MeMoTF Merge and Mosaic Through Folders Une boîte à outils de prétraitements automatiques de grandes bases de données géoréférencées Deux exemples de coopération ingénierie Géomatique et recherche en Géographie

Kévin Bourrand, Ibrahima Diedhiou, Romain Courault, Clélia Bilodeau

MeMoTF (*Merge and MosaicThroughFolders*), une boîte à outils de recomposition automatique de grandes bases de données géoréférencées : 2 exemples de coopération entre ingénierie géomatique et recherche en géographie

Bourrand Kévin (k.bourrand@hotmail.fr), Ingénieur contractuel, UMR Lied, Pôle Image, Université Paris Diderot
Diédhiou Ibrahima (ibrahima.diedhiou85@yahoo.fr), Doctorant UMR Lied, Pôle Image, Université Paris Diderot
Courault Romain (romain.courault@paris-sorbonne.fr), Doctorant UMR ENeC, Pôle Image, Université Paris Sorbonne
Bilodeau Clelia (clelia.bilodeau@gmail.com), Maître de Conférences, UMR Ladyss, Pôle Image, Université Paris Diderot

Public visé: Tout public

Logiciels Esri utilisés :ArcGIS 10.2

Thématique : Tout public

Mots-clés:Gestion de données vectorielles et raster décomposées, Indexage de données géographiques, boîte à outil opérationnelle.

RESUME

La recomposition des données, aussi bien raster que vectorielles, est une tâche qui peut s'avérer fastidieuse. L'utilisation des caractères génériques et des commandes par lots, bien qu'ayant déjà prouvé leur efficacité, présentent l'inconvénient de devoir être programmées. Nous proposons ici un complément à ces méthodes s'appuyant sur le principe suivant : une dénomination efficace des éléments d'un tout permet de rassembler automatiquement ceux-ci. Si plusieurs sources de données partageant une même étendue géographique gardent une même logique de dénomination, il est alors possible d'effectuer un assemblage automatique de toutes ces données en spécifiant

uniquement un dossier contenant l'ensemble des données ainsi que la plage de caractères partagée par les fichiers à assembler. La sortie applicative de cet exposé se matérialise par la création d'une boîte à outil (toolbox) permettant de faire des assemblages vectoriels et raster. Nous présenterons trois cas d'applications scientifique ayant mobilisé ces nouveaux outils (données vectorielles : projet OLIZERO-EST-Lied ; données raster : doctorats Ibrahima Diédhiou et Romain Courault).

.

CADRE DE DEVELOPPEMENT

La création de cet outil de réassemblage des données a pris lieu au sein du Pôle Image (Université Paris-Diderot). C'est au sein de cette structure qui sert de plateforme informatique aux enseignants - chercheurs, étudiants, doctorants pour effectuer leurs travaux personnels nécessitant des logiciels spécifiques à la géographie quantitative (traitements d'images, SIG, calculs de statistiques...etc) que des demandes convergentes concernant l'assemblage/réassemblage des données nous ont incités à développer un outil permettant l'accélération de ces prétraitements.

Deux demandes sont à l'origine de la création de cet outil :

1- L'assemblage de données vectorielles constituant une base de données topographique couvrant la Sierra Magina. Cette demande provient du programme de recherche OLIZERO. Celui-ci cherche à valoriser les déchets de taille des oliviers par la voie de la dégradation fongique et du traitement pyrolytique.

2- L'assemblage de données raster (images MODIS) compressées sur l'Afrique de l'ouest subsaharienne et le nord de la Scandinavie. Cette seconde demande émanait des doctorants Ibrahima Diédhiou et Romain Courault qui cherchaient à réaliser un suivi de la phénologie des formations végétales sur leurs zones d'étude respectives (Afrique de l'ouest subsaharienne et nord de la Scandinavie).

