Étude comparative entre les deux indices de validité WB et RST pour la détermination automatique du nombre de clusters : Application aux images satellitaires

H. MAHI & N. FARHI
Centre des Techniques Spatiales, Département Observation de la Terre
BP 13 Arzew, Oran
Email : hmahi@cts.asal.dz, nfarhi@cts.asal.dz

ملخص : الهدف من هذه الدراسة هو اقتراح مقارنة بين معيارين للجودة النسبية لنمذجة إحصائية ما، المعيارين متمثلان في معيار الجمع التربيعي(WB) و معيار RST. تم مقارنة هذين الأخيرين مع إستعمال الخوارزمية التصنيفية Bisecting K–Means (BKM). أجريت التجارب على مختلف أنواع المعطيات. النتائج المحصلة تثبت أن المعيارين الذي تم مقارنتهما لديهما نفس الوتيرة، مع ملاحظة الأفضلية للمعيار WB فيما يخص معالجة معطيات الإستشعار عن بعد.

الكلمات الأساسية : التصنيف، معيارين للجودة K–Means، Bisecting، النسبية،BW،TSR، معطيات الإستشعار عن بعد.

Résumé : Dans le cadre de ce travail, nous proposons de comparer deux indices de validité à savoir, l'indice des sommes carrées (WB) et l'indice RST. Les deux indices ont été comparés en les associant à l'algorithme Bisecting K-means. Les expérimentations ont été menées sur différents jeux de données. Les résultats obtenus montrent que les deux indices ont un comportement similaire avec un avantage pour l'indice WB lorsqu'il s'agit de traiter des données satellitaires.

Mot clés : Clustering, Indice de validité, WB, RST, Bisecting K-means, données satellitaires.

Abstract : In this work, we propose to compare two validityclustering indices, the WB index and the RST index. The two indices were be compared using the bisectingK-means algorithm; this later is a variant of the classical K-means. The experiments were conducted on different datasets. The obtained results show that both indices have the same behavior with slightly advantage to WB index when dealing with remotely sensed data.

Keywords : Clustering, Validity index, WB, RST, Bisecting K-means, remotely sensed data.

1. Introduction

La classification non supervisée, aussi appelée *clustering* est une technique qui tend à générer à partir d'un ensemble de données non labélisées, des groupes ou des clusters homogènes (*G. Gan et al 2007*). Généralement, un bon clustering est synonyme d'une faible inertie intra-clusters et une grande inertie inter-clusters (*D.T. Larose 2005*). De part sa facilité de mise en œuvre, le clustering est largement utilisé dans de nombreux domaines, tels que la fouille de données, la bio-informatique, la reconnaissance des formes et l'indexation des bases d'images (*Qi. Letaoet al 2013)(A. Hasnat 2014*). En observation de la Terre, le clustering a comme objectif de regrouper les pixels d'une image en clusters en se basant dans la plupart des cas sur leurs réponses spectrales. Chaque cluster correspond en réalité à un thème d'occupation ou d'utilisation des sols.

Parmi les algorithmes de clustering les plus utilisés en télédétection, on trouve l'algorithme K-Means (KM) (*J. McQueen 1967*) et l'algorithme ISODATA (*Iterative Self-Organizing Data Analysis Technique*) (*G. Ball and D. Hall 1965*).

Cependant, la principale limite lors de l'application de ces algorithmes réside dans le nombre de clusters k fixé préalablement. En effet, à chaque initialisation de ce paramètre peut correspondre une solution ou résultat différent. Pour surmonter cette limitation, nous proposons dans cet article, l'utilisation de deux indices de validité de clustering conjointement avec l'algorithme BKM (*M. Steinbach* et al 2000). Notre démarche consiste à :

1) Exécuter l'algorithme BKM sur un intervalle $[k_{min}, k_{max}]$.
2) Evaluer leurs résultats en utilisant les deux indices de validité.
3) Détecter le nombre optimal de clusters correspondant à la valeur maximale ou minimale de l'indice.

