المملكة المغربية
المندوبية السامية للتخطيط
ⵜⴰⵎⵎⵓⵔⵜ ⵏ ⵍⵎⵖⵔⵉⴱ
HAUT-COMMISSARIAT AU PLAN

INSEA

Rapport de projet
Séries Chronologiques

Réalisé par:

Safaa Nidabdellah &
Hafsa Ezzidani

Table de matières

1 Introduction

1.1 Description des données

La série étudiée représente la production mensuelle moyenne de lait par vache aux États-Unis, exprimée en kilogrammes, et couvre intégralement la période allant de janvier 1962 à décembre 1975, soit 168 observations mensuelles consécutives.

1.2 Chargement des données

On importe le fichier csv sur R

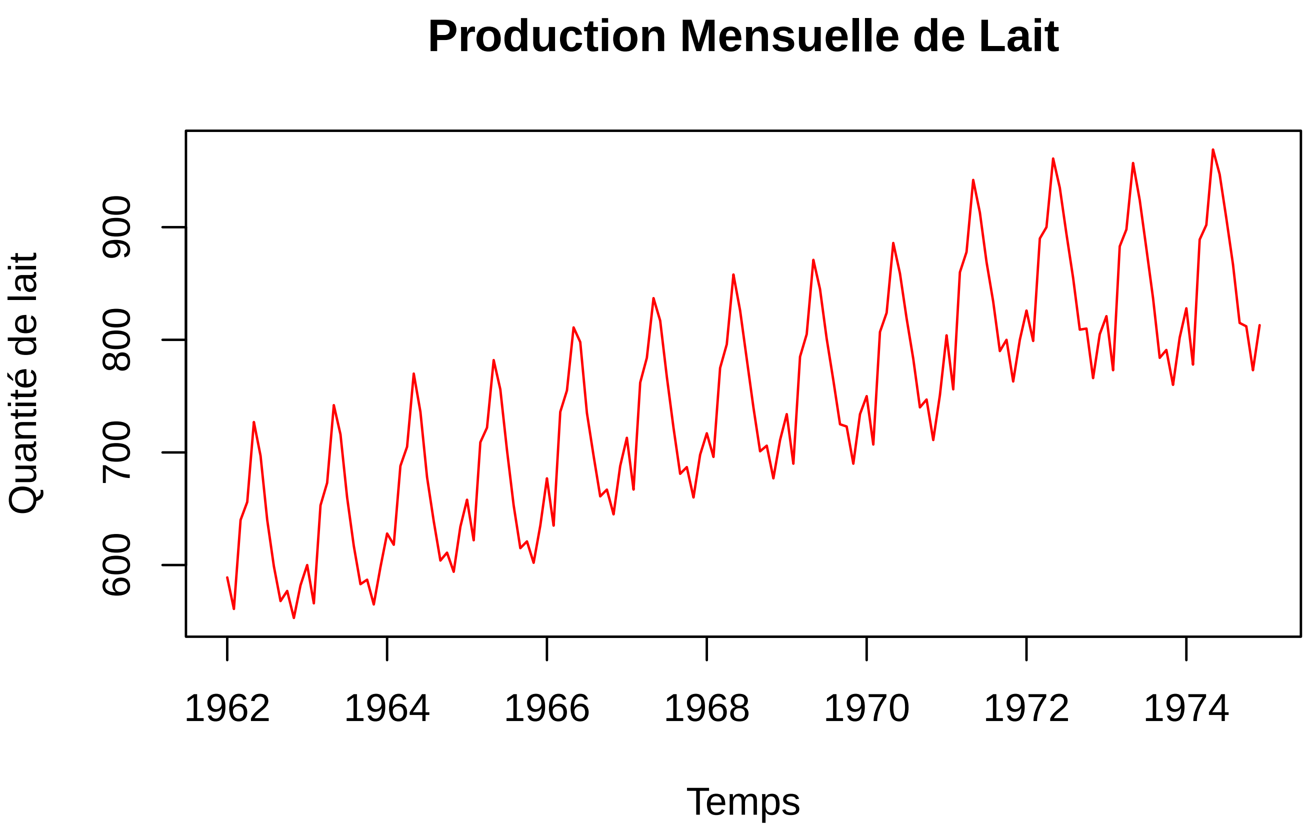
Month	Milk Production
1962-01-01	589
1962-02-01	561
1962-03-01	640
⋮	⋮
1975-10-01	827
1975-11-01	797
1975-12-01	843

