Projet Hydrogen

Abdoulaye Diabakhaté

15 mai 2018

Table des matières

1 Analyse multivariée non paramétrique à partir de matrices de distances **3**
1.1 Analyse multivariée classique 3
1.2 Notations 3
1.2.1 Exemple : 3
1.3 Définitions 3

2 Écriture matricielle du modèle : $Y = X\beta + U$ **3**
2.1 Propriétés 3

3 Analyse multivariée sur base de distances **5**

4 Distances entre valeurs prédites **5**

5 Distances entre résidus **7**

6 Multidimensional scaling (MDS) **8**
6.1 Cas de la matrice Jaccard abundance et de la fraction $0_0.2$ 8
6.2 Stress plot VS dimensions 11
6.3 MDS Solution et MDS solution sur la carte de la France 12

1 Analyse multivariée non paramétrique à partir de matrices de distances

1.1 Analyse multivariée classique

Soit un échantillon de taille n d'observations individuelles, indicées par i=1,...,n réalisations de variables aléatoires (y_i,x_i).

— y_i variable continue prenant ses valeurs dans $\mathbb{R}$.
— x_i variables en nombre K, de type quelconque.

1.2 Notations

— On confond les variables aléatoires et leurs réalisations
— On réserve les majuscules pour des vecteurs

1.2.1 Exemple :

$Y = (y_1, ..., y_n)'$

1.3 Définitions

La variable dépendante y_i s'écrit comme : $y_i = x_i\beta + u_i$
— β est un paramètre à estimer
— Le modèle est linéaire en β

2 Écriture matricielle du modèle : $Y = X\beta + U$

$$\text{Avec X=} \begin{pmatrix} \mathrm{x}_{11} & \cdots & \mathrm{x}_{1K} \\ \vdots & \ddots & \vdots \\ \mathrm{x}_{n1} & \cdots & \mathrm{x}_{nK} \end{pmatrix}$$

2.1 Propriétés

D'après le théorème de Gauss-Markov,qui énonce que dans un modèle linéaire dans lequel les erreurs ont une espérance nulle,sont non corrélées et dont les variances sont égales(homoscédasticité),le meilleur estimateur linéaire non biaisé des coefficients est son estimateur par les moindres carrés.

$\hat{\beta}= \underset{\beta}{\operatorname{argmin}}||Y - X\beta|| = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - x_i\beta)^2$

Donc le meilleur estimateur linéaire en Y et sans biais est :

$$\hat{\beta}=(X'X)^{-1}X'Y$$

En effet :

$||Y - X\beta||^2 = (Y - X\beta)'(Y - X\beta)$

$||Y - X\beta||^2 = Y'Y - Y'X\beta - X'\beta'Y - X'\beta'X\beta$

Avec : $Y = X\beta$ alors $Y' = X'\beta'$.

Par suite : $X'\beta'Y = Y'X\beta$ et $X'\beta'X\beta = \beta'X'X\beta$

$||Y - X\beta||^2 = Y'Y - 2Y'X\beta + \beta'X'X\beta$

Aussi :

$$\frac{\partial}{\partial\beta}Y'Y = 0\,;$$
$$\frac{\partial}{\partial\beta}2Y'X\beta = 2Y'X$$

$$\frac{\partial}{\partial\beta}\beta'X'X\beta = 2\beta'X'X$$

Ainsi :

$\frac{\partial}{\partial\beta}||Y - X\beta||^2 = -2Y'X + 2\beta'X'X = 0$

Donc on a :

$\beta = \hat{\beta} = (X'X)^{-1}X'Y$

La matrice des valeurs prédites est : $\hat{Y} = X\hat{\beta} = HY$

Avec H=X$(X'X)^{-1}X'$

La matrice des résidus est : $R = Y - \hat{Y} = (I - H)Y$

La matrice totale SSCP est alors décomposée par les matrices SSCP prédites et résiduelles de la façon suivante :

$Y'Y = \hat{Y}'\hat{Y} + R'R$

où : $S_T = tr(Y'Y)$; $S_H = tr(\hat{Y'}\hat{Y})$; $S_R = tr(R'R)$

Une statistique appropriée pour tester l'hypothèse nulle de l'absence d'effet des paramètres du modèle est la pseudo statistique F :

$F = \frac{tr(\hat{Y'}\hat{Y})/(m-1)}{tr(R'R)/(n-m)}$

où m est le nombre de paramètres.

