Analyse de données du GIS Sol

David Quérin

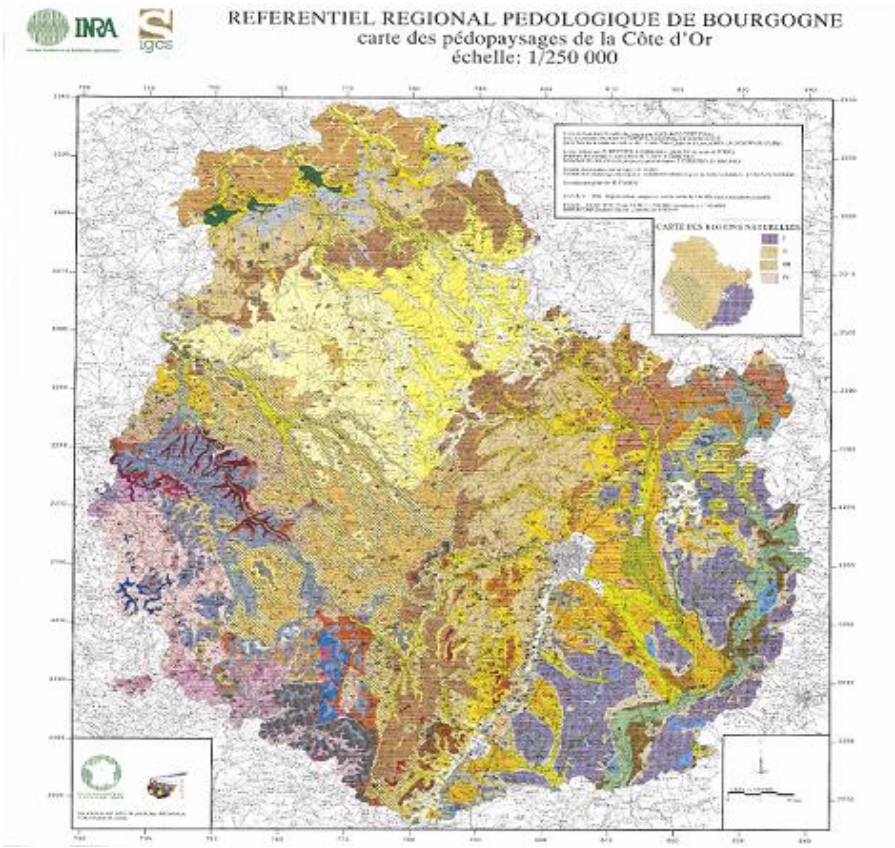
Groupement d'intérêt scientifique Sol

UNIVERSITE D'ORLEANS

INRA SCIENCE & IMPACT

Master Professionnel
M1 Statistiques et Probabilité Appliquée

-Rapport de Stage-
Analyse de données du Gis Sol



Source : J. Chrétien – 1996 – Référentiel Régional Pédologique de Bourgogne – département de Côte d'Or , programme Inventaire Gestion et Conservation des Sols

Quérin David

Mi-Mai - Juillet 2015

Sous la Direction de Marion Bardy
Référents et Tuteur de Stage : Nicolas Saby et Bertrand Laroche (INRA d'Orléans), Diarra Fall

Stage réalisé à l'Unité InfoSol de l'INRA d'Orléans :
Avenue de la Pomme de Pin, Ardon BP 20619, 45160 OLIVET Cedex

Remerciements

Je tiens à remercier Marion Bardy , pour m'avoir permis de faire ce stage et pour son attention durant toute la durée de cette formation.

Ma gratitude va aussi envers Nicolas Saby et Bertrand Laroche afin de m'avoir assisté dans la réalisation de ce projet, pour les astuces techniques qu'ils ont pu m'apporter. Ainsi que Benoit Bertouy, Jean-Philippe Chenu, ainsi que Diarra Fall pour leurs conseils pertinents.

Mes remerciements vont aussi envers l'equipe d'InfoSol et de Sciences du Sol ainsi que pour Joel Moulin, Sophie Maillant, Jean Villerd et à tous les autres membres de l'équipe avec qui j'ai étroitement collaboré.

Un grand merci tout particulièrement à Magalie et à Florent pour leur générosité depuis mon arrivée dans ce service.

Enfin, je remercie tous les membres de l'INRA qui ont permis mon intégration dans l'Unité d'InfoSol à l'INRA d'Orléans ainsi que pour leur incroyable gentilesse .

SOMMAIRE

1 Introduction :

J'ai effectué mon stage au sein de l'unité INFOSOL de l'INRA d'Orléans. Elle réalise ou coordonne l'acquisition des données nécessaires à la constitution d'un système d'information sur les sols de France. Elle assure le contrôle de la qualité de ces données. Elle crée et alimente les bases de données permettant l'archivage et l'exploitation des informations. Elle contribue à leur diffusion et à leur valorisation en assurant notamment l'articulation avec les bases de connaissances sur les sols et les outils d'exploitation thématique produits par les recherches de l'Institut.

Le Gis Sol

Le **Groupement d'intérêt scientifique Sol (Gis Sol)** a été créé en 2001 pour constituer et gérer un système d'information sur les sols de France et répondre aux demandes des pouvoirs publics et de la société au niveau local et national. Il regroupe le ministère de l'Alimentation, de l'Agriculture et de la Pêche, le ministère de l'Écologie, du Développement Durable et de l'Énergie, l'Institut National de la Recherche Agronomique (Inra), l'Agence de l'Environnement et de la Maîtrise de l'Énergie (ADEME), l'Institut de Recherche pour le Développement (IRD) et l'Institut national de l'information géographique et forestière (IGN). Le Gis Sol conçoit, oriente et coordonne l'inventaire géographique des sols, le suivi de leurs propriétés et l'évolution de leurs qualités, et gère le système d'information sur les sols. Le GIS Sol coordonne plusieurs programmes dont un spécifique sur l'inventaire cartographique des sols IGCS (Inventaire Gestion et Conservation des Sols), qui a pour but de représenter la distribution spatiale des sols sur le territoire français. L'unité InfoSol d'Orléans a pour mission de mettre en place ce programme au niveau national et d'en assurer la gestion et la mise à disposition.
Mon travail s'appuie donc sur les données issues du programme IGCS, et spécifiquement sur un de ses volets. Il comporte plusieurs niveaux de représentation cartographique (nationale, petite région naturelle, parcelle). Le volet prioritaire reste la constitution des Référentiels Régionaux Pédologiques (RRP), définis à l'échelle du 1/250000. Ces référentiels sont structures sous forme de bases de données géographiques permettant la représentation des sols sous forme d'Unités cartographiques (pédopaysages) associées à une base de données sémantique.

Le contexte de la demande, le projet TYPTERRE :

L'ensemble des partenaires IGCS se sont organisés en un Réseau Mixte Technologique appelé « RMT Sols et Territoires ». Ce regroupement permet de mutualiser et valoriser les données au sein de projets communs. C'est dans le cadre d'un projet soutenu par le RMT Sols et Territoire que s'est déroulé mon stage. Ce projet, appelé TYPTERRES, vise à valoriser les RRP produits dans les différents

départements sous la forme d'une typologie agronomique, issue d'un regroupement e type de sols. Elle doit être partagée entre tous les acteurs sur les sols et valorisables au sein d'outils d'aide à la décision.
Les RRP sont aujourd'hui disponible sur un grand nombre de territoire, mais leur valorisation par les acteurs extérieurs à IGCS est aujourd'hui freinée par la multiplicité des types de sols et par la complexité de la structure de la base de données. Ces observations ont conduit les opérateurs IGCS à une réflexion sur une valorisation des RRP spécifique aux demandes des agronomes. De plus des typologies de sols existaient par ailleurs dans certains territoires. Ces réflexions ont poussé les membres du RMT à proposer un projet de regroupement de types de sols en fonction de critères agronomiques dans l'objectif de proposer une typologie « digérée » à partir des informations des RRP et directement exploitables pour les agronomes dans les outils d'aides à la décision.
Cette classification peut se faire de manière experte, mais le RMT souhaite une méthode générique ré-applicable à tous les RRP.
Mon stage intervient à ce niveau afin de proposer une méthode de classification automatique.

Ce travail va s'appuyer sur la notion de type de sol : L'Unité Typologique de Sol (UTS) est un volume de la couverture pédologique (ou du territoire) présentant en tous lieux de l'espace les mêmes propriétés pédologiques.

La carte des types de sol : Une carte des sols est une représentation des types de sols sur un secteur donné. Dans le cadre du programme IGCS cette carte est associée à une base de données au format national DONESOL. Cet outil a été conçu en s'appuyant sur un Système de Gestion de Base de Données Relationnelle et un Système d'Information Géographique afin de gérer et de stocker toutes les informations liées aux cartes par un modèle commun de données pédologiques. Il est à noter qu'historiquement, d'un point de vue terminologique, « DoneSol » désigne tout autant un modèle de données que le système d'information sur les sols de France utilisant ce modèle. Le modèle de donnée Donesol permet aussi de stocker en un endroit unique et de façon harmonisé l'ensemble des études pédologiques d'un territoire. Les données de cette étude peuvent être rapidement réutilisées dans d'autres études. Par exemple, le pédologue peut rapidement retrouver d'anciens profils pédologiques existant sur sa zone d'étude.

Mon travail va s'appuyer sur la base de données du RRP de Côte d'Or.

2 Présentation de l'organisme d'accueil

2.1 L'INRA

L'**Institut national de la recherche agronomique**, plus connu sous le sigle **INRA**, est un organisme français de recherche placé sous la double tutelle du Ministère de la Recherche et du Ministère de l'Agriculture. Cet institut de recherche en agronomie a été fondé en 1946. Il est aujourd'hui le premier institut de recherche agronomique en Europe et le deuxième dans le monde pour ses publications en sciences agricoles et en sciences de la plante et de l'animal.

Il bénéficie de partenariats avec des universités et de grands instituts de recherche scientifique dans le monde. Ses recherches et ses chercheurs sont régulièrement distingués par des prix prestigieux. Le centre INRA Val de Loire mène des recherches autour de quatre pôles:

- Dynamique des sols et de gestion de l'environnement.
- Biologie Animale intégrative et gestion durable des productions animales
- Santé animale et sante publique
- Biologie intégrative des arbres et des organismes associés.

Je faisais partie de l'unité de service d'InfoSol qui est dans le pole Dynamique des sols et gestion de l'environnement.

2.2 L'Unité InfoSol

Elle réalise ou coordonne l'acquisition des données nécessaires à la constitution d'un système d'information sur les sols de France. Elle assure le contrôle de la qualité de ces données. Elle crée et alimente les bases de données permettant l'archivage et l'exploitation des informations. Elle contribue à leur diffusion et à leur valorisation en assurant notamment l'articulation avec les bases de connaissances sur les sols et les outils d'exploitation thématique produits par les recherches de l'Institut.

La figure 1 présente l'organigramme de l'unité :

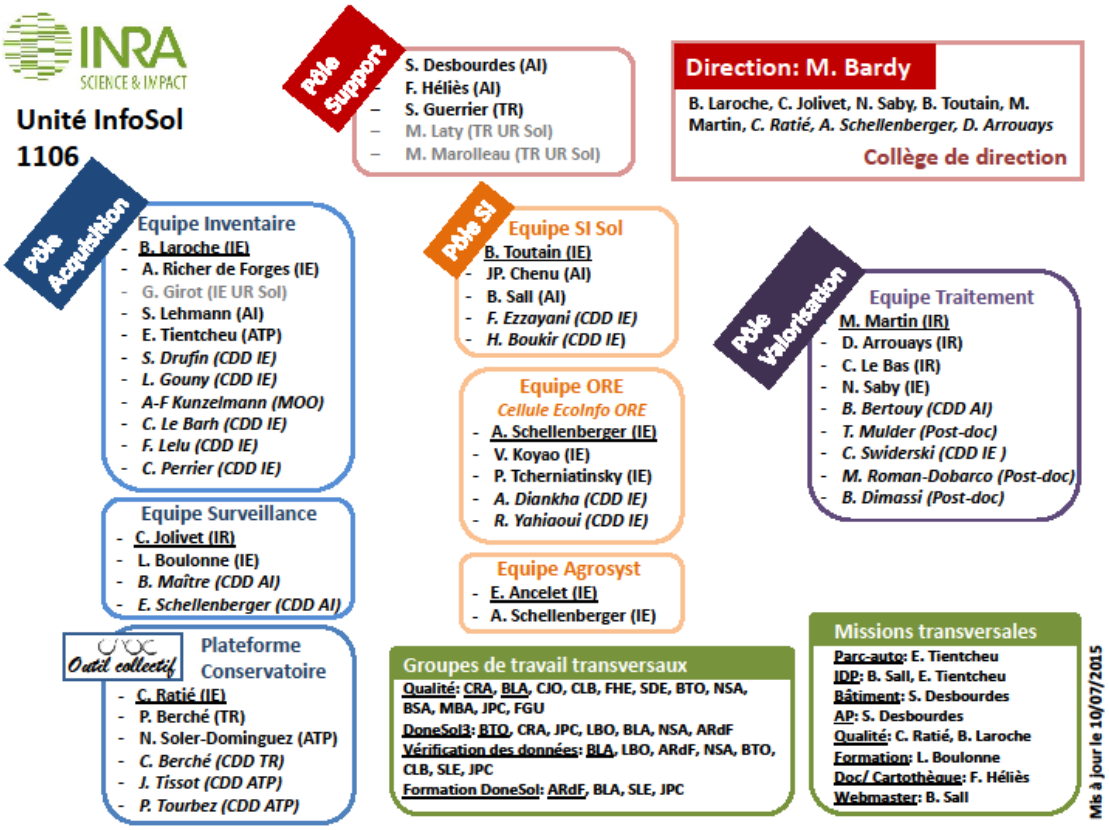


Figure 1 : organigramme de l'unité

2.3 Equipe Traitement

Mon stage s'est déroulé dans le cadre des travaux conduits par une des équipes de l'unité *Infosol* : équipe traitement du pôle valorisation. Ces missions sont principalement la valorisation et le traitement des données acquises dans le cadre du programme du GIS Sol. Durant mon stage, quatre réunions ont eu lieu avec les membres du Pole Valorisation et quelques membres du Pole Acquisition. Le premier étant sur le questionnaire des questions de stockage des données de télédétection et du bilan du congrès EGU. Puis sur la validations des modèles et évaluation de la précision et la troisième sur les vraisemblances et le dernier sur les cartographies de différentes propriétés des sols suivant les spécifications *GlobalSoilMap*

2.4 Copil Typterre

Mon stage s'est déroulé dans le cadre du projet TYPTERRES à l'initiative du RMT « Sols et Territoires ». Un comité de pilotage a été mis en place regroupant des gestionnaires de bases de données, des pédologues, des agronomes, des statisticiens. Il a permis de définir les critères « agronomiques » retenus pour le regroupement des UTS, de définir le choix des méthodes statistiques, d'orienter et de valider mes travaux.

