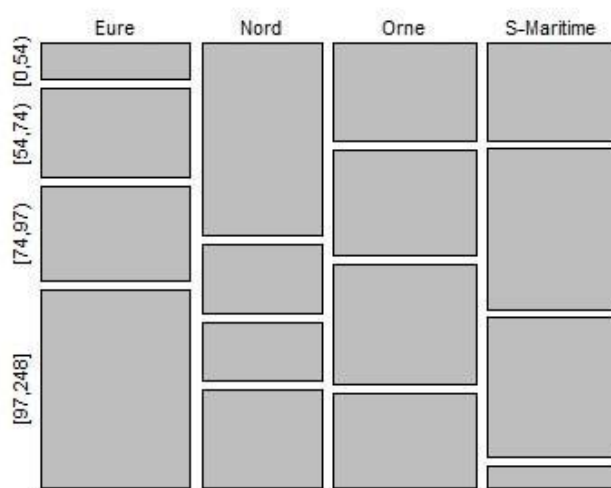
On gardera les labels Surfaces T.petites, Surfaces Petites, Surfaces Moyennes et Grandes Surfaces pour le reste

Ce tableau nous montre que le département d'Eure (27) dispose de plus grandes surfaces d'exploitation parmi les 4 départements.

Nous pouvons aussi matérialiser cela en représentation graphique

**Graphique 3** : Répartition des surfaces par département

### Repartition des surfaces par departement



9. Le tableau suivant nous montre la répartition des départements en fonction des surfaces moyennes

**Tableau 3 :** Répartition des surfaces moyennes par département

	CNTY	HECTARE
Eure	(27)	96.74138
Nord	(59)	70.26879
Orne	(61)	77.43243
Seine-Maritime	(76)	69.21212

La surface moyenne la plus importante est de 96,74138 et elle correspond au département d'Eure (27)

Afin de déterminer le département ayant la dispersion des surfaces plus petite et grande on utilise le l'écart-type.

**Tableau 4 :** Répartition des écarts des surfaces par département

	CNTY	HECTARE
Eure	(27)	38.73990
Nord	(59)	48.23426
Orne	(61)	32.15361
Seine-Maritime	(76)	21.80191

Ce tableau nous montre que le département de Seine-Maritime (76) dispose de la dispersion de surface la plus petite et Nord(59) la plus grande.

Pour quantifier ces dispersions pour chaque département nous utiliserons le coefficient de variation.

**Tableau 5 :** Répartition du coefficient de variation des surfaces par département

	CNTY	HECTARE
Eure	(27)	40.04481
Nord	(59)	68.64251
Orne	(61)	41.52473
Seine-Maritime	(76)	31.50013

10. Représentons la répartition des surfaces d'exploitation selon le statut de l'exploitation (variable Statut).

**Tableau 6 :** Répartition des surfaces par statut

La surface moyenne pour le statut entreprise et individuel est respectivement :  
103,35714 et 74,00656.

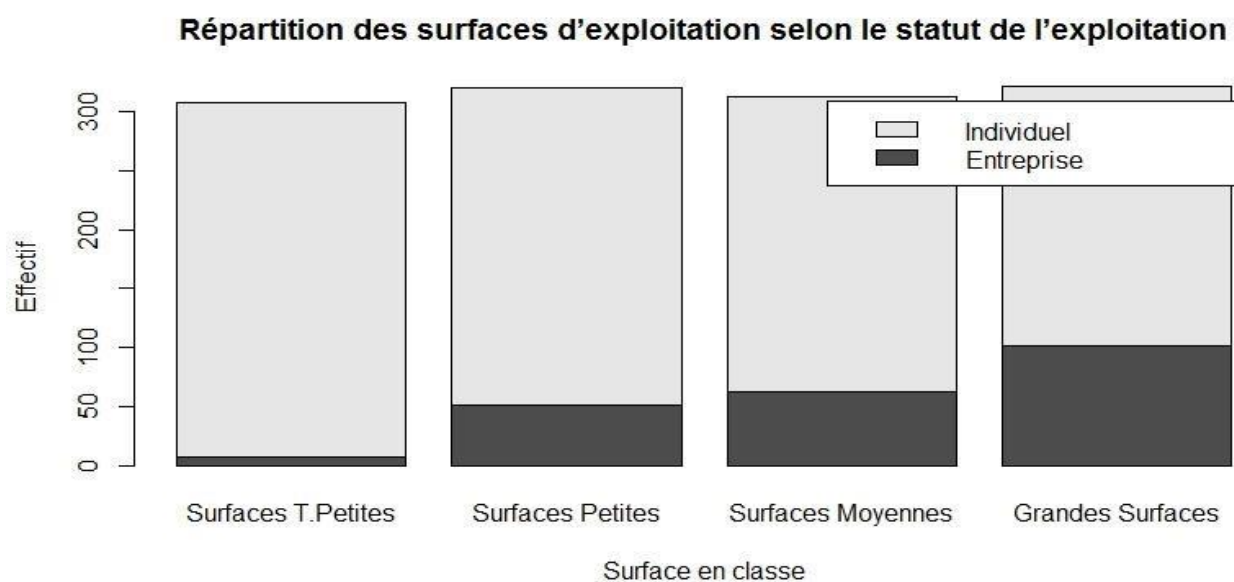
	Surfaces T.Petites	Surfaces Petites	Surfaces Moyennes	Grandes Surfaces
Entreprise	8	51	63	102
Individuel	299	269	249	219

Rappelons le nombre d'exploitations dans chaque statut :

**Tableau 7 :** Nombre d'exploitations dans chaque statut

	Entreprise	Individuel
	224	1036

**Graphique 4 :** Répartition des surfaces d'exploitation selon le statut



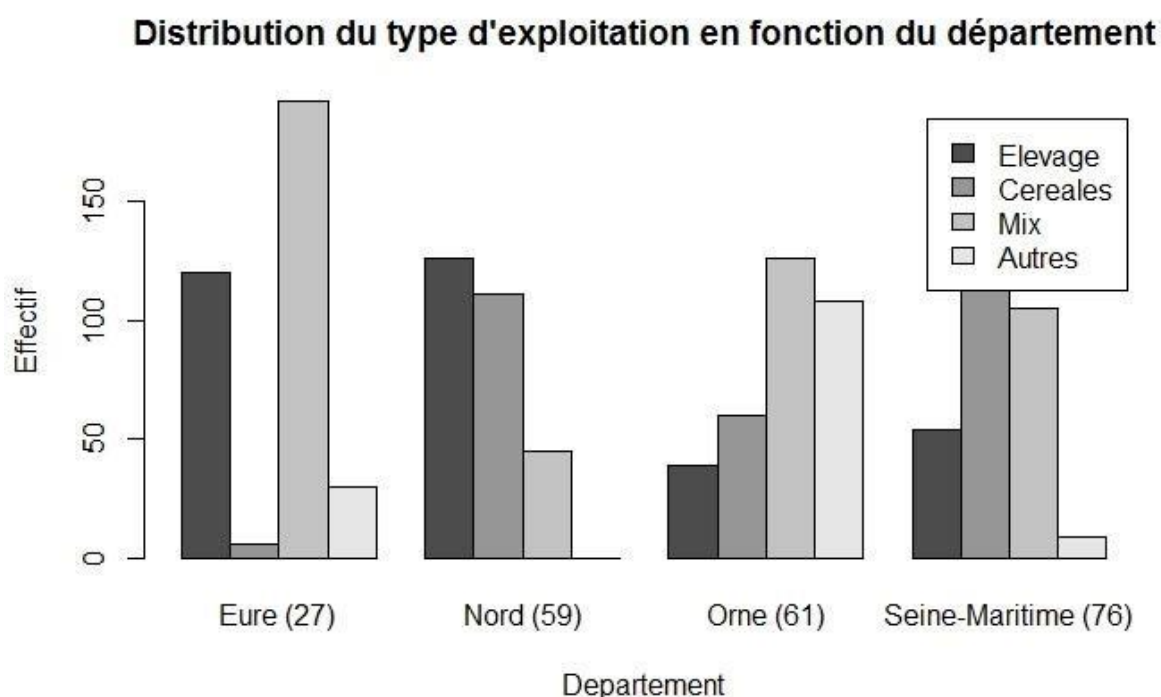
11. Tableau croisé de la variable CNTY (département) et de la variable ToF (Type d'exploitation).

**Tableau 8:** Tableau croisé de la variable CNTY (département) et de la variable ToF (Type d'exploitation)

	Eure (27)	Nord (59)	Orne (61)	Seine-Maritime (76)
Elevage	120	126	39	54
Cereales	6	111	60	129
Mix	192	45	126	105
Autres	30	0	108	9

Distribution conditionnelle du type d'exploitation en fonction du département.

**Graphique 5 :** Répartition du type d'exploitation selon le département



## 12. Taille d'exploitation des propriétaires

**Tableau 9:** Taille des surfaces selon la propriété

	Surfaces T.Petites	Surfaces Petites	Surfaces Moyennes	Grandes Surfaces
Non proprietaire	221	192	195	156
Proprietaire	86	128	117	165

Le tableau 9 nous montre que la classe grandes surfaces possède une modalité plus grande pour les propriétaires que pour les non propriétaires. On peut alors dire que les propriétaires ont de plus grandes surfaces d'exploitation.

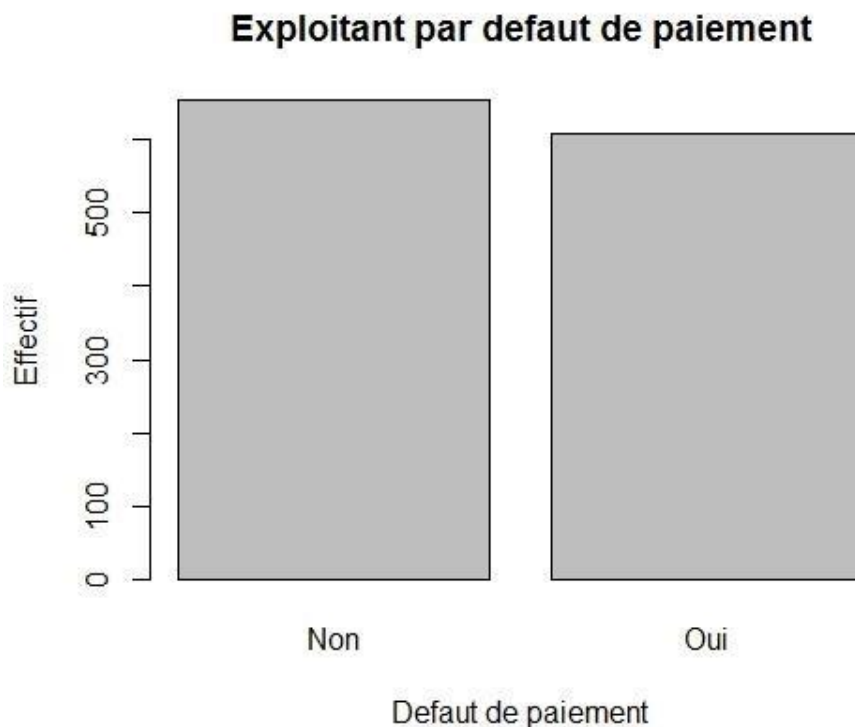
**Tableau 10:** Age des exploitants selon la propriété

	[24,37)	[37,44)	[44,50)	[50,65]
Non propriétaire	221	166	180	197
Propriétaire	87	134	153	122

Le tableau 10 nous montre que sur toutes les 4 classes d'âge les non propriétaires sont plus nombreux que les propriétaires. On peut donc déduire que les non propriétaires sont plus âgés que les propriétaires.

