International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 34 No. 1 August 2022, pp. 367-375

Détection des anomalies des données de démographie par région à Madagascar par la méthode iForest

Koto J.B.[1], Ramahefy T.R.[2], Randrianja S.[3]

École Doctorale Glocalisme, Environnement et Sécurités indienocéaniques (GENESIS) – Université d'Antsiranana
Antsiranana 00201 – Madagascar

Thématique : Intelligence artificielle

[1]koto87@yahoo.com - [2]ramahefy@yahoo.fr – [3]srandrianja@gmaim.com

Abstract – La détection des anomalies a connu dernièrement une grande attention, tant au niveau académique qu'au niveau industriel. La prédiction automatique pour la détection d'anomalie s'applique également aux données démographies à Madagascar. Une approche alternative qui se montre prometteuse, est l'utilisation de la méthode des forêts d'isolation (iForest).

Cet article nous démontre que la détection des anomalies peut être appliquer sur tous les différents domaines, comme sur les données de la démographie par rapport au nombre de bureau de l'état civil à Madagascar. Les résultats de l'algorithme iForest présentent des réponses qui sont difficiles à l'interpréter. Il est toujours difficile de savoir la contribution de chaque variable dans cet algorithme d'isolation et de comprendre pourquoi des observations particulières reçoivent un score plus élevé.

Keywords – Intelligence artificielle ; Anomalies ; iForest ; Démographie de Madagascar.

I. INTRODUCTION

La détection des anomalies a connu dernièrement une grande attention, tant au niveau académique qu'au niveau industriel. Cet intérêt a été d'autant plus important avec la disponibilité des bases de données volumineuses, qui a nettement augmenté la capacité des organismes à suspecter et/ou détecter des points anormaux. Par exemple, dans le secteur bancaire, il est probable que les anomalies au niveau des transactions par cartes de crédit soient des cas de fraude. Dans le domaine de l'assurance santé, une anomalie pourrait être également une fraude commise par un client ou un fournisseur de soins, qui a fait de fausses réclamations afin d'avoir un meilleur remboursement. Pour cet article, la prédiction automatique pour la détection d'anomalie s'applique aux données démographies à Madagascar. Une approche alternative qui se montre prometteuse, est l'utilisation de la méthode des forêts d'isolation (iForest).

II. L'ANOMALIE, LA METHODE NON SUPERVISEE ET L'IFOREST

Les anomalies sont les observations qui sont différentes et loin des points considérés normaux dans un jeu de données. Elles peuvent être générées par un mécanisme différent [1], comme elles peuvent être produites par une erreur de collecte de données [2]. Ce type de données est appelé donnée aberrante, anomalie, etc. Dans la littérature il existe trois catégories d'anomalies : les anomalies ponctuelles, les anomalies contextuelles et les anomalies collectives [3]. Les anomalies ponctuelles sont les points qui s'éloignent et se démarquent par rapport aux autres selon une métrique donnée (par exemple similarité, distance). Les anomalies contextuelles sont les observations qui peuvent être des points normaux dans un contexte, mais anormaux dans un autre. Par exemple une température de 35 degrés pendant l'été est normale alors que pendant l'hiver est une anomalie. Les anomalies collectives se présentent lorsque les valeurs d'un sous-ensemble de données sont très différentes par

Corresponding Author: Koto J.B

rapport aux valeurs du reste de l'ensemble de données. Une observation de ce groupe prise toute seule pourrait ne pas être considérée comme anomalie dans un contexte ponctuel ou contextuel, mais le groupe auquel appartient cette observation indique une anomalie [3]. Dans cet article, nous allons observer les anomalies ponctuelles sur le rapport du nombre de bureaux d'état-civil, rayon moyen d'action théorique et population moyenne desservie par région à Madagascar.

La détection d'anomalies est transversale à tous les domaines qui exploitent les données. Ainsi, elle a des nombreuses applications possibles. Les domaines d'application ayant leur spécificité en fonction des données générées ou exploitées, toutes les méthodes de la détection d'anomalies ne sont pas adaptées à tous les domaines d'application [4][5][6].

