European Journal of Computer Science and Information Technology
Vol.10, No.3, pp.46-58, 2022
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)

Classification de la population de quelques régions à Madagascar par une approche non supervisée à base de K-moyennes

Koto J.B[1] – Randrianja S.[2] – Ramahefy T.R[3]
École doctorale en Cadre de vie, Affirmation identitaire, Développement Durable et ETHIQUE (CADDETHIQUE) – Université de Toamasina
Toamasina 00501 – Madagascar

Citation : Koto J.B., Randrianja S, Ramahefy T.R. (2022) Classification de la population de quelques régions à Madagascar par une approche non supervisée à base de K-moyennes, *European Journal of Computer Science and Information Technology*, Vol.10, No.3, pp.46-58

RÉSUMÉ : *La méthode k-moyennes ou k-means est un moyen très utilisé pour l'analyse des données car elle peut s'appliquer à des différents secteurs d'activés ou d'études. L'algorithme k-means s'adapte aux divers changements des données. Cette méthode nécessite de définir le nombre de cluster (k) pour que la classification soit efficace. L'algorithme k-moyennes peut être exécuté dans tous les données numériques et à un grand nombre d'ensembles de données tel que la « Classification de la population de quelques régions à Madagascar par une approche non supervisée à base de K-moyennes ». Il s'agit de la classification non supervisée les régions le plus semblable selon le sexe qui prend en considération les résidents ruraux et les résidents urbains d'après les données de recueillir RGPH 2018, INSTAT Madagascar. K-means génère des descriptions de cluster sous une forme minimisée pour maximiser la compréhension des données. Les données le plus proche sont groupées et le nombre dans les groupes ne sont pas forcément égaux.*

MOT CLE : Analyse des données, non supervisée, clustering, k-means

INTRODUCTION

De nos jours, les données sont devenues l'un des atouts majeurs qui constituent la richesse des entreprises. Les informations présentes, mais noyées dans la grande masse de données, sont devenues pour ces entreprises un facteur de compétitivité et d'innovation. En général, ces données permettent à l'utilisateur de découvrir et d'expliquer certains phénomènes existants ou bien d'extrapoler des nouvelles connaissances à partir des informations présentes. Pour exploiter ces grandes masses de données, de nombreuses techniques d'apprentissage automatique ont été développées. Dans cet article, nous nous intéressons particulièrement aux techniques de k-moyennes (k-means)., « **Classification de la population de quelques régions à Madagascar par une approche non supervisée à base de k-moyennes** ».
Le « k-means » a été utilisé pour la première fois par James MacQueen en 1967 [1], bien que l'idée originale ait été proposée par Hugo Steinhaus en 1957 [2]. Les k-means sont notamment utilisées en apprentissage non supervisé où l'on divise des observations en k partitions [web01]. Plusieurs domaine utilise cette méthode, actuellement, tel que :

- Marketing : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats, e-commerce [web01].

- Environnement : identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- Assurance : identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- Planification de villes : identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique,
- Autres domaines ou secteurs des activités

Méthode et l'algorithme

Le clustering K-means est l'un des algorithmes de machine learning non supervisés les plus facile à comprendre et à utiliser [web03][web04]. Cet algorithme est l'un des plus répandus. En règle générale, les algorithmes non supervisés font des inférences à partir d'ensembles de données en utilisant uniquement des vecteurs d'entrée sans faire référence à des résultats connus ou étiquetés, appelé aussi algorithme du centre mobile [3]. Les K-means est une technique de classification par apprentissage automatique utilisée pour simplifier des ensembles de données volumineux en ensembles de données simples et plus petits [1].

La méthode K-means

La méthode k-means classe les objets en plusieurs groupes d'où les clusters [4]. Elle les objets au sein du même groupe sont aussi semblables que possible, tandis que les objets provenant de différents groupes sont aussi dissemblables que possible [web02]. Dans le clustering k-means, chaque cluster est représenté par son centre (centroïde) qui correspond à la moyenne des points attribués au cluster : étant donnés des points et un entier k, l'algorithme vise à diviser les points en k groupes, homogènes et compacts, appelés clusters [web03].