#### Analyse des exploitants en défauts de paiement

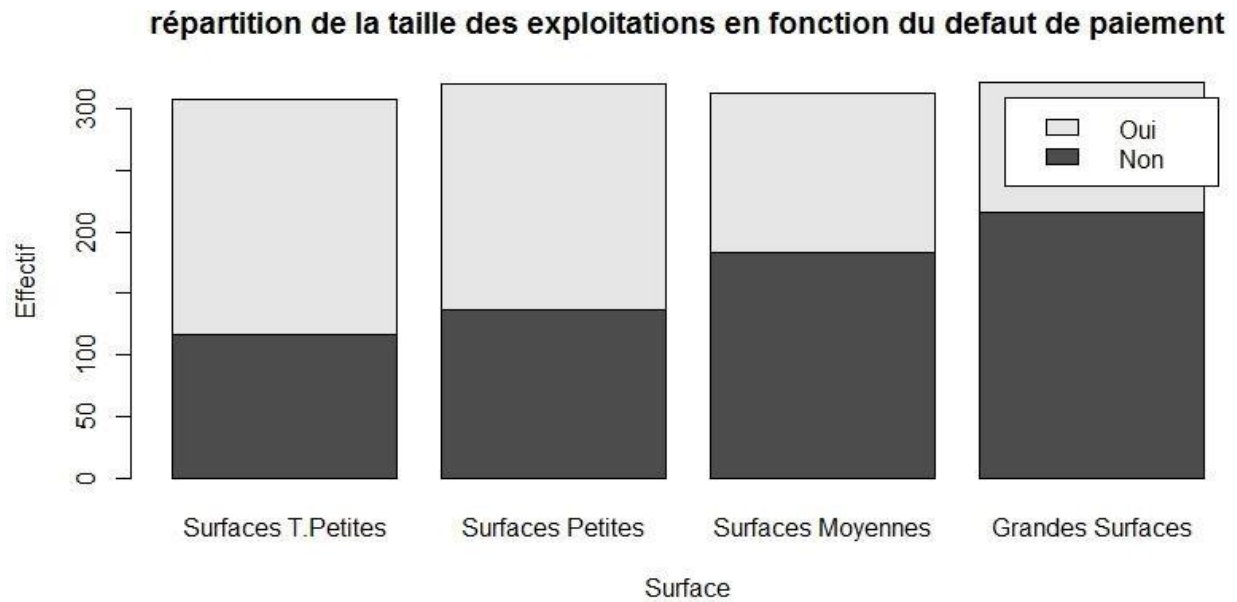
**Graphique 6 :** Répartition des exploitants par défaut de paiement



Le graphique précédent nous montre les exploitants n'ayant pas défaut de paiement sont supérieurs aux exploitants en défaut de paiement.

13. Représentons la répartition de la taille des exploitations en fonction de la variable DIFF.

**Graphique 7:** Répartition de la taille des exploitations selon le défaut de paiement

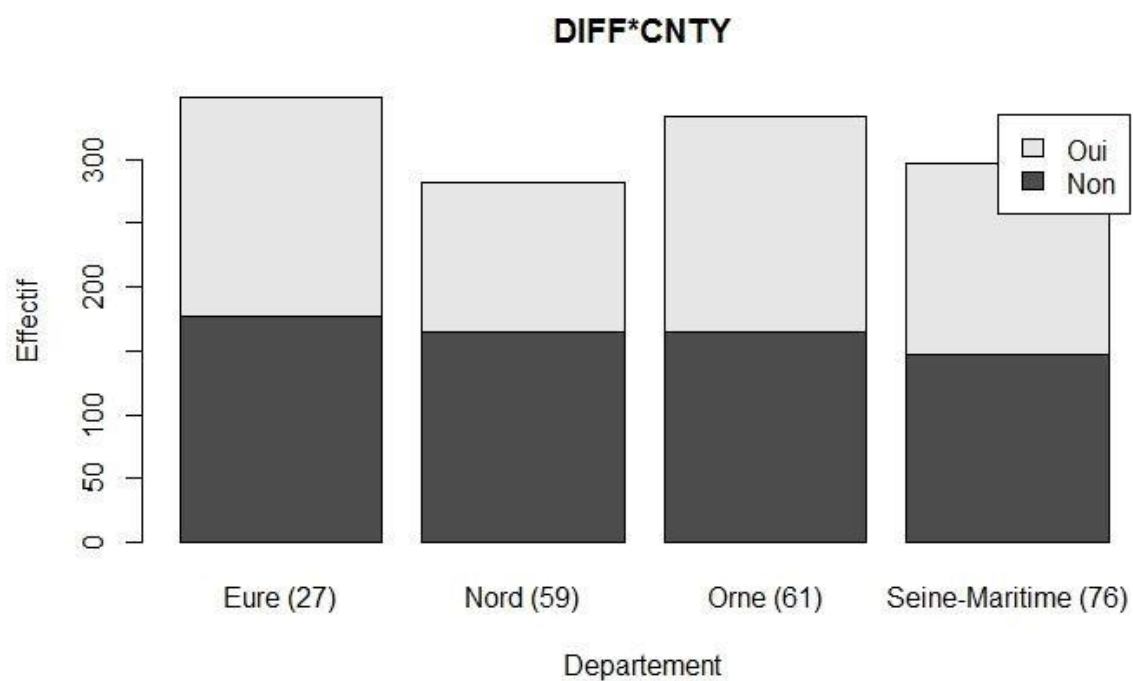


14. Tableau croisé de la variable DIFF et de la variable CNTY

**Tableau 11:** Tableau croisé de la variable DIFF et de la variable CNTY

	Eure (27)	Nord (59)	Orne (61)	Seine-Maritime (76)
Non	177	164	165	147
Oui	171	118	168	150

**Graphique 8 :** DIFF\*CNTY



Le Tableau 11 nous montre que dans les départements d’Orne et de Seine-Maritime les exploitants en défaut de paiement sont plus fréquents.

## Partie II :

# Séries

# Temporelles



## Préambule

Une série chronologique est la réalisation d'un processus aléatoire indice par le temps (jour, mois, année...). L'étude d'un processus aléatoire à partir d'une série chronologique a généralement deux objectifs :

- Expliquer les variations
- Prédire les valeurs futures.

Nous allons dans un premier temps étudier la série de la Côte d'Ivoire sur l'indice des prix industriels puis étudier la série concernant les températures mensuelles d'une série.

L'ensemble du traitement se fera à l'aide du logiciel R.

### □ Méthodologie de travail

Afin d'étudier les séries nous procéderons comme suit :

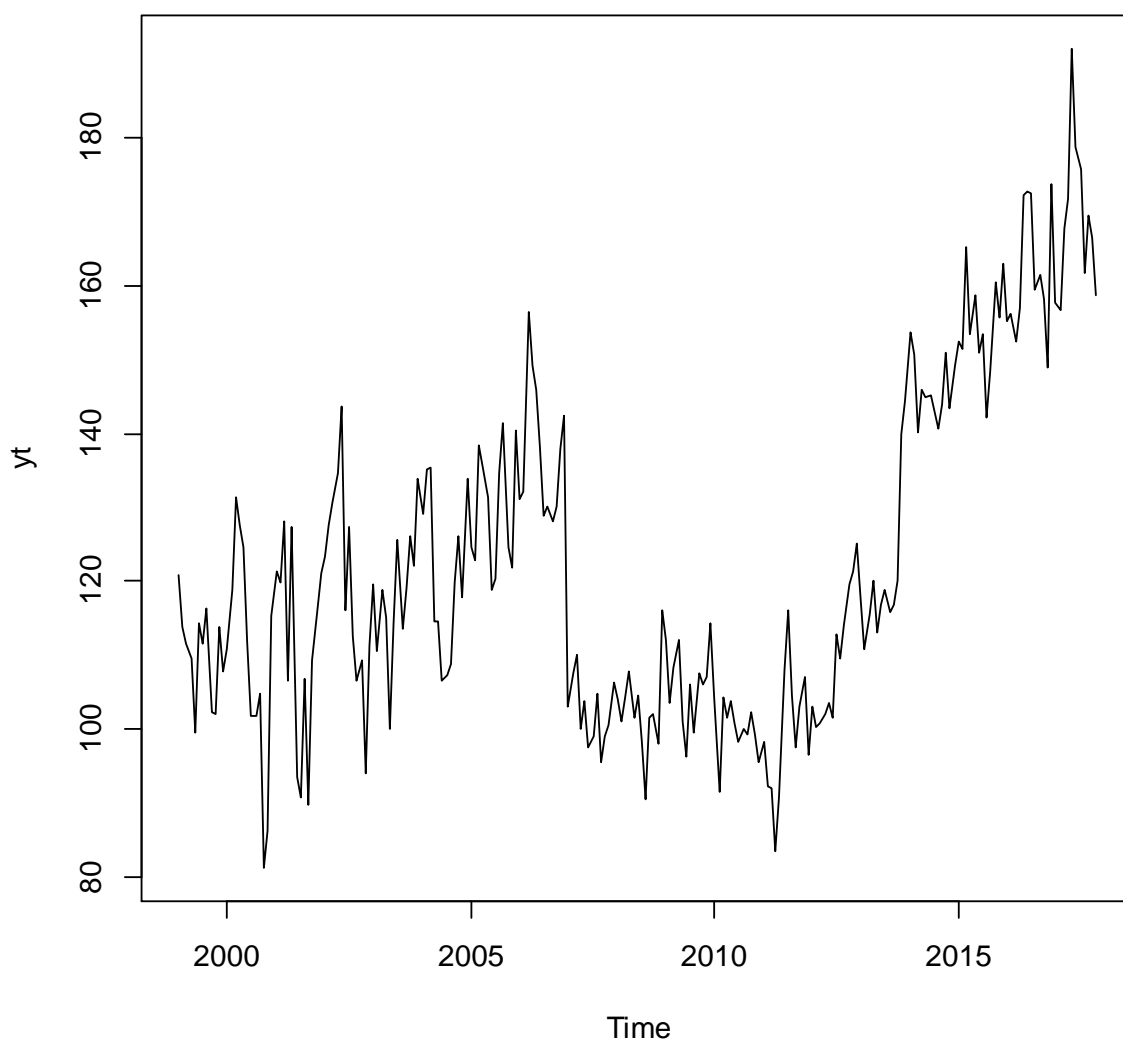
- ☞ Présentation de la série brute
- ☞ Décomposition en tendance, saisonnalité et bruit (résidu).
  - ☞ Tester la stationnarité de la série brute
  - ☞ Différencier la série si elle n'est pas stationnaire
  - ☞ Tester la stationnarité de la série différenciée
- ☞ Estimer un modèle à l'aide des autocorrélations et des critères d'informations (AIC)
  - ☞ Tester la validité du modèle
  - ☞ Faire des prévisions

# I. Indice de la production industrielle de la Cote d'Ivoire

## 1. Présentation de la série

La série mise à notre disposition fera l'objet d'une étude temporelle.  
En effet cette dernière regroupe les données mensuelles de la Cote d'Ivoire à partir de janvier 1999 à décembre 2017.