La méthode non supervisée présente un grand avantage puisqu'elle ne requière pas l'étiquetage des données. Ainsi, elle arrive à détecter les anomalies en isolant les observations qui s'avèrent inhabituelles par rapport aux autres. Ceci permet de détecter de nouveaux types d'anomalies, contrairement aux algorithmes d'apprentissage supervisé qui identifient seulement les anomalies qui sont conformes avec les données étiquetées et le modèle prédictif construit. Plusieurs algorithmes non supervisés de détection des anomalies utilisent des mesures de distance et similarité pour déterminer les observations qui s'éloignent des autres. Quel que soit le domaine de l'application de la détection des anomalies, la majorité des approches non supervisées cherchent tout d'abord à calculer le degré des écarts entre les observations en utilisant une mesure de distance entre les observations, ensuite à affecter un score à chaque observation tel que les observations qui s'éloignent des autres reçoivent les plus hauts scores. Par la suite, un seuil est fixé à partir duquel nous pouvons juger si une observation est une anomalie ou non : un point est une anomalie si son score est supérieur au seuil fixé [7]. Certaines méthodes non supervisées se basent sur la classification ("Clustring"). C'est la méthode utilisée dans cet article. Elles considèrent les anomalies comme les points qui s'éloignent du centre de gravité des classes ("clusters").

L'Isolation Forest ou iForest est une méthode basée sur les arbres de décision et les forêts aléatoires [8][9]. Elle utilise l'isolation d'observations à partir de la construction de plusieurs arbres aléatoires. Quand une forêt d'arbres aléatoires et indépendants produit collectivement un chemin d'accès court pour atteindre une observation depuis la racine, celle-ci a une forte probabilité d'être une anomalie. Le nombre d'arbres utilisés est donc un important paramètre pour iForest. Le seuil de la détection est aussi un paramètre clé, il est donné par le score calculé pour chaque observation relativement aux autres observations. Si ce score est proche de 1 alors l'observation est considérée comme anomalie. Considérons un jeu de données de n observations et une observation x, le score s(x, n) d'aberrance de x est calculé par la formule suivante :

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

avec h(x) la longueur du chemin entre la racine de l'arbre et x. $E(h(x))$ est la moyenne des h(x) de toute la forêt d'arbres. c(n) représente la longueur du chemin moyen d'une observation depuis la racine dans le jeu de données de *n* observations. Isolation Forest fait partie des méthodes de détection d'anomalies les plus récentes et les plus utilisées. Elle donne également de bons résultats pour les jeux de données de grandes dimensions. L'approche de forêt d'isolation est une méthode non supervisée [8]. Elle se basent sur le principe d'isolation sans avoir besoin de faire appel à aucune mesure de similarité ou de distance entre les instances.

III. DESCRIPTION DE DONNEES ET L'ANALYSE IFOREST

Madagascar compte 1700 bureaux d'état-civil répartis entre 1695 Communes de Madagascar [10]. Le tableau ci-après présente la cartographie des bureaux d'état-civil, leur rayon moyen d'action théorique (RMAT) et la population moyenne desservie.

Tableau 1. Nombre de bureaux d'état-civil, rayon moyen d'action théorique et population moyenne desservie par région

Région	Population	Superficie (km2)	Nombre de bureau d'EC	Densité (hab/km2)	RMAT (km)	Population moyenne desservie
ALAOTRA MANGORO	1 013 560	28 061	79	36	10,6	12 830
AMORON'I MANIA	705 594	16 077	55	44	9,6	12 829
ANALANJIROFO	1 021 475	21 692	63	47	10,5	16 214
ANDROY	724 251	18 995	51	38	10,9	14 201
ANOSY	662 942	29 342	64	23	12,1	10 358
ANTSIRANANA	1 658 652	44 563	144	37	9,9	11 518
ATSIMO ANDREFANA	1 299 384	66 628	110	20	13,9	11 813
ATSIMO ATSINANANA	886 845	16 699	90	53	7,7	9 854
ATSINANANA	1 253 916	21 993	88	57	8,9	14 249
BETSIBOKA	289 649	28 366	35	10	16,1	8 276
BOENY	789 125	30 973	44	25	15,0	17 935
BONGOLAVA	451 334	18 211	26	25	14,9	17 359
DIANA	690 785	20 530	65	34	10,0	10 627
HAUTE MATSIATRA	1 183 363	21 208	88	56	8,8	13 447
IHOROMBE	308 187	25 909	26	12	17,8	11 853
ITASY	723 166	6 572	51	110	6,4	14 180
MELAKY	285 774	40 352	37	7	18,6	7 724
MENABE	584 301	48 302	51	12	17,4	11 457
SAVA	967 867	24 034	79	40	9,8	12 251
SOFIA	1 230 585	51 129	108	24	12,3	11 394
VAKINANKARATRA	1 779 516	17 907	91	99	7,9	19 555
VATOVAVY FITOVINANY	1 397 772	20 723	139	67	6,9	10 056