Pour deux matrices $A_{n,p}$ et $B_{n,p}$: tr(AB)=tr(BA)

On a : $YY^{'} = \hat{Y}\hat{Y}^{'} + RR^{'}$

Donc : $tr(YY^{'}) = tr(\hat{Y}\hat{Y}^{'}) + tr(RR^{'})$

La matrice H est symétrique :

En effet, on a : $H = X(X^{'}X)^{-1}X^{'}$

$H^{'} = [X(X^{'}X)^{-1}X^{'}]^{'} = H$

$$\hat{Y}\hat{Y}^{'} = HY(HY)^{'} = H(YY^{'})H$$
$$RR^{'} = (I-H)Y.Y^{'}(I-H)^{'} = (I-H)(YY^{'})(I-H)$$

3 Analyse multivariée sur base de distances

Soit $D = (d_{ij})$, une matrice de distance de taille n*n.

Posons $A = (a_{ij}) = (\frac{-1}{2}d_{ij}^2)$.

Nous pouvons alors calculer la matrice centrée de Gower G en centrant les éléments de A :

$G = (\mathbb{I} - \frac{1}{n}11^{'})A(\mathbb{I} - \frac{1}{n}11^{'})$.

Avec 1 est une colonne de taille n, contenant uniquement des 1.

Ainsi en remplaçant $(\mathrm{YY}^{'})par G, nous avons S_T = tr(G)$ et la pseudo statistique F est :

$F = \frac{tr(HGH)/(m-1)}{tr[(\mathbb{I}-H)G(\mathbb{I}-H)]/(n-m)}$

4 Distances entre valeurs prédites

Soit Y le jeu de données centré contenant toutes nos observations.

Nous nous plaçon dans un cadre linéaire, c'est-à-dire : $Y = X\theta + \epsilon$, où X est la matrice de design contenant les covariables.

Nous disposons uniquement de la matrice de distances $D^2 = (||Y_{i,.} - Y_{j,.}||^2)_{i,j}$, de métrique variable(jaccard,braycurtis,...).

Supposons que D est une matrice de distances euclidiennes. Notre objectif est de calculer une matrice de distances entre valeurs prédites (ou ajustées) à partir de la prédication $\hat{Y}$ de Y :

Supposons que D est une matrice de distances euclidiennes.

Notre objectif est de calculer une matrice de distances entre valeurs prédites(ou ajustées)à partir de la prédication $\hat{Y}$ de Y : $\hat{Y} = X\hat{\theta} = HY$.

Notons par : $\hat{D^2} = (||\hat{Y_{i,.}} - \hat{Y_{j,.}}||^2)_{i,j}$,une telle matrice.

Cette dernière est calculable en fonction de :
— X et $\tilde{Y}$, où $\tilde{Y}$ correspond aux observations placées dans un cadre euclidien et obtenues à partir d'une matrice de distance quelconque à l'aide du MDS,
— X et D

Avec MDS

Nous avons ici la matrice de distances euclidiennes :

$D^2 = (||\tilde{Y_{i,.}} - \tilde{Y_{j,.}}||^2)_{i,j}$

$D^2 = (||(H\tilde{Y})_{i,.} - (H\tilde{Y})_{j,.}||^2)_{i,j}$.

Sans MDS

Calculons $(\hat{D^2})_{i,j}$ en fonction de X et D.

$\hat{D}^2_{i,j} = ||(HY)_{i,.} - (HY)_{j,.}||^2$

$\hat{D}^2_{i,j} = ||(HY)_{i,.}||^2 + ||(HY)_{j,.}||^2 - 2 < (HY)_{i,.}, (HY)_{j,.} >$

$< (HY)_{i,.}, (HY)_{j,.} >=< H_{i,.}Y, H_{j,.}Y >$

$< (HY)_{i,.}, (HY)_{j,.} >= (H_{i,.}Y)(H_{j,.}Y)^{'}$

$< (HY)_{i,.}, (HY)_{j,.} >= H_{i,.}(YY^{'})H^{'}_{j,.}$

Avec $G = YY^{'}$ et $H^{'}_{j,.} = H_{.,j}$

Donc : $< (HY)_{i,.}, (HY)_{j,.} >= H_{i,.}GH_{.,j}$

Par ailleurs on a :

$(HY)_{i,.} = (H_{i,.}Y_{.,1}, ..., H_{i,.}Y_{.,p}) = (Y_{1,.}H_{.,i}, ..., Y_{p,.}H_{.,i})$

5 Distances entre résidus

Cette fois-ci nous cherchons à calculer la matrices de distances entre résidus,c'est-à dire :

$$D_R^2 = (||R_{i,.} - R_{j,.}||^2)_{i,j}$$

Avec : $R = Y - \hat{Y} = (I - H)Y$ est la matrice des résidus.