1.3 Division des données

Pour modéliser et valider les prévisions, le jeu de données a été scindé : les 156 premières observations (janvier 1962 – décembre 1974), soit environ 90 % de la série, constituent le jeu d'apprentissage utilisé pour estimer les paramètres du modèle. Les 12 observations restantes (janvier – décembre 1975), représentant les 10 % les plus récents, forment le jeu test dédié à l'évaluation hors-échantillon.

2 Illustration graphique

2.1 Illustration graphique



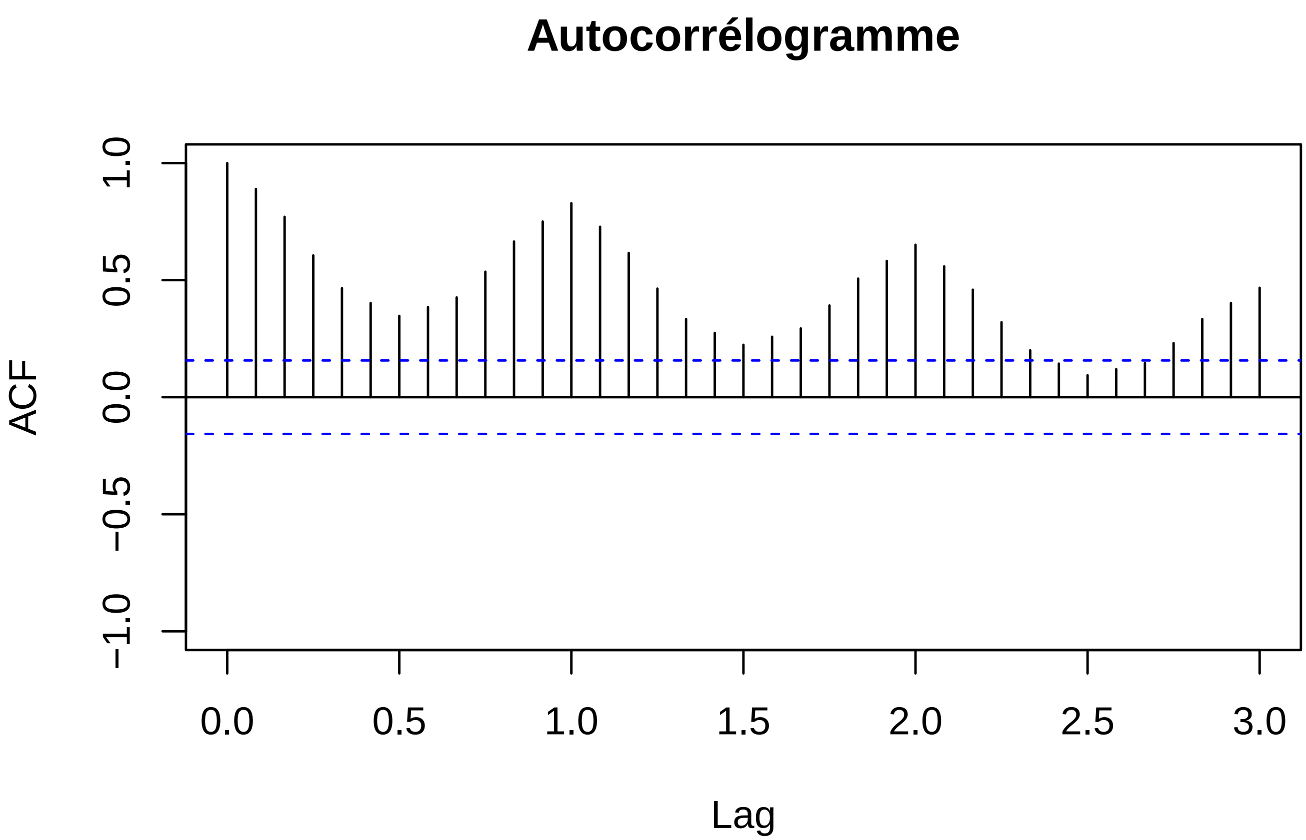
2.2 Analyse Qualitative

- **Tendance**: La série présente une tendance haussière soutenue
- **Saisonnalité**: Chaque année : creux prononcé en hiver (déc.-févr.), pic net en été (juin-août).

3 Test de stationnarité

Dans ce qui suit, on effectue les tests et les calculs sur l'échantillon d'apprentissage

