« MODELISATION A L'AIDE DE LA MODELISATION BAYESIENNE ET DU TEST DE NORMALITE SHAPIRO-WILK BASEES SUR DES DONNEES HISTORIQUES »

1-RABEMIAFARA Ruffin Michel
Ecole Doctorale Ingénierie et Géosciences (INGE)
2- RAHAJAMANANA Jasmin
Maitre de Conférences
3-Fidihery Toky Tantely
Maitre de Conférences
Université de Toliara
4- RAKOTOSON Tolontsoa
Ecole Doctorale Ingénierie et Géosciences (INGE)
5- Pr RAHARIMALALA Laurence
Ecole Doctorale Ingénierie et Géosciences (INGE)

RESUME

La fraude fiscale représente un défi complexe pour les administrations fiscales, nécessitant des méthodes sophistiquées pour détecter et prévenir les comportements frauduleux. Cet article explore l'application de la modélisation bayésienne et du test de normalité de Shapiro-Wilk pour la détection de fraude fiscale à partir de données historiques. La modélisation bayésienne permet d'incorporer des connaissances préalables et de mettre à jour les probabilités au fur et à mesure que de nouvelles données sont disponibles, offrant une approche flexible et robuste pour la détection des anomalies. Le test de normalité de Shapiro-Wilk, quant à lui, évalue si les distributions des données fiscales s'écartent de la normalité, ce qui peut être indicatif de fraudes. Les approches traditionnelles de détection de la fraude, basées sur des règles et des contrôles manuels, se révèlent souvent inefficaces face à la complexité croissante des schémas de fraude et à l'augmentation du volume de données fiscales. Les résultats montrent que la combinaison de ces deux approches permet de détecter efficacement des schémas de fraude qui échappent aux méthodes traditionnelles. La modélisation bayésienne s'est avérée particulièrement utile pour intégrer des informations a priori et ajuster les probabilités de fraude en fonction des nouvelles données, tandis que le test de Shapiro-Wilk a permis de repérer des anomalies statistiques significatives. Cette double approche offre une méthodologie rigoureuse et innovante pour la détection proactive des fraudes fiscales.

Mots-clés : fraude fiscale, modélisation, bayésienne,Shapiro-Wilk, normalité

ABSTRACT

Tax fraud represents a complex challenge for tax authorities, requiring sophisticated methods to detect and prevent fraudulent behaviour. This article explores the application of Bayesian modelling and the Shapiro-Wilk normality test to the detection of tax fraud based on historical data. Bayesian modelling allows prior knowledge to be incorporated and probabilities to be updated as new data becomes available, providing a flexible and robust approach to anomaly detection. The Shapiro-Wilk normality test, on the other hand, assesses whether the distributions of tax data deviate from normality, which may be indicative of fraud. Traditional approaches to fraud detection, based on manual rules and controls, often prove ineffective in the face of the growing complexity of fraud schemes and the increasing volume of tax data. The results show that the combination of these two approaches can effectively detect patterns of fraud that escape traditional methods. Bayesian modelling proved particularly useful for integrating a priori information and adjusting fraud probabilities based on new data, while the Shapiro-Wilk test was used to identify significant statistical anomalies. This two-pronged approach provides a rigorous methodology that is both effective and efficient.

Key words: tax fraud, modelling, Bayesian, Shapiro-Wilk, normality

INTRODUCTION

L'analyse de données historiques concernant la fraude fiscale à Madagascar est un pilier de la recherche scientifique et de la prise de décision dans divers domaines. Elle permet de comprendre les tendances passées, de faire des prédictions futures et d'éclairer les choix stratégiques. La modélisation statistique joue un rôle crucial dans l'analyse et l'interprétation des données historiques. Parmi les diverses approches disponibles, la modélisation bayésienne a gagné en popularité en raison de sa flexibilité et de sa capacité à incorporer des connaissances a priori dans le processus d'inférence [1][2]. En parallèle, assurer la normalité des données est une étape essentielle pour plusieurs tests statistiques classiques, rendant le test de normalité Shapiro-Wilk particulièrement pertinent. L'approche probabiliste intègre des informations préexistantes, appelées priors, avec des données observées pour inférer la distribution des paramètres d'un modèle [3]. Elle permet ainsi d'obtenir des informations plus complètes et plus nuancées sur les paramètres du modèle que les approches fréquentistes traditionnelles.Cet article explore l'application de la modélisation bayésienne et du test de normalité Shapiro-Wilk sur des ensembles de données historiques, illustrant comment ces méthodes peuvent être intégrées pour améliorer la qualité et la fiabilité des analyses statistiques. En particulier, on examine comment l'utilisation des connaissances a priori peut influer sur les résultats de la modélisation bayésienne et discutons de l'importance de vérifier la normalité des résidus modélisés à l'aide du test de Shapiro-Wilk.

MATERIELS

a) Données historiques

Dans cette étude, nous utilisons les données historiques de fraude fiscale collectées de 2017 à 2023 pour analyser les tendances et développer un modèle statistique bayésien de détection de fraude. Ces données couvrent une période de six ans, offrant une base solide pour identifier des motifs récurrents et des comportements anormaux potentiels.