En nous plaçant dans un cadre euclidien, cette dernière est calculable en fonction de :
— X et $\tilde{Y}$, où $\tilde{Y}$ correspond aux observations placées dans un cadre euclidien et obtenues à partir d'une matrice de distance quelconque à l'aide du MDS.
— X et D

Avec MDS

Nous avons ici la matrice de distances euclidiennes :

$$D^2 = (||\tilde{Y}_{i,.} - \tilde{Y}_{j,.}||^2)_{i,j}$$

$D_R^2 = (||R_{i,.} - R_{j,.}||^2)_{i,j}$

$D_R^2 = (||((I - H)\tilde{Y})\{i, . - ((I - H)\tilde{Y})_{j,.}||^2)_{i,j}$

Sans MDS

Calculons $D_R^2(i, j)$ en fonction de X et D :

$D_R^2(i, j) = ||((I - H)Y)_{i,.} - ((I - H)Y)_{j,.}||^2$

$D_R^2(i, j) = ||((I - H)Y)_{i,.}||^2 + ||((I - H)Y)_{j,.}||^2 - 2 < ((I - H)Y)_{i,.}, ((I - H)Y)_{j,.} >$

Or : $((I - H)Y)_{i,.} = ((I - H)_{i,.}Y_{.,1}, ..., (I - H)_{i,.}Y_{.,p}) = (Y_{1,.}(I - H)_{.,i}, ..., Y_{p,.}(I - H)_{.,i})$

Donc : $<$((I-H)Y)$_{i,.}, ((I - H)Y)_{j,.} >= \sum_{k=1}^{p}((I - H)_{i,.}Y_{.}, k)((I - H)_{j,.}Y_{.}, k)$

$<$((I-H)Y)$_{i,.}, ((I - H)Y)_{j,.} >= \sum_{k=1}^{p}(I - H)_{i,.}Y_{.,k}Y_{k,.}(I - H)_{.,j}$

$<$((I-H)Y)$_{i,.}, ((I - H)Y)_{j,.} >= \sum_{k=1}^{p}(I - H)_{i,.}Y_{.,k}Y'_{.,k}(I - H)_{.,j}$

$<$((I-H)Y)$_{i,.}, ((I - H)Y)_{j,.} >= (I - H)_{i,.}(\sum_{k=1}^{p} Y_{.,k}Y'_{.,k})(I - H)_{.,j}$

$<$((I-H)Y)$_{i,.}, ((I - H)Y)_{j,.} >= (I - H)_{i,.}G(I - H)_{.,j}$

Car : $G = YY'$ quand D est euclidienne.

Par suite : $D_R^2(i, j) = (I - H)_{i,.}G(I - H)_{.,i} + (I - H)_{j,.}G(I - H)_{.,j} - 2(I - H)_{i,.}G(I - H)_{.,j}$

6 Multidimensional scaling (MDS)

Le MDS («positionnement multidimensionnel») est un ensemble de techniques statistiques utilisées dans le domaine de la visualisation d'information pour explorer les similarités dans les données.

6.1 Cas de la matrice Jaccard abundance et de la fraction $0_0.2$

MDS en dimension 1 et 2 :

