# Document de formalisation des étapes d’automatisation du calcul des statistiques relatives aux variables quantitatives et qualitatives pour les TypTerres

Objet : *Réalisation d’un script pour automatiser le calcul des moyennes, minimums et maximums des variables quantitatives ; ainsi que la définition de la variables qualitatives dominantes ; des UTS (Unités Typologiques de Sols) regroupées au sein d’une même UTT (Unité TypTerres) pour la strate de surface et les strates sous-jacentes.*

Support : logiciel R

Matériel nécessaire : Extractions csv DONESOL + fichier excel de jointure UTS/Strate-UTT

# Description de la structure des fichiers d’entrée

## Extraction csv des fichiers DoneSol de l’étude : le datamart

Il contient les informations contenues dans les tables : Ucs, Uts, L\_ucs\_uts, Strate, Strate\_quant, Strate\_qual sous forme horizontalisée. Nous avons donc une ligne par liaison uts-ucs et par strate (attention aux UTS qui sont affectées à plusieurs UCS, les informations les concernant sont répétées autant de fois qu’elles sont affectées

Attention les variables du datamart peuvent être présentes dans plusieurs colonnes si elles ont été renseignées par différentes méthodes.

).

*NB : On préfèrera le datamart horizontalisé à l’import des n tables csv. La procédure de jointure étant déjà faite et sûre dans le datamart et les données horizontalisées seront plus faciles à manipuler ensuite.*

## Fichier xls de jointure UTS/Strate-UTT

Prérequis indispensable : ce fichier doit être fourni par la personne qui a réalisé la typologie

Ce fichier permet de rattacher les UTT et les couches définies avec les couples UTS/Strates issus du Datamart. Il se présentera sous la forme d’un fichier excel à 4 colonnes avec autant de lignes que de couples UTS/Strates (cf. exemple ci-dessous).

|  |  |  |  |
| --- | --- | --- | --- |
| ID UTS | N° strate | ID UTT | Couche |
| 101 | 1 | 1 | 1 |
| 102 | 1 | 1 | 1 |
| 101 | 2 | 1 | 2 |
| 101 | 3 | 1 | 2 |
| 101 | 4 | 1 | 2 |
| 102 | 2 | 1 | 2 |
| 104 | 1 | 2 | 1 |
| 104 | 2 | 2 | 2 |
| 104 | 3 | 2 | 2 |
| … | … | … | … |

# Etapes de traitement :

## 2.1- Import de l’environnement de travail et des données

On charge l’environnement de travail (à remplir par l’utilisateur) pour donner le chemin d’accès aux données (type D:/Donnees/…) et le nom de deux fichiers csv contenant les données d’entrées :

* t\_var = le datamart relatif à l’étude
* t\_utt = le fichier de jointure (couples UTS/Strates avec l’horizon du sol TypTerres )

## 2.2- « Nettoyage » des données : cas des horizons O/AH

Dans t\_var, suppression des lignes correspondant à des strates de type O du datamart.

Commentaires : AH (spécifique au fichier de Joël : si deux strates commencent à 0 on prend la deuxième)

Concaténation des variables issues de plusieurs méthodes différentes pour obtenir une seule donnée modale pour chaque variable. On s’intéresse uniquement aux valeurs modales et aux modes PRIN.

## 2.3- Jointure afin d’obtenir un tableau horizontalisé des variables. A chaque ligne du tableau correspond un couple UTS/Strate

On joint les tables t\_var et t\_utt. La jointure se fait sur les champs ID\_UTS et NO\_STRATE. Il s’agira d’une jointure de type RIGHT JOIN (on conserve tous les champs contenus dans la table t\_utt).

## 2.4- Calcul des surfaces de chaque UTS et de chaque UTT

Afin de calculer les modales (moyennes pondérées à la surface par UTT), on calcule la surface des UTS et la surface totale des UTT pour chaque étude.

Pour chaque UTS :

*surf\_uts = pourcent \* surf\_unit* (champs des tables l\_ucs\_uts et ucs).

Même si l’uts est présente plusieurs fois dans l’ucs, le pourcentage englobe la totalité des répétitions.