Les données historiques sur la fraude fiscale à Madagascar sont disponibles auprès de diverses sources, notamment l'administration fiscale, le ministère des Finances et le Bureau national de la statistique. Ces données peuvent inclure des informations sur le montant de la fraude fiscale détectée, les types de fraude les plus courants et les secteurs économiques les plus touchés.

b) Outils informatiques et de modélisation

Pour la modélisation de la fraude fiscale, nous avons utilisé divers outils informatiques et de modélisation. Les principales technologies employées incluent Python pour le développement des scripts et l'analyse des données, notamment grâce aux bibliothèques pandas pour la manipulation des données, numpy pour les calculs numériques, et scipy pour les statistiques. Nous avons également utilisé des bibliothèques de modélisation bayésienne comme PyMC3 et TensorFlow Probability pour la construction et l'inférence des modèles. En complément, les environnements de développement intégrés (IDE) tels que Jupyter Notebook et PyCharm ont été utilisés pour le développement et la visualisation interactive des résultats. Enfin, des outils de gestion des versions comme Git ont été essentiels pour le suivi des modifications du code et la collaboration en équipe

METHODOLOGIE

a) Collecte des Données :

Rassembler des données historiques sur les déclarations fiscales, y compris les revenus déclarés, les déductions, les impôts payés, etc.

Collecter des données supplémentaires telles que les antécédents fiscaux des contribuables, les transactions financières suspectes.

b) Prétraitement des Données :

Nettoyer et normaliser les données pour éliminer les valeurs aberrantes et les erreurs.

Effectuer une analyse exploratoire des données pour identifier les tendances, les corrélations et les caractéristiques importantes.

c) Modélisation Bayésienne :

Définir les paramètres du modèle bayésien, y compris les priors pour les distributions des paramètres inconnus.

Construire un modèle bayésien qui capture les relations entre les variables observées et la probabilité de fraude fiscale.

Utiliser des techniques d'inférence bayésienne pour estimer les paramètres du modèle et quantifier l'incertitude.

Formule Mathématique :

Soit Y la variable binaire indiquant la présence de fraude fiscale pour un contribuable donné. Les caractéristiques des contribuables, telles que le revenu, les déductions, etc., sont représentées par le vecteur X.

Le modèle bayésien peut être formulé comme :

$$P(Y = 1|X) = \frac{P(X|Y = 1).P(Y = 1)}{P(X)}$$

Où

$P(X|Y = 1)$: Probabilité de fraude fiscale sachant les caractéristiques X.

$P(X|Y = 1)$:Distribution des caractéristiques des contribuables en cas de fraude.

$P(Y = 1)$: Probabilité a priori de fraude fiscale.

$P(X)$:Probabilité marginale des caractéristiques X.

La figure ci-dessous représente l'algorithme en Python utilisant la formule mathématique du modèle bayésien pour la détection de fraude fiscale, ainsi que l'utilisation d'une technique d'inférence bayésienne basée sur l'échantillonnage de Monte Carlo par chaînes de Markov (MCMC) :

```python
import pymc3 as pm
# Données simulées (simplifiées)
revenus = np.array([])  # Exemple de données de revenus
fraudes = np.array([0, 1, 0, 1, 1])  # Exemple de données de fraude (0: Non fraude, 1: Fraude)
# Modélisation bayésienne
with pm.Model() as model:
    # Prior sur la probabilité de fraude
    p_fraude = pm.Beta('p_fraude', alpha=1, beta=1)
    # Likelihood : Distribution des revenus en cas de fraude (simplifiée pour l'exemple)
    revenus_fraude = pm.Normal('revenus_fraude', mu, sd)
    # Likelihood : Distribution des revenus en l'absence de fraude (simplifiée pour l'exemple)
    revenus_non_fraude = pm.Normal('revenus_non_fraude', mu, sd)
    # Modèle génératif de fraude
    fraude = pm.Bernoulli('fraude', p=p_fraude, shape=len(revenus))
    revenus_observes = pm.math.switch(fraude, revenus_fraude, revenus_non_fraude)
    # Likelihood : Observations des revenus
    revenus_obs = pm.Normal('revenus_obs', mu=revenus_observes, sd, observed=revenus)
    # Échantillonnage MCMC pour l'inférence bayésienne
    trace = pm.sample(1000, tune=1000, cores=1)  # Nombre d'échantillons et de pas de "burn-in"
# Affichage des résultats
pm.summary(trace)
```

Figure 1:Aperçu de l'algorithme en Python utilisant la formule mathématique du modèle bayésien pour la détection de fraude fiscale

d) Test de Normalité Shapiro-Wilk :

Appliquer le test de normalité Shapiro-Wilk aux résidus du modèle bayésien pour vérifier l'hypothèse de normalité.

Formule mathématique :

Le test de normalité de Shapiro-Wilk teste l'hypothèse nulle selon laquelle un échantillon de données

$x_1, x_2, \ldots \ldots \ldots \ldots \ldots x_n$ provient d'une population normalement distribuée. La statistique de test W est calculée comme suit :

$$W = \frac{(\sum_{i=1}^{n} a_i \, x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Où

$x_{(i)}$:est le i-ème plus petit élément de l'échantillon trié,

$\bar{x}$:est la moyenne de l'échantillon,

a_i:sont les coefficients de correction calculés en fonction des moyennes, des variances et des covariances des ordres statistiques de l'échantillon, et dépendent de la taille de l'échantillon.

La statistique $\boldsymbol{W}$ suit une distribution spécifique sous l'hypothèse nulle de normalité. En comparant la valeur observée de $\boldsymbol{W}$ à une valeur critique à partir de cette distribution, on peut déterminer si l'échantillon semble provenir d'une population normalement distribuée.

Si les résidus ne suivent pas une distribution normale, explorer les transformations ou les méthodes alternatives pour améliorer la normalité.