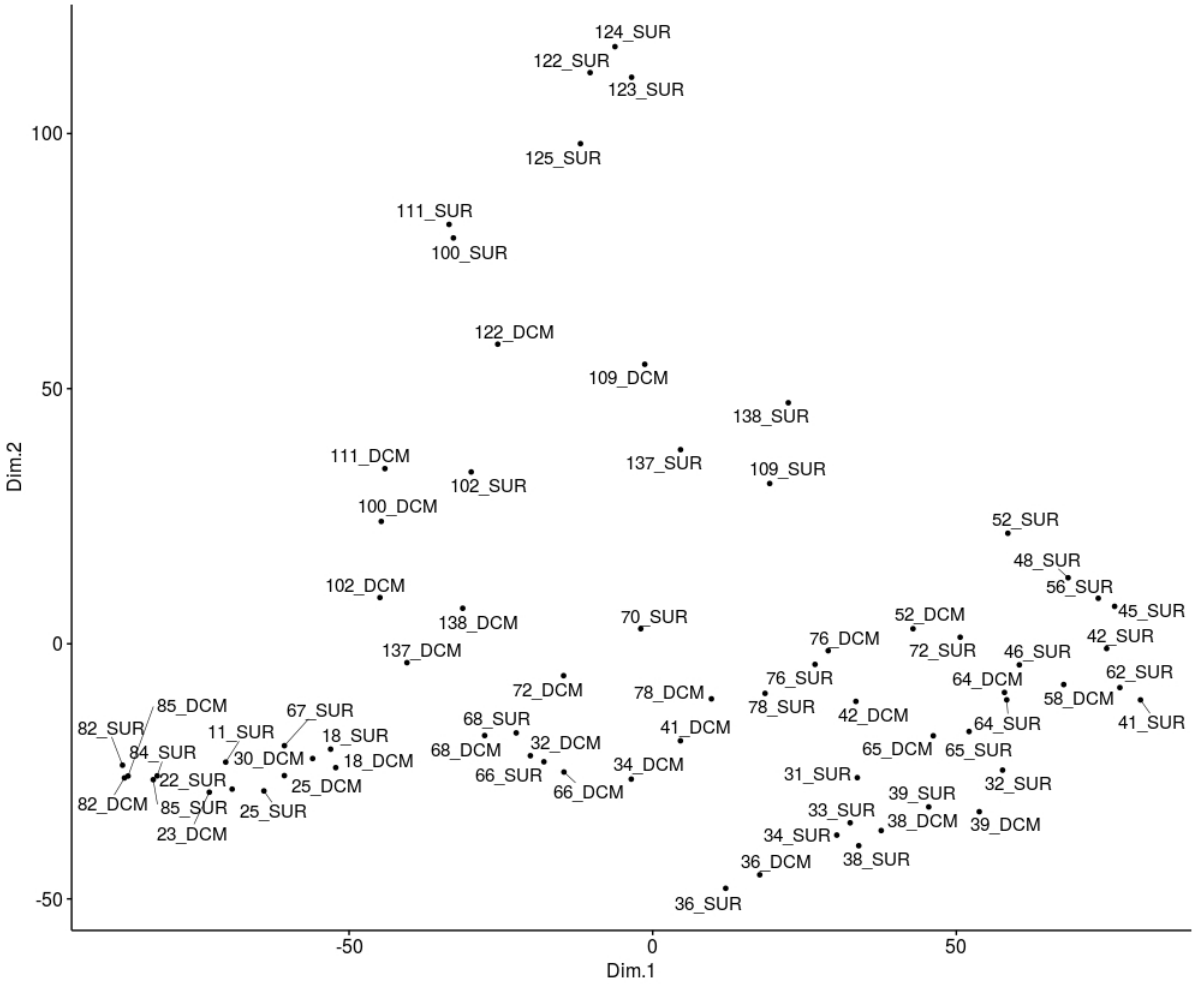


FIGURE 1 – MDS dim1-2

MDS en dimension 1 et 3 :