INTRODUCTION

Une des étapes importantes pouvant précéder le traitement de données géographiques sur une vaste zone d'étude est leur assemblage. En effet, il est courant d'avoir besoin d'une donnée vectorielle et/ou matricielle fusionnée pour opérer directement des traitements sans avoir à répéter l'opération sur chaque partie non assemblées des fichiers. Cet assemblage peut s'avérer fastidieux lorsque la fragmentation des données est importante (découpage spatial, temporel, répartition des données sources dans différents répertoires) et lorsque les données sont multiples (plusieurs couches composant un package de données lié à une zone géographique par exemple). Nous proposons ici un outil capable d'effectuer cet assemblage de manière automatique en fonction du nom des parties composant un même fichier, ou un même jeu de données.

PRINCIPE DE BASE A TRAVERS L'EXEMPLE DES DONNEES VECTORIELLES- projet OLIZERO

La recomposition automatique des données se base sur la logique de nommage des parties décomposées. En effet, si l'on considère que chacune des parties d'un même fichier ont en commun une partie de leurs noms, on peut en théorie aisément les regrouper ensemble puis les assembler.

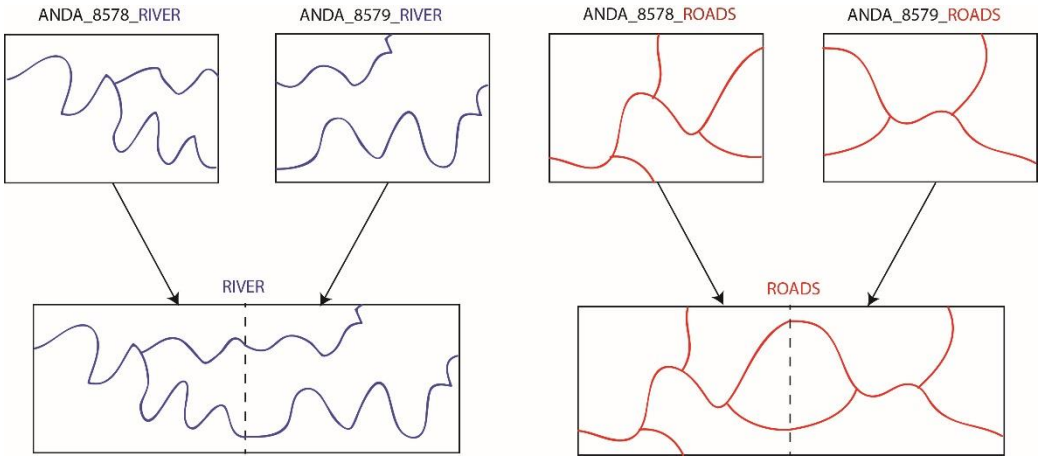

Figure 1 : Schéma de recomposition de données vectorielles selon un caractère générique

Dans la figure1, les classes d'entités sont associées en fonction de leurs noms. Les fichiers correspondant au réseau de rivières (dont les noms finissent toujours par « RIVER ») sont associés ensemble. Il en est de même pour les fichiers correspondant au réseau de route (noms de fichier finissant par « ROADS »).

Cette opération pourrait se faire en utilisant des caractères génériques[1]. Cependant, leur utilisation impose de réécrire celui-ci à chaque fois que l'on change de catégorie d'objet.

L'originalité du programme MeMoTF consiste à remplacer dans certains cas l'utilisation du caractère générique, permettant à ArcGIS de collecter les éléments à assembler par l'identification du positionnement des caractères communs (index).

Dans notre exemple, si l'on spécifie que la partie commune de la chaîne se situe après le 10ème caractère, les fichiers « RIVER » et les fichiers « ROAD » seront regroupés sans que l'on ait besoin de spécifier le caractère générique spécifique à chaque catégorie d'objet comme illustré par la figure 2.

Caractères	A	N	D	A	_	8	5	7	8	_	R	I	V	E	R
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Caractères	A	N	D	A	_	8	5	7	8	_	R	O	A	D	
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Caractères	A	N	D	A	_	8	5	7	9	_	R	I	V	E	R
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Caractères	A	N	D	A	_	8	5	7	9	_	R	O	A	D	
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	

Figure 2 : Identification de la partie commune de la chaîne de caractère

[1]Caractère générique (informatique): chaîne de caractère utilisé pour effectuer une recherche de fichier contenant, commençant ou finissant par cette chaîne de caractère

Dans le cadre de notre exemple, la reconstitution d'une base de données topographique (BCN25/BTN25) sur la Sierra Magina (Andalousie) nécessitait la recomposition de 95 couches vectorielles divisées en 26 dalles (soit 2470 fichiers) provenant de la plateforme de téléchargement du CNIG (Centro Nacional de Información Geográfica). Les fichiers composant une même donnée étaient répartis dans autant de dossier qu'il y avait de dalles, chacun de ces fichiers portant cependant le même nom. La dénomination de ces fichiers suivait toujours le même schéma (figure 3).