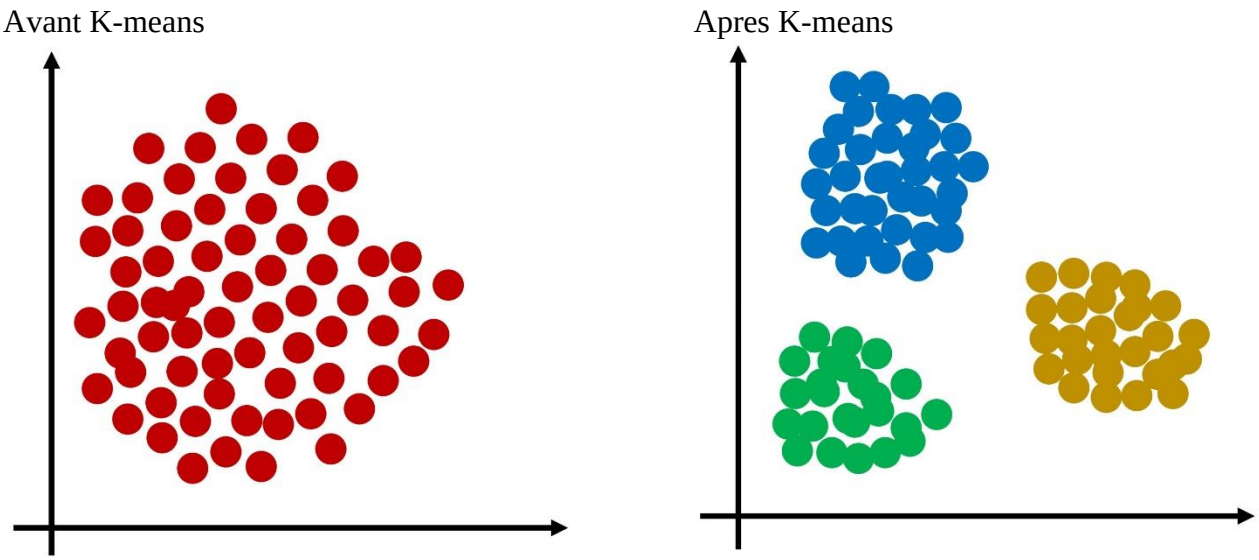


Figure 1. Classification des objets en k-means

La base du clustering k-means consiste à définir des clusters de sorte que la variation intra-cluster totale (connue sous le nom de variation intra-cluster totale) soit minimisée. Il existe plusieurs algorithmes de k-moyennes disponibles. L'algorithme standard est l'algorithme de Hartigan-Wong (Hartigan et Wong 1979) [5], qui définit la variation totale intra-cluster comme la somme des distances au carré.

L'algorithme de k-means

La première étape consiste à définir 3 centroïdes aléatoirement auxquels on associe 3 étiquettes par exemple 0,1,2. Ensuite nous allons pour chaque point regarder leur distance aux 3 centroïdes et nous associons le point au centroïde le plus proche et l'étiquette correspondante. Cela revient à étiqueter nos données. Enfin on recalcule 3 nouveaux centroïdes qui seront les centres de gravité de chaque nuage de points labellisés. On répète ces étapes jusqu'à ce que les nouveaux centroïdes ne bougent plus des précédents. Le résultat final se trouve sur la figure de droite.

Entrée

Ensemble de N données, noté par x
Nombre de groupes souhaité, noté par k

Sortie

Une partition de K groupes {C1, C2, ...C_k}

Début

(1) Choisir k individus au hasard (comme centre des classes initiales) : c_k
(2) Affecter chaque individu au centre le plus proche
$x_i \in c_k \; si \; \forall_j \; |x_i - \mu_k| = min \, |x_i - \mu_j|$
Avec μ_k le centre de la classe k ;
(3) Recalculer le centre de chacune de ces classes

$$\mu_k = \frac{1}{N} \sum_{x \in c_k} x_i$$

(4) Répéter l'étape (2) et (3) jusqu'à stabilité des centres
(5) Éditer la partition obtenue

Fin

La principale limite de cette méthode est la dépendance des résultats des valeurs de départ (centres initiaux). À chaque initialisation correspond une solution différente (optimum local) qui peut dans certain cas être très loin de la solution optimale (optimum global). Une solution naïve à ce problème consiste à lancer l'algorithme plusieurs fois avec différentes initialisations et retenir le meilleur regroupement trouvé. L'usage de cette solution reste limité du fait de son coût et que l'on peut trouver une meilleure partition en une seule exécution [7].