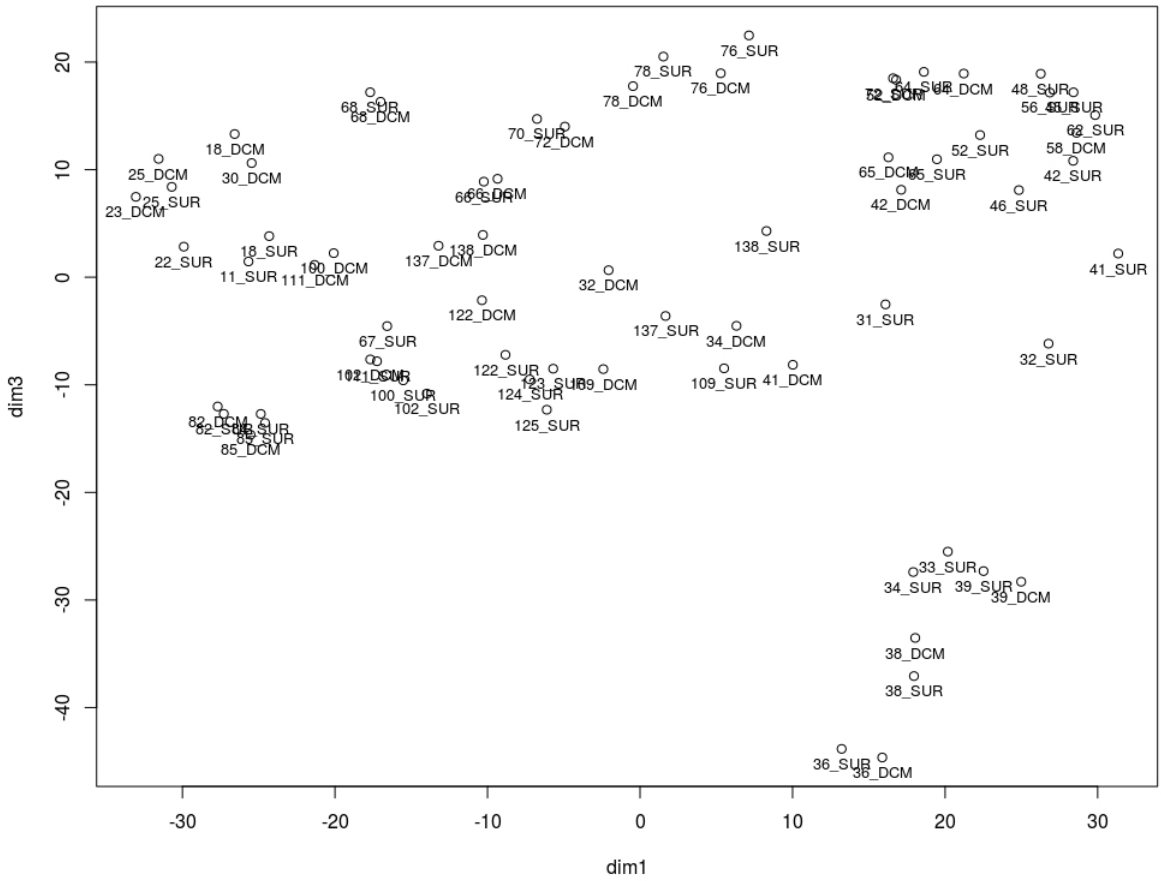


FIGURE 2 – MDS dim1-3

MDS en dimension 2 et 3 :