Source : « Rapport d'évaluation des systèmes d'enregistrement des faits d'état-civil et d'établissement des statistiques de l'état-civil », Ministère de l'intérieur et de la décentralisation, Comité National de Coordination en charge de l'amélioration des systèmes nationaux d'état-civil et d'établissement des statistiques de l'état-civil, Madagascar Novembre 2017, pages 36-37.

Nous allons essayer de trouver des informations aberrantes sur les données dans ce tableau. Pour se faire nous allons utiliser « *scikit-learn* ». Scikit-*learn*, encore appelé *sklearn*, est la bibliothèque la plus puissante et la plus robuste pour la *machine learning* en Python. Elle fournit une sélection d'outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression et le clustering via une interface cohérente en Python. Cette bibliothèque, qui est en grande partie écrite en Python, s'appuie sur *NumPy*, *SciPy* et *Matplotlib*.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest
```

```
data = pd.read_excel('bank.xlsx')
data
```

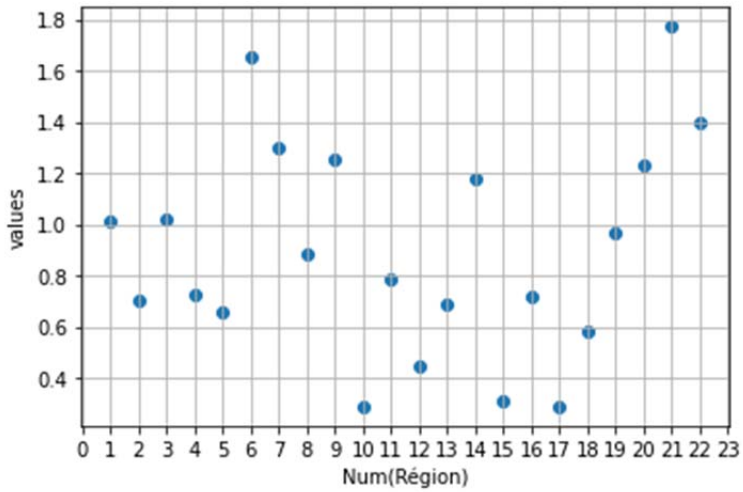
	Num	Region	Population	Superficie (km2)	Nombre de bureau d'EC	Densité (hab/km2)	RMAT (km)	Population moyenne desservie
0	1	ALAOTRA MANGORO	1013560	28061	79	36	10.6	12830
1	2	AMORON'I MANIA	705594	16077	55	44	9.6	12829
2	3	ANALANJIROFO	1021475	21692	63	47	10.5	16214
3	4	ANDROY	724251	18995	51	38	10.9	14201
4	5	ANOSY	662942	29342	64	23	12.1	10358
5	6	ANTSIRANANA	1658652	44563	144	37	9.9	11518
6	7	ATSIMO ANDREFANA	1299384	66628	110	20	13.9	11813
7	8	ATSIMO ATSINANANA	886845	16699	90	53	7.7	9854
8	9	ATSINANANA	1253916	21993	88	57	8.9	14249
9	10	BETSIBOKA	289649	28366	35	10	16.1	8276
10	11	BOENY	789125	30973	44	25	15.0	17935
11	12	BONGOLAVA	451334	18211	26	25	14.9	17359
12	13	DIANA	690785	20530	65	34	10.0	10627
13	14	HAUTE MATSIATRA	1183363	21208	88	56	8.8	13447
14	15	IHOROMBE	308187	25909	26	12	17.8	11853
15	16	ITASY	723166	6572	51	110	6.4	14180
16	17	MELAKY	285774	40352	37	7	18.6	7724
17	18	MENABE	584301	48302	51	12	17.4	11457