3.1 Autocorrélogramme



La fonction d'autocorrélation ne s'éteint pas rapidement, alors la série n'est pas stationnaire.
Pour valider cela on effectue les tests de stationnarité: Dickey-Fuller augmenté et KPSS.

3.2 Tests de stationnarité

3.2.1 Dickey-Fuller augmenté

```
Augmented Dickey-Fuller Test
data:  train_ts
Dickey-Fuller = -9.9714, Lag order = 5, p-value = 0.01
```

La p-value est inférieur à 0.05 ainsi H_0 est rejetée donc d'après le test d'ADF , la série est stationnaire

3.2.2 KPSS

```
KPSS Test for Level Stationarity
data:  train_ts
KPSS Level = 2.7659, Truncation lag parameter = 4, p-value =
  0.01
```

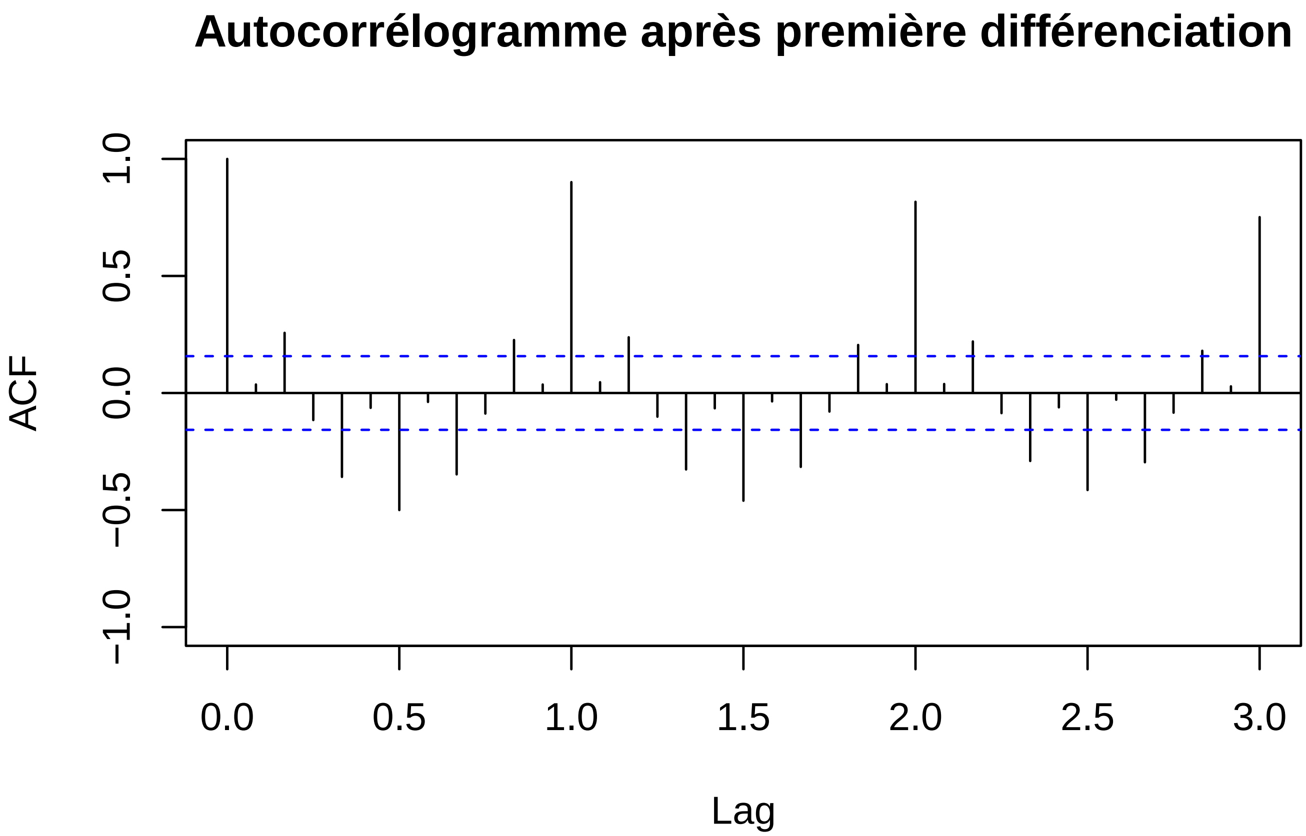
La p-value est inférieure à 0.05 ainsi H_0 est rejetée donc d'après le test de KPSS, la série est non stationnaire