3 Matériels et Méthodes

3.1 Les données de la Côte d'Or

La mise en œuvre des méthodes implémentées dans ce stage repose sur un jeu de données issues d'une base de données décrivant l'organisation des sols dans le département de la côte d'or. Ces données sont issues du programme Inventaire Gestion et Conservation des sols et de son volet 1/250 000ième (RRP). Cette base de données est composée de deux types de données : une couche graphique et des données tabulaires décrivant plusieurs objets. Dans ce stage, 3 objets sont importants à décrire : les Unités Cartographiques de Sols (UCS), les Unités Typologiques de Sols et les Strates (figure 1).

Les UCS = Une Unité Cartographique de Sol est définie comme un morceau de la couverture pédologique. Elle a pour but de représenter l'extension géographique d'un ou de plusieurs types de sol. Elle constitue donc le regroupement d'une ou de plusieurs Unités Typologiques de Sol (UTS) de façon à pouvoir en faire une représentation cartographique à une échelle donnée. (cf dictionnaire *Donesol*)

De façon concrète, cela se traduit par la délimitation de plages cartographiques. Les contours de ces plages cartographiques sont décrits sous la forme de polygones dans l'ensemble « géométrique » de données qui est stocké et géré sous un Système d'Information géographique (SIG). Les UTS est l'Unité Typologique de Sol (UTS) est un volume de la couverture pédologique présentant en tous lieux de l'espace la même succession d'horizons, l'un ou l'autre de ces horizons pouvant être localement absent.

Sans connaître nécessairement l'extension spatiale précise des horizons, il est possible de décrire l'UTS à l'aide des variables qualitatives et quantitatives des horizons identifiés dans les profils et les sondages. L'UTS sera caractérisée par les valeurs représentatives ou modales et la variabilité spatiale sera caractérisée par des valeurs minimales et maximales de ces variables (cf dictionnaire Donesol). Seules, les données modales ont été valorisées durant le stage.

Les strates correspond à l'ensemble des observations ponctuelles (tarière, fossé, talus, fosses) va permettre de définir les types de sol ou Unité Typologique de Sol (UTS), ainsi que leur extension spatiale, et de préciser la variabilité des paramètres pédologiques des horizons qui les composent, et qui forment les strates. Cette variabilité intra-unité est renseignée par l'attribution à chaque paramètre soit d'un intervalle de valeurs s'il s'agit de variables quantitatives, soit d'une valeur modale et de ses valeurs secondaires et mineures dans le cas de variables qualitatives.

Chaque UTS, dans DoneSol, est donc définie par la succession d'une ou de plusieurs strates, et par leur organisation.

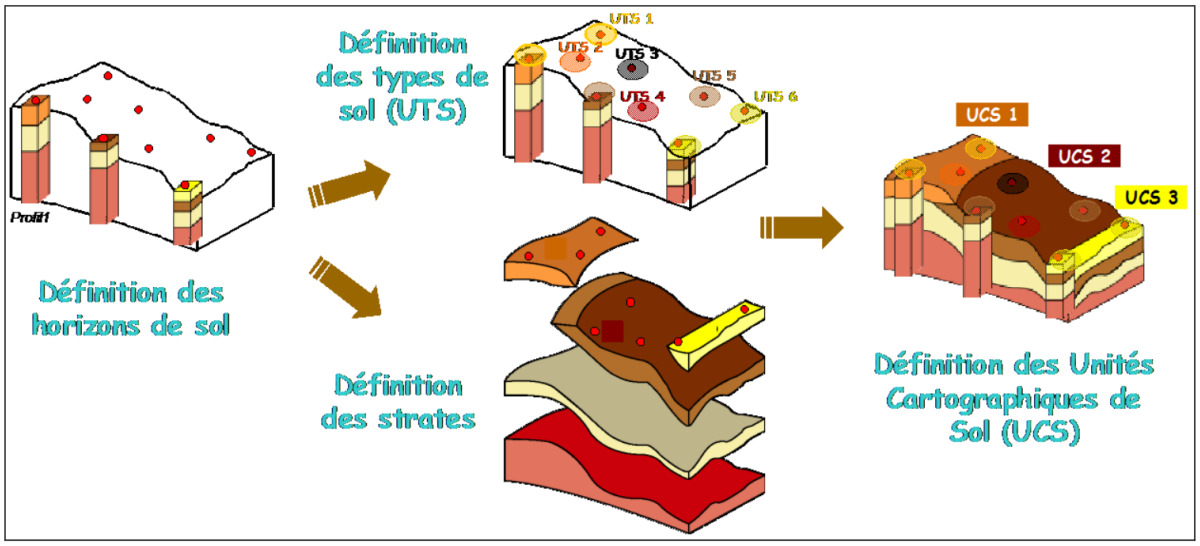


Figure **3-3-1** : organisation des objets dans la base de données manipulées dans ce stage, programme IGCS.

Les données sur ces objets sont organisées dans la base de données selon plusieurs tables. Ces données issues des différentes tables sont exploitées pour construire la typologie. Pour simplifier l'accès aux données, une requête SQL a été développée par un assistant ingénieur de l'équipe SI pour synthétiser l'ensemble des informations nécessaires dans une seule table avec en ligne les individus statistiques et en colonnes les identifiants et les variables descriptives. C'est sur ce dernier tableau que l'analyse a été conduite et qui est décrit dans la suite de ce paragraphe.

Tableau 1: Liste et descriptions des variables décrivant les strates

Variables	Significations du nom des variables	Type de variable	Unité	Nombre de classes regroupées
prof_sol_mod	Profondeur modale du sol		cm	5
taux_argile_sansmeth	Taux d'argile		g/kg	5
taux_sable_sansmeth	Taux de sable		g/kg	4
abondance_eg_sansmeth	Abondance totale des éléments grossiers		%	4
calc_tot_sansmeth	Taux de calcaire		g/kg	5
appar_g_mod	Profondeur modale d'apparition de l'horizon rédoxique	quantitatif	cm	5
appar_gr_mod	Profondeur maximale d'apparition de l'horizon réductique Gr (ou gley)		cm	5

appar_go_mod	**Profondeur modale d'apparition de l'horizon réductique temporaire Go (ou gley réoxydé)**		**cm**	**5**
classe_mat1	**Classe du premier matériau**	**qualitatif**	**absence d'unité**	**0**
drai_nat	**Classe de drainage naturel principale de l'eau au sein du sol**		**absence d'unité**	**0**
org_geol	**Organisation géologique**		**absence d'unité**	**0**
effervescence	**Effervescence**		**absence d'unité**	**0**
abondance_tache_oxy	**Abondance des taches d'oxydation**		**absence d'unité**	**0**

3.2 ACM et Classification Hiérarchique

3.2.1 Mise en contexte du travail :

A partir des paramètres définis par le comité de pilotage, l'objectif du travail est de regrouper ces individus statistiques en tenant compte des variables décrivant ces individus et caractérisant des propriétés du sol. Dans notre cas, ce sont des propriétés agronomiques qui ont été retenues. Les individus statistiques qu'il faut regrouper en classes sont de plus décrits par un ensemble de variables qualitatives et quantitatives. Il convient donc de trouver une méthode de classification qui permet de manipuler ces deux types de variable. Nous avons choisi de tester une classification ascendante hiérarchique sur les composantes principales d'une analyse multivariée. Nous avons retenu, pour l'analyse multivariée, l'analyse des correspondances multiples qui permet d'analyser des variables qualitatives. Les variables quantitatives ont donc été transformées en variables qualitatives en appliquant des règles simples d'expertise. Les règles de transformation sont décrites dans le paragraphe 3.3.1.

3.2.2 Principe de base de l'ACM

L'Analyse des Correspondances Multiples (ACM) est utilisée lorsque 'l'analyse multivariée concerne un jeu de donnée avec des variables qualitatives. Elle a été développée pour l'analyse des tableaux issus de sondage ou de processus d'analyse sensorielle. L'objectif de l'ACM est de mettre en évidence les relations entre les modalités des variables et aussi entre les individus statistiques.

3.2.3 Analyse des Correspondances Multiples

3.2.3.1 Explication théorique

L'ACM, qui est utilisée, est une généralisation de l'Analyse Factorielle des Correspondances (AFC) au cas où l'on observerait plus de deux variables qualitatives. En effet, l'AFC est une analyse destinée au traitement des tableaux de contingences. C'est un tableau d'effectif qui contient à la ligne i et à la colonne k l'individu *cik* (avec *ik* qui est en indice). L'étude de ce tableau se focalise sur la dépendance ou l'indépendance entre les deux caractères. L'ACM s'effectue sur q variables qualitatives. Si nous reprenons l'exemple du questionnaire, un tableau disjonctif complet (TDC) est une matrice de la forme suivante:

$$\mathbf{Z} = [Z1,\ldots,Zj,\ldots,Zq] = \begin{array}{c|ccc} & 1\ldots & s & \ldots m \\ \hline 1 & & & \\ \vdots & & \vdots & \\ i & \ldots & k_{is} & \ldots \\ \vdots & & \vdots & \\ \mathrm{n} & & & \\ \hline & & & \end{array}$$

$$\begin{cases} k_{is} = 1 \text{ si l'individu } i \text{ possède la modalité } s \\ k_{is} = 0 \text{ sinon} \end{cases}$$

Si l'on considère $Zj\,, \forall\, j \in \{1, \ldots, q\}$ comme une matrice issue d'un recodage binaire (0 ou 1) de la *j ième* question. L'ACM consiste à réaliser une AFC du tableau de Burt. Cette table *B* est un croisement deux à deux de toutes les variables c'est-à-dire les tableaux de contingence deux à deux. Elle croise les modalités avec elle-même. Constituée de *m* lignes et *m* colonnes, cette table est une matrice symétrique qui résulte de la multiplication du tableau disjonctif complet et de sa transposée. Elle rassemble les croisements 2 à 2 de toutes les variables :

$$\mathbf{B} = \mathbf{Z}\mathbf{Z}^t = \begin{array}{c|ccc} & 1\ldots & s' & \ldots m \\ \hline 1 & & & \\ \vdots & & \vdots & \\ s & \ldots & b_{ss'} & \ldots \\ \vdots & & \vdots & \\ \mathrm{m} & & & \\ \hline & & & \end{array}$$

3.2.4 Classification Hiérarchique non supervisée

L'objectif de la classification est de constituer des groupes, des classes ou des catégories. Ces nouvelles classes sont des ensembles d'individus qui possèdent des traits de caractères communs, c'est-à-dire que les individus se ressemblent du point de vue de l'ensemble des caractères qui les décrits. Parmi les méthodes de classification, on trouve les méthodes hiérarchiques pour lesquelles on va chercher à construire un arbre hiérarchique. Il est nécessaire de définir une distance entre les individus la plupart du temps il s'agira de la distance euclidienne. Le principe est de regrouper 2 à 2 les individus les plus proches sur le nuage de point des individus c'est-à-dire les plus ressemblants et de créer à chaque étape une partition. Cette classification est représentée par un arbre que l'on appelle un dendrogramme. Cela nécessite d'obtenir les coordonnées des individus afin d'obtenir les distances entres elles ainsi que d'une métrique (distance euclidienne, du chi-deux, distance de Ward…). La distance la plus utilisée est la distance de Ward avec *p* représentant la pondération de l'individu et *G* son barycentre:

$$\Delta(A,B)=\frac{p_A p_B}{p_A+p_B} d^2(G_A,G_B)$$

Elle est la distance entre deux classes est celle de leurs barycentres au carré, pondérée par les effectifs des deux clusters. Elle permet de distinguer plus facilement les classes ce qui n'est pas forcément le cas avec la distance de saut minimale et maximale. En effet le principe de regrouper le nombre de cluster qui baisser le moins l'inertie interclasse. Autrement dit la méthode de Ward permet de séparer au maximum les groupes d'individus.

Les méthodes de classification hiérarchique, construisent des arbres hiérarchiques (ou dendrogramme) afin de représenter l'organisation des individus. Le but de cette classification est de mettre en évidence les liens hiérarchiques entre les individus ou entre des groupes d'individus. Afin d'agréger un individu à un groupe d'individu il est nécessaire d'avoir

Le processus se présente de la manière suivante. Soit les individus A jusqu'à H caractérisant un jeu de donnée. Tout d'abord, nous regardons dans le tableau quel individu est le plus proche de A. Si c'est B alors nous regroupons A et B pour créer un groupe d'individus (AB).Ensuite nous recalculons les distances entres les individus restants et le barycentre du groupe AB crée. On cherche ensuite l'individu le plus proche de ce groupe et ainsi de suite jusqu'à englober tous les individus et ne former qu'un seul groupe

3.2.5 Implémentation de la classification dans *FactoMineR*

Afin d'effectuer cette classification sur les deux premières composantes de l'ACM, nous avons utilisé la commande HCPC du package *FactoMineR*. Nous avons choisi de réaliser la classification non supervisée sans mettre *a priori* de nombre de clusters à regrouper. Pour cela, il faut que le paramètre *nb.clust* soit réglé à -1. Cette outil produit un dendrogramme en utilisant par défaut la méthode de Ward d'utilisé. Il est possible de la modifier en réglant le paramètre *method* par « average » si l'on veut que le regroupement des individus se fasse par rapport à la distance moyenne. Lorsque nous déterminons le nombre de cluster l'algorithme de classification décrit précédemment va itérer ce regroupement jusqu'à arriver au nombre de groupe d'individus voulu. Cela se caractérisera par une coupure de l'arbre avec le nombre de branche coupée. Il faut bien entendu que le nombre de cluster soit inférieur au nombre d'individu que l'on veut classifier. La distance entre deux groupes est déterminée par la hauteur de la branche du dendrogramme. L'inertie représente la dispersion de nos individus. La classification automatisée va déterminer le nombre de cluster optimal c'est-à-dire maximiser l'inertie interclasse ou (variance interclasse) afin que l'on obtienne des groupes homogènes et distincts.