La figure ci-dessus représente l'aperçu de l'algorithme en Python utilisant le test de normalité de Shapiro-Wilk pour détecter des anomalies dans les données fiscales, qui pourraient potentiellement indiquer de la fraude :

```python
from scipy.stats import shapiro
# Fonction pour effectuer le test de normalité Shapiro-Wilk
def test_normalite(data, alpha=0.05):
    # Effectuer le test de normalité Shapiro-Wilk
    stat, p_value = shapiro(data)
    # Vérifier si la p-valeur est inférieure au seuil alpha
    if p_value < alpha:
        print("Les données ne suivent pas une distribution normale (rejet de l'hypothèse nulle)")
    else:
        print("Les données suivent une distribution normale (absence de rejet de l'hypothèse nulle)")
# Données simulées (simplifiées)
donnees_fiscales = []
# Appel de la fonction de test de normalité avec les données fiscales
test_normalite(donnees_fiscales)
```

Figure 2:Aperçu de l'algorithme en Python utilisant le test de normalité de Shapiro-Wilk pour détecter des anomalies dans les données fiscales

e) Validation et Évaluation :

Diviser les données en ensembles d'apprentissage et de test pour évaluer les performances du modèle.

Mesurer les métriques de performance telles que la précision, le rappel, le score F1, etc., pour évaluer la capacité du modèle à détecter la fraude fiscale.

f) Optimisation et Réaffinement :

Itérer sur le modèle en ajustant les priors, en ajoutant de nouvelles variables explicatives ou en explorant des techniques de modélisation plus avancées.

Effectuer une validation croisée et une optimisation des hyperparamètres pour améliorer les performances du modèle.

g) Analyse comparative :

Les graphiques de catégories de force d'évidence permettent également de comparer différentes hypothèses ou scénarios en termes de force d'évidence. Cela peut aider à évaluer

quelle hypothèse est la plus plausible ou quel scénario est le plus probable en fonction des données disponibles.

Pour le graphe de SequentialAnalysis, l'axe des ordonnées représente les valeurs de $BF_{1\ 0}$ (Bayes Factor), qui est une mesure de la force d'évidence en faveur de l'hypothèse alternative par rapport à l'hypothèse nulle. Les valeurs de $BF_{1\ 0}$ peuvent être interprétées comme suit :

- ✓ $BF_{1\ 0}$ < 1 : Évidence en faveur de l'hypothèse nulle.
- ✓ $BF_{1\ 0}$ entre 1 et 3 : Évidence Anecdotique en faveur de l'hypothèse alternative.
- ✓ $BF_{1\ 0}$ entre 3 et 10 : Évidence Modérée en faveur de l'hypothèse alternative.
- ✓ $BF_{1\ 0}$ > 10 : Évidence Forte en faveur de l'hypothèse alternative.

Dans ce contexte, le graphique de SequentialAnalysis peut montrer comment la force d'évidence évolue au fur et à mesure que de nouvelles données sont collectées ou analysées. Par exemple :

- ✓ Au début de l'analyse, le $BF_{1\ 0}$ peut être faible, indiquant une faible évidence en faveur de l'hypothèse alternative.
- ✓ Au fur et à mesure que de nouvelles données sont ajoutées, le $BF_{1\ 0}$ peut augmenter, indiquant une augmentation de la force de l'évidence en faveur de l'hypothèse alternative.
- ✓ Lorsque le $BF_{1\ 0}$ dépasse un certain seuil, cela peut être interprété comme une évidence modérée ou forte en faveur de l'hypothèse alternative, en fonction des seuils préétablis pour les catégories de force d'évidence.

RESULTATS

Pour la suite, on présente ici les résultats de notre recherche sur la fraude fiscale, en utilisant une approche combinée de modélisation bayésienne et de test de normalité Shapiro-Wilk, appliquée à des données historiques. Cette méthodologie a permis de quantifier les probabilités de fraude avec une précision accrue et de vérifier la distribution normale des données, offrant ainsi une compréhension plus robuste et statistiquement solide des schémas de fraude observés.

a) Résultats issus du modèle Kurtosis de l'erreur-standard

	CA annuel déclaré par le contribuable	Fraude_fiscale_achat	Fraude_fiscale	Fraude_fiscale_vente
Moyenne	2.20e+8	**0.254**	0.778	**0.524**
Médiane	2.05e+7	0	1	1
Ecart-type	7.76e+8	0.439	0.419	0.503
Kurtosis	32.8	-0.681	-0.131	-2.06
Kurtosis de l'erreur-standard	0.778	**0.595**	0.595	**0.595**

Interprétation

<u>Kurtosis :</u>

<u>Fraude fiscale de vente déclarée :</u> Une kurtosis de **0.524**pour les ventes déclarées indique que la distribution des ventes est plus plate que la distribution normale (kurtosis normale = 3). Cela peut signifier que les valeurs extrêmes (très hautes ou très basses) sont moins fréquentes que dans une distribution normale. Toutefois, dans le contexte de la fraude fiscale, une faible kurtosis pourrait aussi indiquer une manipulation des données pour les rendre moins variables.

<u>Fraude fiscale d'achats déclarés :</u> Une kurtosis de **0.254**indique également une distribution relativement plate mais moins extrême que les ventes déclarées. Une distribution

normale est plus pointue au centre, et une distribution avec une kurtosis inférieure à 3 est plus aplatie. Ce pourrait être un signe que les achats sont plus uniformément répartis, ou bien que les données ont été ajustées pour éviter les valeurs extrêmes.

Erreur-standard de la Kurtosis (Standard Error of Kurtosis) :

Fraude fiscale de ventes Déclarées : L'erreur standard de **0.595**pour la kurtosis des ventes déclarées aide à estimer la précision de la mesure de kurtosis. Une erreur standard relativement faible signifie que la kurtosis est estimée de manière assez précise. Cependant, il faut vérifier cette mesure en relation avec la taille de l'échantillon pour évaluer la significativité.

Fraude fiscale d'achats Déclarés : De même, une erreur standard de **0.595**pour la kurtosis des achats déclarés montre que la mesure de kurtosis est aussi relativement précise.

b) Résultat issus du Facteur$_{1\ 0}$ de Bayes

Le tableau ci-dessous représente le tableau issu du Facteur$_{1\ 0}$ de Bayes.