Les différentes versions de k-means

Global k-means

- Global k-means [6] est une solution au problème d'initialisation du k-means, elle est fondée sur les données et vise à atteindre une solution globalement optimale. Elle consiste à effectuer un clustering incrémental et à ajouter dynamiquement un nouveau centre suivi par l'application du k-means jusqu'à la convergence.

- Les centres sont choisis un par un de la façon suivante : le premier centre est le centre de gravité de l'ensemble des données (résultat de l'application du k-means avec k=1), les autres centres sont tirés de l'ensemble de données ou chaque donnée est une candidate pour devenir

un centre, cette dernière sera testée avec le reste de l'ensemble, le meilleur candidat est celui qui minimise la fonction objective de l'équation 1

$$J = \sum_{i=1}^{k} \sum_{x_j \epsilon} c_i \, \|x_j - \ c_i\|^2 \qquad (1)$$

L'algorithme suivant permet d'illustrer le principe :

Entrée

Ensemble de N données, notés par x ;
Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes {C1, C2, ...C_k}

Début

1) C 1 = Centre de gravité de l'ensemble des données ;
Répéter
2) Initialiser les centres i-1 par le résultat de l'étape précédente ;
3) Trouver l'i^{éme} centre :
Pour chaque donnée x faire
3.1) Considère x comme étant le i^{éme} centre ;
3.2) Affecter les données aux plus proches centres ;
3.3) Calculer l'erreur quadratique pour C_i=x ;
$J = \sum_{i=1}^{k} \sum_{x_j \epsilon} c_i \, \|x_j - \ c_i\|^2$

Fin faire

3.4) Garder le centre Ci = x qui minimise l'erreur quadratique ;
4) Appliquer le k-means jusqu'à la convergence ;
Jusqu'à obtenir une partition en k groupes ;

Initialisation par le mal classé

Principe :
L'absence d'un signe indiquant si l'optimum global est atteint ou pas fait penser à la possibilité d'améliorer les résultats.
Observant l'équation 02 :

$$x_i \epsilon \, c_k \; si \; \forall_j \; |x_i - \mu_k| = min \, |x_i - \mu_j| \qquad (2)$$

Un objet est affecté à un groupe s'il lui est le plus proche, plus la distance diminue plus la probabilité d'appartenance à ce groupe augmente, dans le cas contraire, l'objet le plus loin de son groupe d'appartenance est considéré comme étant mal classé, il fera certainement un bon candidat afin de former le nouveau centre.

Le global k-means est amorcé par un seul groupe ayant pour représentant le centre de gravité de l'ensemble des données, dans certain cas, cette partie de l'espace est vide ce qui permet de dégrader la classification, nous proposons d'amorcer l'initialisation du k-means avec deux groupes, les centres de ces groupes doivent assurer la séparabilité des données au cours de classification, il est évidant de choisir les deux données les plus éloignées [7].