Le reste de l'article est structuré comme suit : la section 2 présente les aspects théoriques relatifs à la fois à l'algorithme BKM et les deux indices de validité RST (*A. Starczewski 2015*) et WB (*Q. Zhao and P. Fränti 2014*). La section 3 est dédiée aux différentes expérimentations menées à la fois sur des données artificielles et réelles, et des données de télédétection. Enfin, une conclusion est donnée en section 4.

2. Aspects théoriques

Dans cette section, nous allons décrire la méthode de clustering utilisée, à savoir le BKM ainsi que les indices de validité RST et WB.

2.1 Bisecting K-means

Introduit par Steinbach et al. en 2000 (*M. Steinbach et al 2000*), l'algorithme BKM est une version améliorée de

l'algorithme KM classique qui offre un gain en termes de temps de calcul tout en préservant une bonne précision des résultats, en se référant dans quelques situations à des données de référence. Le BKM utilisé dans cette étude opère seulement en quatre étapes, données comme suit :

Étape 1 : affecter l'ensemble des objets à un seul cluster.
Étape 2 : sélectionner le cluster à scinder dont l'Erreur Moyenne Quadratique (MSQ) (*M.I. Malinen et al 2014*) est maximale.
Étape 3 : diviser le cluster choisi en appliquant l'algorithme 2-means avec plusieurs itérations, (*procédure de bissection*).
Étape 4 : répéter les étapes de 2 et 3 jusqu'à obtention du nombre de clusters final.

2.2 Indices de validité

Formellement, un indice de validité est une fonction qui mesure la qualité du résultat final d'un algorithme de clustering (*E.C. Aggarwal and C. Reddy 2013*). Dans cette sous-section, deux indices de validités sont présentés, nommés respectivement, RST et WB.

2.2.1 Indice de validité RST

Introduit par Artur Starczewski (*A. Starczewski 2015*), l'indice de validité RST est donné pour K variant sur un intervalle [k_{min}, k_{max}] par :

$$STR = [E(K)-E(K-1)][D(K+1)-D(K)] \quad (1)$$

Avec *E(K–1) l*a mesure de compacité calculée pour un nombre de clusters égale à *K–1* et est donnée par :

$$E(K-1) = \frac{E_o}{E_{K-1}} \quad (2)$$

Avec,

$$E_0 = \sum_{x \in X} \|x-c\| \quad (3)$$

$$E_{K-1} = \sum_{K=1}^{K-1} \sum_{x \in C_K} \|x-c_K\| \quad (4)$$

Où *C* désigne le centre de l'ensemble de données *X* et C_k celui du cluster *k*.
Le terme *D(K)* désigne le rapport entre la séparabilité maximale et minimale elle est donné par :

$$D(K) = \frac{D_{K\,max}}{D_{K\,min}} \quad (5)$$

Avec

$$D_{K\,max} = \max_{i,k=1}^{K} \|c_i - c_k\| \quad (6)$$

$$D_{K\,max} = \min_{i,k=1}^{K} \|c_i - c_k\| \quad (7)$$

2.2.2 Indice de validité WB

L'indice de validité WB *(Q. Zhao and P. Fränti 2014)* est défini comme le rapport entre la mesure de compacité intra classe (WSS) et celle de la séparabilité inter classes (SSB). Il est donné par :

$$WB = K \times \frac{WSS}{SSB} \quad (8)$$

$$WSS = \sum_{i=1}^{N} \|x_i - c_{ki}\|^2 \quad (9)$$

$$SSB = \sum_{k=1}^{K} n_k \|c_k - c\|^2 \quad (10)$$

Avec *K* le nombre de clusters, *N* la cardinalité de l'ensemble *X* et n_k celle du cluster *k*.

3. Expérimentations

Dans cette section, nous proposons de comparer dans un premier temps, le comportement des deux indices de validité sur des jeux de données artificielles et réelles. Le second banc d'essais est conduit sur des images satellitaires acquises par différents capteurs.

La méthode utilisée consiste premièrement à exécuter l'algorithme du BKM sur un intervalle [k_{min}, k_{max}] avec k_{min} = 2. Le nombre maximal de clusters k_{max} est donné par $[|X|/2]^{0.5}$(*K.V. Mardia et al 1979*), dans le cas des données artificielles et réelles, et est fixé à $|X\_app|^{0.5}$, avec X_app étant l'ensemble d'apprentissage construit à partir d'un échantillonnage systématique aléatoire sur les données satellitaires. La dernière étape du processus consiste à déterminer le nombre optimal de clusters k* par évaluation des résultats avec les indices de validité. Le nombre optimal de clusters k* correspond à la valeur maximale de l'indice RST et à la valeur minimale pour l'indice WB.