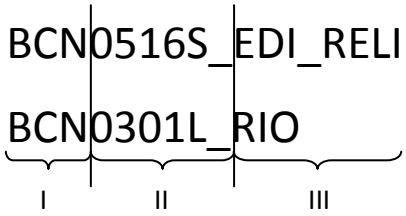

Figure 3: Schéma de découpage de la chaîne de caractère composant le nom d'une couche vectorielle de la base « BCN25/BTN25 » où I: Partie constante de la chaîne de caractère représentant l'identifiant de la base de données, ici : « BCN » pour la base « BCN25/BTN25 » , II : Partie variante en fonction de la couche représentée, elle fait toujours 6 caractères, III : Partie variante en fonction de la couche représentée, son nombre de caractères est variant. Cette partie décrit la nature de la donnée (ici : « RIO » -> rivière)

Afin de réunir les parties composant une même couche de donnée, nous spécifions que le positionnement des caractères en commun aux parties à assembler commence à partir du premier caractère afin de capter l'ensemble de la chaîne de caractère. Ainsi, tous les fichiers trouvés portant le même nom sont assemblés.

TRAME DE FONCTIONNEMENT DE L'OUTIL MeMoTF

1) Définition de la partie commune de la chaîne de caractère composant le nom des éléments à assembler (figure 2)
2) Collection de tous les fichiers trouvés dans un dossier défini et répondant à un caractère générique d'extension spécifique aux éléments à assembler (*.shp, *.tif,*.hdf, etc)
3) Création d'une liste de premier rang dont chaque élément est une liste de second rang constituée par des fichiers partageant la partie de la chaîne de caractère spécifié lors de l'étape 1 (figure 4).
4) Assemblage raster ou vecteur sur chacune des listes de second rang

[[Première Partie du premier fichier à reconstituer, Deuxième partie du premier fichier à assembler,….],[Première partie du deuxième fichier à reconstituer,…],…]

Figure 4 : Représentation des deux rangs de listes. La liste de premier rang est représentée par les crochets et virgules rouges, les listes de second rang sont représentées par les crochets et virgules bleues.

PRESENTATION DE LA TOOLBOX MeMOTF

Afin de mettre notre méthode d'assemblage en application, nous avons créé une toolbox regroupant deux outils permettant l'assemblage des données vectorielles et raster. Ces outils présentent tous une trame commune. La saisie de l'emplacement où chercher les fichiers à assembler et la saisie de la partie nominale commune. Cette deuxième saisie se décompose en deux booléens et deux champs (figure 5).

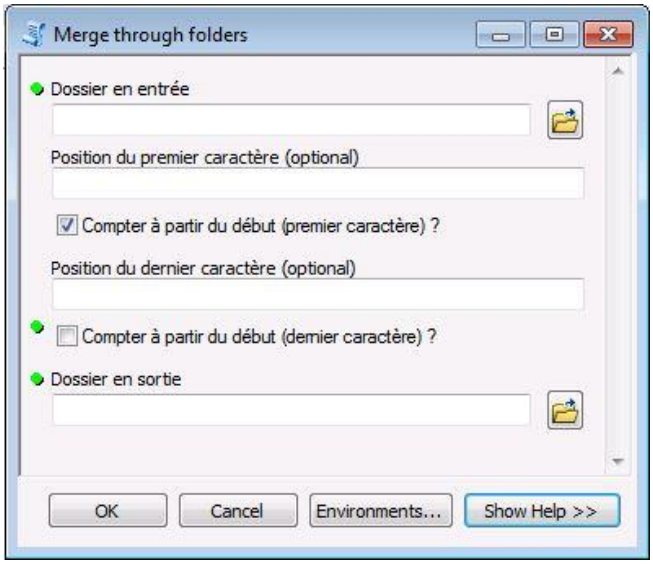

Figure 5 : Capture d'écran de l'outil « Merge through folders »