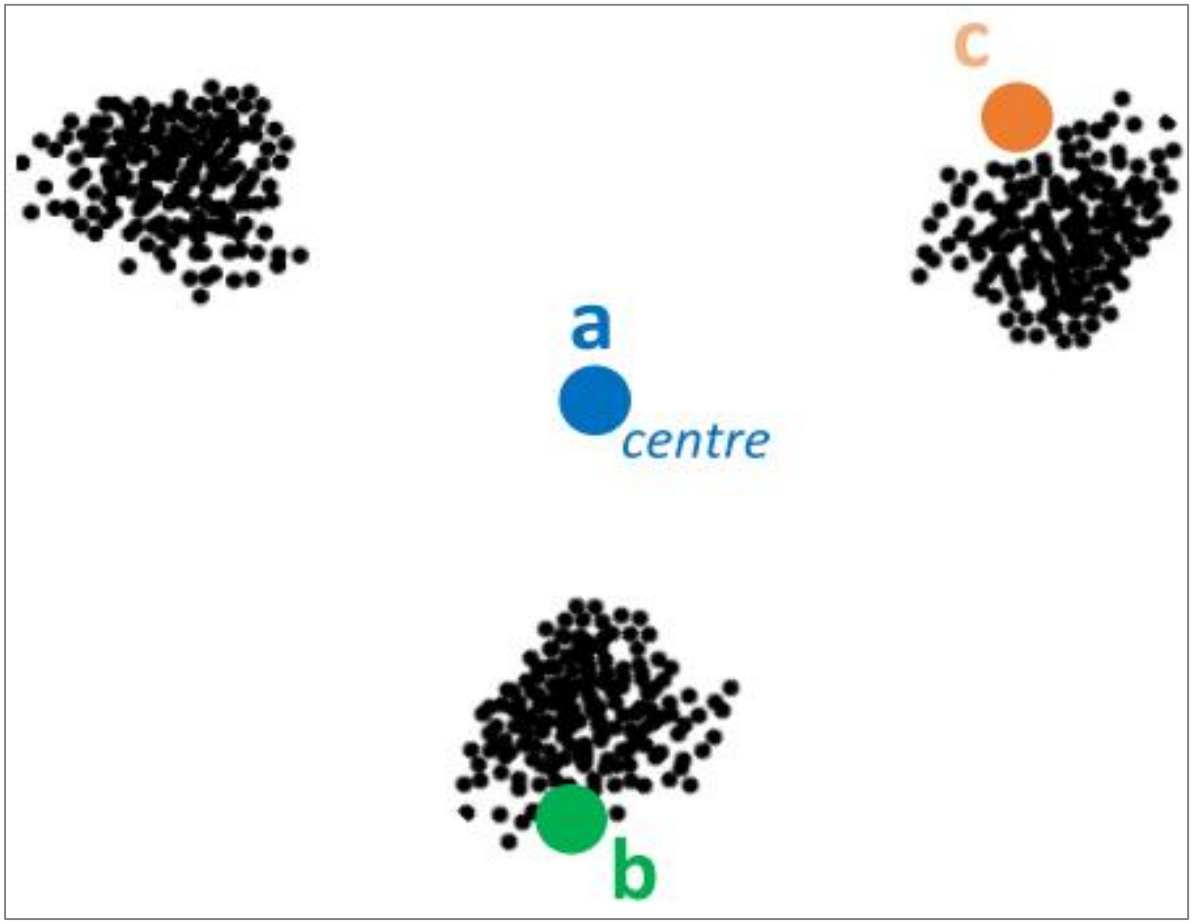


Figure 2. Classification par le principe d'initialisation par le mal classé

Classification par le principe d'initialisation par le mal classé
(a) Le centre des données en rouge,
(b) Le bleu et le vert représentent les deux objets les plus éloigné
L'algorithme se présente comme suit :

Début
1) Création d'une matrice de distance
2) Choisir les deux éléments les plus éloignés (ils représentent les deux premiers centres);

TANT QUE le nombre de classes souhaité n'est pas atteint ***Faire***
3) Affecter les individus aux noyaux disponibles ;
4) Sélectionner un élément mal classé (celui qui possède la plus grande distance de son centre le plus proche) ;
5) Ajouter cet individu à l'ensemble des noyaux ;
6) Augmenter le nombre des noyaux ;

Fin TANTQUE

Fin

L'approche incrémental (ou Modified Fast Global K-means)

Cette approche incrémentale de classification est similaire à celle du globale k-means, la différence entre elles réside dans les points suivants :

- le nombre de points initiaux, dans notre cas deux au lieu de un seul dans le global k-means.
- la recherche du nouveau centre ce limite à la recherche de l'élément le mal classé au lieu de testé toutes les données.

Son algorithme se présente :

Entrée

Ensemble de N données, notés par x ;
Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes {C1, C2, ...C_k}

Début

1) C1 = X1 ;
C2 = X2 ;

avec $d(x_1, x_2) = max_{i,j \in \{1,\dots,N\}} \left(d(x_i - x_j)\right) \quad , i \neq j$

Répéter

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;
3) Trouver l'i éme centre C_i : $C_i = x : x = max_{i \in [1,n]} \left(d_{k-1}^i\right)$
$avec\ d_{k-1}^i$ la distance entre $x_i\ et\ son\ plus\ proche\ centre\ parmi\ les\ k-1\ centre$
4) Appliquer le k-means jusqu'à la convergence ;
Jusqu'à obtenir une partition en k groupes ;

Fin

Grâce au faible cout de la stratégie de choix du nouveau centre, il est clair que l'approche proposée est plus rapide que le global k-means [web05].