18	19	SAVA	967867	24034	79	40	9.8	12251
19	20	SOFIA	1230585	51129	108	24	12.3	11394
20	21	VAKINANKARATRA	1779516	17907	91	99	7.9	19555
21	22	VATOVAVY FITOVINANY	1397772	20723	139	67	6.9	10056

Les 22 lignes et les 8 colonnes sont en lecture. Pour avoir plus de charité sur la lecture et l'utilisation des données, nous enlevons la colonne « Region » qui n'aura pas d'utilité puis que nous avons la colonne « Num » qui est le numéro d'identification de la région, pour cet article.

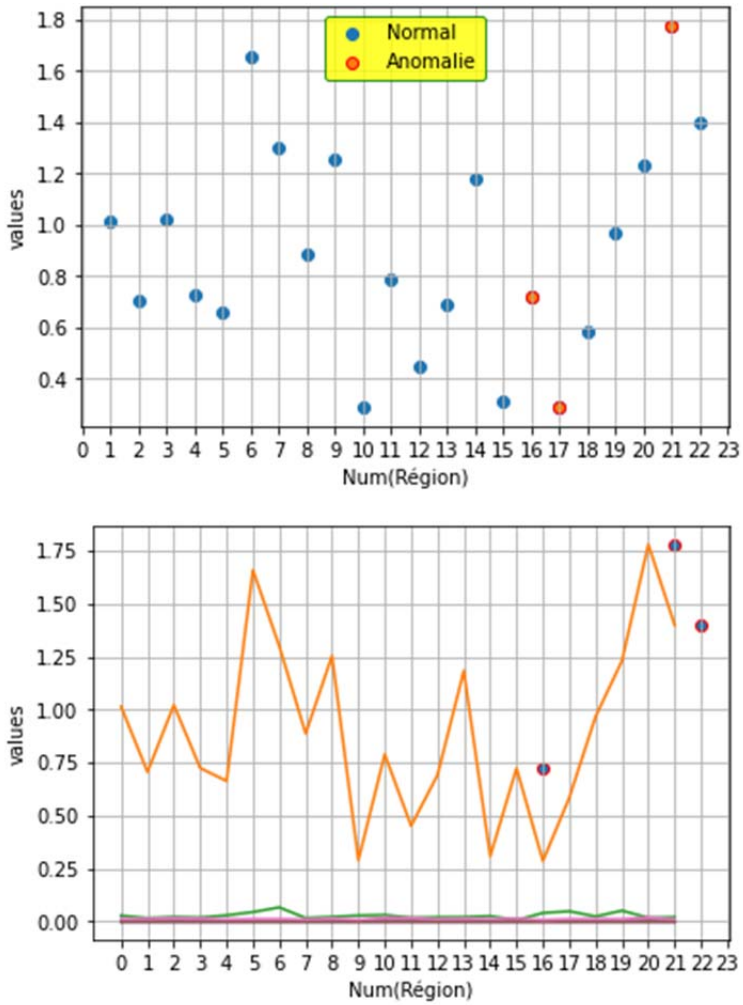
	Num	Population	Superficie (km2)	Nombre de bureau d'EC	Densité (hab/km2)	RMAT (km)	Population moyenne desservie
0	1	1013560	28061	79	36	10.6	12830
1	2	705594	16077	55	44	9.6	12829
2	3	1021475	21692	63	47	10.5	16214
3	4	724251	18995	51	38	10.9	14201
4	5	662942	29342	64	23	12.1	10358
5	6	1658652	44563	144	37	9.9	11518
6	7	1299384	66628	110	20	13.9	11813
7	8	886845	16699	90	53	7.7	9854
8	9	1253916	21993	88	57	8.9	14249
9	10	289649	28366	35	10	16.1	8276
10	11	789125	30973	44	25	15.0	17935
11	12	451334	18211	26	25	14.9	17359
12	13	690785	20530	65	34	10.0	10627
13	14	1183363	21208	88	56	8.8	13447
14	15	308187	25909	26	12	17.8	11853
15	16	723166	6572	51	110	6.4	14180
16	17	285774	40352	37	7	18.6	7724
17	18	584301	48302	51	12	17.4	11457
18	19	967867	24034	79	40	9.8	12251
19	20	1230585	51129	108	24	12.3	11394
20	21	1779516	17907	91	99	7.9	19555
21	22	1397772	20723	139	67	6.9	10056

Après avoir effectué la lecture de nos données, on crée des graphiques pour représenter visuellement les données de notre ensemble de données, ce qui nous aidera à découvrir d'autres informations cachées. Python dispose de nombreuses bibliothèques qui fournissent des fonctions permettant de réaliser des visualisations de données sur des ensembles de données. Nous pouvons utiliser le Pandas pour créer un nuage de points des caractéristiques ou des champs de notre ensemble de données les uns par rapport aux autres, et le *matplotlib* qui fournit une API orientée objet pour intégrer les graphiques dans les applications.