3.2.3 Conclusion

Les tests ne sont pas unanimes sur le résultat de la stationanrité. La série n'est donc pas stationnaire. On procède à la stationnariser par différenciation.

4 Stationnarisation de la série

4.1 Première différenciation



Une première différenciation a été appliquée pour supprimer la tendance. Toutefois, la présence persistante de motifs saisonniers visibles dans l'ACF impose une différenciation saisonnière supplémentaire.

4.2 Différenciation saisonnière

Après différnciation saisonnière la série devrait être stationnaire. Pour s'en assurer, on effectue les tests de stationnarité.

```
        Augmented Dickey-Fuller Test

data:  train_dif1s
Dickey-Fuller = -5.8483, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary


        KPSS Test for Level Stationarity

data:  train_dif1s
KPSS Level = 0.040142, Truncation lag parameter = 4, p-value
  = 0.1


        Phillips-Perron Unit Root Test

data:  train_dif1s
Dickey-Fuller Z(alpha) = -189.4, Truncation lag parameter =
  4, p-value = 0.01
alternative hypothesis: stationary
```

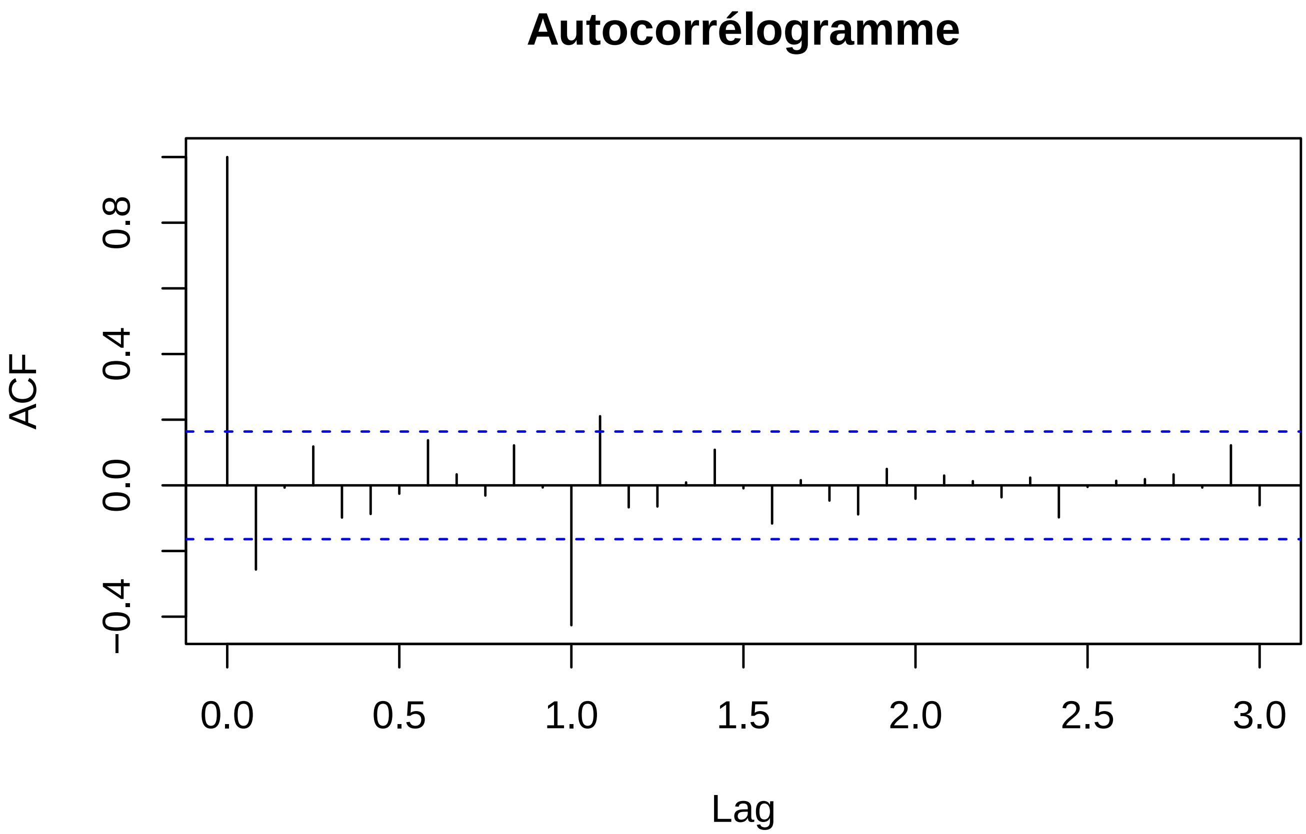
Tous les tests montrent que la série est stationnaire.
Alors on procède à la méthode de Box-Jenkins.