3.4. Illustration et représentation du résultat

Tableau 1. Effectif de la population recensée dans les ménages ordinaires par région et sexe

Région	HU	FU	EU	HR	FR	ER	EH	EF	EHF
Analamanga	663 702	707 433	1 371 135	1 114 167	1 138 623	2 252 790	1 777 869	1 846 056	3 623 925
Vakinankaratra	152 748	160 233	312 981	888 162	878 516	1 766 678	1 040 910	1 038 749	2 079 659
Itasy	75 405	76 026	151 431	378 063	369 055	747 118	453 468	445 081	898 549
Bongolava	21 908	22 553	44 461	318 933	307 599	626 532	340 841	330 152	670 993
Haute Matsiatra	117 772	128 841	246 613	596 712	601 262	1 197 974	714 484	730 103	1 444 587
Amoron'IMania	52 194	55 525	107 719	360 298	369 099	729 397	412 492	424 624	837 116
Vatovavy Fitovinany	64 497	72 078	136 575	639 282	664 800	1 304 082	703 779	736 878	1 440 657
Ihorombe	18 998	20 558	39 556	188 879	188 877	377 756	207 877	209 435	417 312
Atsimo Atsinanana	34 195	39 018	73 213	468 195	488 996	957 191	502 390	528 014	1 030 404
Atsinanana	192 897	214 461	407 358	531 749	539 365	1 071 114	724 646	753 826	1 478 472
Analanjirofo	86 995	94 988	181 983	479 508	488 598	968 106	566 503	583 586	1 150 089
AlaotraMangoro	85 901	89 360	175 261	537 012	537 658	1 074 670	622 913	627 018	1 249 931
Boeny	159 484	173 612	333 096	296 555	299 661	596 216	456 039	473 273	929 312
Sofia	88 372	93 669	182 041	655 255	670 295	1 325 550	743 627	763 964	1 507 591
Betsiboka	24 774	26 125	50 899	171 266	171 113	342 379	196 040	197 238	393 278
Melaky	15 944	17 680	33 624	137 176	138 144	275 320	153 120	155 824	308 944
Atsimo Andrefana	121 593	133 400	254 993	760 265	782 636	1 542 901	881 858	916 036	1 797 894
Androy	41 639	44 678	86 317	389 949	423 969	813 918	431 588	468 647	900 235
Anosy	62 191	68 409	130 600	335 113	343 338	678 451	397 304	411 747	809 051
Menabe	54 440	57 778	112 218	289 096	291 149	580 245	343 536	348 927	692 463
Diana	140 747	161 491	302 238	292 230	295 494	587 724	432 977	456 985	889 962
Sava	100 070	108 520	208 590	454 614	460 568	915 182	554 684	569 088	1 123 772

Source : MDG, INSTAT - RGPH2018

HU : Homme Urbain FU : Femme Urbain EU : Ensemble Urbain

HR : Homme Rural FR : Femme Rural ER : Ensemble Rural

EH : Ensemble Homme EF : Ensemble Femme EHF : Ensemble Homme et Femme

L'objectif est de classifier les régions le plus semblable selon le sexe qui prend en considération les résidents ruraux et les résidents urbains.

Tableau 2. Statistiques descriptives

Variable	Minimum	Maximum	Moyenne	Ecart-type
Homme en Urbain (HU)	15944	663702	108021	133634
Femme en Urbain (FU)	17680	707433	116656	142834
Homme et Femme en Urbain (HFU)	33624	1371135	224677	276447
Homme en Rural (HR)	137176	1114167	467385	239339
Femme en Rural (FR)	138144	1138623	474946	244119
Homme et Femme en Rural (HFR)	275320	2252790	942332	483335
Ensemble Homme (EH)	153120	1777869	575407	348156
Ensemble Femme (EF)	155824	1846056	591602	360469
Ensemble Homme et Femme (EHF)	308944	3623925	1167009	708509