La dernière partie des expérimentations sera consacrée à une étude comparative entre l'algorithme KM et sa variante le BKM.

3.1 Cas des données artificielles et réelles

Dans cette sous-section, nous proposons d'évaluer et de comparer le comportement des deux indices de validité cités dans la section 2. Pour y parvenir, nous avons choisi dix ensembles de données artificielles, à savoir :
S1, S2, S3, S4 (différents taux de recouvrement), a1, a2, a3 (différents nombre de clusters), dim032, dim064 et dim128 (différentes dimensions). Nous avons aussi testé les deux indices de validité sur trois ensembles de données réels, à savoir : Wine, Iris et Glass. **La Figure 1** présente l'ensemble S2. L'intégralité des données peuvent être obtenues à partir la base UCI (*A. Frank and A. Asuncion2010*).

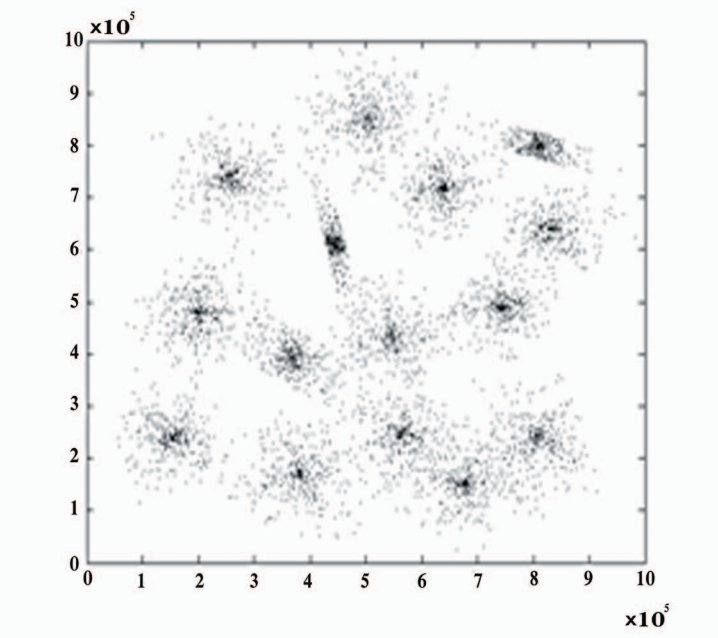


Fig. 1 Données artificielles S2.

Le tableau 1 reporte à la fois les caractéristiques des données utilisées ainsi que le nombre de clusters retourné par les deux indices de validité.

Tab 1 . Caractéristiques des données et résultats

Données	Nb. de clusters	Dimension	Nb. Points	Nb. de clusters obtenu	
				RST	WB
S1	15	2	5000	**15**	16
S2				**15**	**15**
S3				18	20
S4				4	21
a1	20	2	3000	19	28
a2	35	2	5250	**35**	39
a3	50	2	7500	**50**	56
dim032	16	32	1024	**16**	**16**
dim064	16	64	1024	**16**	**16**
dim128	16	128	1024	**16**	**16**
Wine	3	13	178	**3**	**3**
Iris	3	4	150	2	**3**
Glass	7	9	214	4	**7**

A travers les résultats obtenus, nous constatons que les deux indices RST et WB retournent le nombre exact de clusters pour les ensembles S1, S2, dim032, dim064, dim128 et Wine. Pour les trois ensembles a1, a2 et a3, l'indice RST retourne le nombre exact de clusters pour a2 et a3 et se rapproche de la valeur exacte pour a1. L'indice WB échoue, par contre, pour ces trois ensembles mais retourne le nombre exact de clusters pour les deux derniers ensembles réels. De façon générale, le comportement des deux indices de validité RST et WB est sensible au degré de recouvrement, cas des deux ensembles S3 et S4.