Le paramètre *metric* définit la métrique de notre classification. La distance euclidienne est la métrique définie par défaut.

3.3 Préparation du jeu de donnée

3.3.1 Le regroupement en classe des variables quantitatives

Les variables quantitatives ne peuvent être intégrées en l'état dans le modèle. Elles sont transformées sous forme de variables qualitatives via des regroupements en classes. Cela concerne les variables liées à la profondeur modale d'apparition des strates (*appar_g_mod*), à la profondeur maximale de leur apparition (*appar_gr_mod*) ainsi que la profondeur modale d'apparition réductique (*appar_go_mod).*

Pour les variables quantitatives comme la profondeur modale du sol, le taux de calcaire, de sable, d'argile ou encore pour l'abondance en éléments grossiers nous avons regroupé les valeurs selon les règles suivantes définies par le comité de pilotage :

- La variable *prof_sol_mod* a été regroupé en 5 modalités : 0 à10 cm ,10 à 30cm, 30 à 60 cm, 60 à 90cm, >90cm,
- la variable *abondance_eg_sansmeth* en 4 modalités : 0 - 5%, 5% - 15%, 15% - 30%, >30%,
- *taux_sable_sansmeth* en 4 modalités : <40%, 40% - 60%, 60% - 80%,>80%
- *taux_argile_sansmeth* en 4 modalités : 0%-12%,12%-18%,18%-30%,30%-45%,

- *calc_tot_sansmeth* en 5 modalités : 0%-1%,1%-5%,5%-25%,25%-50%,>50%
- La variable rp_95_ger désigne les types de sols défini précédemment et classé en 3 modalités: *Diff++ (sols différentiés), Par (sols particuliers) et Diff- (sols peu différenciés)*, cf. tableau 2

Tableau 2 regroupement des noms de sols

Classe	Nom	Numéro dans Donesol
Sols particuliers *Par*	COLLUVIOSOL, FLUVIOSOL, THALASSOSOL, SODISOL	34 47-50 141 136-139
Sol très différencié: Diff++	LUVISOL, PLANOSOL, PODZOSOL, PELOSOL	69-75, 99-102, 103-111, 86-92
Sol peu différencié: Diff-	RANKOSOL,…	Tous les autres

(Même s'ils ont un double rattachement, prise en compte en priorité de l'aspect sol très différencié).
Exemple : double rattachement LUVISOL-REDOXISOL -> sol très différencié

Photo 1 : Luvisol : sol très différencié

Photo 2 : Fluviosol: sol particulier

Photo 3 : Rankosol: sol peu différencié

Source: http://global.britannica.com/science/Luvisol
Source : http://solsetpaysages.canalblog.com/archives/2012/09/26/25193577.html

Source : http://solsetpaysages.canalblog.com/archives/2010/01/12/16499815.html

3.3.2 Imputation des valeurs manquantes

Les calculs de l'ACM ne tolèrent pas la présence de valeurs manquantes dans le tableau de données. Il n'est en effet pas possible de calculer le tableau disjonctif complet en l'absence d'une ou plusieurs modalités sur un groupe d'individu. Un algorithme itératif a été développé par Josse et al. (2010) permettant d'estimer les valeurs manquantes à partir d'un nombre à fixer de composantes principales calculées sur les individus ne présentant pas de données manquantes. La méthode d'imputation

utilisée ici par la fonction *imputeMCA* permet d'estimer les valeurs manquantes à partir des données existantes. Cette imputation utilise par défaut soit la méthode régularisée soit l'algorithme EM (Expectation-Maximisation). Cet algorithme itératif dû à Dempster et Rubin est une méthode d'estimation paramétrique notamment utilisée pour le maximum de vraisemblance. Cette commande provenant du package FactoMineR permet d'imputer le tableau disjonctif qui peut ensuite être utilisé dans la fonction MCA. En sortie de la fonction *imputeMCA*, nous obtenons le tableau disjonctif complet imputé ainsi que le jeu de donnée avec les données manquantes qui sont remplacées.

3.4 Algorithme général

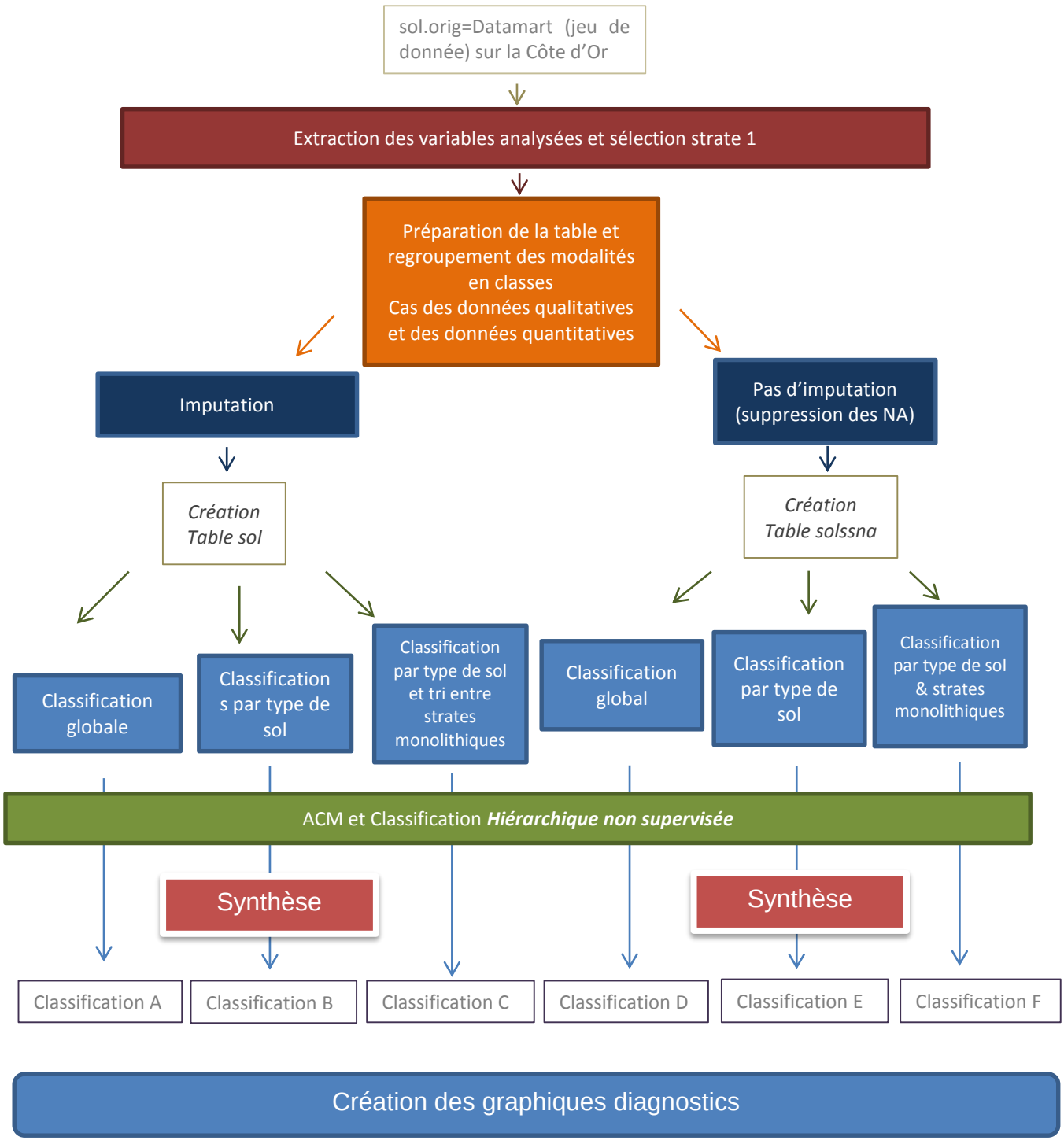
L'algorithme général utilisé pour la production d'une classification des UTS selon la liste des variables retenues lors du comité de pilotage est défini de la manière suivante à partir du datamart c'est-à-dire le jeu de donnée provenant de Donesol:

1. Ne garder que les informations de surface elles ont contenues dans les strates de surface = le champ *no_strate est égal à 1*.
2. Classer les strates selon le tableau 2
3. Créer le tableau sol0 en éliminant les colonnes n'intervenant pas dans la classification hiérarchique non supervisée.
4. 1ère Méthode: Eliminer les lignes de sol0 avec des données manquantes pour produire le tableau *solssna*
5. 2ème Méthode : Imputer les valeurs manquantes de la table sol0 avec 5 composantes pour produire le tableau *sol (cas pour l'imputation avec la commande imputMCA ou les valeurs manquantes sont remplacées)*
   Remarque : les deux tables obtenues ne possèdent plus de valeurs manquantes l'une a été supprimé et l'autre remplacé. (cf. Annexe 1)
6. Effectuer des HCPC sur
   a. le tableau solssna => D
   b. le tableau sol => A (cf. Annexe 2)
   c. les 3 tableaux issus de sol distingués selon le type de sol puis synthétiser les clusters en ajoutant le nom de sol => B (cf. Annexe 3)
   d. les 3 tableaux issus de solssna distingués selon le type de sol puis synthétiser les clusters en ajoutant le nom de sol => E
   e. les 6 tableaux issus de sol distingués selon le type de sol et le type de strate puis synthétiser les clusters en ajoutant le nom de sol => C (cf. Annexe 4)
   f. les 6 tableaux issus de solssna distingués selon le type de sol et le type de strate puis synthétiser les clusters en ajoutant le nom de sol => F
7. Construire les graphiques des ACM, des clusters et de caractérisation des clusters

8. Faire la synthèse des classifications dans le tableau sol0
9. Faire la carte des 6 classifications

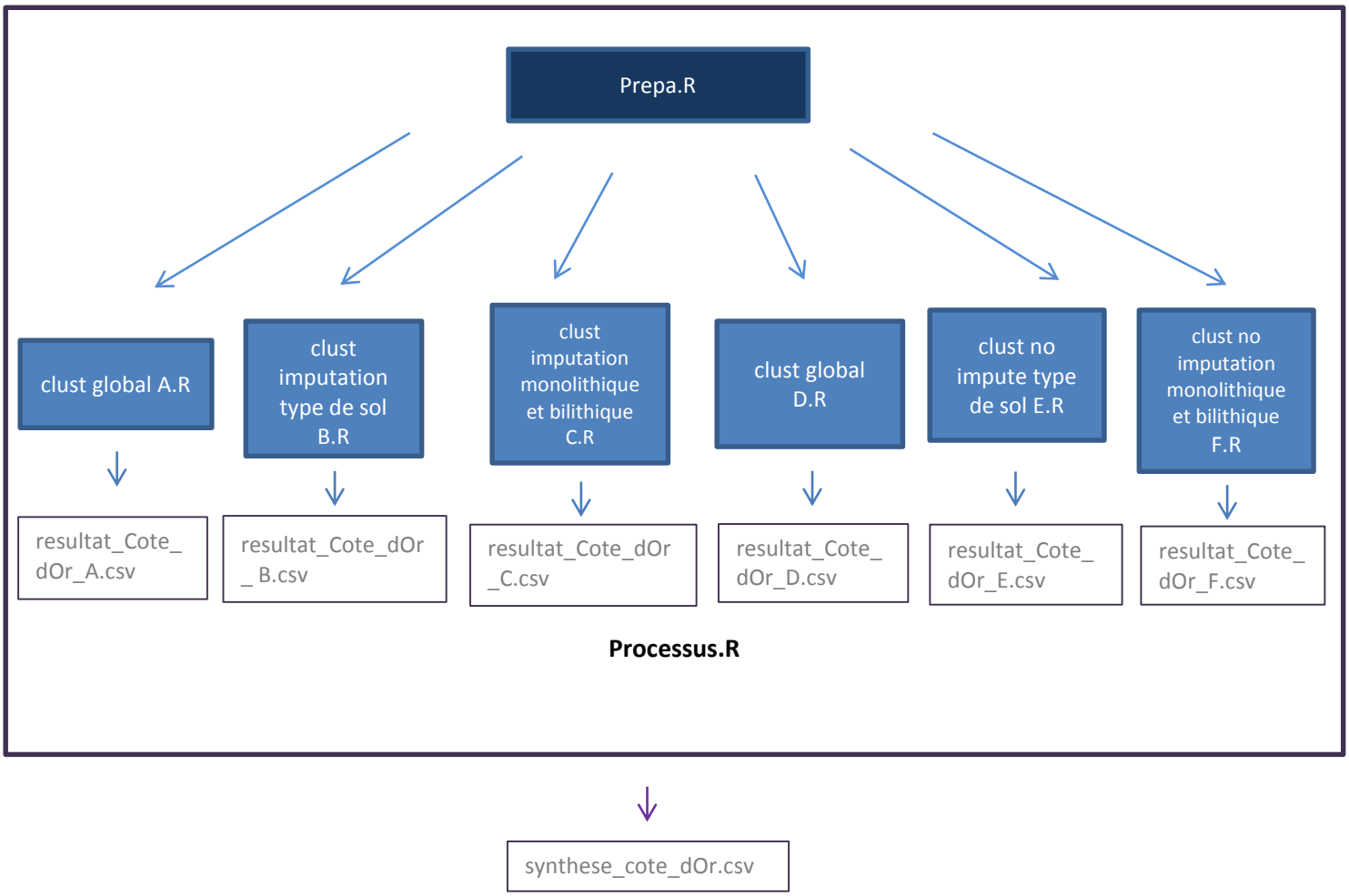
Schématisation de l'algorithme :

Processus utilisé pour la classification



3.5 Scripts et mise en œuvre

Script utilisé pour la classification



Le fichier *Processus.R* nous effectue tout le cheminement afin d'arriver à ces six fichiers csv. Il faut noter que ces fichiers csv nous donnent le jeu de donnée traité mais avec l'ajout d'une colonne cluster en nous explicitant à quels clusters appartiennent ces individus. Les clusters sont numérotés de façon à ce que nous puissions savoir d'où il provient.