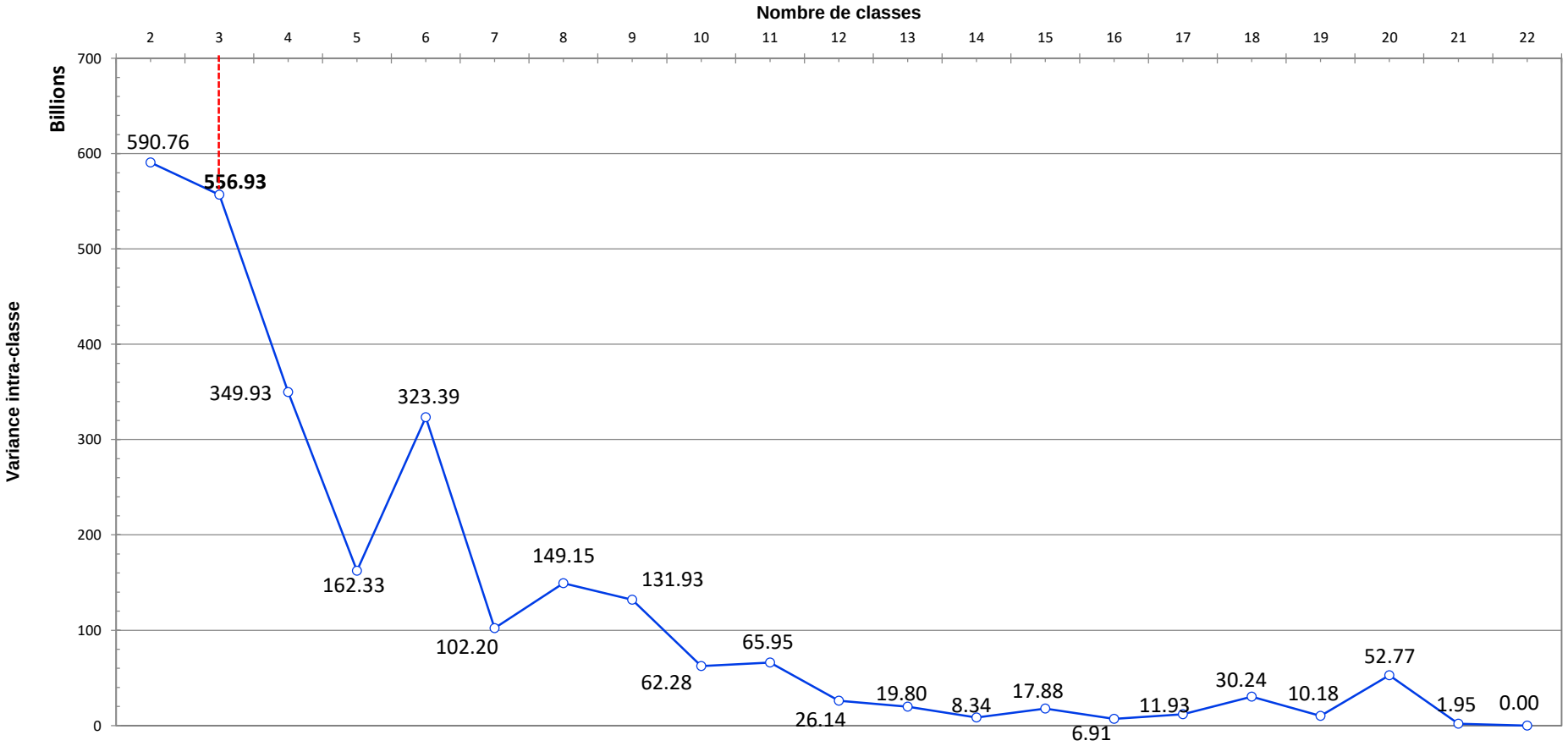
Tableau 3. Évolution des variances

Variance\Classes	2	3	4	5	6	7	8	9
Intra-classe	590755623891	556927015531	349932491240	162326421180	323390479215	102197941128	149153899775	131934323120
Inter-classes	627552594353	661381202713	868375727004	1055981797064	894917739029	1116110277116	1069154318468	1086373895124
Totale	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244

10	11	12	13	14	15	16	17	18
62284197515	65954202475	26143119810	19804967237	8340294921	17875947902	6911031887	11933791029	30236439938
1156024020729	1152354015769	1192165098434	1198503251007	1209967923323	1200432270342	1211397186357	1206374427215	1188071778306
1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244	1218308218244

18	19	20	21	22
30236439938	10178816627	52769693698	1951985420	0
1188071778306	1208129401617	1165538524546	1216356232824	1218308218244
1218308218244	1218308218244	1218308218244	1218308218244	1218308218244