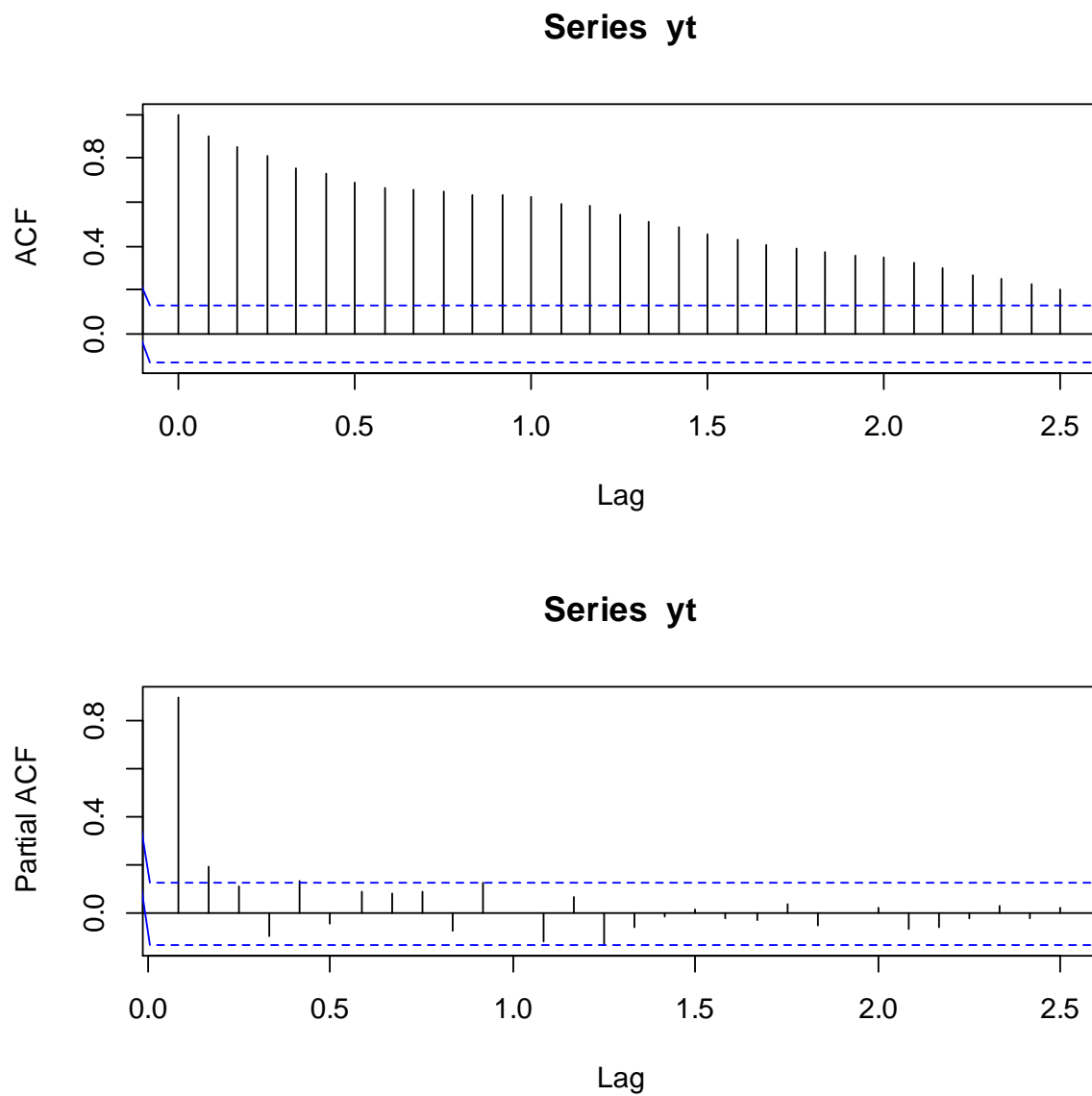
**Graphique 11** : Evolution de la série dans le temps



## 2. Décomposition de la série brute

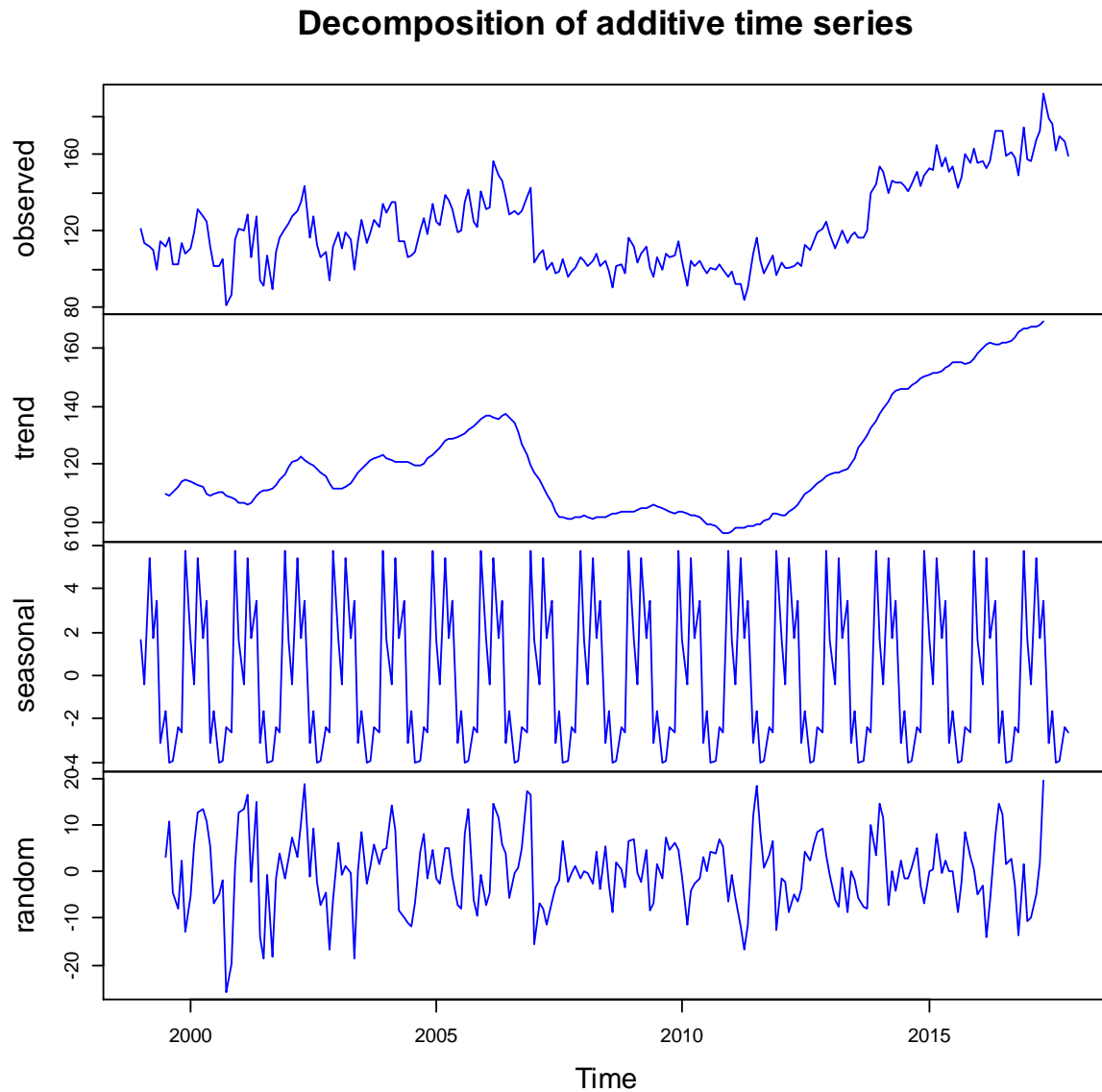
La décomposition de la série fait ressortir la tendance, la saisonnalité et l'aléa.

**Graphique 13** : Décomposition de la série brute



## 2. Décomposition de la série brute

La décomposition de la série fait ressortir la tendance, la saisonnalité et l'aléa.



Les observations du graphique (1ere figure) montrent que la série est multiplicative. En effet on a différente amplitude.

### 3. Stationnarité de la série brute

Pour tester la stationnarité de la série on utilisera le test de Dickey et Fuller

#### ✱ Le test de Dickey et Fuller

Ce test compare la valeur de la p-value à 5%. L'hypothèse alternative de ce test est la stationnarité de la série donc si la valeur de la p-value est supérieure à 5% rejette l'hypothèse alternative en faveur de l'hypothèse nulle.

Le test nous fournit les résultats suivants :

Augmented Dickey-Fuller Test

```
data: yt
Dickey-Fuller = -1.6055, Lag order = 6, p-value = 0.7416
alternative hypothesis: stationary
```

Le test nous donne une p-value de 0,6043 qui est supérieure à 5%. On peut donc conclure que la série brute n'est pas stationnaire. Donc cela vient confirmer l'hypothèse émise dans le graphique 11.