Type de sol					
Nom	Sol très différencié	Sol particulier	Sol peu différencié	Sol monolithique	Sol bilithique
Numérotation	1	2	3	1	2

Lorsque qu'un cluster est numéroté 2.1.2 par exemple alors on a affaire au deuxième groupe des types de sols particulières monolithiques.

Le fichier synthese_cote_dor.csv nous permet de regrouper le jeu de donnée traité avec toutes les colonnes cluster affiché à chaque méthode noté de A à F. Cela nous donne donc une comparaison sur la répartition des clusters en fonction de la méthode utilisée. Grace à ce fichier nous pouvons spatialiser nos clusters en affichant leur répartition sur une carte représentant la région étudiée ici la Côte d'Or.

En même temps que la création de ces fichiers csv, nous obtenons les résultats de l'ACM avec les résultats paramètres analysés. Concernant la séparation qui est faite entre les strates monolithiques et bilithique, on les note respectivement 1et 2 dans le titre des fichiers afin que puissions savoir à quoi correspondent ces résultats.

4 Résultats

4.1 Synthèse de la création des modalités

Lors de la séparation par type de sol on constate une faible présence de strates appartenant à des sols très différencié (8.55% des individus soit 46 individus). Leur présence en des sols particuliers est peu présente (15% des individus soit 81 individus) En revanche, pour les sols peu différencié on a une très forte quantité de strate (411 individus sur 538 soit 76.45% des individus)

4.2 ACM

Sur la figure 1 est représenté un nuage de points des individus appartenant aux sols peu différencié et monolithiques sans imputation. Nous constatons que la plupart des individus gravitent autour du centre du plan factoriel et que 6 individus séparent du groupe en longeant le premier axe. Notre représentation du nuage de points nous

permet de capter environ 17.6% de l'information de notre jeu de donnée. Ce pourcentage est assez faible car nous utilisons une ACM. La figure 2 nous montre que le premier facteur est fortement lié à la variable drai_nat (drainage naturel) ainsi que plus légèrement à la variable abondance_tache_oxy, appar_go_mod et appar_gr_mod. Pour le deuxième facteur la classe des matériaux ainsi que le taux d'argile et de sable sont très fortement liés à cet axe. Concernant les axes à conserver le critère de Kaiser nous dit de conserver les facteurs ayant une valeur propre étant supérieur à 1. Cependant nous ne pouvons pas utiliser cette condition car toutes nos valeurs propres sont inférieures à 1. Une deuxième méthode est de chercher un point d'inflexion dans l'éboulis des valeurs propres autrement appelée la règle du coudé. Par conséquent nous allons donc conserver les deux premiers facteurs. La figure 3 nous montre quelle modalité active a discriminé le plus cette individu des autres. En effet, l'individu 113 par exemple se distingue des autres par un drainage très pauvre (drai_nat_7) et un phénomène d'oxydoréduction très marqué à la surface. Concernant les contributions, nous avons que la modalité active appar_go_mod_4 (correspondant aux profondeurs d'apparition de l'horizon réductique inférieur à 40 cm) ainsi que drai_nat_8 (drainage très pauvre) et abondance_tache_oxy_4 (tache d'oxydation nombreuses) contribuent le plus à la construction du premier axe. Pour le second axe, ce sont le taux de sable entre 40 et 60 %, le taux d'argile entre 12% et 18% qui contribuent le plus à sa construction. Ces résultats se retrouvent aussi sur le graphe du carré des liaisons (cf. figure 2).

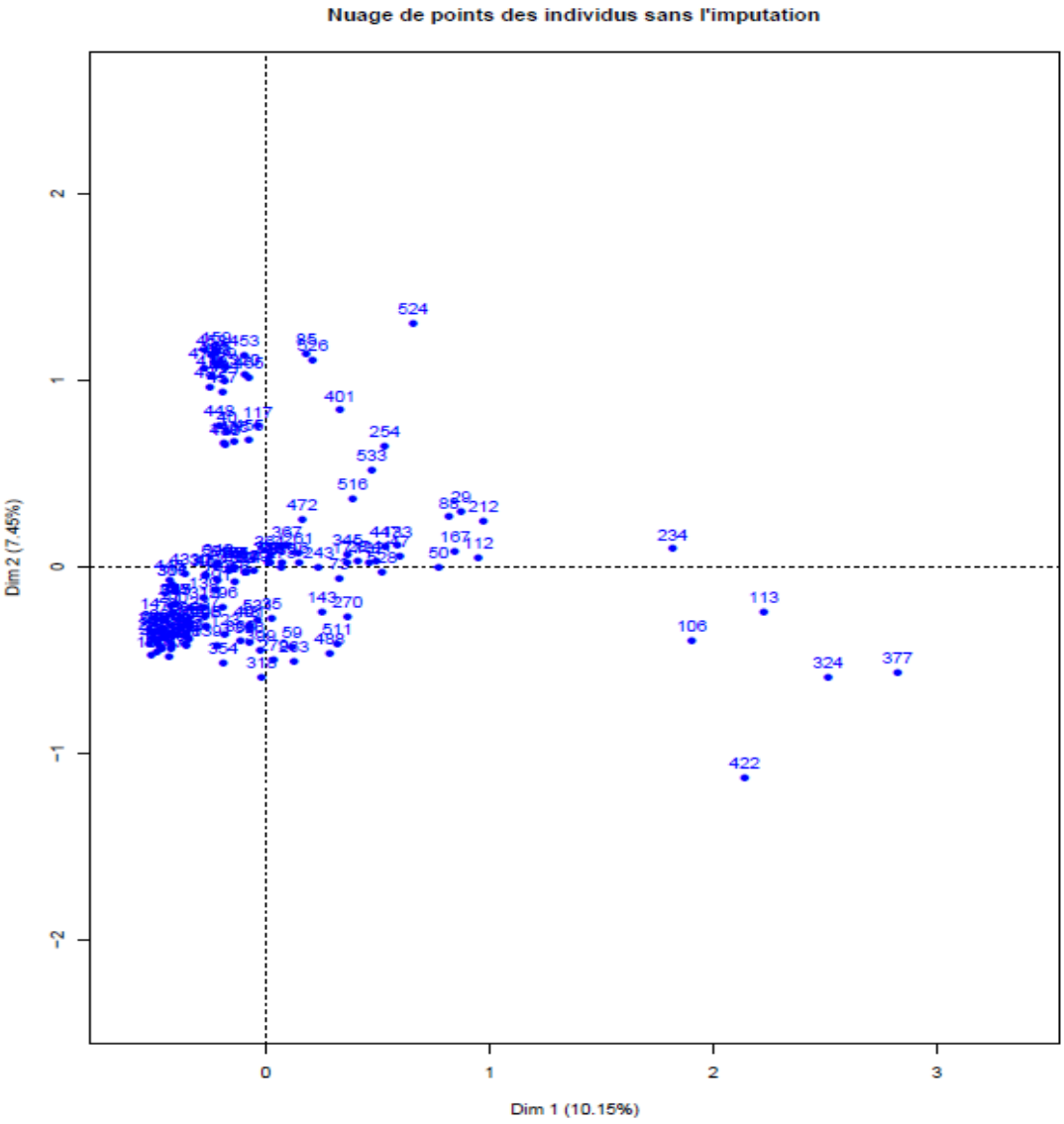


Figure 1: Nuage de points des individus appartenant aux sols peu différenciés sans imputation