Partant de ce constat, nous pouvons conclure que les deux indices RST et WB donnent des résultats satisfaisant dans la

plupart des cas étudiés avec un avantage pour l'indice RST qui se rapproche avec plus de précision du résultat optimal. Néanmoins, les deux indices perdent considérablement en précision au fur et à mesure que le taux de recouvrement augmente.

Les expérimentations menées sur les données artificielles et réelles ont permis d'étudier le comportement des deux indices de validité par rapport aux : taux de recouvrement, le nombre de clusters et le nombre de dimensions. Dans ce qui suit, nous proposons d'appliquer les deux indices sur des données d'observation de la terre.

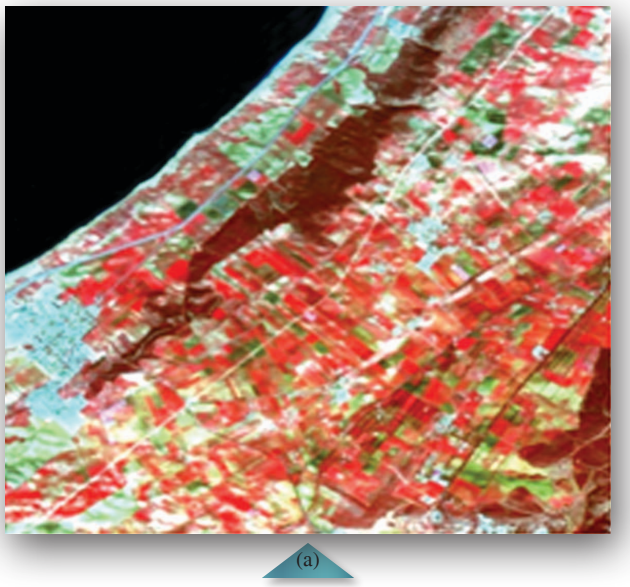
3.2 Cas des données satellitaires

Afin d'évaluer l'apport des deux indices en clustering des données de télédétection, nous les avons appliqués sur des images satellitaires issues de différents capteurs. La première (**Figure 2.a**) et la seconde scène (**Figure 2.b**) sont issues du satellite Spot-5 et Landsat-5 respectivement, et représente la région de Mostaganem, Algérie. La troisième scène (**Figure 2.c**) représente une portion de la ville d'Oran en Algérie et est acquise par le satellite Terra. Les détails relatifs à ces scènes sont reportés dans le tableau 2.

Tab 2. Caractéristiques des images satellitaires

	Scène 1	**Scène 2**	**Scène 3**
Taille	600 x 600	700 x 700	700 x 600
Résolution (m)	20	30	15
Capteur	HRG 2	TM	Aster
Satellite	Spot 5	Landsat-5	Terra
Bandes	3-2-1	4-3-2	3-2-1
Date acquisition	23-03-2012	27-04-2011	28-08-2004

Sachant que chaque scène contient un nombre important de pixels (360 000 pixels pour la première scène), deux techniques ont été adoptées pour fixer le nombre maximal de clusters k_{max}. La première technique, comme cité précédemment, consiste à extraire un ensemble d'apprentissage à partir de la donnée initiale avec des pas fixés empiriquement à 5 et 10, puis à exécuter l'algorithme du BKM sur un intervalle de [2, k_{max}].



(a)