Figure 3. Le courbe d'apprentissage



On désigne la partition K, le premier coude du courbe, ici le nombre de classe est 3

La classification k-means (Nombre de classes = 3)

Tableau 4. Statistiques pour chaque itération

Itération	Variance intra-classe	Trace(W)	ln(Déterminant(W))	Lambda de Wilks
0	1280541991545,610	24330297839366,60	-Inf	0,012
1	556927015530,543	10581613295080,30	-Inf	0,002

Tableau 5. Décomposition de la variance pour la classification optimale

	Absolu	Pourcentage
Intra-classe	556927015530,543	45,71%
Inter-classes	661381202713,405	54,29%
Totale	1218308218243,950	100,00%

Tableau 6. Distances entre les barycentres des classes

	1	2	3
1	0	1841099,098	1130321,532
2	1841099,098	0	716908,636
3	1130321,532	716908,636	0

Tableau 7. Objets centraux

Classe	HU	FU	EU	HR	FR	ER	EH	EF	EE
1 (Atsimo Andrefana)	121593	133400	254993	760265	782636	1542901	881858	916036	1797894
2 (Menabe)	54440	57778	112218	289096	291149	580245	343536	348927	692463
3 (Analanjirofo)	86995	94988	181983	479508	488598	968106	566503	583586	1150089

Tableau 8. Les distances entre les objets centraux

	1 (Atsimo Andrefana)	2 (Menabe)	3 (Analanjirofo)
1 (Atsimo Andrefana)	0	1803956,783	1064564,789
2 (Menabe)	1803956,783	0	739723,763
3 (Analanjirofo)	1064564,789	739723,763	0

Tableau 9. Résultats par classe

Classe	1	2	3
Objets	7	12	3
Somme des poids	7	12	3
Variance intra-classe	1464844149975,860	159859860722,288	17044963640,000
Distance minimale au barycentre	247873,586	60972,577	38320,016
Distance moyenne au barycentre	857338,206	334470,092	97154,928
Distance maximale au barycentre	2576181,311	668540,243	143607,154

Tableau 10. Résultat par objet

Observation	Classe	Distance au barycentre
Analamanga	1	2576181,311
Vakinankaratra	1	413279,154
Itasy	2	268166,477
Bongolava	2	123838,563
Haute Matsiatra	1	707424,926
Amoron'IMania	2	196669,809
Vatovavy Fitovinany	1	708903,229
Ihorombe	2	488649,027
Atsimo Atsinanana	2	565119,331
Atsinanana	1	740805,085
Analanjirofo	3	38320,016
AlaotraMangoro	3	143607,154
Boeny	2	354696,001
Sofia	1	606900,150
Betsiboka	2	534921,458
Melaky	2	668540,243
Atsimo Andrefana	1	247873,586
Androy	2	328934,109
Anosy	2	127721,916
Menabe	2	60972,577
Diana	2	295411,592
Sava	3	109537,614