Afin de compléter cette saisie, il faut donner la position du premier caractère à être capté dans le premier champ puis indiquer si celui-ci doit être pris à partir du début de la chaîne de caractère ou de la fin à l'aide d'un opérateur booléen. De même, on indique dans le deuxième champ la position du dernier caractère à être capté puis si celui-ci doit être pris à partir du début ou de la fin de la chaîne de caractère. Il est à noter que ces champs sont facultatifs, il est envisageable de ne pas renseigner la position du premier caractère et/ou celui du dernier. Cela permet de capter des parties de chaîne dont le nombre de caractère serait susceptible de varier. Les champs supplémentaires éventuels sont spécifiques aux outils assemblant les données vectorielles ou raster. Ces champs sont liés à des caractéristiques que les deux types de données ne partagent pas.

APPLICATION DE L'ASSEMBLAGE RASTER AUTOMATIQUE A DEUX ETUDES DE CAS, EN CASAMANCE (SENEGAL) ET EN LAPONIE (NORVEGE, SUEDE)

> Contexte

Le travail de suivi phénologique régional engagé par Romain Courault (doctorant UMR ENeC, Université Paris Sorbonne) dans la moitié nord de la Scandinavie entre en résonnance avec celui d'Ibrahima Diédhiou (doctorant UMR LIED, Université Paris-Diderot). Un premier travail de réflexion nous a permis d'identifier des besoins communs, aussi bien du point de vue théorique, méthodologique que pratique et technique.

Ainsi, une méthodologie commune entre suivi phénologique de la couverture arborée au Sénégal et dans le nord de la Scandinavie semble se justifier. En effet, la densité des strates arborées et arbustives y sont dépendants des facteurs climatiques, respectivement les précipitations de

mousson pour l'Afrique sub-saharienne et les températures estivales en ce qui concerne le couvert arboré et arbustif dans le nord de la Scandinavie.

L'utilisation synchronique de l'indice de végétation EVI (*EnhancedVegetation Index*) apporte ainsi des éléments de réponse intéressants quant à nos deux problématiques de thèse. Précisément, l'EVI évite les effets de saturation radiométrique liés à une couverture herbacée particulièrement active (en termes de répartition spatiale et de biomasse ; ainsi qu'en termes de réponse radiométrique) lors des saisons phénologiques favorables subsahariennes et scandinaves. De plus, la sensibilité de l'EVI aux effets de canopée pour les strates arbustives et arborées peut justifier de son utilisation pour la toundra arbustive majoritaire dans le nord de la Scandinavie et des mosaïques forestières plus ou moins lâches en ce qui concerne le Sénégal. Une interrogation commune était d'identifier les tendances, cycles et extrêmes phénologiques pluri- et intra- annuels pour des occupations de sols données, en particulier pour les formations végétales.

> Implications de la *toolbox* MeMOTF dans la méthodologie

Concrètement, les besoins techniques liés au programme MeMOTF sont axés sur le prétraitement des sous-produits issus de l'imagerie satellitale MODIS (produits 13Q1, résolution spatiale : 250m, résolution temporelle : 16j). Après le téléchargement de 15 années d'images (2000-2015), découpées temporellement en quinzaine de jours, cette opération initiale a dû être multipliée par 4, correspondant au nombre de sous-scènes adjacentes (figures 6a et 6b) délimitant une échelle régionale commune.

