Cartographie des surfaces naturelles : méthodes supervisée et non supervisée de classification d'images de télédétection

L'Association internationale de cartographie définit la cartographie comme la discipline qui s'occupe de la conception, de la production, de la diffusion et de l'étude des cartes. La cartographie est aussi la représentation – la carte. La cartographie englobe donc tout le processus de la création de cartes.

La ***cartographie*** d'aujourd'hui possède deux caractéristiques essentielles. Premièrement, les cartes jouent un rôle fondamental et indispensable comme pilier de la civilisation. Plusieurs activités reliées à la surface de la Terre comme l'utilisation des terres, le cadastre et la localisation de propriétés, les prédictions météorologiques, la construction de routes, les analyses de localisation, les mesures d'urgence, l'aménagement des forêts, la prospection minière, la navigation, – la liste est sans fin – seraient impraticables sans les cartes. Et ce rôle n'a jamais été aussi essentiel qu'aujourd'hui. L'humanité fait face à plusieurs problèmes sévères dont plusieurs concernant l'environnement et la cartographie s'avère un outil crucial dans la recherche de solutions. Associée à la grande capacité de collecte de données et de capacité d'analyse de la télédétection et des systèmes d'information géographique (SIG) la cartographie, dans plusieurs cas, fournit la clé à la découverte de solutions.

Une seconde caractéristique primordiale de la cartographie d'aujourd'hui est sa nature dynamique. La discipline de la cartographie est au cœur d'une révolution engendrée par les avancées technologiques de l'informatique.

La ***télédétection*** est l'ensemble des techniques qui permettent, par l'acquisition d'images, d'obtenir de l'information sur la surface de la Terre, sans contact direct avec celle-ci. La télédétection englobe tout le processus qui consiste à capter et enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et analyser l'information qu'il représente, pour ensuite mettre en application cette information. (Site Web du Centre Canadien de Télédétection : http://www.ccrs.nrcan.gc.ca).

I. Méthodes supervisée et non supervisée de classification d'images de télédétection

I.1. Les classifications d'images pixel à pixel

Les méthodes de classification d'images les plus courantes utilisent ***l'information radiométrique*** d'une ou plusieurs bandes spectrales pour ***classifier chaque pixel individuellement***. Ce type de classification standard est appelé ***reconnaissance de regroupements spectraux***. L'image issue de la classification est composée de pixels appartenant chacun à un thème particulier (exemple : eau, forêt, bâti, etc.).

Lorsqu'on parle de ***classes***, il faut faire la distinction entre des ***classes d'information*** et des ***classes spectrales*** :

- Les ***classes d'information*** sont des ***catégories d'intérêt*** que l'analyste cherche à identifier dans les images, comme différents types de cultures, d'espèce d'arbres, différents types de caractéristiques géologiques, etc.
- Les ***classes spectrales*** sont des groupes de pixels qui ont les ***mêmes caractéristiques spectrales*** (ou presque) en ce qui a trait à leur valeur radiométrique dans les différents canaux.

I.2. Objectif de la classification

L'objectif ultime de la classification est de faire la ***correspondance entre les classes spectrales et les classes d'information.*** Dans ce contexte, l'analyste a le rôle de déterminer de l'utilité des différentes classes spectrales et de valider leur correspondance à des classes d'informations utiles.

I.3. Démarche de classification

La démarche de classification comporte ainsi trois étapes fondamentales :

- ✓ Établissement de classes de signatures ou classes spectrales,
- ✓ Classification des pixels en fonction des classes de signatures,
- ✓ Vérification de la classification (fiabilité par rapport aux classes thématiques envisagées).

C'est la manière d'établir les classes de signatures (1ère étape) qui détermine les deux grandes approches en classification numérique :

- ✓ ***L'approche non dirigée*** (***non supervisée***) : pas de connaissance a priori ; les classes sont créées automatiquement par le logiciel. Les classes sont alors nommées, ***étiquetées a posteriori*** ;
- ✓ ***L'approche dirigée*** (***supervisée)*** : les connaissances a priori sont utilisées pour la création des classes et la saisie des échantillons (***zones (parcelles) d'entraînement = données test***).

Les méthodes de classifications par pixels sont délicates à mener car les ***confusions radiométriques*** entre classes augmentent avec ***la résolution des capteurs*** et ***l'hétérogénéité spatiale*** des milieux.