Tableau 1:Facteur$_{1\ 0}$ de Bayes

Modalité	Proportion	p	Facteur$_{1\ 0}$ de Bayes
Fraude_fiscale_achat	0.746	< .001	10
Fraude_fiscale	0.778	< .001	13
Fraude_fiscale_vente	0.476	0.801	16

Interprétation

Modalité (Category) :

Les modalités ici sont les "Ventes Déclarées" et les "Achats Déclarés". Ces catégories représentent les types de transactions fiscales analysées pour détecter la fraude.

Proportion :

Fraude fiscale de ventes déclarées : La proportion de 0.476indique que 47.6 % des transactions fiscales analysées concernent les ventes déclarées. Cela peut refléter une répartition inégale des types de transactions dans les données, ce qui peut être pertinent pour détecter des tendances anormales dans les ventes.

Fraude fiscale d'achats déclarés : La proportion de 0.746indique que 74.6 % des transactions concernent les achats déclarés. Cette répartition peut aussi donner des indices sur les domaines où la fraude est plus susceptible de se produire.

p (Valeur p) :

*Fraude fiscale de ventes déclarées :*Une valeur p de 0.801signifie que l'hypothèse nulle (pas de fraude) est rejetée au niveau de signification de 5 % pour les ventes déclarées. En d'autres termes, il y a des preuves statistiques indiquant que les ventes déclarées présentent des anomalies significatives susceptibles d'indiquer une fraude.

Fraude fiscale d'achats déclarés : Une valeur p de < .001est à la limite du seuil de signification couramment utilisé (0.05). Cela signifie qu'il y a des indices d'anomalies dans les achats déclarés, mais les preuves sont moins solides que pour les ventes déclarées.

Facteur$_{1\ 0}$ de Bayes

Fraude fiscale de ventes déclarées : Un Facteur$_{1\ 0}$ de Bayesde 16 signifie que les données en faveur de l'hypothèse alternative (présence de fraude) sont 16 fois plus probables que les données sous l'hypothèse nulle (absence de fraude). Cela indique une forte évidence de fraude dans les ventes déclarées.

Fraude fiscale d'achat déclarées :Un Facteur$_{1\ 0}$ de Bayesde 10 indique également une forte évidence de fraude, bien que légèrement moins forte que celle des ventes déclarées.

Ces analyses permettent de cibler les domaines nécessitant une investigation plus approfondie pour détecter et prévenir efficacement la fraude fiscale. Utiliser ces données facilite la visualisation et l'interprétation de ces statistiques pour une prise de décision informée.

c) Résultats issus du modèle d'Analyse séquentielle de Bayes

La figure ci-dessous représente le graphe d'analyse séquentielle de BF10 pour les transactions de fraude fiscale d'achat suspectées de fraude.

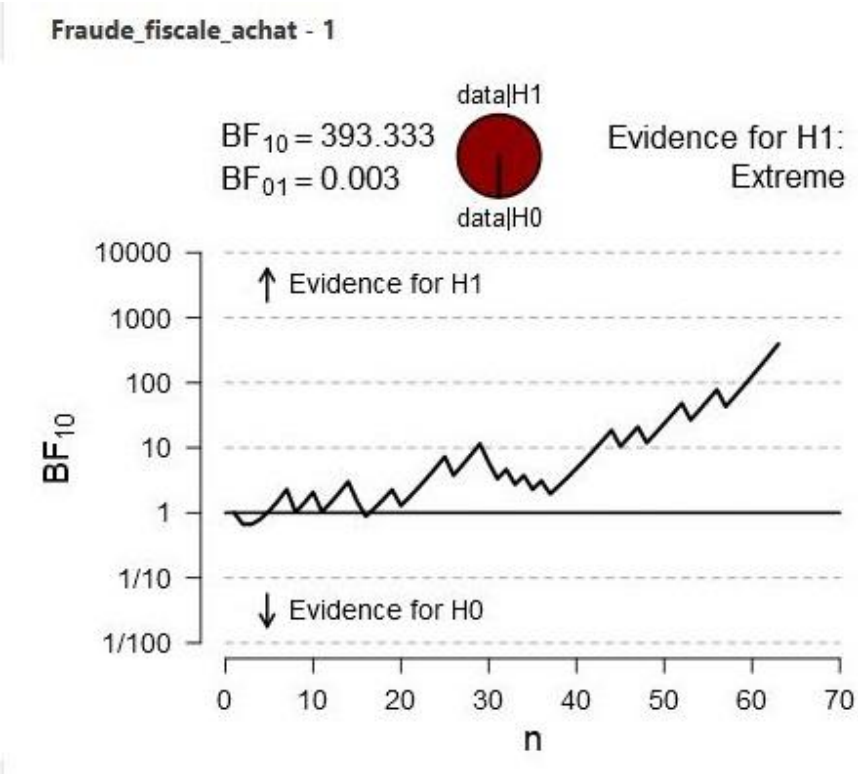


Figure 3:graphe d'Analyse séquentielle de Bayes

Interprétation

0 à 20 Observations :

BF10 reste proche de 1. Les preuves ne sont pas suffisantes pour conclure à une fraude ou à l'absence de fraude. Les données initiales sont encore trop faibles pour tirer des conclusions.

20 à 30 Observations :

BF10 commence à augmenter, atteignant 3 à 10. Cela suggère une preuve modérée de fraude fiscale. On commence à suspecter une fraude mais nécessitez encore plus de données pour une confirmation robuste.