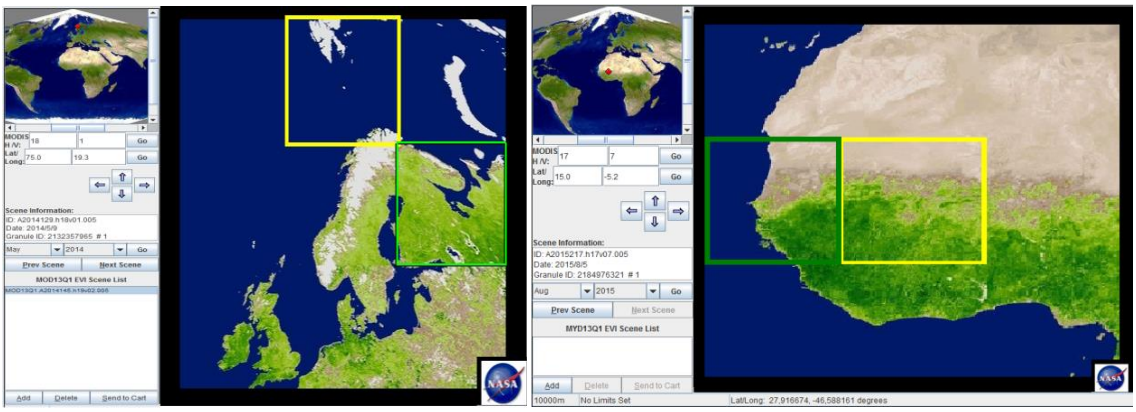

Figure 6 : Plate-forme web USGS GLOVIS utilisée pour le téléchargement des 1440 rasters MODIS 13Q1 (24 images par an, 2000-2015, 4 dalles par région, 2 régions étudiées – Nord Scandinavie, figure 6a / Sénégal, figure 6b)

Un travail préliminaire est représenté par l'extraction des fichiers sources, de choix de répertoire et d'identification des canaux MODIS à même de nous fournir les produits EVI(figure 7). Les étapes suivantes ont été effectuées à l'aide de l'outil « *Mosaic through Folders* ». Une deuxième tâche est représentée par le mosaïquage des dalles rasters, limitant les différences de contrastes entre des sous-scènes voisines et donc subséquemment les biais liés aux traitements des séries temporelles.

Une troisième étape consiste à re-projeter des images, passant du système de coordonnées

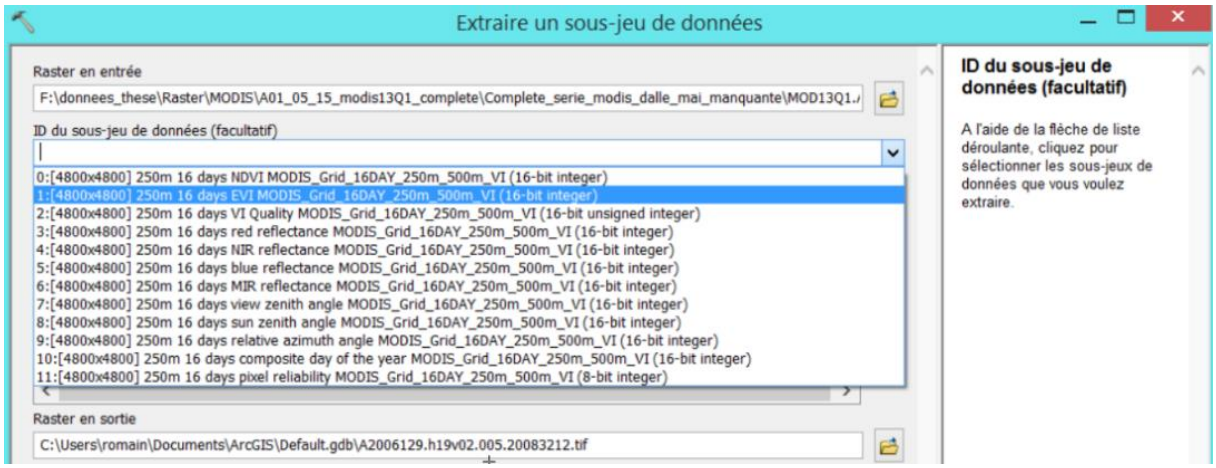

Figure 7: ToolboxArcGIS 10.2 "Extraire un sous-jeu de données [raster]", étapes cruciales d'import du canal MODIS souhaité (ici EVI) et de conversion matricielle (*.hdf vers *.tif): comment automatiser la tâche pour des données massives dans le temps et dans l'espace géographique?

polaire du satellite Terra à un système projeté régional (WGS84 / UTM 34N pour le cas scandinave).