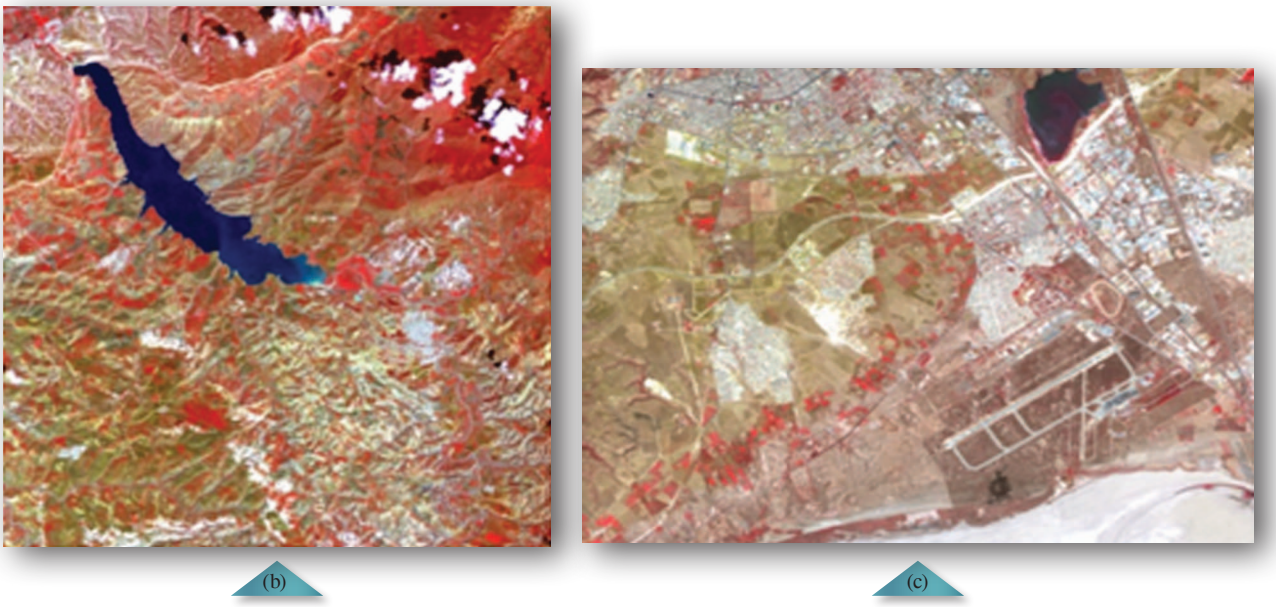


Fig. 2 Images satellitaire utilisées – de haut en bas : scène 1, scène 2 et scène 3.

Les résultats des expérimentations sont présentés dans le tableau 3.

Tab 3. Résultats des deux indices sur des données satellitaires avec variation du pas d'échantillonnage

	Scène 1		Scène 2		Scène 3	
Pas	5	10	5	10	5	10
K_WB	7	7	5	5	6	6
K_RST	3	3	10	5	2	10

A partir des résultats du tableau 3, nous remarquons que l'indice WB retourne le même nombre de clusters pour différentes valeurs du pas d'échantillonnage. Tandis que l'indice RST, présente une instabilité des résultats pour chaque pas utilisé. Ceci s'explique par le principe du RST dont la fonction est calculée pour k = i puis k = i+1 et k = i-1, la variation des paramètres d'entrée tel que les points pris en compte ou le k_{max} influent donc sur les résultats finaux de l'indice.

L'inspection visuelle des images de classes (**Figure 3**) obtenues en comparaison avec les données brutes atteste que l'indice WB est nettement plus adapté que le RST à ce type de données. En effet, les images satellitaires possèdent des tailles importantes, l'échantillonnage s'avère une étape nécessaire pour optimiser le temps de traitement.

Cependant, quelques confusions ont été constatées, plus particulièrement, entre le cluster «Eau» et «Ombres des nuages» dans l'image des classes (**Figure 3.b**) relative à la scène 2 et le cluster «Eau» et «Végétation naturelle» dans l'image des classes (**Figure 3.c**) relative à la scène 3. Ces confusions sont dues à la similarité entre les réponses spectrales des pixels dans les données initiales.