```
x = data.values
plt.scatter(x[:,0], x[:,1])
plt.show()
```

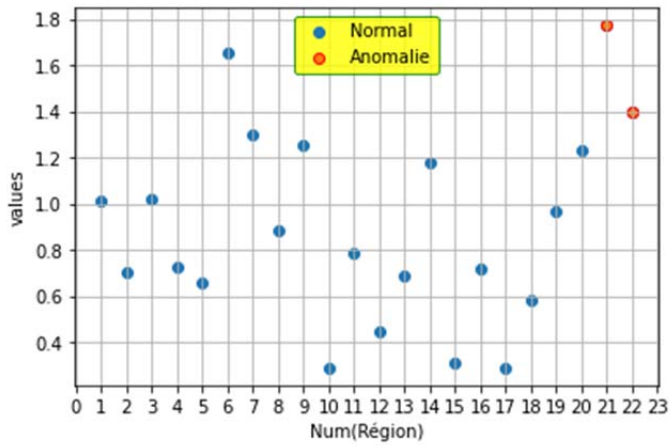


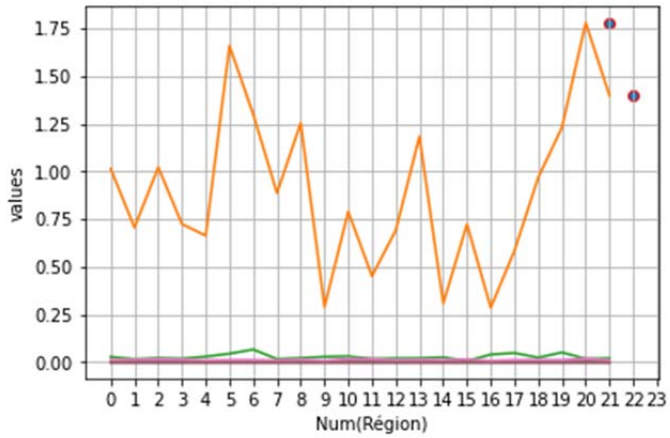
Maintenant, commençons à adapter cela à un modèle de forêt d'isolement avec un paramètre de contamination défini à 10 %, qui la part (suspectée) des valeurs aberrantes dans l'ensembles de données.

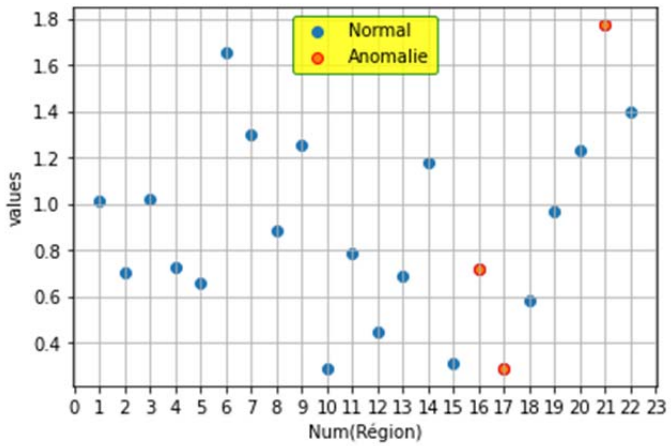


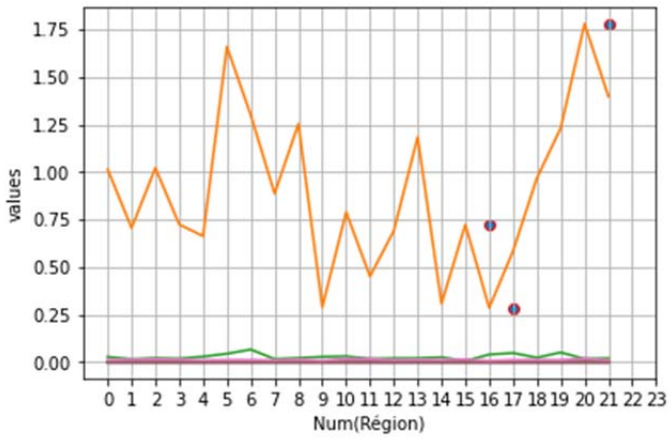
A partir de la visualisation du graphique, la région numéro 16, 21, et 22 qui sont Melaky, Menabe et Vatovavy fitovinany, ont des anomalies à 10% par rapport autres régions. En modifiant la valeur de contamination, nous avons des résultats suivants :

contamination = .05	**contamination = .1**

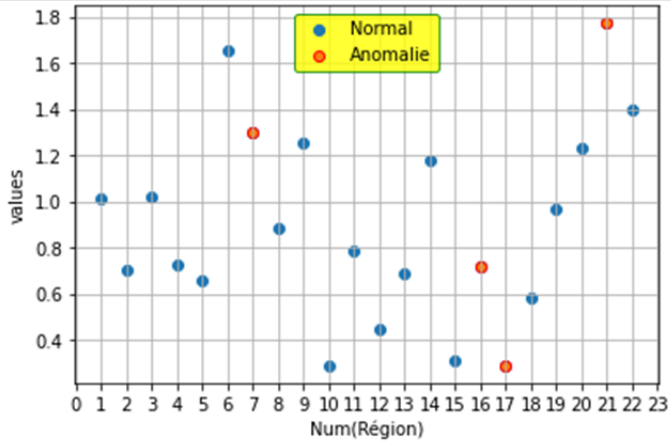


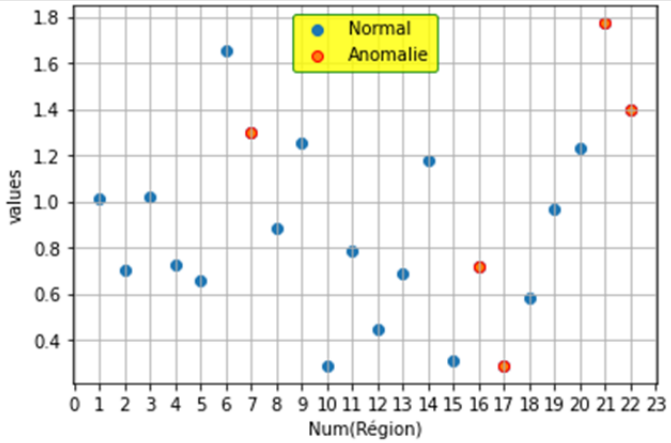


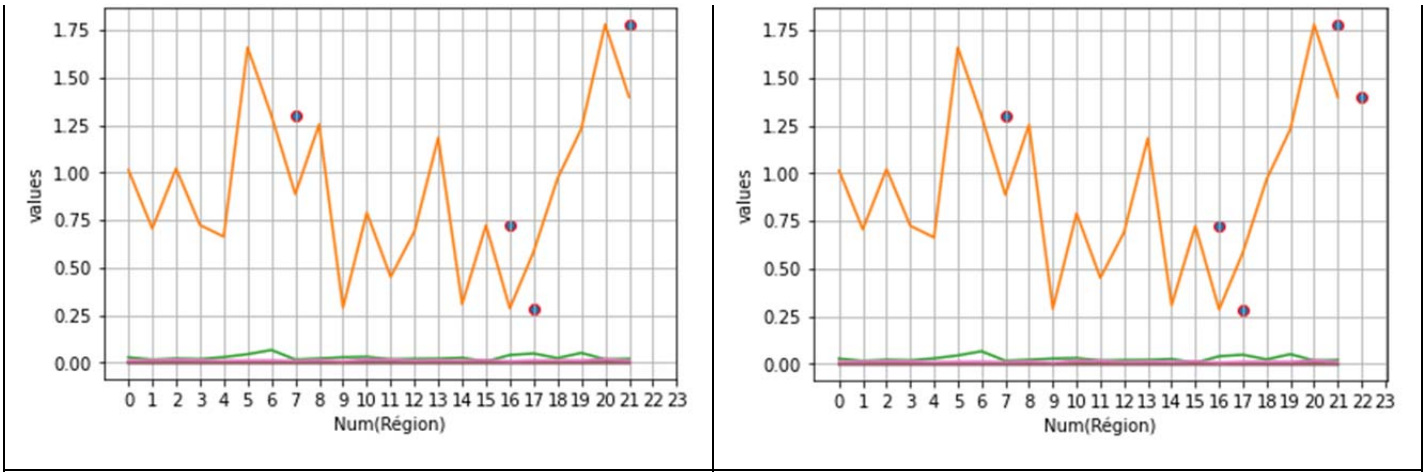




contamination = .15	contamination = .2







IV. CONCLUSION

Pour conclure, la détection d'anomalies est l'une des applications les plus intéressantes des algorithmes d'apprentissage non supervisés. Cet article nous démontre que la détection des anomalies peut être appliquer sur tous les différents domaines, comme sur les données de la démographie par rapport au nombre de bureau de l'état civil à Madagascar. Pour ce faire, nous avons appliqué la méthode iForest qui est très utilisée dans plusieurs travaux de détection des anomalies. L'analyse de la concordance