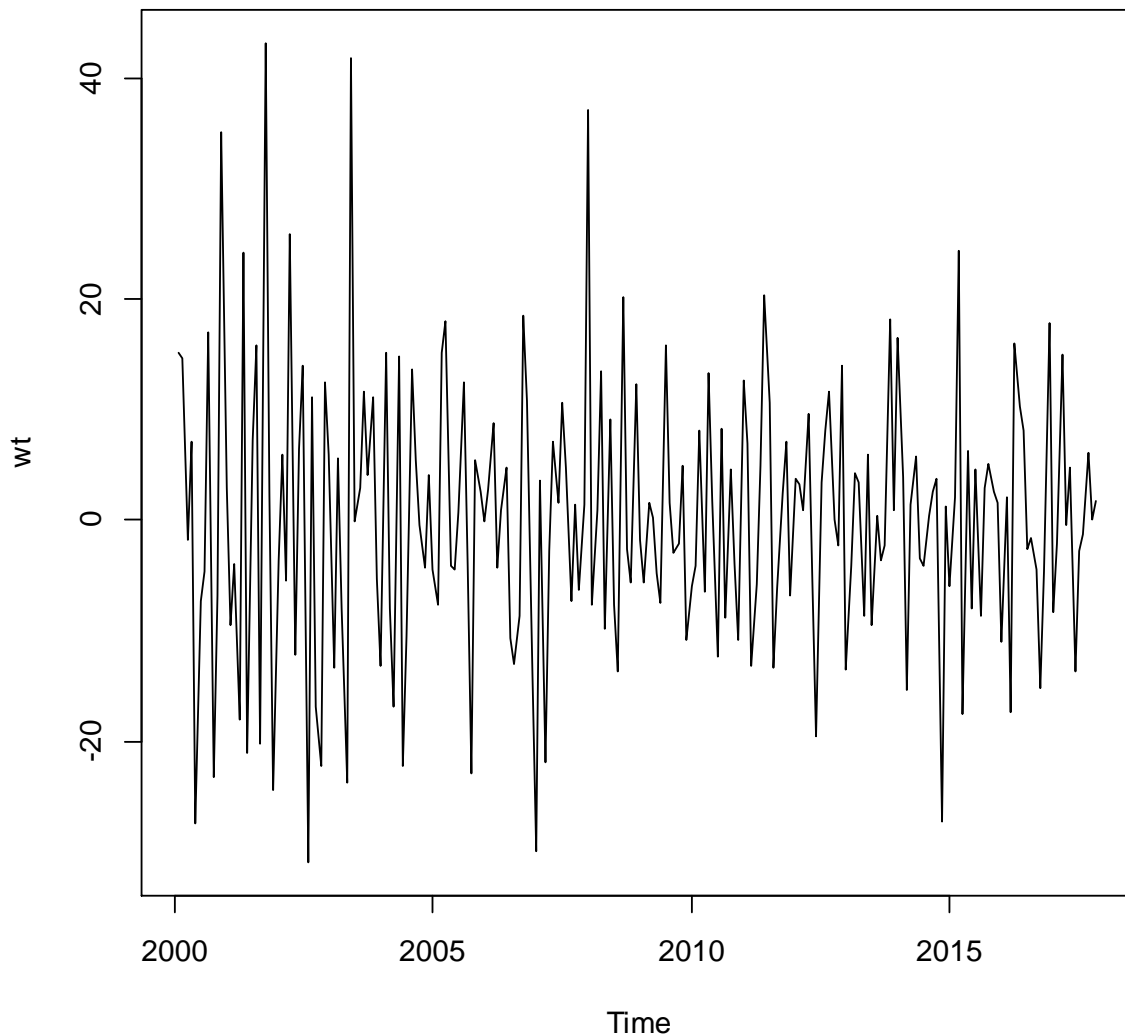
### 4. *Stationnarité de la série différenciée*

L'estimation des modèles ARIMA suppose que l'on travaille sur une série stationnaire.

Ceci signifie que la moyenne de la série est constante dans le temps, ainsi que la variance. La meilleure méthode pour éliminer toute tendance est de différencier, c'est-à-dire de remplacer la série originale par la série des différences adjacentes. Une série temporelle qui a besoin d'être différenciée pour atteindre la stationnarité est considérée comme une version intégrée d'une série stationnaire (d'où le terme *Integrated*).

Une différenciation d'ordre 1 suppose que la différence entre deux valeurs successives de  $y$  est constante.

**Graphique 14 :** Evolution de la série différenciée dans le temps



Le graphique 14 nous montre que la série différenciée tourne autour d'une constante  $t$ . on peut alors émettre l'hypothèse que l'opérateur de différenciation a stationnarité la série. Nous vérifierons cela à l'aide de différents tests.

### 5-Estimation d'un modèle

Call:

```
arima(x = wt, order = c(2, 0, 2), seasonal = list(order = c(2, 0, 1), period = 12),  
      include.mean = T, method = "ML")
```

Coefficients:

	ar1	ar2	ma1	ma2	sar1	sar2	sma1	intercept
	0.2769	0.6357	0.4476	-0.3657	0.1824	0.0369	-0.9999	2.4034
s.e.	0.1298	0.1108	0.1390	0.0816	0.0790	0.0801	0.4135	1.2777

sigma^2 estimated as 72.68: log likelihood = -781.39, aic = 1580.77

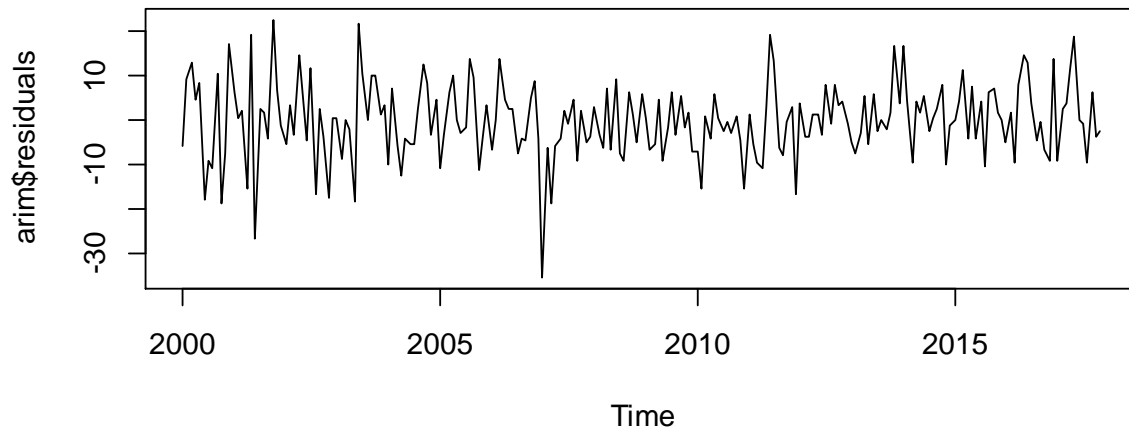
### *6. Validation du modèle*

Afin de vérifier la validation du modèle nous allons analyser les résidus du modèle choisi. Pour cela nous allons en premier lieu :

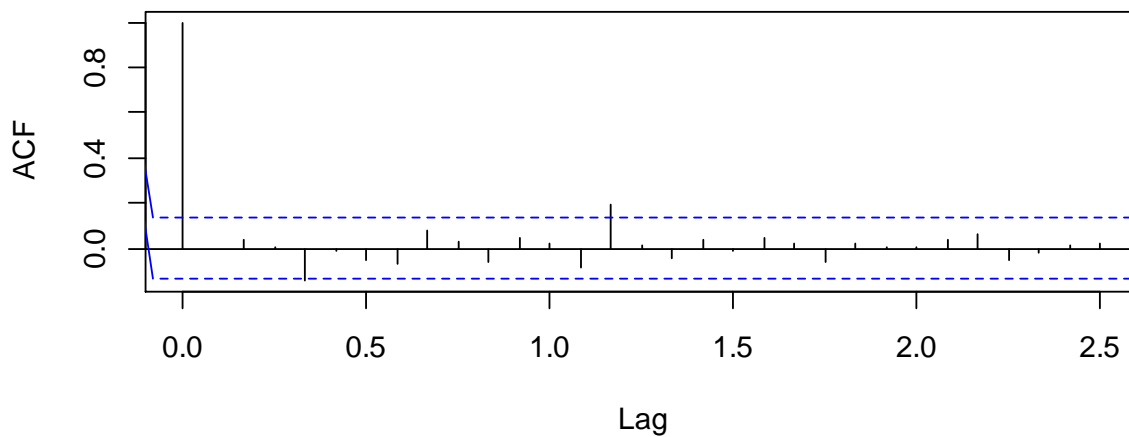
#### ○ Vérifier l'absence d'autocorrélation

Pour vérifier l'absence d'autocorrélation nous allons procéder comme tel : Représenter les résidus, tracer l'acf et utiliser le test de Box-Pierce.

Graphique 16 : Résidus et ACF du modèle ARIMA (2,0,2) (2,0,1)



### Series residus



Box-Pierce test

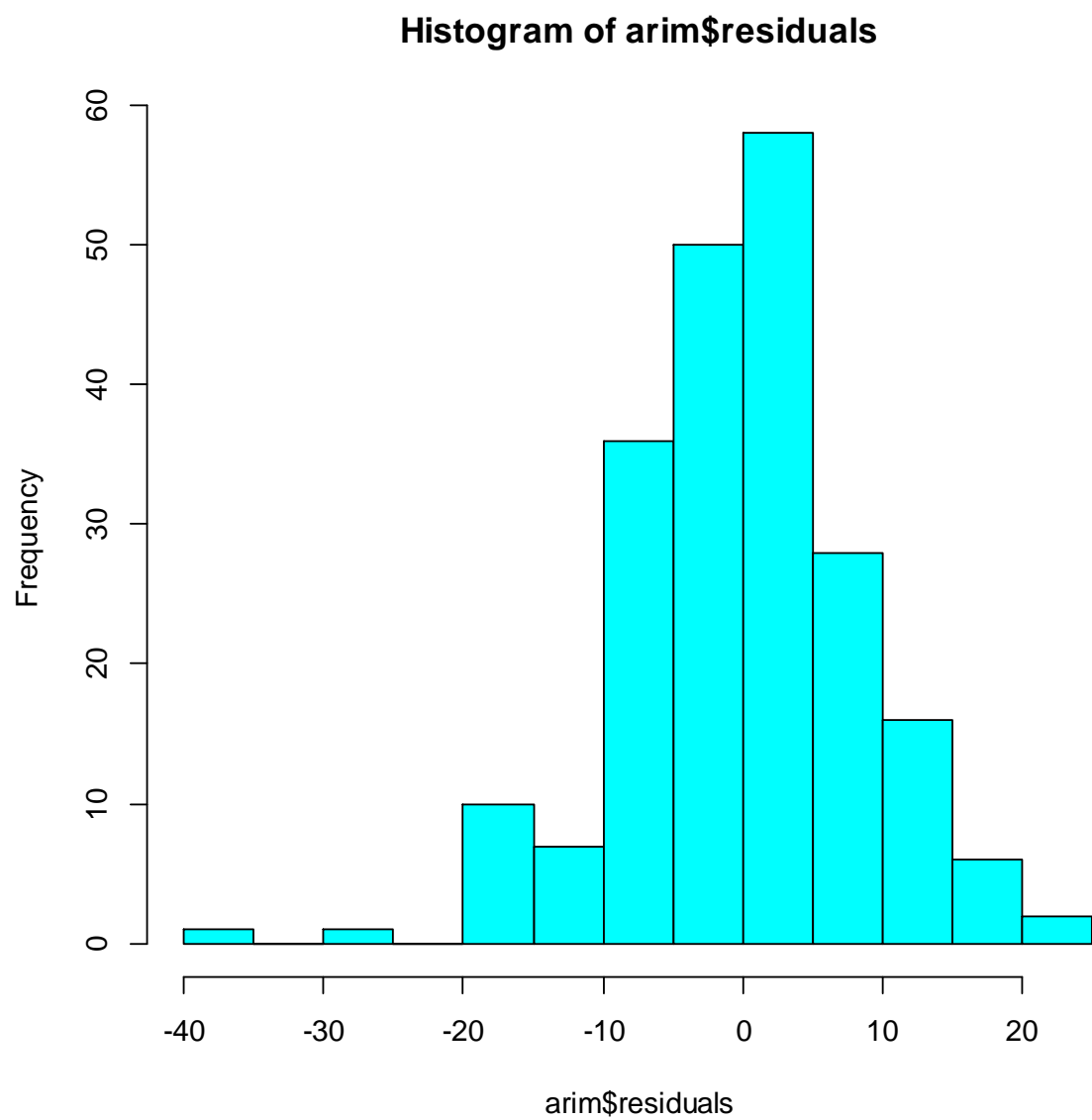
data: arim\$residuals

X-squared = 5.3536, df = 4, p-value = 0.2529

Le test de Box-Pierce nous donne une p-value=0,6806 supérieur a 5%. On rejette donc l'hypothèse alternative en faveur de l'hypothèse nulle, alors on en déduit que les résidus ne sont pas corrélés.