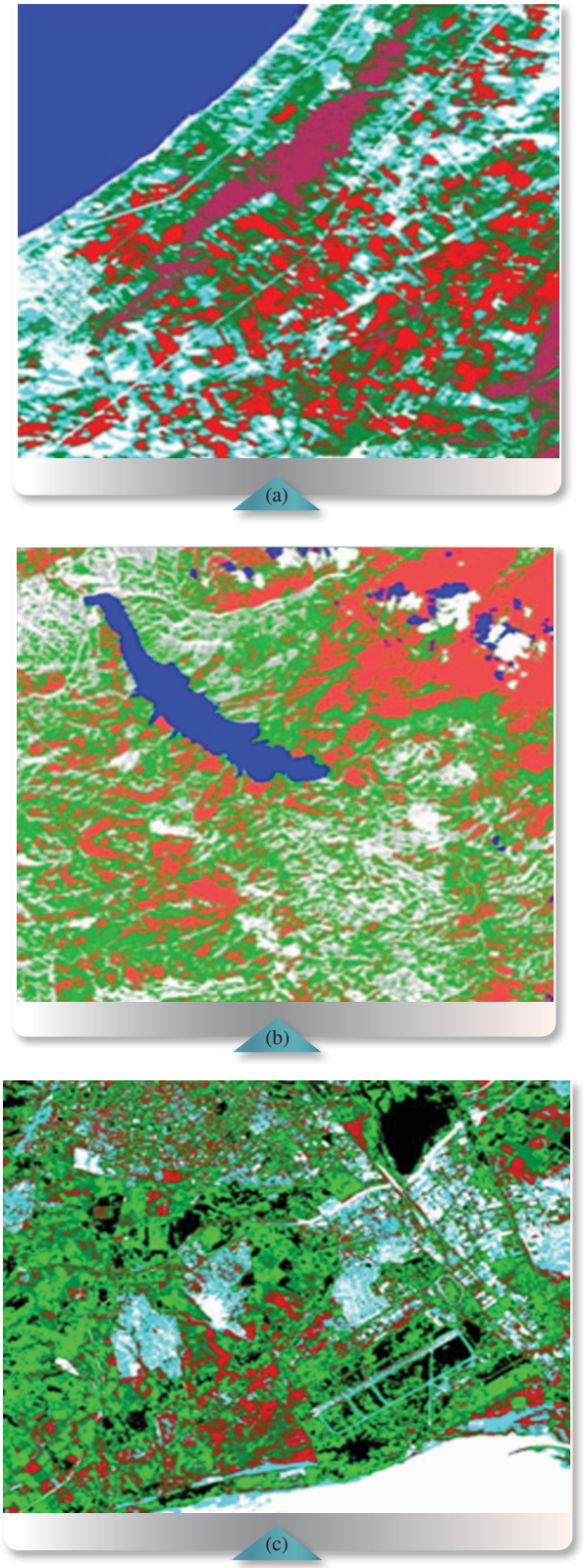


Fig. 3 Images des classes obtenues par BKM et WB – de haut en bas : Sept clusters, cinq clusters et six clusters.

A partir de ces résultats, il apparaît clairement que l'indice WB représente le choix le plus adapté pour la détection automatique du nombre de clusters relatif aux données de télédétection. La **Figure 3** montre les images classifiées par le BKM avec le nombre de clusters retourné par l'indice WB.

La seconde technique utilisée consiste quant à elle à extraire les pics de l'histogramme de l'image en entrée après sa conversion en niveaux de gris. Ces derniers correspondent à k_{max}. **La Figure 4** illustre l'histogramme ainsi que les 57 pics relatifs à la première scène. Le nombre de clusters retourné par l'indice WB avec le nombre de pics est le même que celui trouvé avec la technique d'échantillonnage. Cette variante mérite d'être mieux explorée.

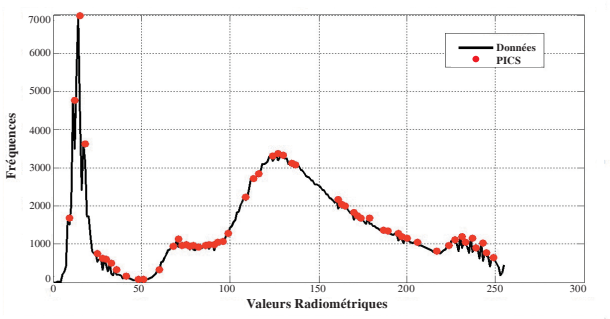


Fig. 4 Histogramme et pics de la scène 1.

3.3 K-means versus Bisecting K-means

Dans le but de montrer les avantages du BKM par rapport à l'algorithme KM, nous les avons comparés sur les quatre données synthétiques S1, S2, S3 et S4. Les critères externes de validation utilisés sont : la F-Mesure, la pureté (*A.K Alok et al 2014*) et l'entropie (*K. Murugesan and J. Zhang 2011*). L'ensemble des résultats obtenus sont reportés dans le tableau 4.

Tab 4. Comparaison entre KM et BKM.