des variables entre les données, nous a permis d'identifier les régions qui sont différents par rapport aux autres régions, concernant le rapport au nombre de bureaux d'état-civil, rayon moyen d'action théorique et population moyenne desservie par région. Le nombre des régions suspectés dépend de la valeur de la contamination. Plus on augmente la valeur de la contamination, le nombre des régions suspectés s'augmente aussi.

Bien que l'algorithme iForest soit efficace dans l'isolation des données aberrantes, il présente d'inconvénient : la difficulté à l'interprétation. Il est toujours difficile de savoir la contribution de chaque variable dans cet algorithme d'isolation et de comprendre pourquoi des observations particulières reçoivent un score élevé.

Dans un travail futur, nous allons appliquer l'intelligence artificiel, qui s'emble être l'outils très important à utiliser pour connaitre les données, pour analyser de données.

REFERENCES

[1] « Identification of Outliers », Irad Ben-Gal, 16 pages, 2005.

[2] « Controlling quality and amount of mitochondria by mitophagy: insights into the role of ubiquitination and deubiquitination. » Biol Chem, 2016, 397(7):637-47. Tan T, Zimmermann M, Reichert AS

[3] « Body stiffness in orthogonal directions oppositely affects worm-like robot turning and straight-line locomotion », A Kandhari, Y Huang, K A Daltorio, H J Chiel and R D Quinn, 2018, 17 pages.