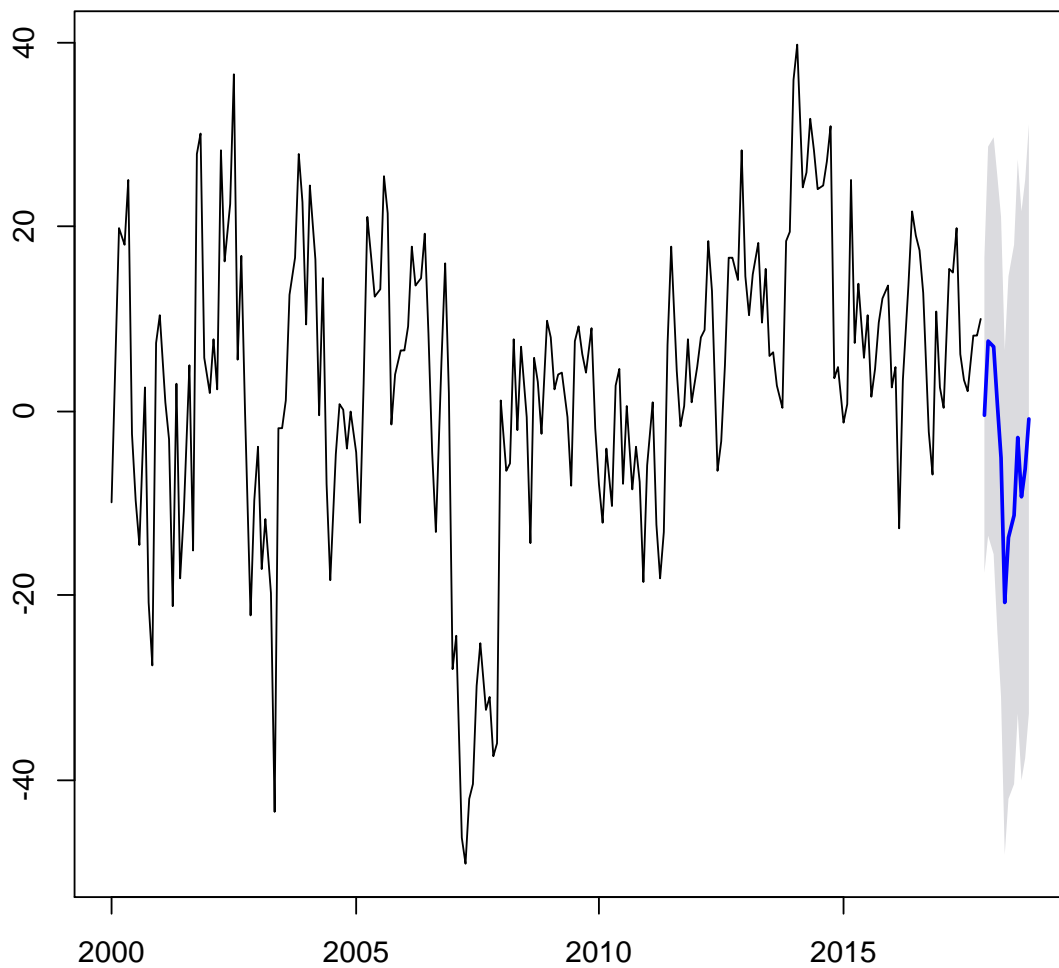


Graphique 17 : Histogramme des résidus du modèle



## 7. Prévision

### Forecasts from ARIMA(2,0,2)(2,0,1)[12] with non-zero mean



## **SECTION 2 ANALYSES AVEC LE LOGICIEL SAS**

# Partie I:

# Exploitation des données &

# Théorie des Sondages

Results Viewer - SAS Output

La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum
Num	1260	630.5000000	363.8749785	1.0000000	1260.00
CNTY	1260	54.6976190	18.2688812	27.0000000	76.0000000
DIFF	1260	1.4817460	0.4998651	1.0000000	2.0000000
STATUS	1260	0.8222222	0.3824774	0	1.0000000
HECTARE	1062	80.3549906	35.0350191	0	245.0000000
ToF	1260	2.4261905	1.1864123	1.0000000	6.0000000
OWNLAND	1260	0.3936508	0.4887529	0	1.0000000
AGE	1260	43.5396825	8.7621023	24.0000000	65.0000000
HARVEST	1260	90.4222222	1.8088516	88.0000000	94.0000000
r1	1260	0.5837619	0.3321035	0.1200000	3.4900000
r2	1260	0.5216349	0.2882747	0	1.0000000
r3	1260	0.3733968	0.2153027	-1.5100000	1.0000000
r4	1260	0.2253889	0.1949813	-0.4900000	1.1100000
r5	1260	0.3563810	0.2225569	0.000060000	2.7900000
r6	1260	1.0473889	0.6914834	0.1800000	4.3600000
r7	1260	0.6563096	0.4662306	0.000050000	3.1200000
r8	1260	0.3837143	0.3391428	-0.9300000	1.8500000
r11	1260	0.2711746	0.4245406	-0.9100000	2.3100000
r12	1260	0.4399206	0.6452338	-1.1600000	3.9700000
r14	1260	0.6959683	0.6480365	-0.6900000	5.1700000
r17	1260	0.0642540	0.0255985	0.0100000	0.1900000
r18	1260	0.0710397	0.0563684	0	0.3000000
r19	1260	0.1826111	0.1273837	0.0200000	1.6500000
r21	1260	0.2805238	0.3642325	0.0100000	5.0800000
r22	1260	0.6989683	0.8337390	0.0500000	13.6700000
r24	1260	0.1055229	0.0884155	0.0100000	0.7500000

Sortie - (Sans titre) | Journal - (Sans titre) | Script\_Sas | Results Viewer - SAS ...

Results Viewer - SAS Output

La procédure MEANS

Variable	N	Moyenne	Ec-type	Minimum	Maximum
Num	1260	630.5000000	363.8749785	1.0000000	1260.00
CNTY	1260	54.6976190	18.2688812	27.0000000	76.0000000
DIFF	1260	1.4817460	0.4998651	1.0000000	2.0000000
STATUS	1260	0.8222222	0.3824774	0	1.0000000
HECTARE	1062	80.3549906	35.0350191	0	245.0000000
ToF	1260	2.4261905	1.1864123	1.0000000	6.0000000
OWNLAND	1260	0.3936508	0.4887529	0	1.0000000
AGE	1260	43.5396825	8.7621023	24.0000000	65.0000000
HARVEST	1260	90.4222222	1.8088516	88.0000000	94.0000000
r1	1260	0.5837619	0.3321035	0.1200000	3.4900000
r2	1260	0.5216349	0.2882747	0	1.0000000
r3	1260	0.3733968	0.2153027	-1.5100000	1.0000000
r4	1260	0.2253889	0.1949813	-0.4900000	1.1100000
r5	1260	0.3563810	0.2225569	0.000060000	2.7900000
r6	1260	1.0473889	0.6914834	0.1800000	4.3600000
r7	1260	0.6563096	0.4662306	0.000050000	3.1200000
r8	1260	0.3837143	0.3391428	-0.9300000	1.8500000
r11	1260	0.2711746	0.4245406	-0.9100000	2.3100000
r12	1260	0.4399206	0.6452338	-1.1600000	3.9700000
r14	1260	0.6959683	0.6480365	-0.6900000	5.1700000
r17	1260	0.0642540	0.0255985	0.0100000	0.1900000
r18	1260	0.0710397	0.0563684	0	0.3000000
r19	1260	0.1826111	0.1273837	0.0200000	1.6500000
r21	1260	0.2805238	0.3642325	0.0100000	5.0800000
r22	1260	0.6989683	0.8337390	0.0500000	13.6700000
r24	1260	0.1055229	0.0884155	0.0100000	0.7500000

Sortie - (Sans titre) | Journal - (Sans titre) | Script\_Sas | Results Viewer - SAS ...

# Le Système SAS

## La procédure FREQ

DIFF	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
NON	653	51.83	653	51.83
OUI	607	48.17	1260	100.00

Results Viewer - SAS Output

**Le Système SAS**

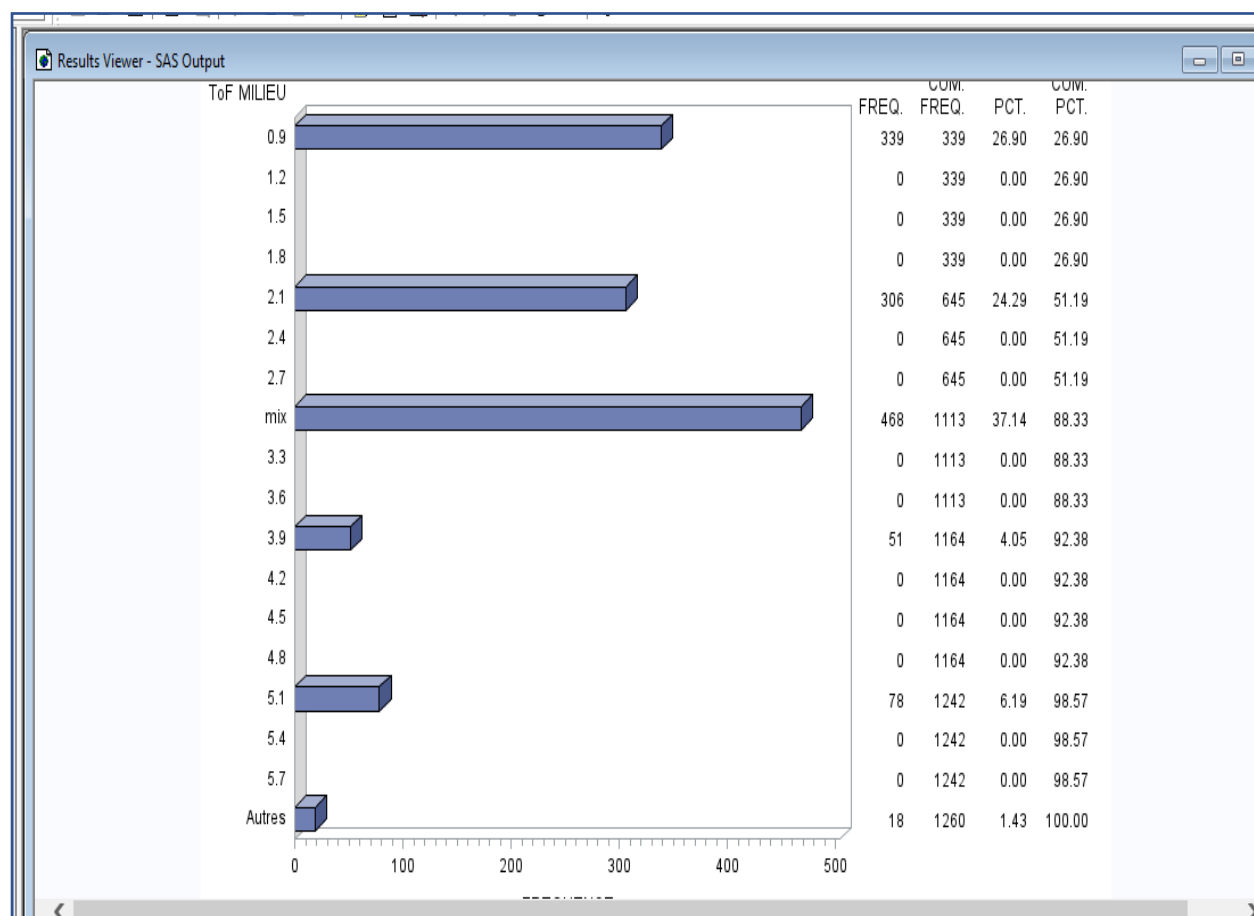
La procédure UNIVARIATE  
Variable : AGE