	F-mesure		Pureté		Entropie	
	KM	*BKM*	*KM*	*BKM*	*KM*	*BKM*
S1	*0.72*	**0.98**	*0.72*	**0.98**	*1.60*	*0.60*
S2	*0.79*	**0.94**	*0.82*	**0.94**	*3.17*	**2.11**
S3	*0.69*	**0.73**	*0.72*	**0.74**	**5.56**	*6.37*
S4	*0.74*	**0.75**	*0.75*	*0.74*	*6.37*	*6.57*

D'après le tableau 4, on remarque que le BKM donne de bien meilleurs résultats que le KM en terme de F-Mesure, de pureté et d'entropie quand la donnée possède des clusters bien séparés et donc un taux de recouvrement assez bas. Cependant, l'algorithme perd en précision au fur et à mesure que le taux de recouvrement augmente. Ceci s'explique du fait que le BKM par sa méthode de bissection effectue une séparation linéaire peu précise. Dans le cas

des images satellitaires utilisées dans cet article, seul les bandes couleurs (RGB) sont exploitées, par conséquent, le clustering se fait sur une base purement radiométrique/ spectrale qui n'inclue pas les propriétés géométriques des images. Le BKM est de ce fait nettement plus adapté que le KM en vue du gain de temps et de la précision qu'il offre pour ce type d'images.

4. Conclusions

Dans cet article, nous nous sommes intéressés à l'utilisation de l'algorithme BKM conjointement avec les indices de validité RST et WB. Les résultats obtenus à l'issue de ce travail montrent l'efficacité de la méthode adoptée en vue de détecter le nombre optimal de clusters dans un ensemble de données et plus particulièrement, les données de télédétection. De futurs travaux consisteraient à proposer une version modifiée du BKM qui améliorerait d'avantage le processus de bissection en appliquant par exemple le KM sur l'ensemble des centres et en le comparant avec l'algorithme X-means.

Références Bibliographiques

G. Gan, C. Ma, and J. Wu (2007) "Data Clustering Theory, Algorithms, and Applications", ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alex-Andria, VA.

D. T. Larose (2005), "Discovering Knowledge in Data : An Introduction to Data Mining", JohnWiley & Sons, Inc., Hoboken, New Jersey.

Qi. Letao, H.T. Lin and V. Honavar (2013) "Clustering remote RDF data using SPARQL update queries", Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference, 236-242.

A. Hasnat (2014), "Unsupervised 3D image clustering and extension to joint color and depth segmentation", Signal and Image Processing. Université Jean Monnet, Saint-Etienne.

J. McQueen (1967), "Some methods for classification and analysis of multivariate observations", In Proc.5th Berkeley Symp. Mathematics, statistics and probability, pp. 281-296.

G. Ball and D. Hall (1965) "A novel method of data analysis and pattern classification", In Technical report, Stanford Research Institute, Menlo Park, CA,USA.

M. Steinbach, G. Karypis and V. Kumar (2000) "A comparison of document clustering techniques", Work-shop on Text Mining, KDD.

A. Starczewski (2015) "A new validity index for crisp clustering". Pattern Anal Applications.

Q. Zhao and P. Fränti (2014), "WB-index: a sum-of-squares based index for cluster validity", Knowledge and Data Engineering, Vol.92, pp.77-89.

M.I. Malinen, R. Mariescu-Istodor, and P. Fränti (2014) "K-means* : Clustering by gradual data transformation", Pattern Recognition, 47(10):3376–3386.

E.C. Aggarwal and C. Reddy (2013) "Data Clustering Algorithms and Applications", CRC Press.

K.V. Mardia, J.T. Kent and J.M. Bibby (1979), "Multivariate Analysis", Academic Press.

A. Frank and A. Asuncion (2010) "UCI machine learning repository".

A.K.Alok, S. Saha and A. Ekbal (2014) "Development of An External Cluster Validity Index using Probabilistic Approach and Min-max Distance", International Journal of Computer Information Systems and Industrial Management Applications, ISSN 2150-7988, 6:494-504.

K.Murugesan and J. Zhang (2011) "Hybrid Bisect K-Means Clustering Algorithm", Business Computing and Global Informatization (BCGIN), 2011 International Conference, 216-219.