La surface de l’UTT correspond donc à la somme des surfaces des UTS qui la compose, on aura donc pour chaque UTT :

On définit donc un pourcentage de l’UTS par UTT : Surf\_uts / Surf\_utt Ce pourcentage sera repris pour le calcul des valeurs.

*Exemple :*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Champs fournis* | | | | | *Champs calculés* | | |
| *ID\_UCS* | *Surface\_UCS* | *ID\_UTS* | *pourcent* | *ID\_UTT* | *Surf\_UTT* | *Surf\_UTS* | *%uts\_utt* |
| *1* | *500* | *101* | *50* | *1* | *750* | *250* | *33.3* |
| *1* | *500* | *102* | *30* | *1* | *750* | *150* | *20* |
| *1* | *500* | *103* | *20* | *1* | *750* | *100* | *13.3* |
| *2* | *250* | *205* | *40* | *1* | *750* | *100* | *13.3* |
| *2* | *250* | *210* | *45* | *1* | *750* | *112.5* | *15* |
| *2* | *250* | *100* | *15* | *1* | *750* | *37.5* | *5* |

## 2.5- Calcul des statistiques sur les variables quantitatives pour chaque couple UTT/Horizon

Dans le script, on définit les champs correspondants aux variables quantitatives des strates. On prendra toujours le champ « moyenne» ou « modale » de Donesol.

Boucle 1 : Pour les variables i définies de 1 à m :

Boucle 2 : Pour chaque couple UTT/Horizon j de 1 à n :

* + On fait la moyenne pondérée (MP) par la surface des UTS qui composent l’UTT de la variable i mesurée dans la strate de chaque UTS correspondantes.
  + On calcule le min et le max pour le couple UTT/Horizon : on prend le minimum et le maximum des valeurs modales (Donesol) prise par chaque couple UTS/strate

### Exemple 1 : cas d’une couche composée d’une unique strate par UTS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| UTS | strate | UTT | H | argile | Pourcent  Surf\_uts / Surf\_utt |
| 101 | 1 | 1 | 1 | 26 | 10 |
| 102 | 1 | 1 | 1 | 28 | 25 |
| 208 | 1 | 1 | 1 | 35 | 45 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| UTT | couche | strate | UTS | **Surf\_UTT** | **Surf\_UTS** | Pourcent  Surf\_uts / Surf\_utt | argile |
| 1 | 1 | 1 | 101 | **500** | **50** | 10 | 26 |
| 1 | 1 | 1 | 102 | **500** | **125** | 25 | 28 |
| 1 | 1 | 1 | 208 | **500** | **325** | 65 | 35 |

Résultat : valeur moyenne : 32.35

Valeur minimale 26

Valeur max : 35

Le tableau final :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| UTT | H | Taux argile modal | Taux argile min | Taux argile max |
| 1 | 1 | 32.35 | 10 | 35 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Exemple 2 : cas d’une couche composée par plusieurs strates d’une même UTSUTT | couche | UTS | strate | Surf\_UTT | Surf\_UTS | Mod\_Var |
| 4 | 2 | 45 | 2 | 400 | 250 | 25 |
| 4 | 2 | 45 | 3 | 400 | 250 | 12 |
| 4 | 2 | 33 | 2 | 400 | 150 | 20 |

### On obtient :

### Valeur modale = 18.85 =

Valeur min = 12

Valeur max =25

On stocke le résultat dans un tableau nommé « stats\_UTT\_horiz ». Ce tableau contient x lignes et 3y colonnes (x = nombre de couples UTT/horizon distincts et y = nombre de variables quantitatives (3 car modale/min/max)).

### Cas particulier des épaisseurs :

On prend les champs prof\_appar\_min, prof\_appar\_moy, prof\_appar\_max.

**Question : mieux vaut-il calculer les stats sur les valeurs moyennes (même pour le min et le max… dans ce cas c’est la même méthode que décrit plus haut)**

**Ou alors pour le min et le max, je prend le min des prof\_appar\_min et le max des prof\_appar\_max ?**

Pour moi, on fait le même calcul que des valeurs quantitatives, sauf si je n’ai pas compris. On prend la moyenne des valeurs moyennes pondérées par le pourcentage et on prend les extrêmes des valeurs moyennes.