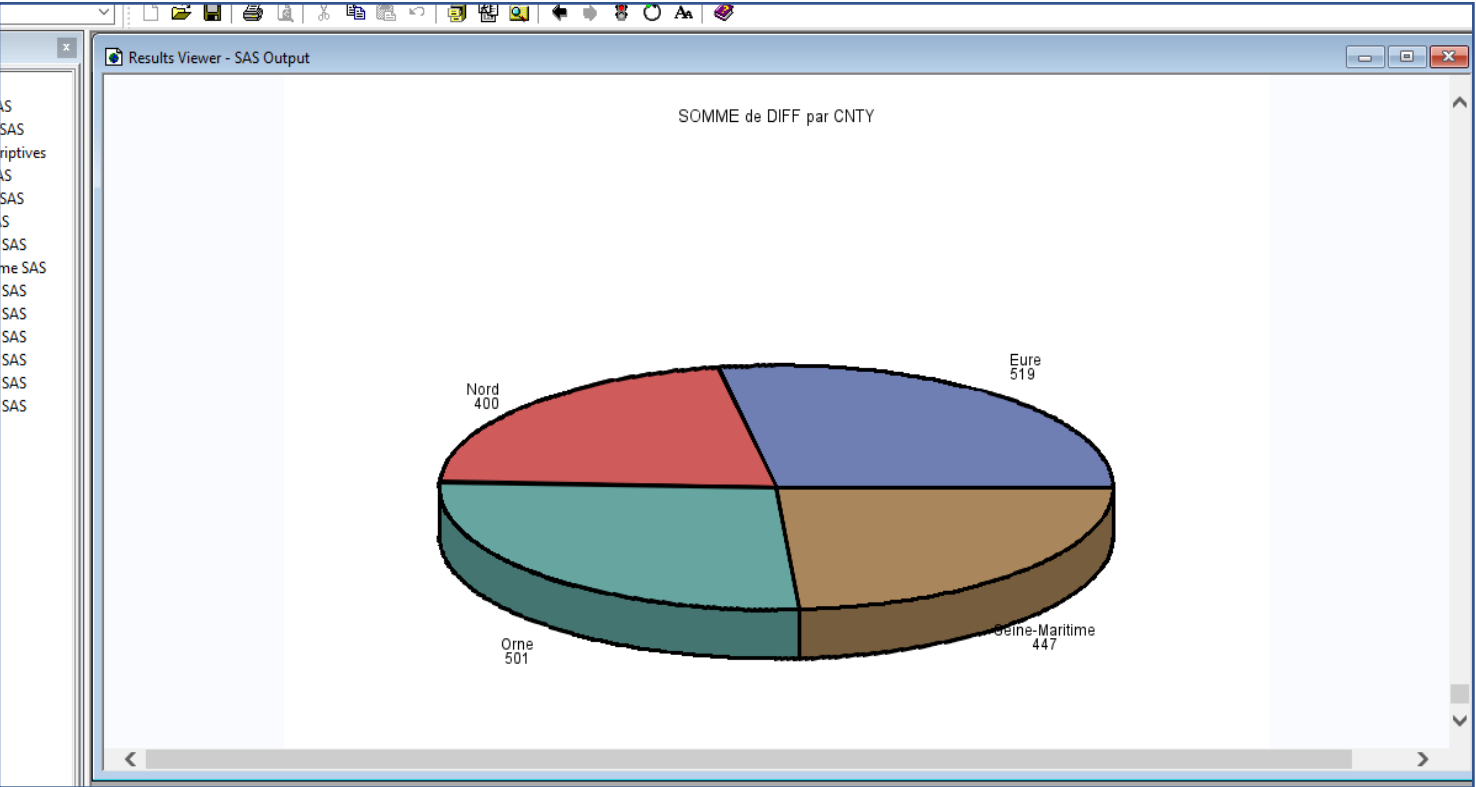
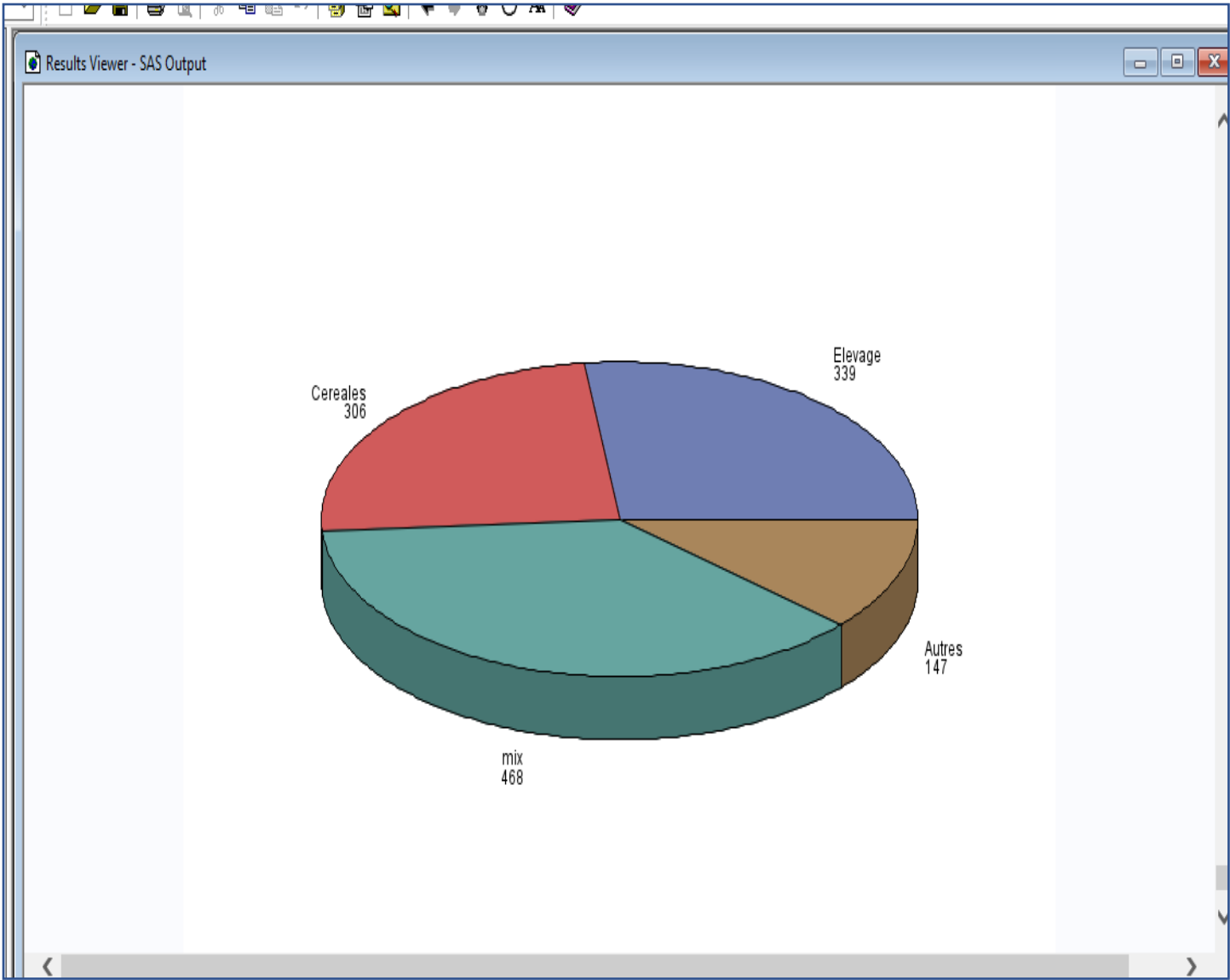
Moments			
N	1260	Somme des poids	1260
Moyenne	43.5396825	Somme des observations	54860
Ecart-type	8.7621023	Variance	76.7744368
Skewness	0.0902828	Kurtosis	-0.7488795
Somme des carrés non corrigée	2485246	Somme des carrés corrigée	96659.0159
Coeff Variation	20.1244056	Std Error Mean	0.24684427

Mesures statistiques de base			
Location		Variabilité	
Moyenne	43.53968	Ecart-type	8.76210
Médiane	44.00000	Variance	76.77444
Mode	45.00000	Intervalle	41.00000
		Ecart interquartile	13.00000

Tests de tendance centrale : $\mu_0=0$			
Test	Statistique		p-value
t de Student	t	176.3852	Pr >  t  <.0001
Signe	M	630	Pr >=  M  <.0001
Rang signé	S	397215	Pr >=  S  <.0001

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	65
99%	62
95%	59
90%	56
75% Q3	50
50% Médiane	44
25% Q1	37
10%	32
5%	30
1%	27
0% Min	24