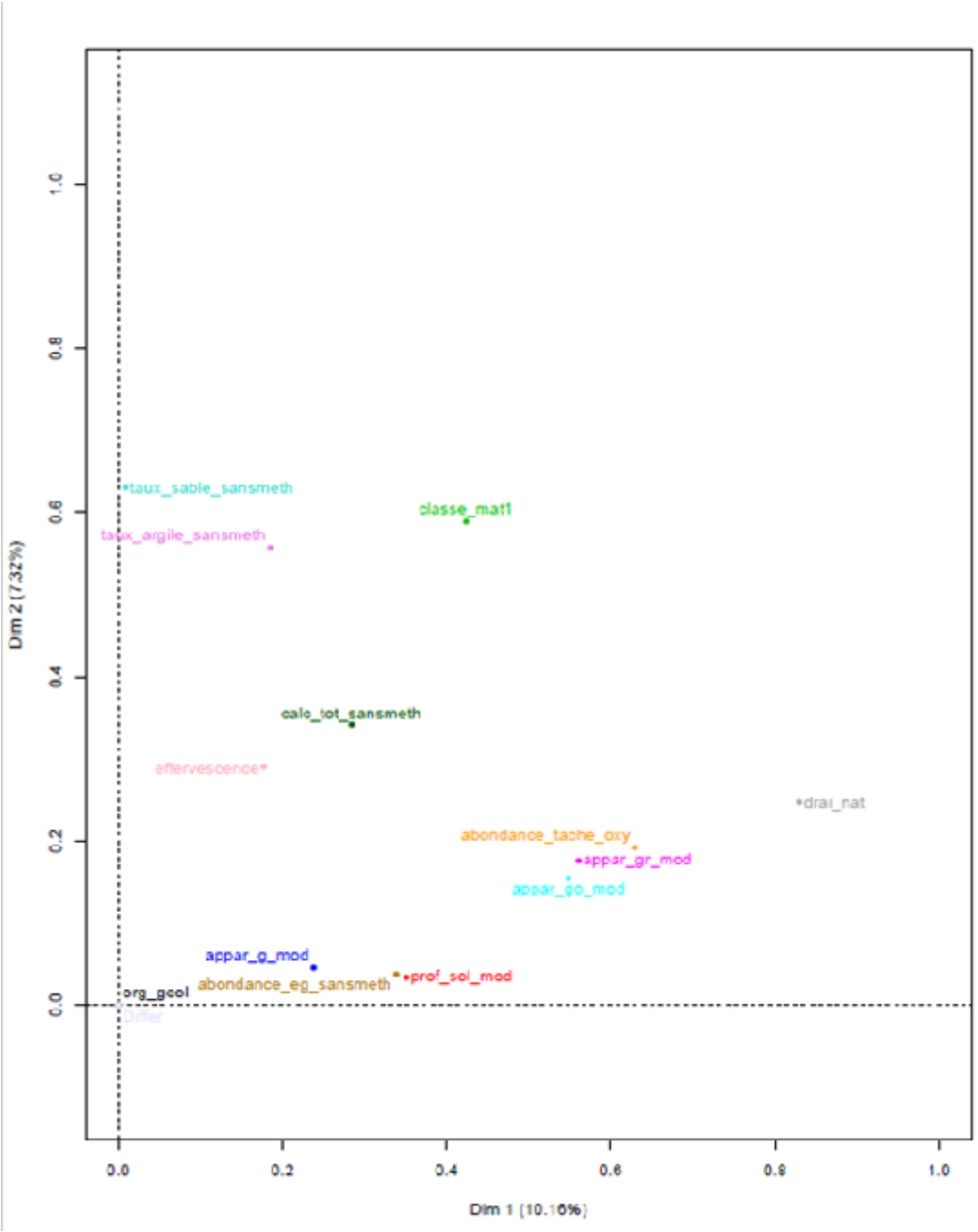


Figure 2: Graphe du carré des liaisons

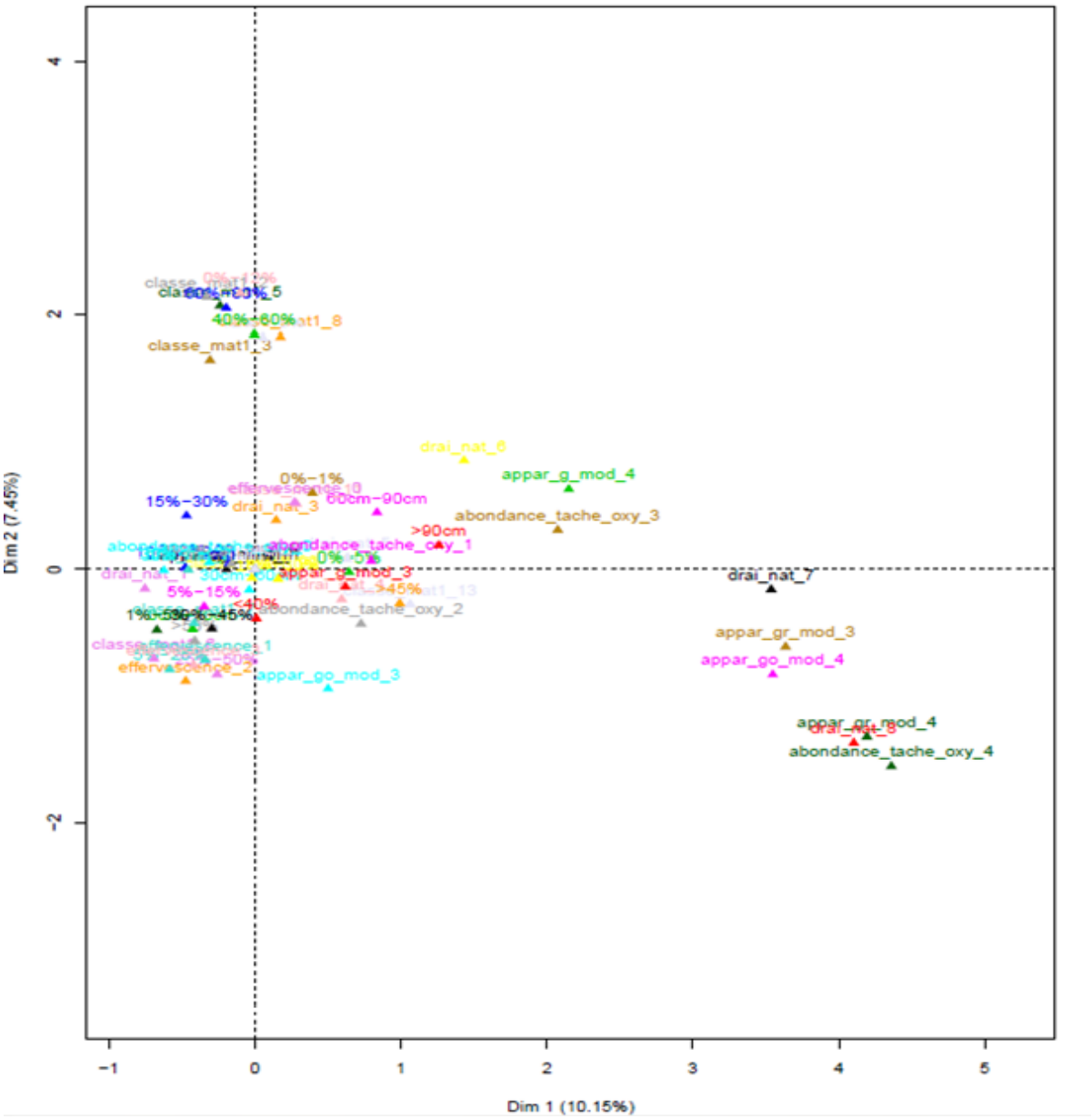


Figure 3: Nuage de points représentant les modalités actives

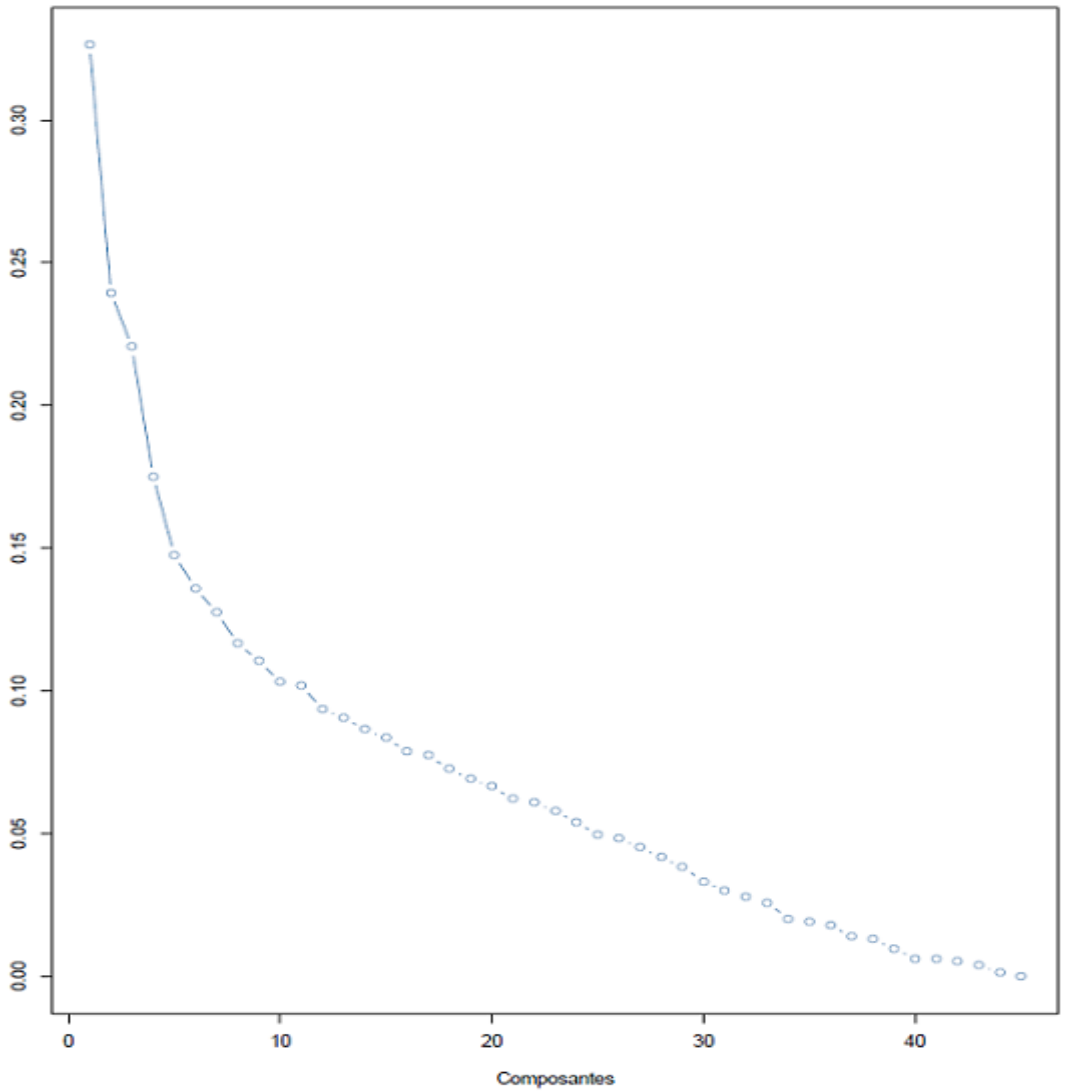


Figure 4 : Eboulis des valeurs propres

4.3 Résultat de la classification

Lors de la classification de nos sols monolithiques et peu différenciés, nos individus se sont formés en quatre classes (figure 5). Le premier groupe se caractérise par la présence d'un drainage très favorable de l'eau (évacuation très rapide de l'eau dans le sol : cf. figure 7) ce qui conduit à une absence d'oxydoréduction (absence de tache) et une très faible présence de sable. Elle est proche de la surface avec une profondeur modale entre 10 et 30cm. Cette classe correspond aux roches sédimentaires cohérentes calcaires (calcaires, craie,…). Le deuxième groupe possède les même caractéristiques que le groupe précédent cependant elle détient un taux de sable relativement élevée (entre 60 à 80 %), un très faible taux de calcaire (0%-1%) ainsi qu'un taux d'argile entre 30% et 45%. Ce groupe englobe en majorité quatre types de roches (32% de roches cristallines grenues, 28% de roches métamorphiques, 20% de roches

sédimentaires cohérentes silico-alumineuses (grès,…) ainsi que 16% de roche volcanique massive). La troisième classe possède un faible niveau de taches d'oxydation et un drainage naturel de l'eau moins favorable. Elle a très peu de sable et d'éléments grossiers et se situe en profondeur (tous >30 cm) ce qui explique la difficulté de l'eau à s'infiltrer. Cette classe correspond aux roches sédimentaires meubles. Le dernier groupe formé de 6 individus plus en retrait que les autres se distingue par la particularité d'avoir un taux de sable très bas (<40 %), une présence d'éléments grossiers très faible aussi ainsi que l'appartenance à la même classe de matériau (classe_mat_13) c'est-à-dire aux roches sédimentaires meubles.