Si tu as une seule UTS, tu répercutes les valeurs min, moy et max de cette seule UTS, pourquoi se compliquer …

## 2.6- Calcul de la dominante pour les variables qualitatives pour chaque couple UTT/Horizon

On définit les champs correspondants aux variables qualitatives que l’on veut analyser.

Boucle 1 : Pour les variables i définies de 1 à m :

Boucle 2 : Pour chaque couple UTT/Horizon j de 1 à n :

* On va chercher les valeurs que prennent les variables i de chaque couple UTS/strate qui composent l’UTT/Horizon.
* A chaque valeur on associe la surface de l’UTS et on somme les surfaces par valeur distincte.
* On affecte la valeur comptabilisant la plus grande surface au couple UTT/Horizon

On stocke les résultats à la suite dans le tableau stats\_UTT\_horiz (x lignes et 3y + z ; avec z le nombre de variables qualitatives)

Exemple 3 : On reprend le cas de la couche 2 de l’UTT 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| UTT | couche | UTS | strate | Surf\_UTT | Surf\_UTS | Var qual |
| 4 | 2 | 45 | 2 | 400 | 250 | « A » |
| 4 | 2 | 45 | 3 | 400 | 250 | « B » |
| 4 | 2 | 33 | 2 | 400 | 150 | « B » |

Total des surfaces concernées par A : 250

Total des surfaces concernées par B : 400 (250+150)

Tableau final : :

|  |  |  |
| --- | --- | --- |
| UTT | couche | Var\_qual dominante |
| 4 | 2 | B |

Se pose tout de même deux cas particuliers qui montrent les limites de cette méthode et sur lesquels il nous faut trancher :

Cas A : Avec la somme des surfaces, on obtient le même maximum pour deux valeurs (ou plus). La solution serait de fournir les valeurs dominantes et c’est l’expert qui tranche ?

|  |  |  |
| --- | --- | --- |
| surface | Valeur qual | Somme des surfaces |
| 50 | A | 140 |
| 10 | B | 50 |
| 30 | A | 140 |
| 40 | B | 50 |
| 140 | C | 140 |
| 60 | A | 140 |

Quelle valeur dominante fournir A ou C ?

Cas B : dans le cas d’une utt composée de plusieurs strate d’une même UTS, on accorde forcément plus de poids à cette UTS si la valeur des différentes strates qui la composent prenne la même valeur (et ceux quelle que soit l’épaisseur de ces strates)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| UTT | couche | UTS | strate | Surf\_UTT | Surf\_UTS | Var qual |
| 6 | 3 | 78 | 2 | 100 | 30 | A |
| 6 | 3 | 78 | 3 | 100 | 30 | A |
| 6 | 3 | 78 | 4 | 100 | 30 | A |
| 6 | 3 | 23 | 3 | 100 | 60 | C |
| 6 | 3 | 23 | 4 | 100 | 60 | B |
| 6 | 3 | 12 | 2 | 100 | 10 | B |

La valeur dominante fournie par le script sera « A » (90 contre B :70 et C :60) alors que l’uts 78 ne représente que 30% de l’UTT.

## 2.7- Calcul des statistiques pour les variables propres au « profil-type » de chaque UTT : Drainage et profondeur

Il s’agit ici de calculer des statistiques sur les variables quantitatives à l’ensemble des x couches de l’UTT.

Ces deux variables sont affectées à l’UTS (pas de valeur par strates), on procèdera comme dans l’exemple 1.

*Nb : On peut exporter ce résultat sous forme d’un tableau à part : stats\_UTT\_profil*

*Ou bien on agrège au tableau de résultat précédent*

**Question  pour les profondeurs :**

**1/faut-il toujours raisonner avec la modale (champs prof\_sol de la table UTS)? Dans ce cas, cela ne sera pas cohérent avec les valeurs calculées par UTT/Horizon ?**

**2/Faut-il calculer la somme des min, des moy et des max de chaque épaisseur des strates ?**

## 2.9- Export des résultats dans l’environnement de travail

On sauve le (ou les) tableau(x) de résultats au format csv dans le répertoire de travail défini en 2.1.

Possibilité d’avoir deux tableaux : stats\_UTT\_profil et stats\_UTT\_horiz