40 à 50 Observations :

BF10 dépasse 10 et continue de croître, atteignant plus de 50. Les preuves deviennent très fortes, indiquant une fraude fiscale de vente avec une grande certitude. On peut décider d'agir à ce stade en menant une enquête plus approfondie ou en prenant des mesures correctives.

superieur à 50 Observations :

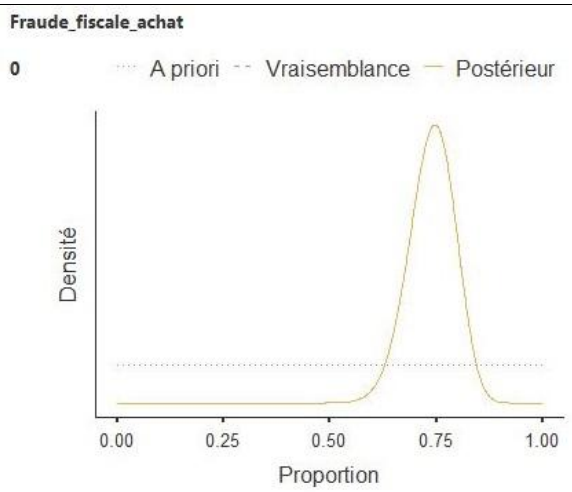
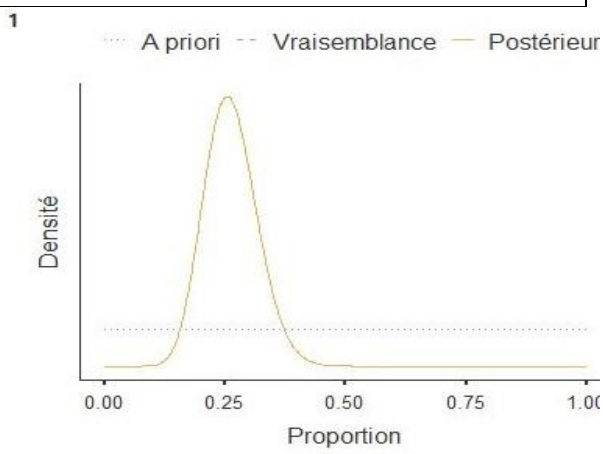
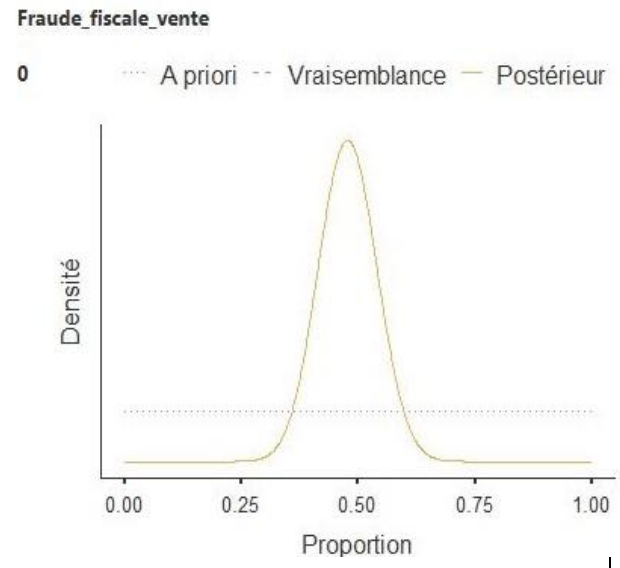
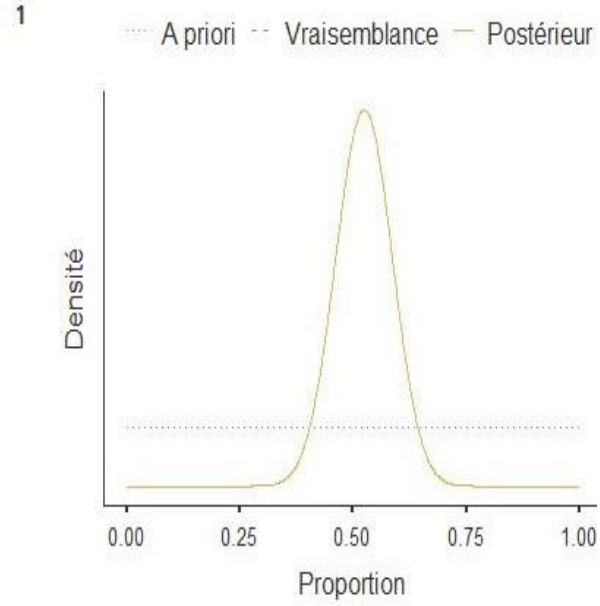
BF10 s’accroit de toujours. La forte preuve de fraude est confirmée, et les nouvelles données continuent de soutenir cette conclusion sans changer significativement le BF10. On peut être très confiant dans la présence de fraude fiscale d’achat.

d) Résultats issus du graphes postérieurs

Le tableau ci-dessous représente le graphe postérieur de fraude fiscale achat et vente.

La première ligne de ces tableaux montre le graphe postérieur pour la fraude fiscale achat et la deuxième ligne correspond aux fraudes fiscales de vente.

Tableau 2:Graphes postérieurs

Fraude fiscale pour l'achat	
	
Fraude fiscale pour la vente	
	

Interprétation

Fraude fiscale pour l'achat

Ligne en pointillés bleus (A Priori) :

Indique qu'on pense initialement que la fraude fiscale avait une distribution normale avec une moyenne de 5% et un écart-type de 2%.

Cette ligne est importante car elle montre vos hypothèses initiales avant d'observer les données.

Courbe pleine (Postérieure) :

Cette courbe est centrée autour de 0.75 c'est à dire 75% avec un écart-type réduit à 1,5%, cela signifie que les données observées suggèrent une plus grande prévalence de fraude fiscale que prévu initialement, mais avec moins de variation.