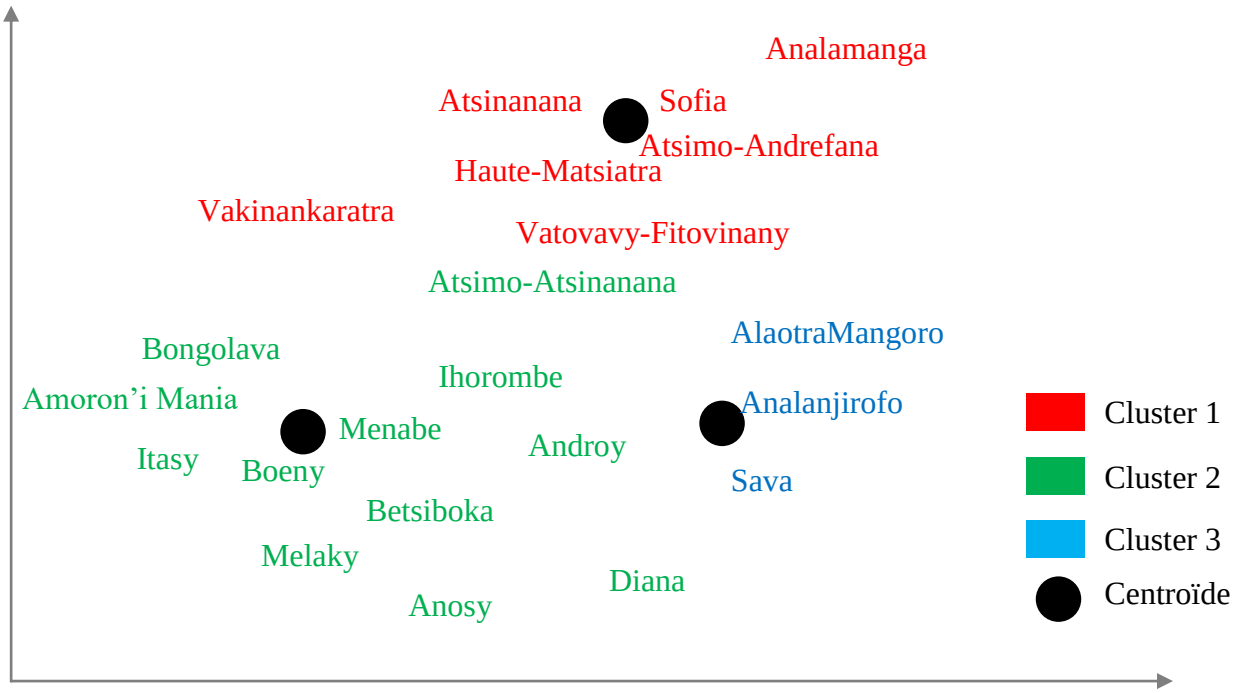


Figure 4. Représentation par classe des régions

CONCLUSION PERSPECTIVE

Il est facile d'implémenter k-means et d'identifier des groupes de données inconnus à partir d'ensembles de données complexes, et les résultats sont généralement présentés de manière rapide. L'algorithme k-means s'adapte aux divers changements des données.

Pour que la classification par k-moyennes soit efficace, le nombre de clusters (k) doit être défini au début de l'algorithme.

Il est difficile de prévoir les valeurs k ou le nombre de clusters et également, c'est difficile de comparer la qualité des clusters produites. Mais l'analyse par k-means améliore la précision de la classification et garantit que des informations sur un domaine de problème particulier sont disponibles. La modification de l'algorithme k-means basé sur ces informations améliore la précision des clusters.

L'algorithme k-moyennes peut être exécuté dans tous les données numériques et à un grand nombre d'ensembles de données et est calculé beaucoup plus rapidement que le plus petit.
Les résultats sont très faciles à interpréter. K-means génère des descriptions de cluster sous une forme minimisée pour maximiser la compréhension des données. Les données semblables sont groupées mais le nombre dans les groupes ne sont pas forcément égale. Une autre analyse pourra être effectué afin de détecter les événements rares ou plus généralement des observations qui sont aberrantes et différentes de la majorité des données. Ce phénome sera le sujet du prochain article, « Détection d'anomalies non supervisé du donné de l'état civil »

Bibliographie

[1] James MACQUEEN. « Some methods for classification and analysis of multivariate observations ». Dans Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297
[2] H. Steinhaus, « Sur la division des corps matériels en parties », *Bull. Acad. Polon. Sci.*, vol. 4, n°12, 1957, p.801–804
[3] Likas A., Vlassis M. & Verbeek J., ‘‘he global k-means clustering algorithm, Pattern Recognition’’, 36, p. 451-461.,2003
[4] Jacob Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, Cambridge, 2007.
[5] J. A. Hartigan et M. A. Wong, « Algorithm AS 136: A K-Means Clustering Algorithm », *Journal of the Royal Statistical Society, Series C*, vol. 28, n°1, 1979, p.100–108.
[6] Likas A., Vlassis M. & Verbeek J., ‘‘he global k-means clustering algorithm, Pattern Recognition’’, 36, pp. 451-461.,2003
[7] Z.Guellil et L.Zaoui ,‘‘Proposition d'une solution au problème d’initialisation cas du K-means’’, Université des sciences et de la technologie d’Oran MB .

Webographie

[web01] https://www.upgrad.com/blog/k-means-clustering-matlab/, visité le 12 octobre 2021
[web02] https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/, visité le 15 novembre 2021
[web03] https://datascientest.com/algorithme-des-k-means, visité le 15 novembre 2021
[web04] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1, visité le 17 novembre 2021
[web05] https://analyticsinsights.io/k-means/, visité le 17 novembre 202