[4] « Anomaly Detection: A Survey », Varun Chandola, Arindam Banerjee, Vipin Kumar, September 2009, 72 pages.

[5] In Managing and Mining Sensor, Charu C Aggarwal and Tarek Abdelzaher Data, pages 237–297, 2017.

[6] « Alternative polyadenylation diversifiespost-transcriptional regulation by selective RNA–protein interactions », Ishaan Gupta, Sandra Clauder-Münster, Bernd Klaus, Aino I Järvelin, Raeka S Aiyar1, Vladimir Benes, Stefan Wilkening, Wolfgang Huber, Vicent Pelechano & Lars M Steinmetz. 2014, 11 pages.

[7] « Isolation and identification of fungal communities in compost and vermicompost », Antonella Anastasi, Giovanna Cristina Varese, Valeria Filipello Marchisio, Pages 33-44, 2017

[8] « A study of radiative properties of fractal soot aggregates using the superposition T-matrix method », Journal of Quantitative Spectroscopy & Radiative Transfer 109 (2008) 2656– 2663

[9] « Impact of declining Arctic sea ice on winter snowfall », Jiping Liu, Judith A. Curry, Huijun Wang, Mirong Song, and Radley M. Horton, 6 pages, 2012.

[10] « Rapport d'évaluation des systèmes d'enregistrement des faits d'état-civil et d'établissement des statistiques de l'état-civil », Ministère de l'intérieur et de la décentralisation, Comité National de Coordination en charge de l'amélioration des systèmes nationaux d'état-civil et d'établissement des statistiques de l'état-civil, Madagascar Novembre 2017, 83 pages.