Enfin, une quatrième étape se concentre sur la conversion en un format générique (*.tiff), nous affranchissant des contraintes liées au format initial (*.hdf) des sources de données (poids des fichiers, sélection systématique du canal EVI lors de l'import des données raster…). Cette étape de conversion était nécessaire à l'importation des données sur des logiciels plus spécifiques de traitement d'images satellites. Enfin, la désignation et le nommage des images prétraitées dans des répertoires identifiés nous ont permis de limiter la tâche fastidieuse de l'import de données Raster dispersées dans différentes arborescences/sous-répertoires (figure 8).

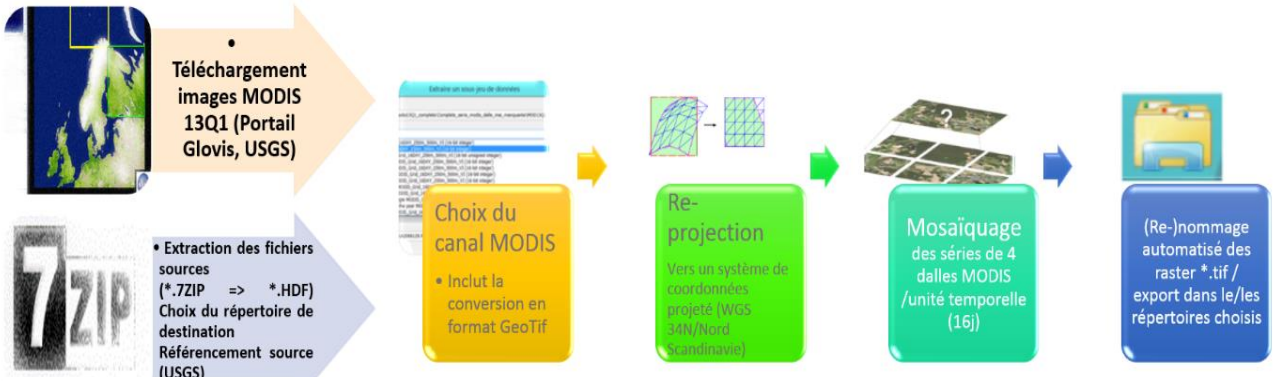

Figure 8: Chaîne de traitement découpant les différentes étapes nécessaires au prétraitement des images satellite grand champs. MeMOTF permet de rassembler les 4 dernières étapes, ce qui constitue un gain de temps considérable.

CONCLUSION

Le développement de la boîte à outil « MeMoTF » a permis dans les deux cas d'études présentés un gain de temps important, temps qui a pu être réinvestit pour l'analyse des données. Le nombre d'utilisateurs potentiellement intéressé par la boîte à outil nous semble grand, étant donné que cet outil est applicable à un grand nombre de données de nature différente (raster ou vecteur)

et que la recomposition des données est un problème récurrent lorsque l'on travaille avec des données spatialisées sur de grandes zones.

Les apports de MeMoTF aux différents travaux de recherche présentés ici sont particulièrement riches. De son utilisation, on retiendra en particulier sa flexibilité (combinaisons faciles avec d'autres outils) et sa relative facilité d'usage qui conviennent à un grand nombre de situations et de besoins en matière de traitements de données géographiques massives: assemblage de données topographiques vectorielles (programme OLIZERO, LIED), prétraitements de produits satellite (Doctorat Université Paris-Sorbonne, UMR ENeC ; Doctorat Université Paris-Diderot, UMR LiED) pour les rasters.

Si la création de cet outil s'est révélé être relativement simple (bien qu'ayant nécessité de programmer en python), nous avons d'ores et déjà pu constater son efficacité d'un point de vue pratique. De plus, le développement de cette boîte à outil est à poursuivre. Nous pensons ainsi développer un « fragmenteur de données géographiques » permettant la dénomination logique selon un schéma prédéterminé (grille…), ceci étant fait de façon à ce que les données puissent être facilement recomposées en aval. Cela permettrait de boucler la boucle en quelque sorte.