II. Indices de végétation radiométriques (NDVI)

L'Indice NDVI (**N**ormalized **D**ifference **V**egetation **I**ndex) est un indice normalisé permettant de générer une image affichant la couverture végétale (biomasse relative). Cet indice repose sur le contraste des caractéristiques de deux canaux d'un jeu de données raster multispectral : l'absorption du pigment chlorophyllien dans le canal rouge et la haute réflectivité des matières végétales dans le canal proche infrarouge (NIR).

L'indice NDVI est utilisé dans le monde entier pour surveiller la sécheresse, contrôler et prévoir la production agricole, aider à la prévention des incendies et cartographier la désertification. L'indice NDVI est privilégié pour l'observation globale de la végétation car il permet de compenser les changements de conditions d'éclairage, de pente de surface, d'exposition et d'autres facteurs exogènes (Lillesand 2004).

Cet indice génère des valeurs comprises entre -1.0 et 1.0, représentant principalement la couverture végétale, où les valeurs négatives sont essentiellement générées par les nuages, l'eau et la neige et les valeurs proches de zéro essentiellement générées par la roche et le sol nu. Les très faibles valeurs de l'indice NDVI (0,1 et inférieures) correspondent aux surfaces stériles de roche, de sable ou de neige. Les valeurs intermédiaires (0,2 à 0,3) représentent des zones d'arbustes et de prairies, alors que les valeurs élevées (0,6 à 0,8) indiquent des forêts tempérées ou tropicales humides.

L'équation NDVI par défaut documentée se présente comme suit :

$$NDVI = ((IR - R)/(IR + R))$$

- IR = valeurs de pixel du canal infrarouge
- R = valeurs de pixel du canal rouge

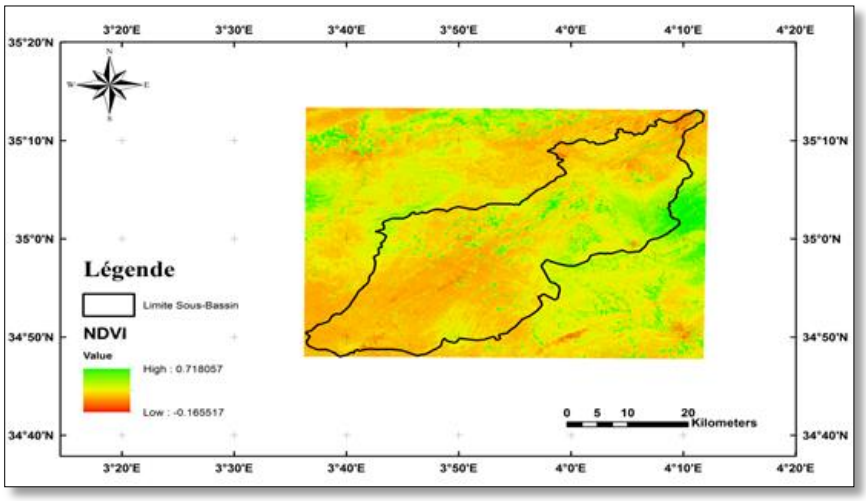


Fig.1: Carte de l'Indice de végétation NDVI d'oued Bou Saâda

III. Méthodes de classification

La classification a pour but de regrouper (partitionner, segmenter) nn observations en un certain nombre de groupes ou de classes homogènes. Il existe deux principaux types de classification:

- La ***classification supervisée***, souvent appelée simplement classification ;
- La ***classification non supervisée***, parfois appelée partitionnement, segmentation ou regroupement (***Clustering*** en anglais).

III.1. Méthode supervisées (classification supervisée)

Soit $D = \{d_1, d_2, \ldots, d_i, \ldots, d_m\}$ un ensemble de documents représentés chacun par une description $\vec{d_1}, \vec{d_2}, \ldots, \vec{d_m}$, et $C = \{C_1, C_2, \ldots, C_k, \ldots, C_c\}$ un ensemble de classes, la classification supervisée suppose connues deux fonctions. La première fait correspondre à tout individu ***di*** une classe C_K. Elle est définit au moyen de couples (d_i, C_K) donnés comme exemples au système. La deuxième fait correspondre à tout individu ***di*** sa description $\vec{d_i}$.

$$C^f : \vec{d_i} \rightarrow C_k$$

Qui à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description. De plus, si l'on veut obtenir un système d'apprentissage, la procédure de classification doit permettre de classer efficacement tout nouvel exemple (pouvoir prédictif).