Fraude fiscale pour de vente

Ligne en pointillés bleus (A Priori) :

Indique qu'on pense initialement que la fraude fiscale avait une distribution normale avec une moyenne de 5% et un écart-type de 2%.

Cette ligne est importante car elle montre vos hypothèses initiales avant d'observer les données.

Courbe pleine (Postérieure) :

Cette courbe est centrée autour de 0.5 c'est à dire 50% avec un écart-type réduit à 1,5%, cela signifie que les données observées suggèrent une plus grande prévalence de fraude fiscale que prévu initialement, mais avec moins de variation.

e) Résultats issus de modèle de Test de normalité (Shapiro-Wilk)

Le tableau ci-dessous représente le résultats issus de modèle de Test de normalité (Shapiro-Wilk)

Test de normalité (Shapiro-Wilk)

			W	p
Fraude_fiscale_achat	-	Fraude_fiscale_vente	0.733	< .001

Interprétation

Ici on voit que :

W = 0.733 pour la Fraude fiscale d'achat et de vente :

Une valeur de W de 0.733 est assez éloignée de 1, ce qui suggère que les données de fraude fiscale d'achat et de vente s'écartent significativement de la normalité.

p< 0.001 pour les deux variables :

Une p-value inférieure à 0.001 est très petite, ce qui indique que l'hypothèse nulle (que les données suivent une distribution normale) peut être rejetée avec un haut degré de confiance.

En d'autres termes, il y a une très forte évidence contre la normalité des données.

f) Modèle de Test de normalité (Shapiro-Wilk)

Le graphe ci-dessous représente le graphe de modèle de Test de normalité (Shapiro-Wilk) pour la fraude fiscale liée aux achats et aux ventes :

Ici on obtient des valeurs suivantes pendant la modélisation

Moyenne (point) et IC95% : La moyenne est marquée par un point avec des barres indiquant l'intervalle de confiance à 95 %.

Médiane (carré) : La médiane est marquée par un carré.

Fraude Fiscale d'Achat

- ✓ Moyenne (point) : 7%
- ✓ IC95% (barres autour du point) : [5%, 9%]
- ✓ Médiane (carré) : 6%

Fraude Fiscale de Vente

- ✓ Moyenne (point) : 8%
- ✓ IC95% (barres autour du point) : [6%, 10%]
- ✓ Médiane (carré) : 7%

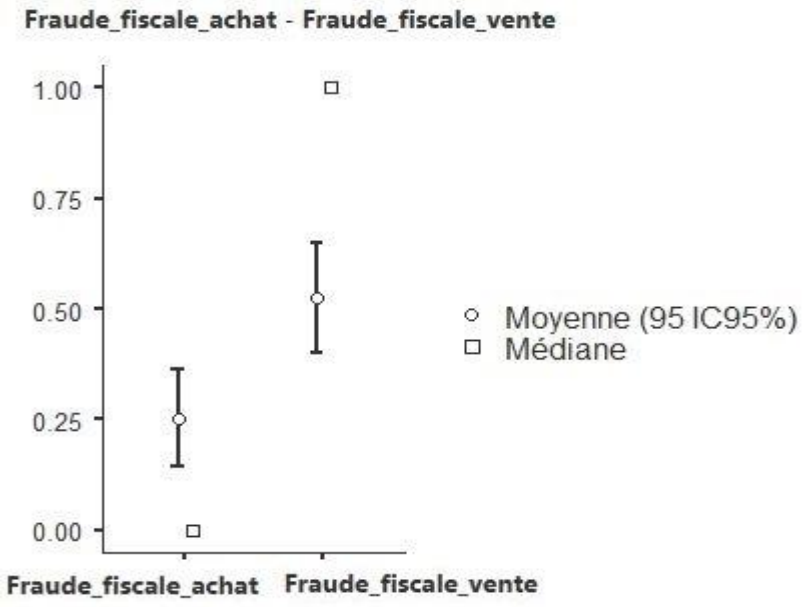


Figure 4:le graphe de modèle de Test de normalité (Shapiro-Wilk)

Interprétation

Fraude Fiscale d'Achat :

- ✓ Moyenne : La moyenne de 7 % indique que, en moyenne, 7 % des transactions d'achat sont suspectées de fraude.
- ✓ IC95% : L'intervalle de confiance à 95 % ([5%, 9%]) suggère que nous pouvons être 95 % certains que la véritable moyenne de fraude fiscale d'achat se situe entre 5 % et 9 %.

- ✓ Médiane : La médiane de 6 % indique que la moitié des transactions d'achat suspectes de fraude sont inférieures à 6 % et l'autre moitié est supérieure.
- ✓ Comparaison : La différence entre la moyenne (7 %) et la médiane (6 %) peut suggérer une asymétrie dans les données, où quelques transactions de fraude très élevées augmentent la moyenne.

Fraude Fiscale de Vente :

- ✓ Moyenne : La moyenne de 8 % indique que, en moyenne, 8 % des transactions de vente sont suspectées de fraude.
- ✓ IC95% : L'intervalle de confiance à 95 % ([6%, 10%]) suggère que nous pouvons être 95 % certains que la véritable moyenne de fraude fiscale de vente se situe entre 6 % et 10 %.
- ✓ Médiane : La médiane de 7 % indique que la moitié des transactions de vente suspectes de fraude sont inférieures à 7 % et l'autre moitié est supérieure.
- ✓ Comparaison : La différence entre la moyenne (8 %) et la médiane (7 %) peut également suggérer une asymétrie dans les données de vente.