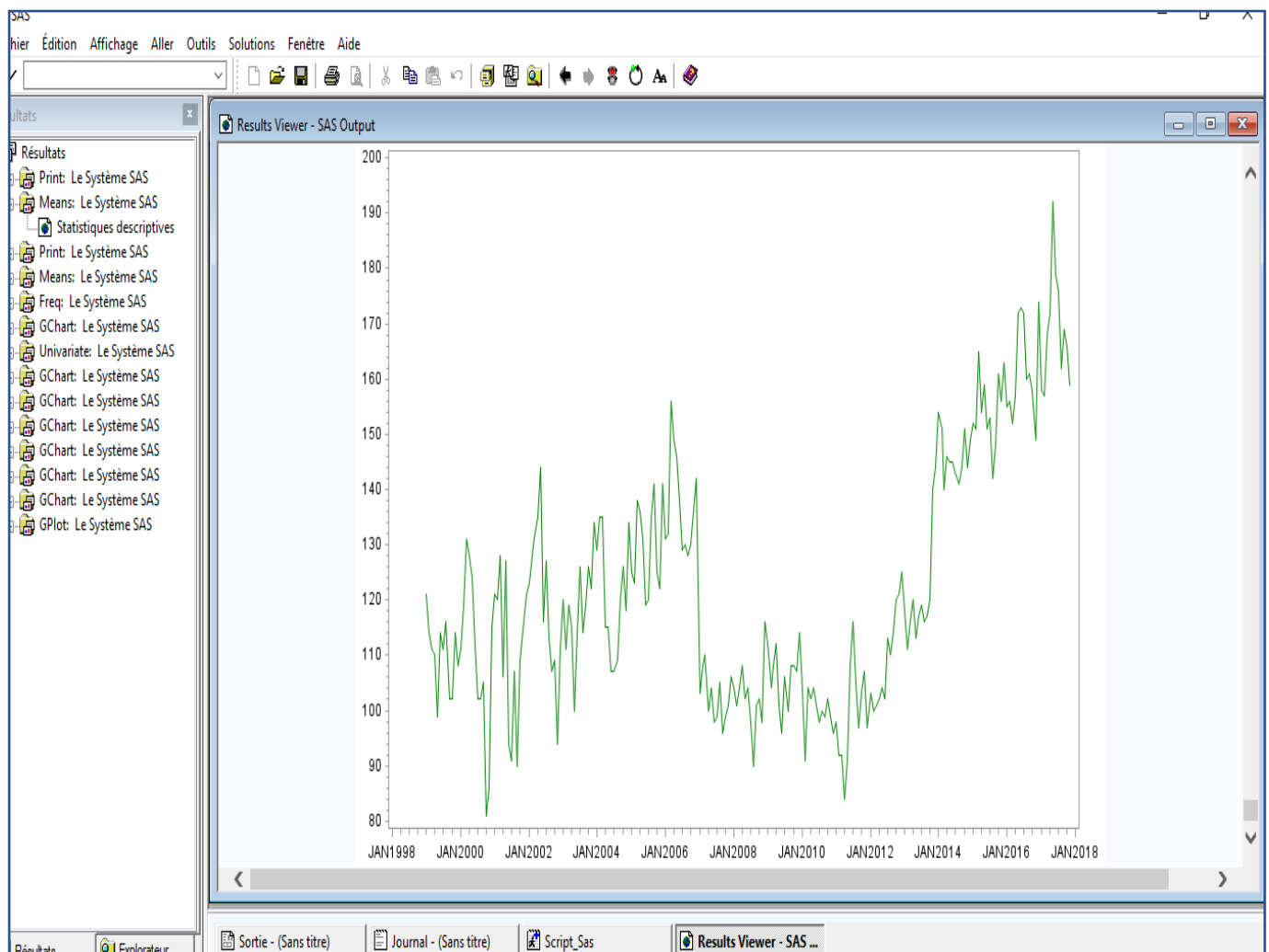






## Partie II :

# Séries Temporelles



```

120 /*test de stationarite*/
121 %dfctest( projet.data, zt, ar=6 );
122     %put p=&dfctest;
p=0.7776790369
123     /*ou*/
124     ods output stationaritytests=StationarityTests ;

```

Le Système SAS

La procédure ARIMA

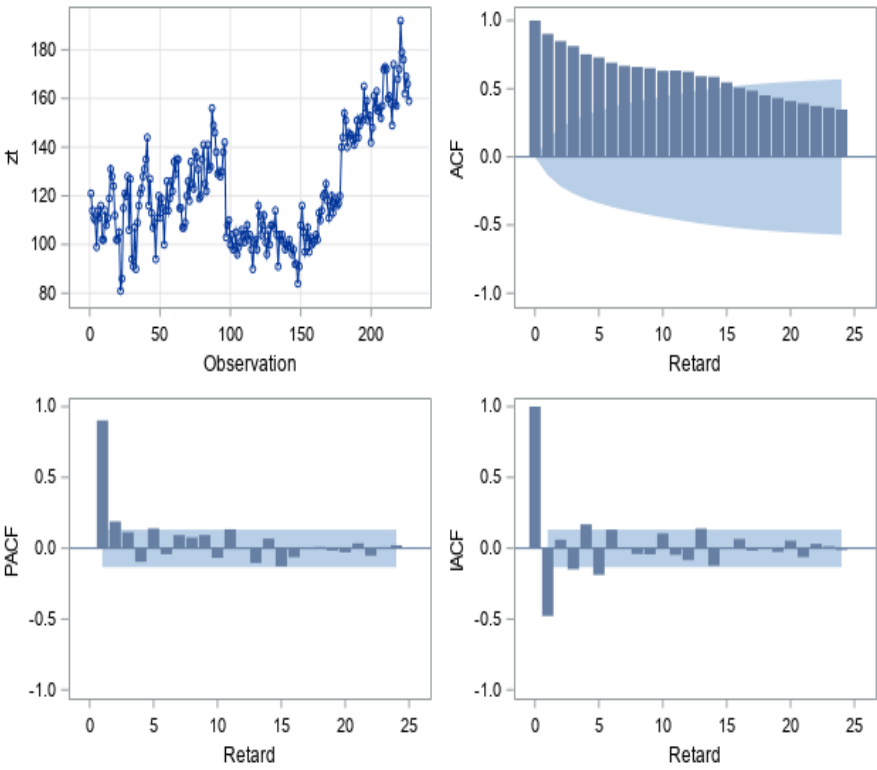
Nom de la variable = zt	
Moyenne des séries de travail	122.3568
Ecart-type	22.53437
Nombre d'observations	227

Vérification de l'autocorrélation pour le bruit blanc									
Jusqu'au retard	Khi-2	DDL	Pr > khi-2	Autocorrélations					
6	876.92	6	<.0001	0.903	0.850	0.813	0.753	0.730	0.691
12	1473.47	12	<.0001	0.668	0.660	0.652	0.631	0.633	0.625
18	1888.69	18	<.0001	0.592	0.587	0.547	0.508	0.486	0.452
24	2116.21	24	<.0001	0.432	0.410	0.393	0.374	0.362	0.348

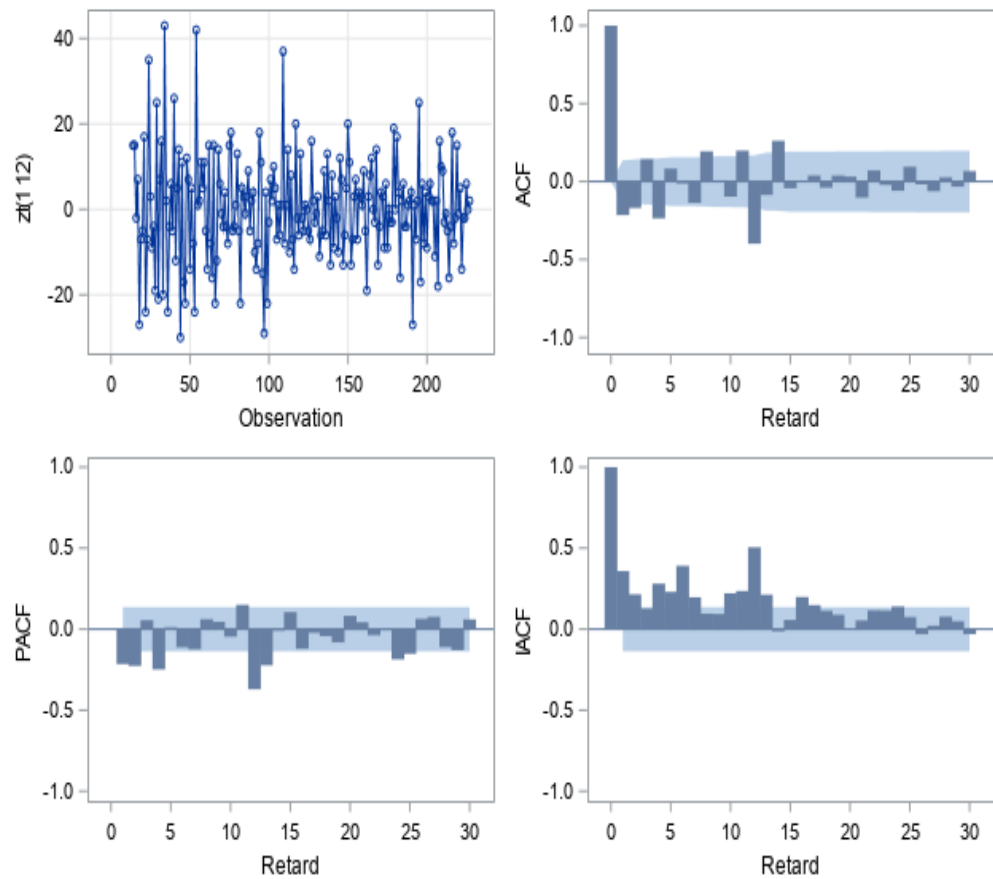
Tests de racine unitaire de Dickey-Fuller augmentés							
Type	Retards	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Moyenne zéro	0	-0.3393	0.6051	-0.29	0.5808		
	1	-0.0014	0.6819	-0.00	0.6817		

	2	-13.0999	0.2511	-2.45	0.3530	3.15	0.5472
--	---	----------	--------	-------	--------	------	--------

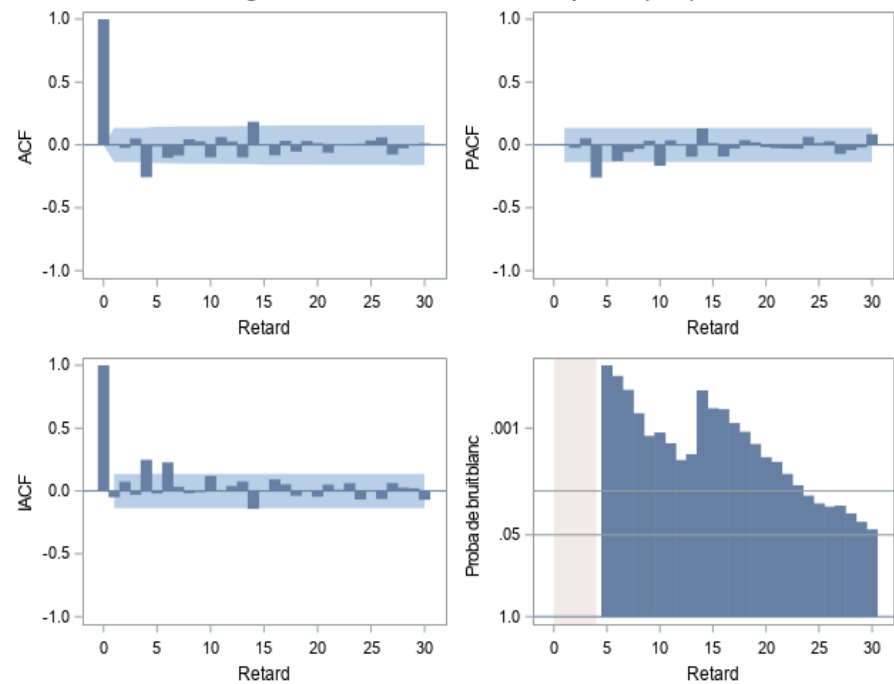
Analyse des tendances et de la corrélation pour zt