La classification supervisée consiste alors à déterminer une procédure de classification :

- On connaît déjà le nombre de groupes qui existent dans la population;
- On connaît le groupe auquel appartient chaque observation de la population;
- On veut classer les observations dans les bons groupes à partir de différentes variables.

On peut ensuite utiliser une règle de classification pour prédire les groupes auxquels appartiennent de nouvelles observations. Des exemples classiques d'applications sont:

- Identifier si une transaction bancaire est frauduleuse ou pas;
- Reconnaître des chiffres écrits à la main;
- Identifier le type de cancer dont souffre un patient.

III.2. Méthode non supervisées (classification non supervisée / clustering)

La classification non-supervisée est utilisée lorsque que l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification. A la fin du processus de classification non-supervisée, les documents doivent appartenir à l'une des classes générées par la classification. On distingue deux catégories de classifications non-supervisées : hiérarchiques et non-hiérarchiques.

Dans la **classification hiérarchique(CH)**, les sous-ensembles créés sont emboîtés de manière hiérarchique les uns dans les autres. On distingue la ***CH descendante (ou divisive)*** qui part de l'ensemble de tous les individus et les fractionne en un certain nombre de sous-ensembles, chaque sous-ensemble étant alors fractionné en un certain nombre de sous-ensembles, et ainsi de suite. Et la ***CH ascendante (ou agglomérative)*** qui part des individus seuls que l'on regroupe en sous-ensembles, qui sont à leur tour regroupés, et ainsi de suite. Pour déterminer quelles classes on va fusionner, on utilise le critère d'agrégation.

Dans la **classification non-hiérarchique**, les individus ne sont pas structurés de manière hiérarchique. Si chaque individu ne fait partie que d'un sous-ensemble, on parle de *partition*. Si chaque individu peut appartenir à plusieurs groupes, avec la probabilité ***Pi*** d'appartenir au groupe i, alors on parle de *recouvrement*.

La classification non supervisée consiste alors à déterminer une procédure de classification :

- On ne connaît souvent pas le nombre de groupes qui existent dans la population;
- On ne connaît pas le groupe auquel appartient chaque observation de la population;
- On veut classer les observations dans des groupes homogènes à partir de différentes variables.

Les applications typiques sont nombreuses. Par exemple:

- En biologie : l'élaboration de la taxonomie animale;
- En psychologie : la détermination des types de personnalités présents dans un groupe d'individus;
- En *text mining* : le partitionnement de courriels ou de textes en fonction du sujet traité.

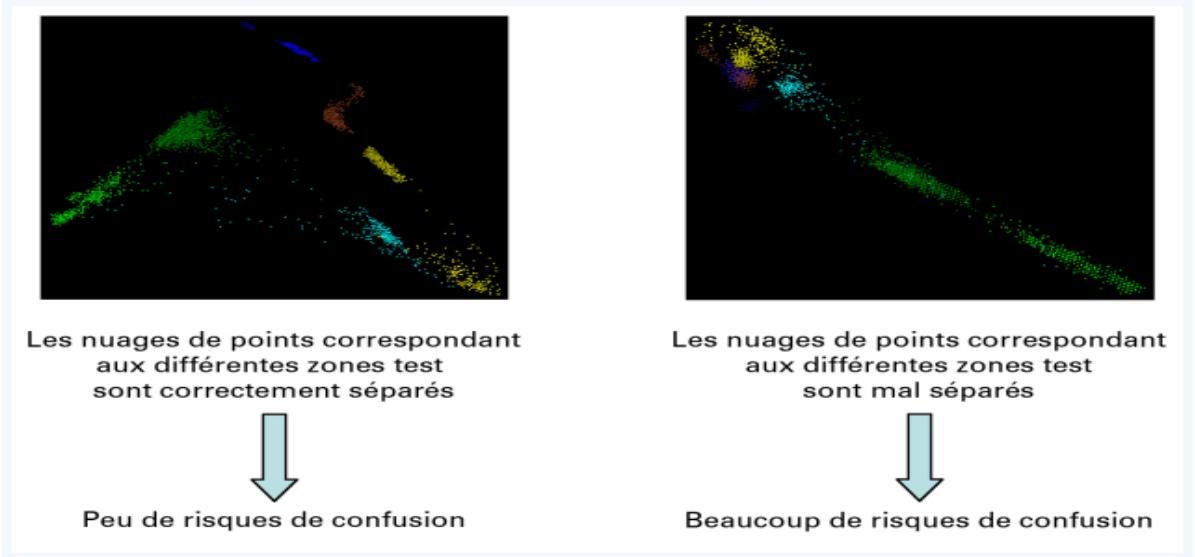
Il existe plusieurs familles de méthodes de classification non supervisée. Les plus communes sont:

- La classification hiérarchique;
- La classification non hiérarchique, par exemple la méthode des k-moyennes (*k-means*);
- La classification basée sur une densité;
- La classification basée sur des modèles statistiques/probabilistes, par exemple un mélange de lois normales.

IV. Evaluation de la qualité d'une classification :

IV. 1. Visualisation des zones test dans l'espace des attributs.

Le but est de visualiser le nuage de points des zones test pour vérifier leur niveau d'homogénéité et leur degré de séparabilité. Cette visualisation s'effectue dans l'espace des attributs les plus performants choisis.



IV. 2. Evaluation d'un système de classification

Nous présentons ici une méthode permettant d'évaluer une classification supervisée, et des techniques classiques pour mesurer et comparer des systèmes de classifications non-supervisées.

Corpus de test (cas supervisé)

Pour tester la qualité d'une procédure de classification supervisée, on sépare aléatoirement les éléments classés entre une base de référence(R) et une base de test(T). Ensuite, on détermine la procédure de classification Cf à partir des exemples de la base de référence. Puis, on utilise Cf pour retrouver la

classe des éléments de la base de test. Enfin, on estime l'erreur de la procédure de classification.

Pour estimer le taux d'erreur TE d'une procédure de classification C^f, une méthode simple est de calculer le nombre d'éléments mal classés sur le nombre d'éléments à classer :

$$TE(C^f) = \frac{1}{card(T)} \sum_{t=1}^{card(T)} (C^f(\vec{d_t}) \neq C_{d_t})$$

où C_{d_t} est la classe d'origine de d_t.

Dans les cas de classifications simples, on peut être amené à calculer l'erreur résultant d'une classification purement aléatoire C^a pour la comparer avec l'erreur faite par notre procédure C^f afin de vérifier la performance de notre système.

Soit P_k la fréquence (ou probabilité à priori) de la classe k dans la base de test, on appelle erreur TE_a du système aléatoire :

$$TE_a = 1 - \sum_{k=1}^{c} (P_k)^2 = 1 - \sum_{k=1}^{c} \left(\frac{card(C_k|T)}{card(T)} \right)^2$$

où c est le nombre de classes et $card(C_k|T)$ est le nombre d'éléments de T qui sont dans la classe C_k.

L'erreur apparente $TE(C^f)$ est dépendante de l'échantillon considéré. Cependant, plus le nombre d'éléments de l'échantillon est grand, plus l'erreur mesurée tend vers l'erreur réelle de C^f.

- ***Cas non-supervisé***

Dans le cas non-supervisé, on peut évaluer la classification par rapport à certaines de ces caractéristiques. On distingue d'une part, les caractéristiques numériques : le nombre de classes obtenues, le nombre d'éléments par classe, le nombre moyen d'éléments par classe, l'écart-type des classes obtenues, et d'autre part, les caractéristiques sémantiques. Par exemple, si à un document est associé un ensemble de mots clés, la sémantique associée à une classe pourra se composer des mots les plus fréquents dans la classe.

Pour évaluer l'homogénéité du nombre d'images par classe, on peut utiliser la variance :

$$V = \sigma^2 = \frac{1}{c} \sum_{k=1}^{c} (card(C_k) - moy)^2$$

Où $moy = \frac{1}{c} \sum_{k=1}^{c} card(C_k)$ est le nombre moyen d'éléments par classe et c est le nombre de classes obtenues. L'écart-type $\sigma = \sqrt{V}$ permet d'exprimer la dispersion dans la même unité que la moyenne.

Références

- Cherel J.P. (2010). Support de cours M1 SIIG3T - Traitement d'images - classification d'images de télédétection. p 13.
- Tollari S. (2003). Rehaussement de la classification textuelle d'une base de données photographiques par son contenu visuel. Mémoire de DEA. Sous la direction de Hervé Glotin et Jacques Le Maitre. Laboratoire SIS - Equipe Informatique.