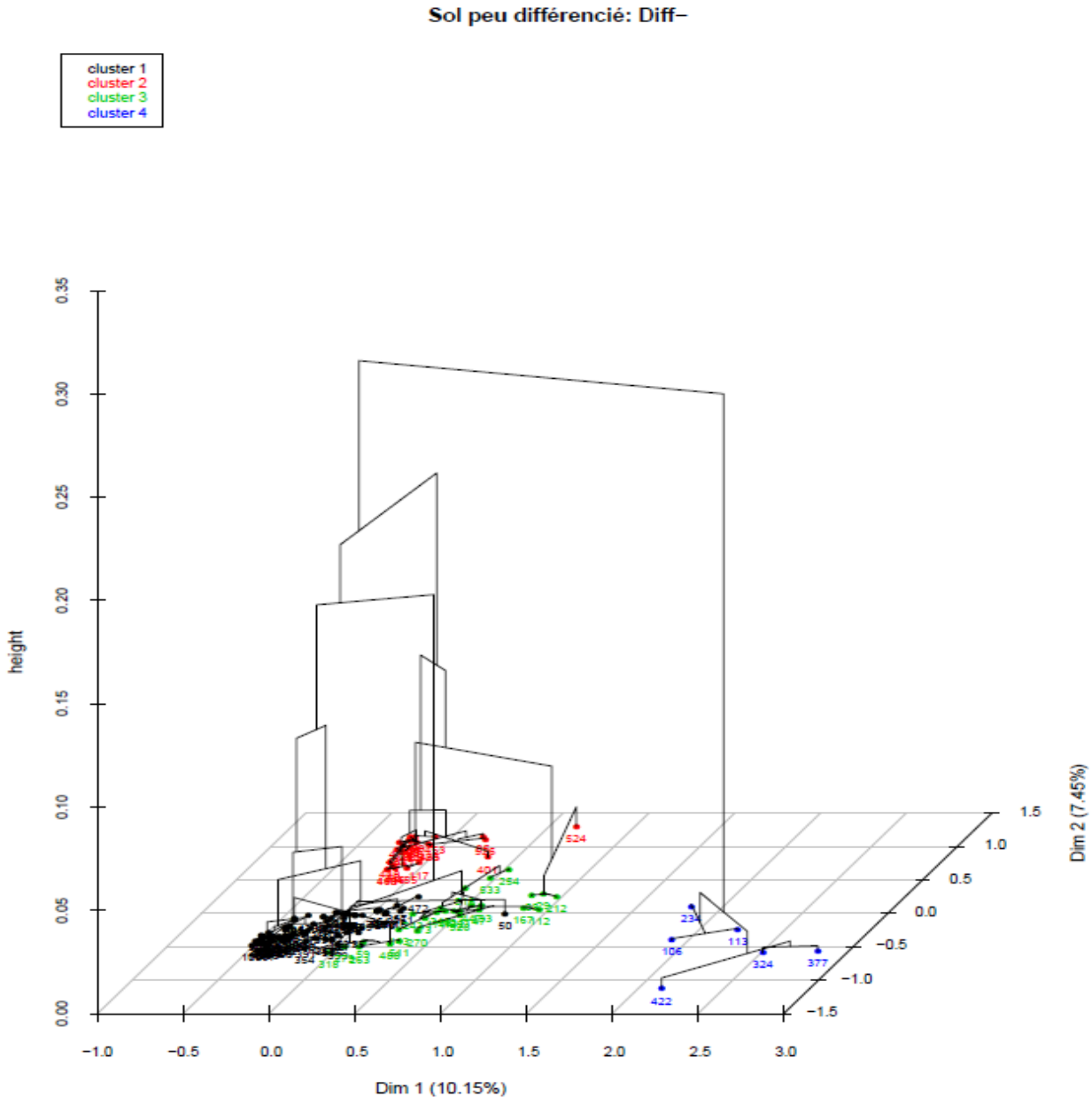


Figure 5 : Représention de l'arbre hiérarchique sur le plan factorielle

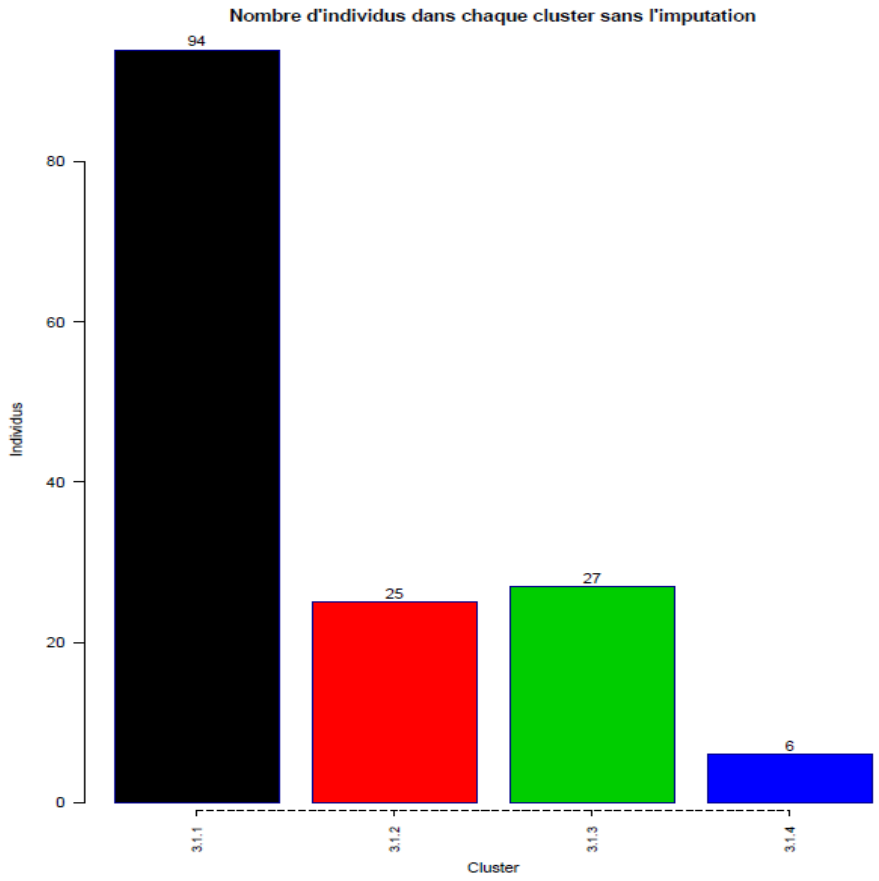


Figure 6 : Répartition des individus dans le groupe des sols peu différenciés

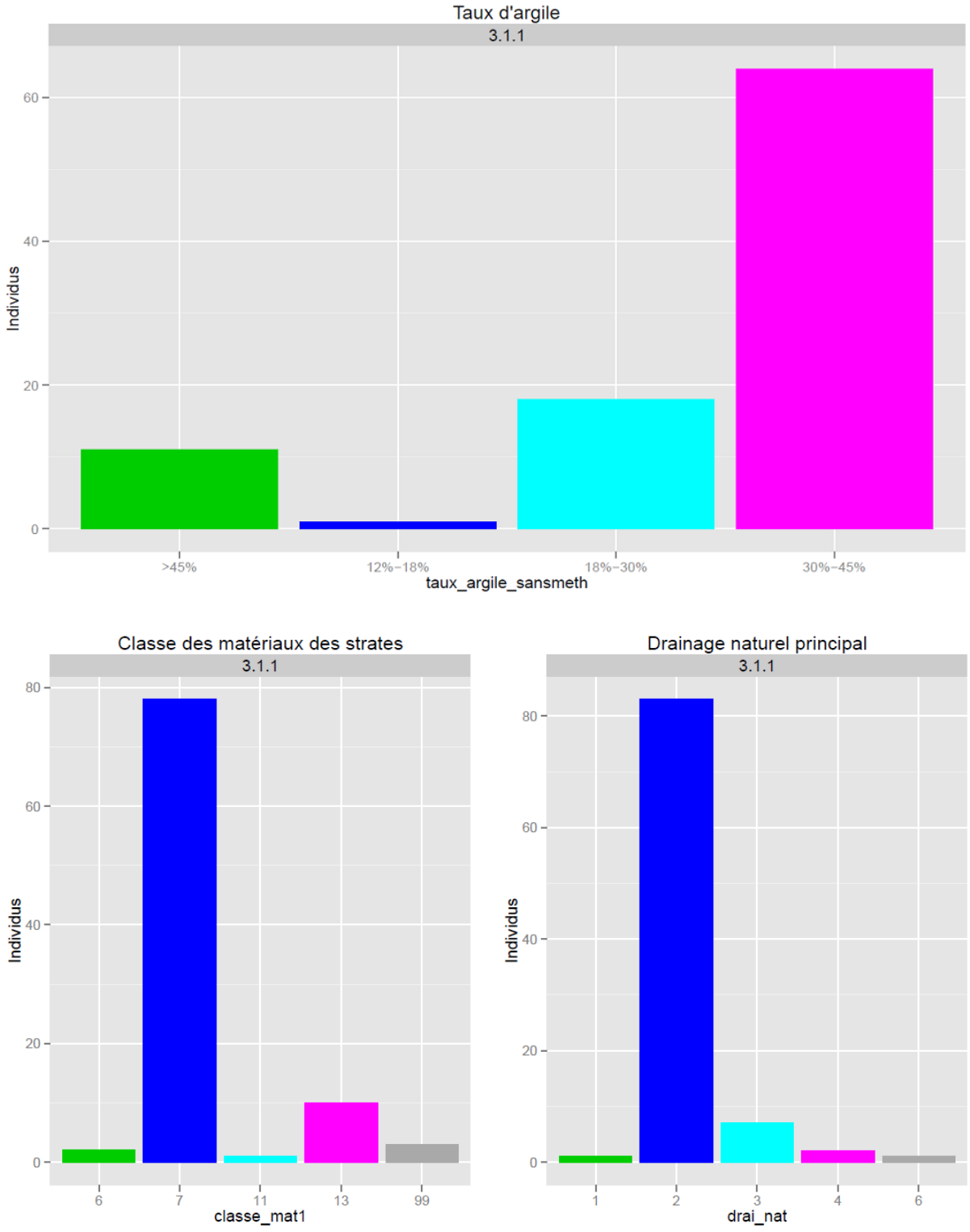


Figure 7: Résultats de la classification sur les variables analysées

4.4 Comparaison des méthodes imputation/no imputation

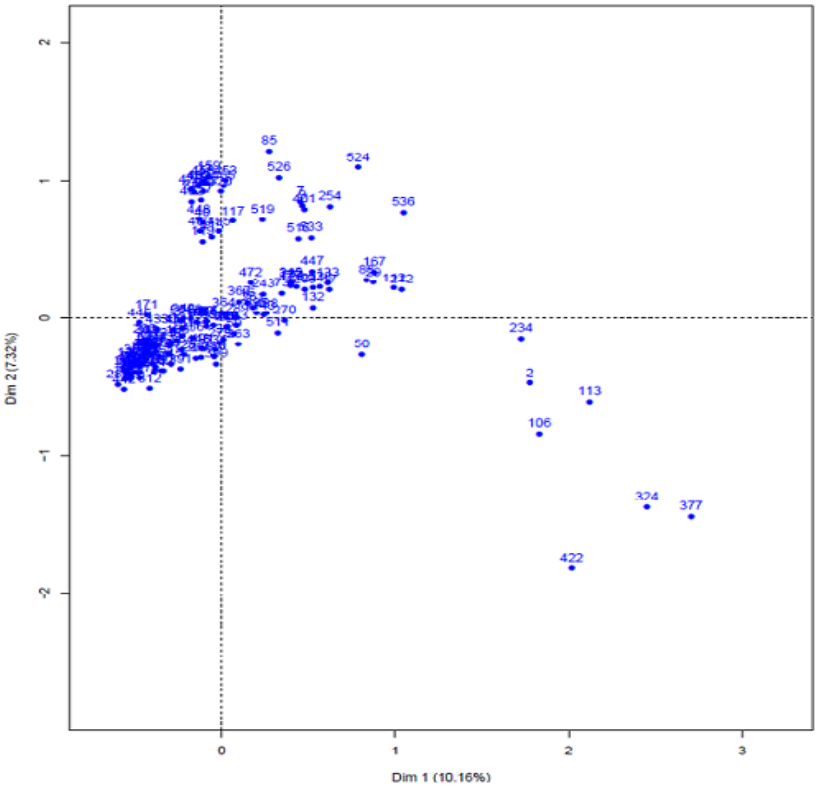


Figure 8 : Nuage de points des individus des sols monolithiques peu différenciés avec l'imputation

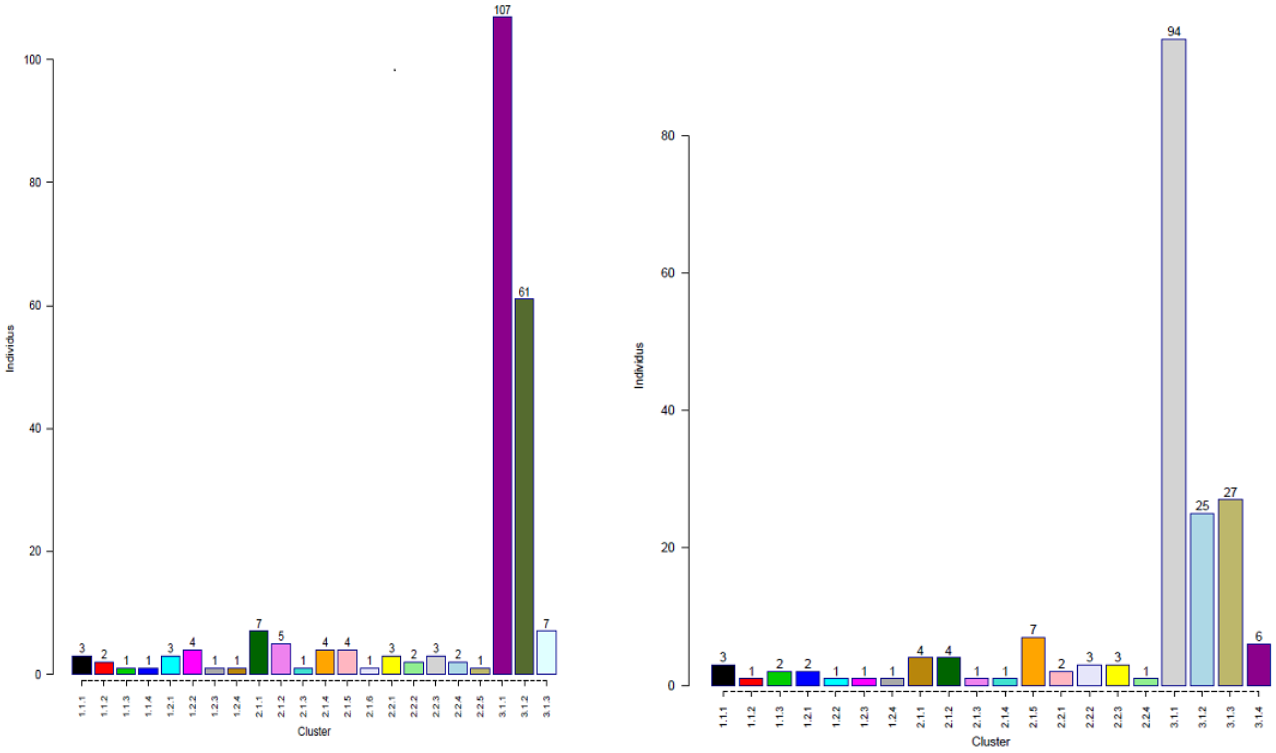


Figure 9: Bilan du nombre d'individu par cluster avec l'imputation (à gauche) et sans imputation (à droite)

Nous constatons à partir de la figure 8 et 9 que l'imputation a permis de regrouper des individus qui étaient dans des classes différentes. En effet, l'ajout des individus manquants a permis de rectifier des classes qui ont été définis. La classe 3.1.1 correspondant au premier groupe des sols peu différenciées et monolithiques est celle qui possèdent le plus d'individus mais qui est aussi la plus stable. Cela nous permet donc d'obtenir une réalisation de la classification qui se rapproche le plus possible de celle avec un jeu de donnée complet.

4.5 Représentation des résultats

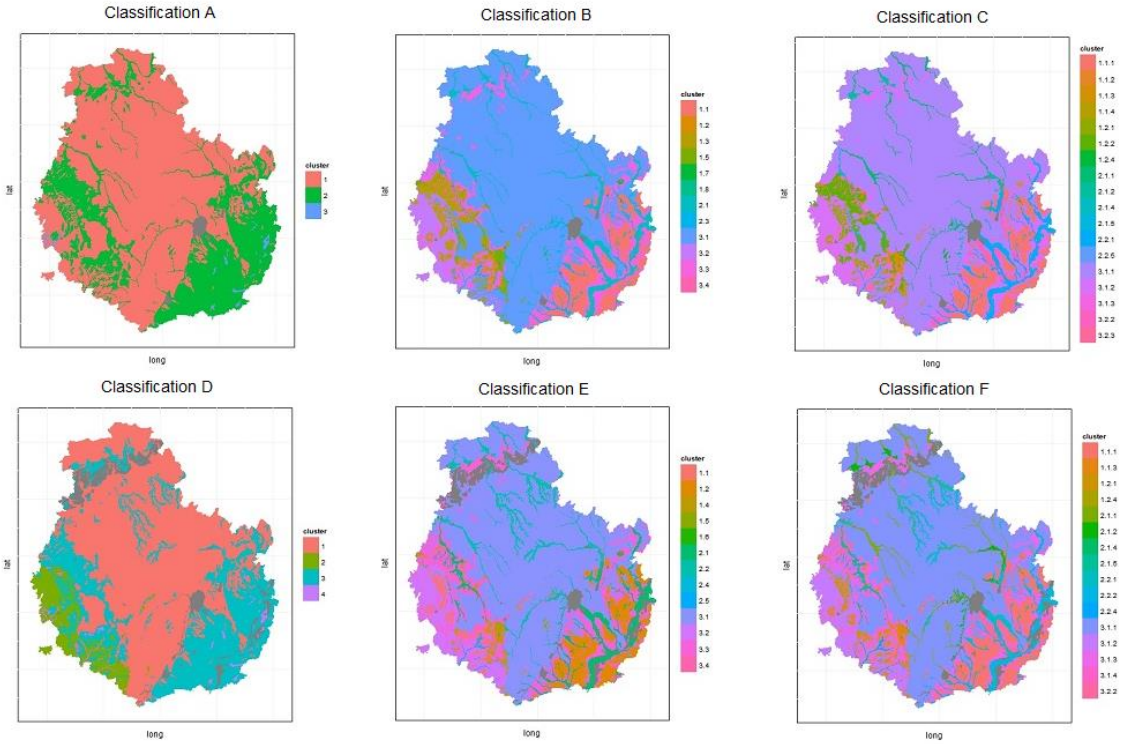


Figure 10: Cartes décrivant la spatialisation de nos clusters

En figure 10, on remarque que la classe 3.1.1 (le premier groupe des sols peu différencié et monolithiques pour la carte C et F) est le groupe le plus stable et ayant le plus d'individus dans les deux méthodes utilisées. On remarque qu'avec l'imputation que la classe 3.1.2 et 3.1.3 furent regroupé dans le groupe 3.1.2. Les cartes A, D sont moins précises que celle des cartes B et E en raison de la stratification faite sur le jeu de données. Ces modifications sont aussi visibles sur les cartes. La partie centrale de la région de la Côte d'Or dans la classification C et F correspondent aux roches sédimentaires calcaires faisant partie des sols peu différenciés et monolithiques. Les sols particuliers de ces deux classifications se retrouvent à proximité des fleuves. Ce sont des roches sédimentaires meubles (galets fluviatiles,…). A l'ouest de la région nous voyons la présence de sols très différenciés et bilithique (en vert sur la classification C) ayant la même classe que celle décrite précédemment. Cependant cette zone est argileuse il s'agit donc d'argile ou encore d'argilites. A proximité de cette zone en violet ainsi qu'au sud-ouest de la région, nous avons des sols peu différenciés monolithiques qui correspondent aux roches cristallines grenues (gabbros, granites,…) mais aussi des roches métamorphiques et volcaniques massives (basaltes, andésites) et des roches sédimentaires silico-alumineuses (grès, conglomérats siliceux). Ces zones correspondent au début du

Morvan qui est un massif des hautes collines. Dans le même secteur nous avons la présence de sols très différenciés monolithiques en orange avec un faible drainage et très profond (>90 cm) ainsi qu'une présence d'argile. Ce lieu correspond aux roches sédimentaires meubles.

La méthode d'imputation (cf. Classification A, B, C de la figure 10) nous permet d'obtenir des cartes avec la nature des strates observées ce qui n'est pas possible sans l'imputation (cf. Classification D, E, F de la figure 10). De plus la présence d'individus supplémentaires a permis de faire apparaitre des groupes qui n'apparaissaient pas forcement lors d'une classification globale ou encore avec une stratification des données sans imputation (classification E). Cependant nous devons souligner le fait que cette analyse n'a capté qu'une faible partie de l'information totale (17.6%) ce qui pourrait remettre en cause l'exactitude de nos informations récoltés. Cela est dû à notre jeu de donnée à traiter d'une nature assez complexe.