5 Méthode de Box-Jenkins

5.1 Identification du modèle

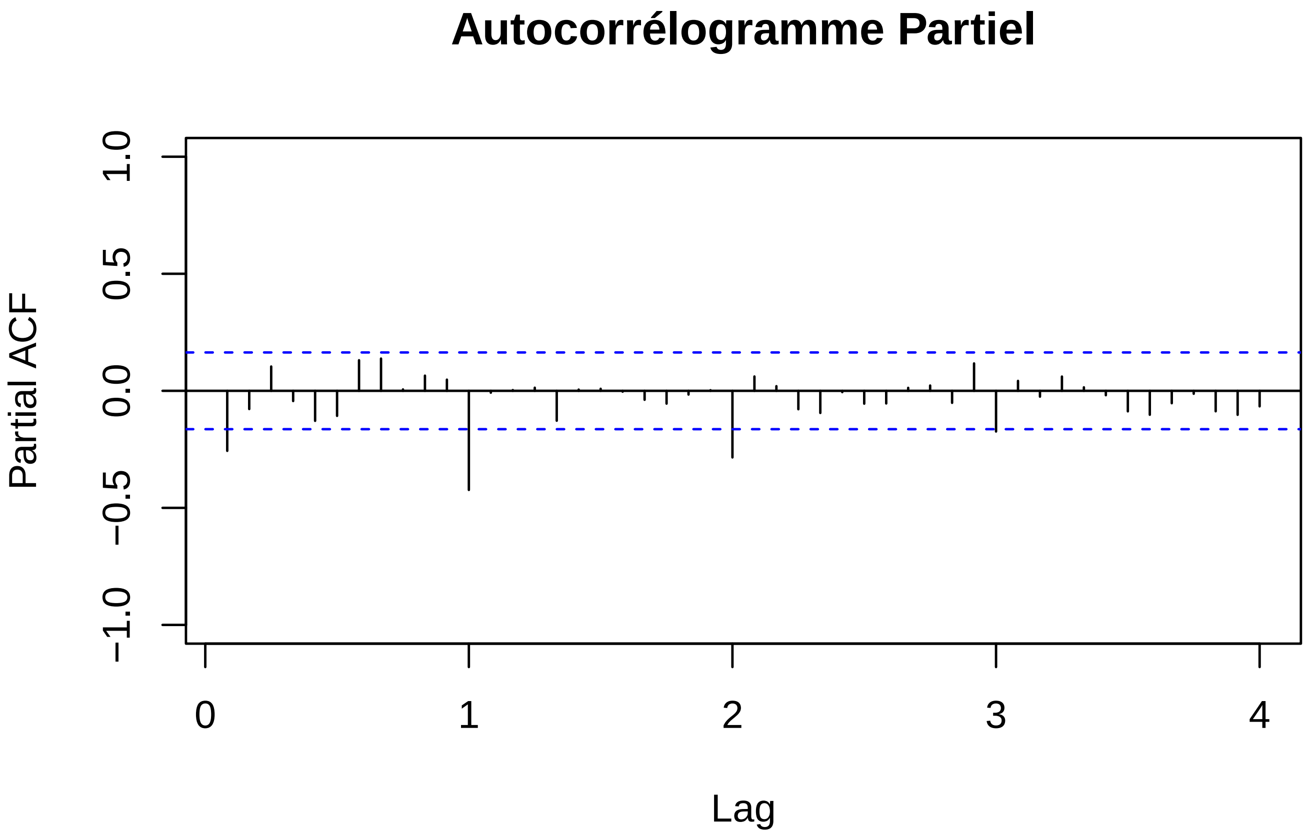
$$SARIMA(p,1,q)(P,1,Q)_{12}$$

5.1.1 Autocorrélogramme



Alors: $q_{max} = 2$ et $Q_{max} = 1$

5.1.2 Autocorrélogramme partiel



Alors: $p_{max} = 1$ et $P_{max} = 2$

5.2 Estimation

Pour l'estimation, nous avons procédé à une recherche systématique des spécifications SARIMA possibles :

- Ordres réguliers : $p \in \{0, 1\}$ et $q \in \{0, 1, 2\}$
- Ordres saisonniers : $P \in \{0, 1, 2\}$ et $Q \in \{0, 1\}$
- Différenciations fixées : $d = 1$ (tendance) et $D = 1$ (saisonnalité), période $s = 12$

Nous n'avons conservé qu'une spécification lorsqu'aucun de ses coefficients (hors variance) n'était non-significatif : critère $|z| > 1,96$ (niveau 5 %)