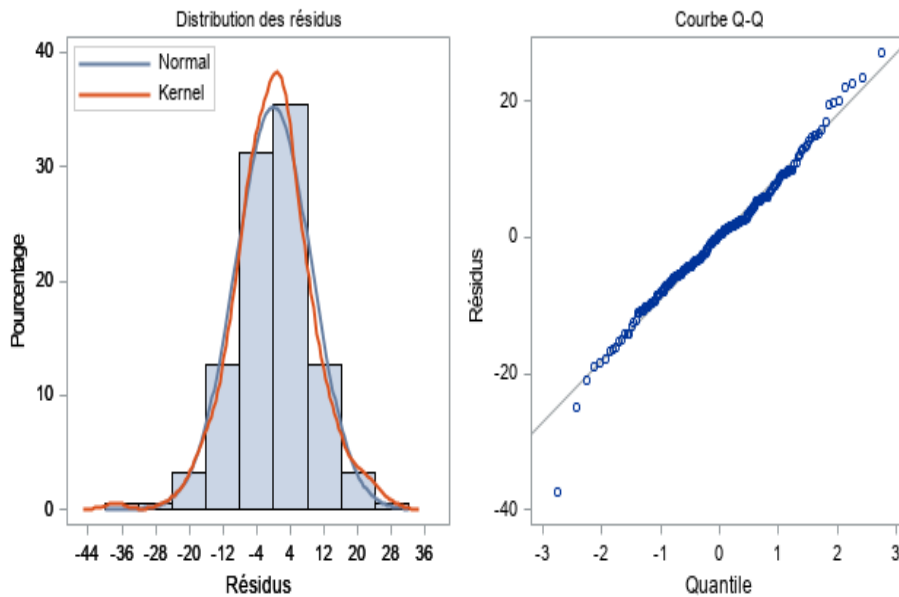
### Analyse des tendances et de la corrélation pour $zt(1\ 12)$



### Diagnostic de corrélation résiduelle pour $zt(1\ 12)$



### Diagnostic de normalité résiduelle pour zt(1 12)



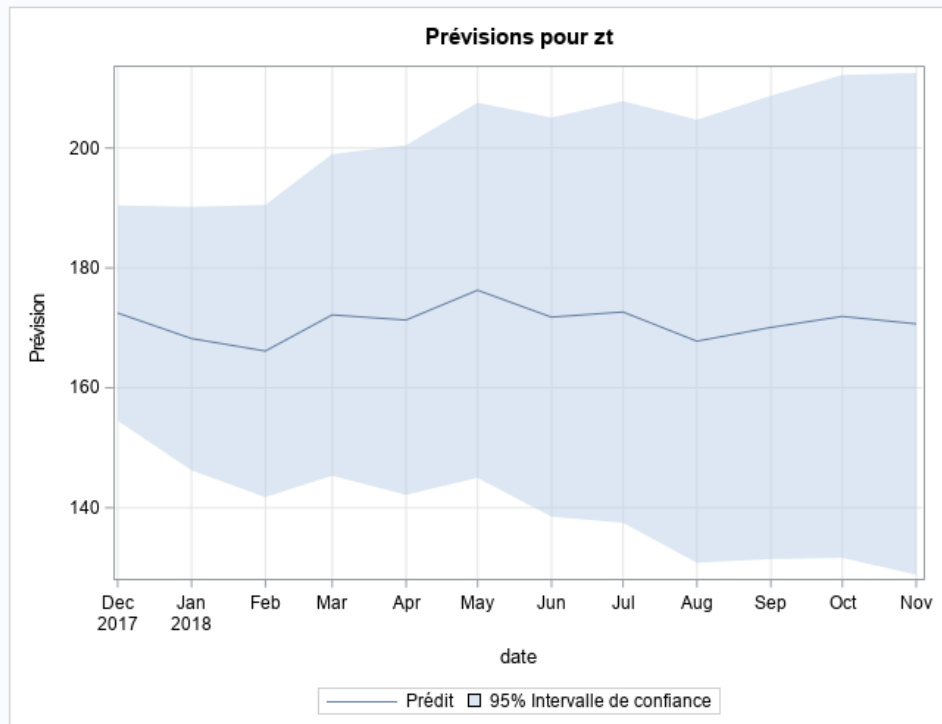
#### Modèle pour la variable zt

Moyenne estimée	0.049644
Période(s) de différenciation	1,12

Facteur Z:  $1 - 0.29295 B^{**}(1)$

#### Prévisions pour la variable zt

Obs.	Prévision	Erreur type	Intervalle de confiance à 95%	
228	172.4953	9.1506	154.5605	190.4301
229	168.2367	11.2068	146.2717	190.2017
230	166.1499	12.4355	141.7768	190.5230
231	172.1757	13.6821	145.3592	198.9922
232	171.3211	14.8739	142.1688	200.4734
233	176.2919	15.9634	145.0042	207.5795
234	171.8057	16.9780	138.5294	205.0819
235	172.6630	17.9367	137.5077	207.8183
236	167.7906	18.8473	130.8506	204.7306
237	170.0857	19.7157	131.4437	208.7278
238	171.9280	20.5474	131.6559	212.2001
239	170.6838	21.3467	128.8451	212.5226



### SECTION 3 ANALYSE AVEC LE LOGICIEL PYTHON

# **Exploitation des données de sondages avec python**

**Les individus et les variables qui sont dans le jeu de données**

```
print(df.shape)
```

```
(1260, 31)
```

**Le type de chacune des variables**

```
print(df.shape)
```

```
print(df.dtypes)
```

```
Num      int64
```

```
CNTY      int64
```

```
DIFF      int64
```

```
STATUS    int64
```

```
HECTARE   object
```

```
ToF       int64
```

```
OWNLAND   int64
```

```
AGE        int64
```

```
HARVEST   int6
```

```
r22        object
```

```
r24        object
```

```
r28        object
```

```
r30        object
```

```
r32        object
```

```
r36        object
```

```
r37        object
```

```
dtype: object
```

**4.le nombre d'exploitants étudiés dans l'échantillon**

```
print(df['Num'].describe())
```

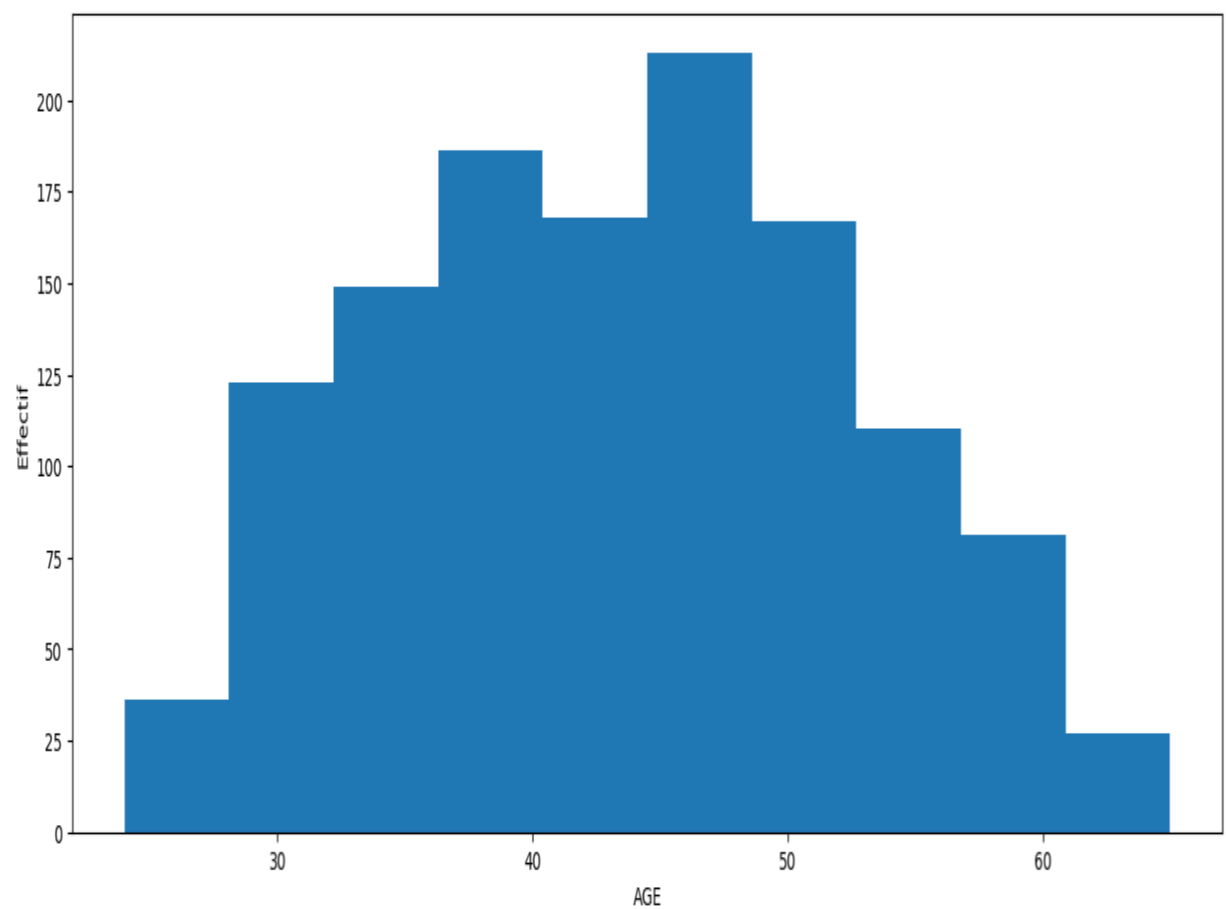


Count 1260.000000

## 5.Representation de la répartition de l'âge des exploitants

```
_=plt.hist(df["AGE"])
_=plt.xlabel( "AGE")
_=plt.ylabel( "Effectif")
_=plt.col="gray")
plt.show()
print(df['AGE'].describe())
```

---



## 6. Répartition des surfaces d'exploitation

```
_ = plt.hist(df["HECTARE"])  
_ = plt.xlabel( "surface")  
_ = plt.ylabel( "Effectif")  
_ = plt.col="gray")  
plt.show()  
print(df['HECTARE'].describe())
```

## QUESTION 11##

```
df.boxplot(column='ToF',by='CNTY')
```

```
boxplot.show()
```