5 Conclusion

L'objectif de ce travail était de produire un regroupement d'une vingtaine de classes à partir de nos strates par le biais d'une méthode de partitionnement afin que le jeu de donnée de départ puisse être transformé en plusieurs groupes. Cette transformation permettra aux pédologues de distinguer plus facilement la nature des strates regroupées. Au cours de ce stage de 2 mois et demi, j'ai construit et mis en œuvre un algorithme de traitement de données en vue d'établir une classification sur plusieurs variables. Cette algorithme implémente de la manipulation de données, de l'analyse multivariée avec un analyse multiple des correspondances, un clustering de type classification hiérarchique et une représentation graphique automatique. Ces résultats peuvent ensuite être spatialisé afin d'avoir un point de vue géographique sur la répartition de nos classes. En raison d'une durée de stage de 2 mois et demi la méthode n'a pas pu être validée correctement en l'appliquant sur toutes les régions. Afin de l'appliquer sur toutes les régions de France il faudra un travail préliminaire afin de formaliser les jeu de données plus particulièrement au niveau des noms des variables qui sont différentes en fonction des méthodes appliquées afin d'obtenir nos observations.

Annexe 1 : Préparation de la table

```
library(FactoMineR)
library(ggplot2)
library(missMDA)

source('F:/Stage/Script/Cote dOr/fonctionsStage.R')

sol.orig <- read.table("F:/Stage/Table/25021.csv",header=T,sep="|")

#Transformation des sols differenciés (du rp_95_ger)

diff <- rep("Diff-",times = length(sol.orig$rp_95_ger))

#luvisol: très différencié
diff[(sol.orig$rp_95_ger>=69)&(sol.orig$rp_95_ger<=75)&(is.na(sol.orig$rp_9
5_ger)==F)] <-  "Diff++"
#luvisol: très différencié

#planosol: très différencié
diff[(sol.orig$rp_95_ger>=99)&(sol.orig$rp_95_ger<=102)&(is.na(sol.orig$rp_
95_ger)==F)] <- "Diff++"
#planosol: très différencié

diff[(sol.orig$rp_95_ger>=103)&(sol.orig$rp_95_ger<=112)&(is.na(sol.orig$rp
_95_ger)==F)] <- "Diff++"
#podzosol: très différencié

diff[(sol.orig$rp_95_ger%in%c(123,150))&(is.na(sol.orig$rp_95_ger)==F)] <-
"Diff++"


diff[(sol.orig$rp_95_ger>=47)&(sol.orig$rp_95_ger<=50)&(is.na(sol.orig$rp_9
5_ger)==F)] <- "Par"
#fluviosol: particulier

diff[(sol.orig$rp_95_ger>=136)&(sol.orig$rp_95_ger<=139)&(is.na(sol.orig$rp
_95_ger)==F)] <- "Par"
#sodisol: particulier


diff[(sol.orig$rp_95_ger>=78)&(sol.orig$rp_95_ger<=83)&(is.na(sol.orig$rp_9
5_ger)==F)] <- "Par"
#organosol: particulier

diff[(sol.orig$rp_95_ger%in%c(34,141,70))&(is.na(sol.orig$rp_95_ger)==F)] <
- "Par"

# test
table(diff)
```

```
#pour sadisol : pas present dans rp_95_ger

sol.orig$Differ <- diff

# test
cbind(sol.orig$rp_95_ger,diff)

sol.orig$appar_g_mod <-classe_appar_mod(sol.orig$appar_g_mod)

sol.orig$appar_go_mod <-classe_appar_mod(sol.orig$appar_go_mod)

sol.orig$appar_gr_mod <-classe_appar_mod(sol.orig$appar_gr_mod)

#selection de la colonne no_strate afin de trier les strates 1

strate1 <- split(sol.orig,sol.orig$no_strate)$`1`#selection des individus e
tant lie a la strate 1.

sol.orig <- strate1

#selection des colonnes

v<-0;

c1=c("no_strate","no_uts","no_ucs","surf_unit","etage_geol1","org_geol","cl
asse_mat1","prof_sol_mod","nom_mat","drai_nat","rp_95_ger","Differ","taux_s
able_sansmeth","taux_argile_sansmeth","effervescence","abondance_eg_sansmet
h","abondance_tache_oxy","calc_tot_sansmeth","appar_g_mod",
    "appar_go_mod","appar_gr_mod")

  v <-  which(colnames(sol.orig) %in% c1)

sol.orig <- sol.orig[,v]

#sol.orig$prof_sol_mod <- classe(sol.orig$prof_sol_mod,c(10,30,60,90,90),c(
"0-10","10-30","30-60","60-90",">90",">90")) #Test


seuileg <- c(5,15,30,30,30)

seuilprof <- c(10,30,60,90,90)

seuilcalc <- c(10,50,250,500,500)

seuilsab <- c(400,600,800,800,800)

seuilarg <- c(120,180,300,450,450)

groupeg <-c("0%-5%","5%-15%","15%-30%","30%<","30%<","30%<")

groupcalc <- c("0%-1%","1%-5%","5%-25%","25%-50%",">50%",">50%")

groupsab <- c("<40%","40%-60%","60%-80%","80%<","80%<","80%<")
```

```
groupprof <- c("0cm-10cm","10cm-30cm","30cm-60cm","60cm-90cm",">90cm",">90c
m")

grouparg <-  c("0%-12%","12%-18%","18%-30%","30%-45%",">45%",">45%")

seuil1 <- matrix(c(seuilprof,seuilsab,seuilarg,seuileg,seuilcalc),5)

group1 <- matrix(c(groupprof,groupsab,grouparg,groupeg,groupcalc),6)
#selection des champs à transformer

v<-0;

c1=c("prof_sol_mod","taux_sable_sansmeth","taux_argile_sansmeth","abondance
_eg_sansmeth","calc_tot_sansmeth")

colnames(group1) <- c1

colnames(seuil1) <- c1

for (j in 1:length(c1)){

  v[j] <-  which(colnames(sol.orig)==c1[j])

}

for (i in v){

  sol.orig[,i] <- classe(sol.orig[,i],seuil1[,which(colnames(sol.orig)[i]==
c1)],group1[,which(colnames(sol.orig)[i]==c1)])

}

for (i in 1:ncol(sol.orig)) {
  sol.orig[,i] <- factor(sol.orig[,i])
}

sol0 <- sol.orig

sol <- sol.orig[,-which(colnames(sol.orig) %in% c("no_strate","rp_95_ger"))
]#suppression de la variable no_strate

solssna <- sol[complete.cases(sol),]

#Récupération des données manquantes
complete <- imputeMCA(sol[,-which(colnames(sol)%in%c("no_uts","surf_unit"))
],ncp=5)

sol <- complete$completeObs #table de départ sol.orig complété
sol$no_uts <-  sol.orig$no_uts
sol$surf_unit <- sol.orig$surf_unit

sol <- sol[,colnames(solssna)]
```

Annexe 2: Classification A: Classification Globale par imputation

```
clust_global_impute_A <- function(dir_prepa,dir_setwd,titre_pdf,dir_resulta
t){

source(dir_prepa)#traitement de notre jeu de donnée issue du chemin dir_pre
pa amenant au fichier

resu <- sol # table faisant le bilan de la répartition des strates (crée po
ur avoir nos résultats en fichier csv).

no_strate <- sol.orig$no_strate

resu <- cbind(no_strate,resu)#regroupement du jeu de donnée avec la variabl
e no_strate en raison de sa suppresion pour l'imputation

resu <- resu [,-which(colnames(resu)=="Differ")] # faisant du clustering gl
obal il est inutile exprimer la variable Differ car aucune stratification n
'est effectué

resu$cluster <- "1" #création de la colonne cluster avec des 1 placé par dé
faut

clustfinal <- c() #création de clustfinal pour le bilan des cluster vide pa
r défaut

  # Mise à jours des modalités

  for (l in 1:ncol(sol)) {
    sol[,l] <- factor(sol[,l])
  }

  table1 <- sol[,-which(colnames(sol) %in% c("no_uts","no_ucs","etage_geol1
","no_strate","nom_mat","no_strate","surf_unit"))]#suppression des variable
s ayant trop de modalités
  #pour l'analyse multivariée

  res.mca <- MCA(table1,graph=F)# ACM

  res.hcpc <- HCPC(res.mca,nb.clust=-1,graph=F)#classification hiérarchique
sur les résultats de l'ACM

  table1$cluster <- paste(res.hcpc$data.clust$clust)#création d'une colonne
cluster sur la table analysée avec ses résultats obtenus

  resu$cluster <- table1$cluster#remplacement de la colonne cluster obtenu
dans la table resu

  clustfinal <- c(clustfinal,table1$cluster) # sauvegarder les clusters

  setwd(dir = dir_setwd)#définition de la direction de notre espace de trav
```

```
ail

  pdf(paste(titre_pdf,"pdf",sep="."),width = 9.45,height = 12.6)#ouvrir un
pdf en mode d'écriture A4 (24*32) 1cm = 0.3937 inches

  plot.MCA(res.mca,choix="var",col.var=1:ncol(table1),title="Représentation
des variables :Graphe du carré des liaisons")#explication dans le titre

  plot.MCA(res.mca,choix="ind",invisible="var",title="Nuage de points des i
ndividus avec l'imputation")#idem

  plot.MCA(res.mca,invisible="ind",col.var=1:ncol(table1),title="Nuage de p
oints des modalités actives")#idem

  plot.HCPC(res.hcpc,title="Classification hierarchique avec l'imputation")
#idem

  plot((res.mca$eig[,1]),type="b",col="steelblue",main="Eboulis des valeurs
propres",xlab="Composantes",ylab="Valeurs Propres")#idem

  par(xaxt = "s")
  par(las=2)
  par(mar = par("mar") + c(4, 1, 0, 0))
  barplot(res.mca$var$contrib[,1],col="indianred",main="Contributions des v
ariables pour le 1er axe",border="dark red",axis.lty=5,
          xlab="",ylab="Contributions (en %)",beside=T,cex.names=0.8)#idem

  par(xaxt = "s")
  par(las=2)
  par(mar = par("mar") + c(4, 1, 0, 0))
  barplot(res.mca$var$contrib[,2],col="brown",main="Contributions des varia
bles pour le 1er axe",border="dark red",axis.lty=5,
          xlab="",ylab="Contributions (en %)",beside=T,cex.names=0.8)#idem


  plot((res.mca$var$cos2[,1]),type="b",col="steelblue",main="Cos 2 des vari
ables pour le 1er axe",xlab="Modalités",ylab="Cos2")#idem

  plot((res.mca$var$cos2[,2]),type="b",col="indianred",main="Cos 2 des vari
ables pour le 2ème axe",xlab="Modalités",ylab="Cos2")#idem

  #début création du bilan du nombre d'individus par cluster

  abs <- barplot(table(table1$cluster),plot=F)

  colnames(abs) <- "x"

  barplot(table(table1$cluster),col=1:length(table1$cluster),main="Nombre d
'individus dans chaque cluster",border="dark blue",cex.names = 0.8,axis.lty
=5,xlab="Cluster",ylab="Individus")
  par(xpd=T)
  text(cbind(abs,table(table1$cluster)),labels=table(table1$cluster),pos=3,
offset=0.2)
```

```
  #fin création du bilan du nombre d'individus par cluster

  # début boucle analyse des clusters par groupe
  for (j in 1:max(as.numeric(levels(res.hcpc$data.clust$clust)))){

    tableclusti<- table1[table1$cluster==paste(j),]

    Analyse_clust(tableclusti)

  } # fin boucle analyse des clusters par groupe

  dev.off()

  setwd(dir = dir_resultat)#precision du chemin pour obtenir le fichier csv

  return(resu)#retour de la fonction afin de la retrouver pour le fichier P
rocessus.R

}

resu <- clust_global_impute_A('F:/Stage/Script/Cote dOr/Prepa.R',"F:/Stage/
output/Cote dOr/clust global"," Global Clustering A","F:/Stage/output/Cote
dOr/Resultats")#Application de la fonction clust_global_impute_A
#sur le jeu de donnée de la Cote d'Or

write.csv(resu,file="resultat_Cote_dOr_A.csv",row.names =  FALSE)#création
du fichier csv avec les résultats (jeu de donnée traité et résultats des cl
usters)
```