Cette procédure aboutit à 13 modèles plausibles, tous compatibles avec la structure $(p, 1, q)(P, 1, Q)_{12}$. On obtient notamment :

$$
\begin{gathered}
SARIMA(0,1,0)(0,1,1)_{12} \\
SARIMA(0,1,0)(1,1,0)_{12} \\
SARIMA(0,1,0)(2,1,0)_{12} \\
SARIMA(0,1,1)(0,1,0)_{12} \\
SARIMA(0,1,1)(0,1,1)_{12} \\
SARIMA(0,1,1)(1,1,0)_{12} \\
SARIMA(0,1,1)(2,1,0)_{12} \\
SARIMA(1,1,0)(0,1,0)_{12} \\
SARIMA(1,1,0)(0,1,1)_{12} \\
SARIMA(1,1,0)(1,1,0)_{12} \\
SARIMA(1,1,0)(2,1,0)_{12} \\
SARIMA(1,1,2)(0,1,0)_{12} \\
SARIMA(1,1,2)(0,1,1)_{12}
\end{gathered}
$$

5.3 Validation

5.3.1 Test de Box-Pierce

Après l'estimation, chacun des 13 modèles jugés "significatifs" a été soumis au test de Box–Pierce.
Seuil : p-value>0,05 $\rightarrow$ absence d'autocorrélation résiduelle, modèle accepté.