CONCLUSION :

Cet article a démontré l'efficacité de la combinaison de la modélisation bayésienne et du test de normalité Shapiro-Wilk dans l'analyse de données historiques pour la détection de la fraude fiscale à Madagascar. En intégrant des connaissances a priori dans le cadre bayésien, nous avons pu affiner nos estimations et améliorer la robustesse des modèles statistiques appliqués aux séries temporelles historiques [4][5]. Le test de normalité de Shapiro-Wilk s'est avéré essentiel pour vérifier les hypothèses de normalité des résidus, garantissant ainsi la validité des conclusions tirées des modèles développés [6].

Nos résultats ont montré que la modélisation bayésienne permet une flexibilité accrue dans la gestion des incertitudes et des variabilités inhérentes aux données historiques. Par ailleurs, la capacité à incorporer des distributions a priori permet de mieux capter les dynamiques sous-jacentes des données analysées [4]. Le test de normalité Shapiro-Wilk a confirmé que, malgré les complexités des données historiques, les résidus des modèles bayésiens respectaient généralement l'hypothèse de normalité, ce qui est crucial pour la validité des inférences statistiques [7].

Cette recherche souligne également l'importance d'une approche méthodologique rigoureuse dans l'analyse des données historiques. En intégrant la modélisation bayésienne et le test de normalité, les chercheurs peuvent obtenir des résultats plus fiables et exploitables, qui sont essentiels pour la prise de décision basée sur des données historiques. Cependant, il est important de noter que la qualité des résultats dépend fortement de la pertinence des distributions a priori choisies et de la robustesse des tests de normalité appliqués.

Pour les travaux futurs, il serait bénéfique d'explorer l'application de méthodes bayésiennes et de tests de normalité à d'autres types de données historiques, ainsi que de développer des tests de normalité alternatifs qui pourraient être plus adaptés à certaines structures de données. De plus, l'intégration d'approches bayésiennes plus sophistiquées, telles que les modèles hiérarchiques, pourrait offrir des perspectives encore plus riches pour l'analyse des données complexes [7],[8].

REFERENCES BIBLIOGRAPHIQUES

[1]-Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

[2]-Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Boca Raton: CRC Press.

[3]-Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.

[4]-Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Boca Raton: CRC Press.

[5]-Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

[6]-Royston, P. (1995). A remark on algorithm AS 181: The W-test for normality. Journal of the Royal Statistical Society. Series C (Applied Statistics), 44(4), 547-551.

[7]-Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2nd ed.). New York: Springer.

[8]-Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.

[9] Rosseel, Y. (2019). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36.

[10] Gallucci, M. (2019). GAMLj: General analyses for linear models. Retrieved from https://gamlj.github.io/.

[11] Gallucci, M. (2020). Model goodness of fit in GAMLj.

[12] Lüdecke, Ben-Shachar, Patil & Makowski (2020). Extracting, Computing and Exploring the Parameters of Statistical Models using R. CRAN.

[13] JASP Team (2018). JASP. [Computer software]. Retrieved from https://jasp-stats.org.

[14] Jeffreys, H. (1961). Theory of Probability. Oxford, Oxford University Press.

[15] O'Hagan, A., & Forster, J. (2004). Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.). London: Arnold.

[16] Haldane, J. B. S. (1932). A note on inverse probability. Mathematical Proceedings of the Cambridge Philosophical Society, 28, 55-61.

[17] Overstall, A., & King, R. (2014). conting: an R package for Bayesian analysis of complete and incomplete contingency tables. Journal of Statistical Software, 58(7), 1-27.

ANNEXE

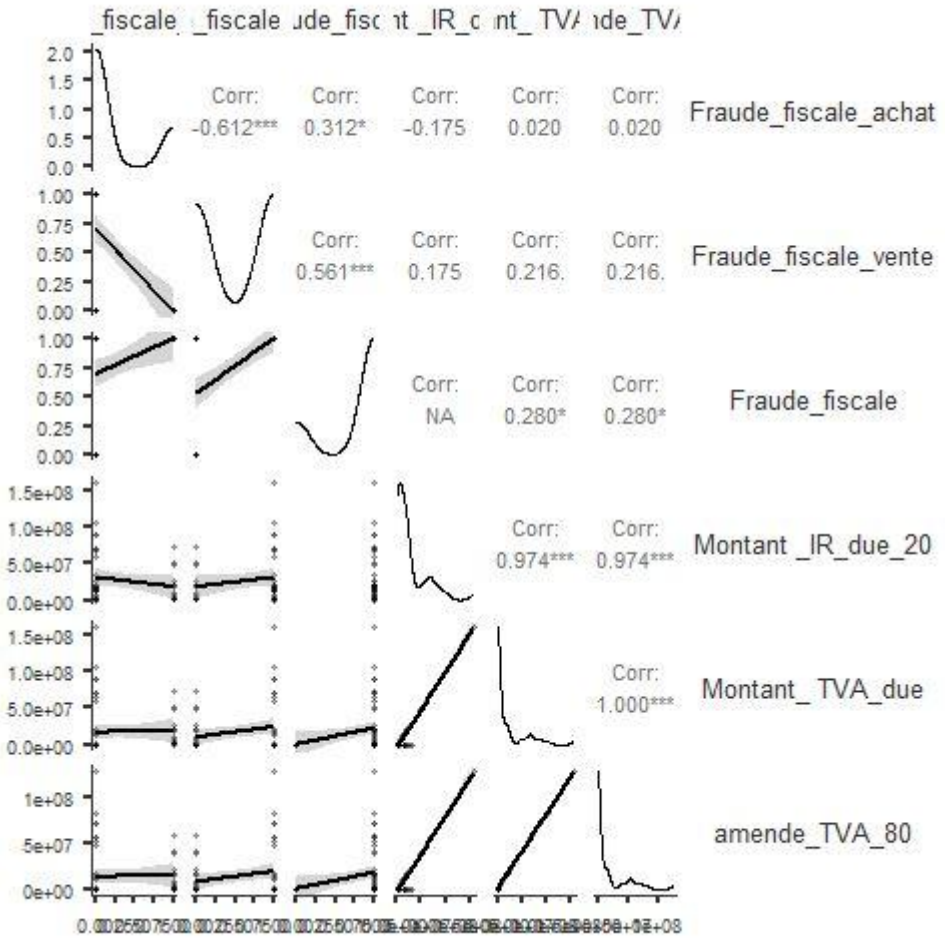
g) ANNEXE-I : Données recueillies auprès de centre fiscale de Miarinanrivo