Annexe 3: Classification B: par type de strate et imputation

```
clust_impute_B <- function(dir_prepa,dir_setwd,dir_resultat){

#Analyse sur toutes les classes (pas de tri entre les strates monolithiques
et bilithiques)

differ <- c("Diff++","Par","Diff-")

t <- c("Sol très différencié: Diff++","Sol particulier: Par","Sol peu diffé
rencié: Diff-")#definition du vecteur t pour les titres de l'HCPC

t1 <- c("Sol très différencié","Sol particulier","Sol peu différencié")#def
inition du vecteur t1 pour les titres du pdf

source(dir_prepa)#traitement de notre jeu de donnée issue du chemin dir_pre
pa amenant au fichier

resu <- sol # table faisant le bilan de la répartition des strates (crée po
ur avoir nos résultats en fichier csv).

no_strate <- sol.orig$no_strate

resu <- cbind(no_strate,resu)#regroupement du jeu de donnée avec la variabl
e no_strate en raison de sa suppresion pour l'imputation

resu$cluster <- "1" #création de la colonne cluster avec des 1 placé par dé
faut

clustfinal <- c() #création de clustfinal pour le bilan des cluster vide pa
r défaut

for( i in 1: length(differ)){

  table1 <- sol[sol$Differ == differ[i],] #extraction des  strates soit très
différencié particulier ou peu différencié (type de sol)

  # Mise à jours des modalités

  for (l in 1:ncol(table1)) {
    table1[,l] <- factor(table1[,l])
  }

  # suppression de la colonne no_uts

  table1 <- table1[,-which(colnames(table1) %in% c("no_uts","no_ucs","etage
_geol1","no_strate","nom_mat","surf_unit"))]#suppression des variables ayan
t trop de modalités

  res.mca <- MCA(table1,graph=F)# ACM
```

```
  res.hcpc <- HCPC(res.mca,nb.clust=-1,graph=F)#classification hiérarchique
sur les résultats de l'ACM


  table1$cluster <- paste(i, res.hcpc$data.clust$clust,sep=".")#création d'
une colonne cluster sur la table analysée avec ses résultats obtenus

  resu$cluster[sol$Differ == differ[i] ] <- table1$cluster#remplacement de
la colonne cluster obtenu dans la table resu

  clustfinal <- c(clustfinal,table1$cluster) # sauvegarder les clusters

  setwd(dir = dir_setwd)#definition de la direction de notre espace de trav
ail

  pdf(paste(t1[i],"pdf",sep="."),width = 10,height = 12.6)#ouvrir un pdf en
mode d'ecriture A4 (24*32) 1cm = 0.3937 inches
  # graphique de la carte factorielle

  plot.MCA(res.mca,choix="var",col.var=1:ncol(table1),title="Representation
des variables :Graphe du carré des liaisons")

  plot.MCA(res.mca,choix="ind",invisible="var",title="Nuage de points des i
ndividus avec l'imputation")

  plot.MCA(res.mca,invisible="ind",col.var=1:ncol(table1),title="Nuage de p
oints des modalités actives")

  plot.HCPC(res.hcpc,title="Classification hierarchique avec l'imputation")
  plot((res.mca$eig[,1]),type="b",col="steelblue",main="Eboulis des valeurs
propres",xlab="Composantes",ylab="Valeurs Propres")

  par(xaxt = "s")
  par(las=2)
  par(mar = par("mar") + c(4, 1, 0, 0))
  barplot(res.mca$var$contrib[,1],col="indianred",main="Contributions des v
ariables pour le 1er axe",border="dark red",axis.lty=5,
          xlab=" ",ylab="Contributions (en %)",beside=T,cex.names=0.8)

  par(xaxt = "s")
  par(las=2)
  par(mar = par("mar") + c(4, 1, 0, 0))
  barplot(res.mca$var$contrib[,2],col="brown",main="Contributions des varia
bles pour le 1er axe",border="dark red",axis.lty=5,
          xlab=" ",ylab="Contributions (en %)",beside=T,cex.names=0.8)

  plot((res.mca$var$cos2[,1]),type="b",col="steelblue",main="Cos 2 des vari
ables pour le 1er axe",xlab="Modalités",ylab="Cos2")

  plot((res.mca$var$cos2[,2]),type="b",col="indianred",main="Cos 2 des vari
ables pour le 2ème axe",xlab="Modalités",ylab="Cos2")

  #début création du bilan du nombre d'individus par cluster
```

```
  abs <- barplot(table(table1$cluster),plot=F)

  colnames(abs) <- "x"
  barplot(table(table1$cluster),col=1:length(table1$cluster),main="Nombre d
'individus dans chaque cluster",border="dark blue",cex.names = 0.8,axis.lty
=5,
          xlab="Cluster",ylab="Individus")
  par(xpd=T)
  text(cbind(abs,table(table1$cluster)),labels=table(table1$cluster),pos=3,
offset=0.2)

  #fin création du bilan du nombre d'individus par cluster

  # début boucle analyse des clusters par groupe

  for (j in 1:max(as.numeric(levels(res.hcpc$data.clust$clust)))){

    tableclusti<- table1[table1$cluster==paste(i,j,sep="."),]

    Analyse_clust(tableclusti)
  } # fin boucle analyse des clusters par groupe
  #début création du bilan du nombre d'individus par cluster sur toutes les
analyses
  if(i==3){
    abs <- barplot(table(clustfinal),plot=F)
    colnames(abs) <- "x"
    barplot(table(clustfinal),col=1:length(clustfinal),main="Bilan du nombr
e d'individus dans chaque cluster",border="dark blue",cex.names = 0.8,axis.
lty=5,xlab="Cluster",ylab="Individus")
    par(xpd=T)
    text(cbind(abs,table(clustfinal)),labels=table(clustfinal),pos=3,offset
=0.2)
  }
  dev.off()
  #fin création du bilan du nombre d'individus par cluster sur toutes les a
nalyses
}
setwd(dir = dir_resultat)#précision du chemin pour obtenir le fichier csv
return(resu)#retour de la fonction afin de la retrouver pour le fichier Pro
cessus.R
}
resu <- clust_impute_B('F:/Stage/Script/Cote dOr/Prepa.R',"F:/Stage/output/
Cote dOr/Imputation","F:/Stage/output/Cote dOr/Resultats")#Application de l
a fonction clust_impute_B
#sur le jeu de donnée de la Cote d'Or


write.csv(resu,file="resultat_Cote_dOr_B.csv",row.names= FALSE)#création du
fichier csv avec les résultats (jeu de donnée traité et résultats des clust
ers)
```

Annexe 4: Classification C: par type de strate et type de sol et imputation

```
clust_impute_C <- function(dir_prepa,dir_setwd,dir_resultat){

#Extraction et Analyse des matrices contenant les strates monolithique ou b
ilithiques ainsi que leurs différenciations (tdiff,par,pdiff)

differ <-c("Diff++","Par","Diff-")

t <- c("Sol très différencié: Diff++","Sol particulier: Par","Sol peu diffé
rencié: Diff-")#definition du vecteur t pour les titres de l'HCPC

t1 <- c("Sol très différencié","Sol particulier","Sol peu différencié")#def
inition du vecteur t1 pour les titres du pdf

source(dir_prepa)#traitement de notre jeu de donnée issue du chemin dir_pre
pa amenant au fichier

resu <- sol # table faisant le bilan de la répartition des strates (crée po
ur avoir nos résultats en fichier csv).

no_strate <- sol.orig$no_strate

resu <- cbind(no_strate,resu)#regroupement du jeu de donnée avec la variabl
e no_strate en raison de sa suppresion pour l'imputation

resu$cluster <- "1" #création de la colonne cluster avec des 1 placé par dé
faut

clustfinal <- c() #création de clustfinal pour le bilan des cluster vide pa
r défaut


for( i in 1: length(differ)){

  for (k in 1:2){

    tables <- sol[sol$org_geol == k,] #table avec les strates monolithiques
ou bilithiques

    table1 <- tables[tables$Differ == differ[i],] #extraction des strates s
oit monolithiques soit bilithiques (type de strates)
    #mais aussi très différencié particulier ou peu différencié (type de so
l)

    # Mise à jours des modalités

    for (l in 1:ncol(table1)) {

      table1[,l] <- factor(table1[,l])
```

```
    }


    table1 <- table1[,-which(colnames(table1) %in% c("no_uts","etage_geol1"
,"no_strate","no_ucs","nom_mat","surf_unit"))] # suppression de la colonne
no_uts, etage_geol1,
    #no_strate, nom_mat,surf_unit,no_ucs,surf_unit(trop de modalités)

    res.mca <- MCA(table1,graph=F)#ACM sur la table

    res.hcpc <- HCPC(res.mca,nb.clust=-1,graph=F)#classification hiérarchiq
ue sur les résultats de l'ACM

    table1$cluster <- paste(i,k, res.hcpc$data.clust$clust,sep=".")

    resu$cluster[(sol$org_geol == k) & (sol$Differ == differ[i])] <- table1
$cluster

    clustfinal <- c(clustfinal,table1$cluster) # sauvegarder les clusters

    setwd(dir = dir_setwd)#définition de la direction de notre espace de tr
avail

    pdf(paste(t1[i],k,"pdf",sep="."),width = 10,height = 12.6)#ouvrir un pd
f en mode d'ecriture A4 (24*32) 1cm = 0.3937 inches
    # graphique de la carte factorielle

    plot.MCA(res.mca,choix="var",col.var=1:ncol(table1),title="Representati
on des variables: Graphe du carré des liaisons")

    plot.MCA(res.mca,choix="ind",invisible="var",title="Nuage de points des
individus avec l'imputation")

    plot.MCA(res.mca,invisible="ind",col.var=1:ncol(table1),title="Nuage de
points des modalités actives")

    plot.HCPC(res.hcpc,title=t[i])# représentation Classification hierarchi
que sur le plan factorielle

    plot((res.mca$eig[,1]),type="b",col="steelblue",main="Eboulis des valeu
rs propres",xlab="Composantes",ylab="Valeurs Propres")

    par(xaxt = "s")
    par(las=2)
    par(mar = par("mar") + c(4, 1, 0, 0))
    barplot(res.mca$var$contrib[,1],col="indianred",main="Contributions des
variables pour le 1er axe",border="dark red",axis.lty=5,
            xlab="",ylab="Contributions (en %)",beside=T,cex.names=0.8)

    par(xaxt = "s")
    par(las=2)
    par(mar = par("mar") + c(4, 1, 0, 0))
    barplot(res.mca$var$contrib[,2],col="brown",main="Contributions des var
```

```
iables pour le 1er axe",border="dark red",axis.lty=5,
            xlab="",ylab="Contributions (en %)",beside=T,cex.names=0.8)


    plot((res.mca$var$cos2[,1]),type="b",col="steelblue",main="Cos 2 des va
riables pour le 1er axe",xlab="Modalités",ylab="Cos2")

    plot((res.mca$var$cos2[,2]),type="b",col="indianred",main="Cos 2 des va
riables pour le 2ème axe",xlab="Modalités",ylab="Cos2")

    #début création du bilan du nombre d'individus par cluster

    abs <- barplot(table(table1$cluster),plot=F)

    colnames(abs) <- "x"

    barplot(table(table1$cluster),col=1:length(table1$cluster),main="Nombre
d'individus dans chaque cluster",border="dark blue",cex.names = 0.8,axis.lt
y=5,xlab="Cluster",ylab="Individus")
    par(xpd=T)
    text(cbind(abs,table(table1$cluster)),labels=table(table1$cluster),pos=
3,offset=0.2)

    #fin création du bilan du nombre d'individus par cluster

    # début analyse des clusters par groupe

    for (j in 1:max(as.numeric(levels(res.hcpc$data.clust$clust)))){

      tableclusti<- table1[table1$cluster==paste(i,k,j,sep="."),]

      Analyse_clust(tableclusti)

    }
    # fin boucle analyse des clusters par groupe


    #début création du bilan du nombre d'individus par cluster sur toutes l
es analyses
    if((i==3)&(k=2)){

      abs <- barplot(table(clustfinal),plot=F)

      colnames(abs) <- "x"

      barplot(table(clustfinal),col=1:length(clustfinal),main="Bilan du nom
bre d'individus dans chaque cluster",border="dark blue",cex.names = 0.8,axi
s.lty=5,xlab="Cluster",ylab="Individus")
      par(xpd=T)
      text(cbind(abs,table(clustfinal)),labels=table(clustfinal),pos=3,offs
et=0.2)

    }
```

```
    dev.off()

    #fin création du bilan du nombre d'individus par cluster sur toutes les
analyses
  }

}

setwd(dir = dir_resultat)# chemin pour placer le fichier csv

return(resu) #retour de la fonction afin de la retrouver pour le fichier Pr
ocessus.R

}

resu <- clust_impute_C('F:/Stage/Script/Cote dOr/Prepa.R',"F:/Stage/output/
Cote dOr/Imputation","F:/Stage/output/Cote dOr/Resultats")#Application de l
a fonction clust_impute_C
#sur le jeu de donnée de la Cote d'Or

write.csv(resu,file="resultat_Cote_dOr_C.csv",row.names = F)#création du fi
chier csv provenant de nos résultats (jeu de donnée traitée et cluster obte
nus)
```

Annexe 5: Processus des classifications

```
source('F:/Stage/Script/Cote dOr/clust global A.R')#table complétant les NA et faisant un clustering global avec imputation

result <- cbind(sol0,resu$cluster)#regroupement de la colonne cluster avec le jeu de donnée traité

names(result)[names(result)=="resu$cluster"] <- "A"#changement du nom de la colonne pour la nommer A pour(méthode A)

source('F:/Stage/Script/Cote dOr/clust imputation type de sol B.R')#table avec les données manquantes complétées et faisant une imputation par type de sol

result <- cbind(result,resu$cluster)#regroupement de la colonne cluster avec le jeu de donnée traité et le resultat du cluster A

names(result)[names(result)=="resu$cluster"] <- "B"#changement du nom de la colonne pour la nommer B pour(méthode B)

source('F:/Stage/Script/Cote dOr/clust imputation monolithique et bilithique C.R')#table remplacant les NA non complétée par leur valeur et faisant une imputation

result <- cbind(result,resu$cluster)#regroupement de la colonne cluster avec le jeu de donnée traité et les resultats du cluster A et B

names(result)[names(result)=="resu$cluster"] <- "C"#changement du nom de la colonne pour la nommer C pour(méthode C)

source('F:/Stage/Script/Cote dOr/clust global D.R')#table supprimant les NA et faisant une classification global par imputation

result$cluster[rownames(result) %in% rownames(resu)] <- resu$cluster#regroupement de la colonne cluster avec le jeu de donnée traité et les resultats du cluster A et B et C

names(result)[names(result)=="cluster"] <- "D"#changement du nom de la colonne pour la nommer D pour(méthode D)

source('F:/Stage/Script/Cote dOr/clust no impute type de sol E.R')#table supprimant les NA non complétée sans imputation par type de sol

result$cluster[rownames(result) %in% rownames(resu)] <- resu$cluster#regroupement de la colonne cluster avec le jeu de donnée traité et les resultats du cluster A et B et C et D
```

```r
names(result)[names(result)=="cluster"] <- "E"#changement du nom de la colo
nne pour la nommer E pour(méthode E)

source('F:/Stage/Script/Cote dOr/clust no imputation monolithique et bilith
ique F.R')#table supprimant les NA non complétée

result$cluster[rownames(result) %in% rownames(resu)] <- resu$cluster#regrou
pement de la colonne cluster avec le jeu de donnée traité et les resultats
du cluster A et B et C et D et E

names(result)[names(result)=="cluster"] <- "F"#changement du nom de la colo
nne pour la nommer F pour(méthode F)

tableauFinal <- result# création de la table tableauFinal

setwd(dir = "F:/Stage/output/Cote dOr/Resultats")#précision de la direction
de notre répertoire de travail

write.csv(tableauFinal,file="synthese_cote_dor.csv",row.names=FALSE)# créat
ion du fichier synthèse
```

Références bibliographiques :

http://factominer.free.fr/

http://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf

http://www.math.univ-angers.fr/~labatte/enseignement%20UFR/masterTVPS/cm2010M2.pdf

http://iml.univ-mrs.fr/~reboul/ADD4-MAB.pdf

http://www.jybaudot.fr/Analdonnees/agregcah.html

http://pbil.univ-lyon1.fr/R/pdf/tdr521.pdf

http://www.foad-mooc.auf.org/IMG/pdf/M05-3.pdf