7 modèles sur 13 présentent des résidus assimilables à un bruit blanc (p > 0,05). Les six autres sont rejetés pour autocorrélation résiduelle significative.

$$SARIMA(0,1,1)(0,1,1)_{12}$$
$$SARIMA(0,1,1)(1,1,0)_{12}$$
$$SARIMA(0,1,1)(2,1,0)_{12}$$
$$SARIMA(1,1,0)(0,1,1)_{12}$$
$$SARIMA(1,1,0)(1,1,0)_{12}$$
$$SARIMA(1,1,0)(2,1,0)_{12}$$
$$SARIMA(1,1,2)(0,1,1)_{12}$$

5.3.2 Critère de l'AIC

Les sept modèles qui avaient passé le test de Box-Pierce ont été comparées au moyen de l'Akaike Information Criterion (AIC)

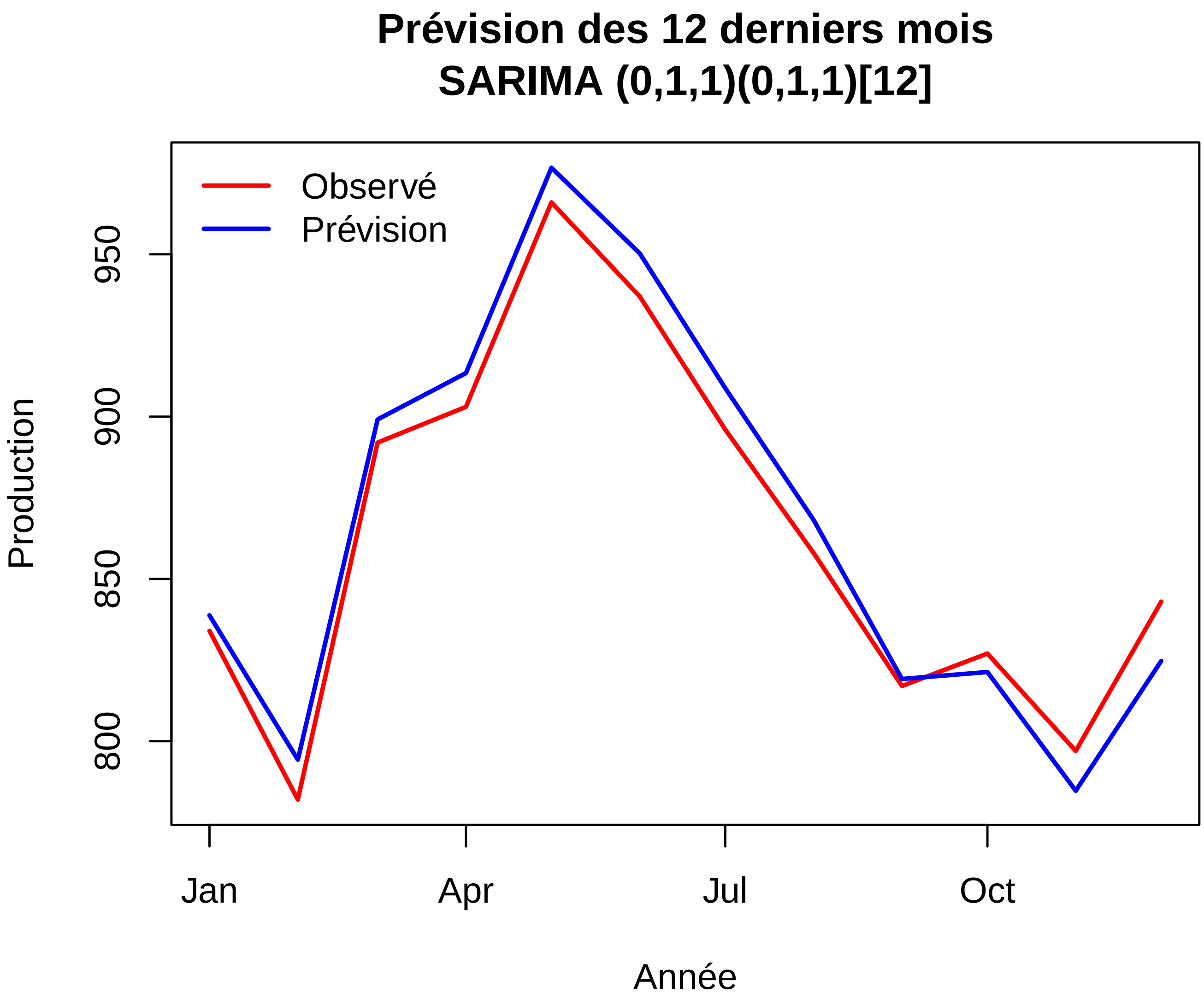
Modèle	p-value(BP)	AIC
$SARIMA(0,1,1)(0,1,1)_{12}$	0.5330	988.2
$SARIMA(1,1,0)(0,1,1)_{12}$	0.4617	988.5
$SARIMA(1,1,2)(0,1,1)_{12}$	0.5007	989.1
$SARIMA(0,1,1)(2,1,0)_{12}$	0.3180	994.9
$SARIMA(1,1,0)(2,1,0)_{12}$	0.2674	995.1
$SARIMA(1,1,0)(1,1,0)_{12}$	0.2107	1003.1
$SARIMA(0,1,1)(1,1,0)_{12}$	0.2509	1003.3