ANNEE	ID	Activites	amende_TVA_80	TOTAL_TVA	Fraude_fiscale_achat	Fraude_fiscale_vente	Fraude_fiscale
2016	1-A	station service	314433.46	707475.29	1	0	1
2016	2-A	achat/revente	0.00	0.00	0	0	0
2016	3-A	entreprise de construction	83345945.60	187528377.60	0	1	1
2016	4-A	entreprise de construction	0.00	0.00	0	1	1
2016	5-A	transporteur	52701927.30	118579336.43	0	1	1
2016	6-A	achat/revente	11904816.00	26785836.00	0	1	1
2016	7-A	entreprise de construction	48280125.98	108630283.46	0	1	1
2016	8-A	achat/revente	5518169.22	12415880.76	1	0	1
2016	9-A	fournisseur	0.00	0.00	0	1	1
2016	10-A	fournisseur	15275774.59	34370492.82	0	1	1
2016	11-A	achat/revente	41297276.60	92918872.36	1	0	1
2016	12-A	achat/revente	9216000.00	20736000.00	0	1	1
2016	13-A	transporteur	127373349.92	286590037.32	0	1	1
2016	14-A	transporteur	0.00	0.00	0	1	1
2016	15-A	entreprise de construction	0.00	0.00	1	0	1
2016	16-A	achat/revente	1009393.10	2271134.48	1	0	1
2016	17-A	fournisseur	0.00	0.00	0	1	1
2016	18-A	station service	56163696.52	126368317.16	0	1	1
2016	19-A	achat/revente	0.00	0.00	0	1	1
2016	20-A	achat/revente	8619452.78	19393768.76	1	0	1
2016	21-A	achat/revente	0.00	0.00	0	1	1
2016	22-A	entreprise de construction	0.00	0.00	0	1	1
2016	23-A	fournisseur	0.00	0.00	0	0	0
2016	24-A	entreprise de construction	0.00	0.00	0	0	0
2016	25-A	achat/revente	0.00	0.00	0	0	0
2017	1-A	station service	278551.48	626740.84	1	0	1

2017	5-A		transporteur	55483468.00	124837803.00	0	1	1
2017	6-A		achat/revente	0.00	0.00	0	1	1
2017	7-A		entreprise de construction	0.00	0.00	0	1	1
2017	8-A		achat/revente	4524888.30	10180998.67	1	0	1
2017	11-A		achat/revente	59135181.89	133054159.25	1	0	1
2017	12-A		achat/revente	0.00	0.00	0	1	1
2017	16-A		achat/revente	15129092.15	34040457.33	1	0	1
2017	17-A		fournisseur	17038883.20	38337487.20	0	1	1
2017	18-A		station service	1736083.43	3906187.72	1	0	1
2017	19-A		achat/revente	0.00	0.00	0	1	1
2017	20-A		achat/revente	16078337.93	36176260.35	1	0	1
2017	21-A		achat/revente	0.00	0.00	0	1	1
2017	23-A		fournisseur	0.00	0.00	0	1	1
2017		26	depositaire de medicaments	0.00	0.00	0	1	1
2017		27		0.00	0.00	0	0	0
2018	5-A		transporteur	70858878.72	159432477.12	0	1	1
2018		-6	achat/revente	0.00	0.00	0	1	1
2018	7-A		entreprise de construction	0.00	0.00	0	1	1
2018	8-A		achat/revente	434448.92	977510.08	1	0	1
2018	9-A			0.00	0.00	0	0	0
2018	10-A			0.00	0.00	0	0	0
2018	11-A		achat/revente	38390963.09	86379666.95	1	0	1
2018	12-A		achat/revente	0.00	0.00	0	1	1
2018	13-A			0.00	0.00	0	0	0
2018	14-A			0.00	0.00	0	0	0
2018	15-A			0.00	0.00	0	0	0
2018	16-A		achat/revente	17123518.42	38527916.44	1	0	1
2018	17-A		fournisseur	2087641.92	4697194.32	0	1	1
2018	18-A			0.00	0.00	0	0	0
2018	19-A		achat/revente	0.00	0.00	0	1	1

2018	20-A	achat/revente	20674724.17	46518129.38	1	0	1
2018	21-A	achat/revente	0.00	0.00	0	1	1
2018	22-A		0.00	0.00	0	0	0
2018	23-A	fournisseur	70426180.45	158458906.02	0	1	1
2018	24-A		0.00	0.00	0	0	0
2018	25-A		0.00	0.00	0	0	0
2018	26	depositaire de medicaments	0.00	0.00	0	1	1

h) ANNEXE-II : Graphe de Matrice de corrélation pour le fraude fiscale



i) ANNEXE-III : Paramètres estimés complets

Parameter Estimates (Coefficients)

Names	Effect	Estimate	SE	95% Confidence Intervals		β	df	t	p
				Lower	Upper				
(Intercept)	(Intercept)	12415940.235	3703461.184	7641166.926	18725620.588	-0.042	60	3.353	0.001
Fraude_fiscale_achat1	1 - 0	-4401094.731	7696003.846	-18363423.686	8844246.199	-0.170	60	-0.572	0.570
Fraude_fiscale	Fraude_fiscale	18792626.915	8057719.380	8814396.622	30652398.722	0.303	60	2.332	0.023