5.3.3 Interprétation

Le modèle $SARIMA(0,1,1)(0,1,1)_{12}$ possède l'AIC le plus faible; il est donc retenu comme modèle final.

5.4 Prévision

Pour évaluer la capacité prédictive du modèle sélectionné $SARIMA(0,1,1)(0,1,1)_{12}$, nous avons généré une pré à 12 mois (janvier – décembre 1975) et nous l'avons comparée à l'échantillon de validation.



6 Évaluation des prévisions

Les performances du modèle final $SARIMA(0, 1, 1)(0, 1, 1)_{12}$ ont été quantifiées à l'aide de deux indicateurs classiques :

- **RMSE** (Root Mean Square Error): 10,85 kg
- **MAPE** (Mean Absolute Percentage Error): 1.15%

La production mensuelle se situe en moyenne autour de 750 kg par vache ; une erreur quadratique moyenne de 10,9 kg représente donc 1,4 % du niveau moyen.
Le MAPE confirme ce diagnostic : à 1,16 %, l'erreur absolue relative est quasi négligeable.

7 Vérification de la normalité

On effectue le test de Shapiro-Wilk sur les résidus du modèle

```
Shapiro-Wilk normality test
data:  std_residuals
W = 0.96115, p-value = 0.0002294
```

p = 0,00023 < 0,05 on rejette l'hypothèse nulle de normalité.

8 Conclusion

L'analyse de la production mensuelle de lait par vache (1962-1975) a montré une tendance haussière soutenue et une saisonnalité annuelle marquée. Après une double différenciation $(1-B)(1-B^{12})$, la série est devenue stationnaire, ce qui a permis d'identifier puis d'estimer un ensemble de modèles SARIMA. Le filtrage successif par :

1. significativité des coefficients,
2. test de Box-Pierce (résidus indépendants)
3. critère d'information d'Akaike,

a conduit à retenir le modèle $SARIMA(0, 1, 1)(0, 1, 1)_{12}$. Sur l'année de validation (1975), ce modèle atteint :

- RMSE = 10,85 kg (1,4 % du niveau moyen)
- MAPE = 1,16 %

ce qui représente une performance largement meilleure qu'un simple modèle naïf basé sur la saison précédente. Bien que le test de Shapiro-Wilk rejette la normalité (p 0,0002), les écarts observés aux extrêmes restent mineurs et n'altèrent pas la pertinence des prévisions ponctuelles.
En pratique, le modèle fournit donc des prévisions mensuelles fiables