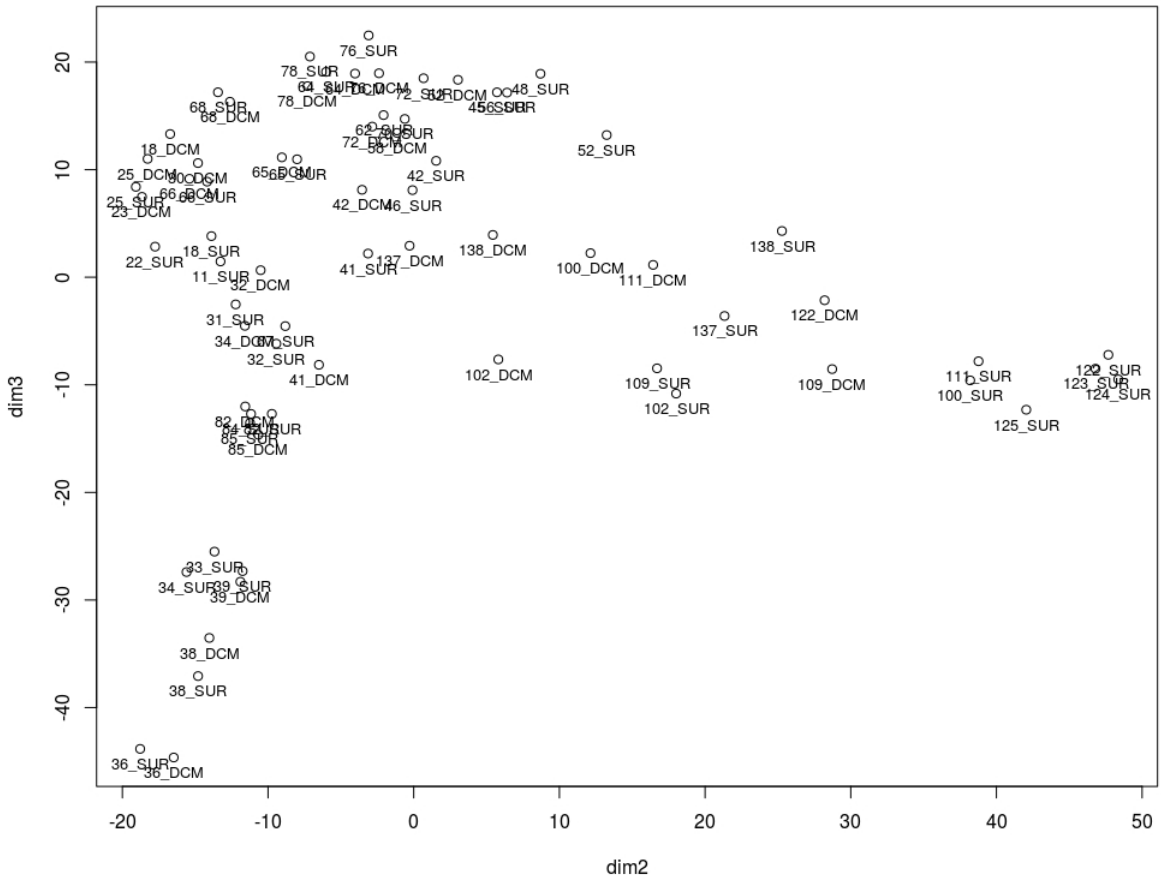


FIGURE 3 – MDS dim2-3

6.2 Stress plot VS dimensions

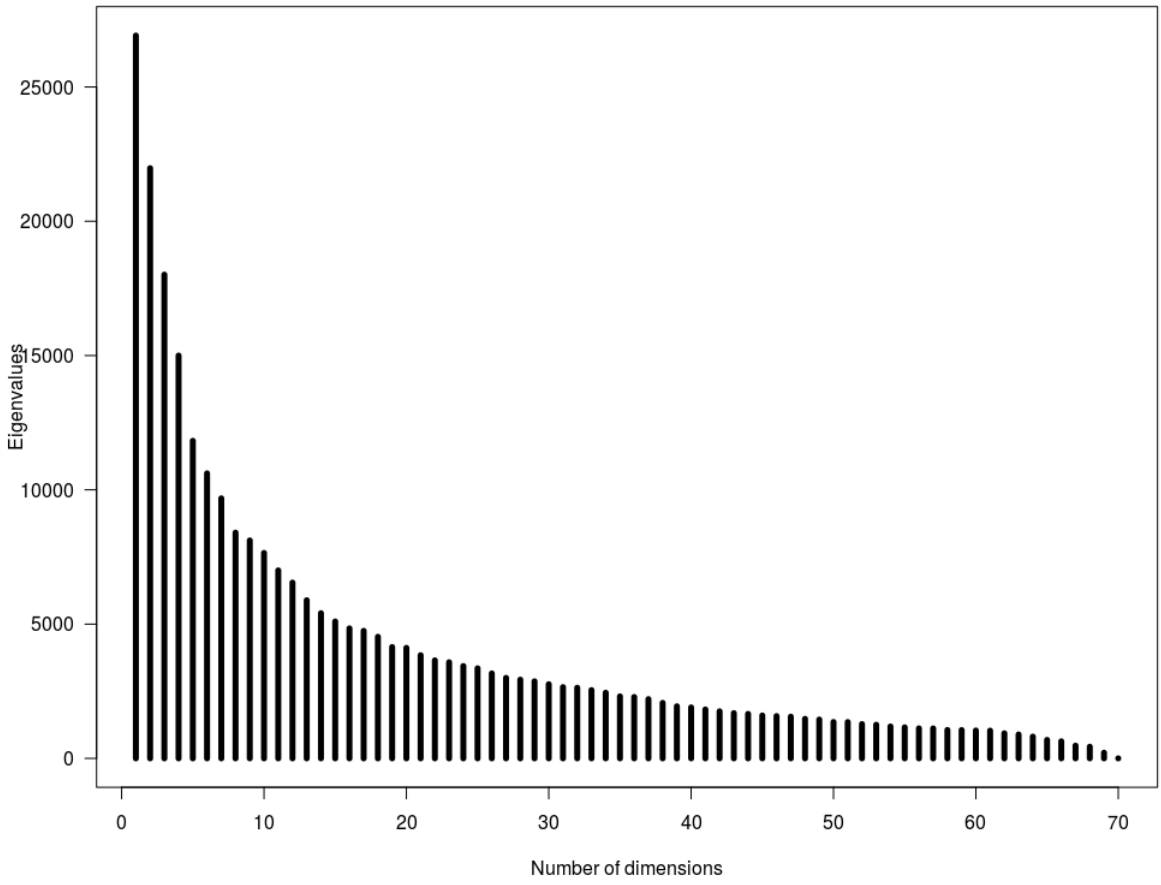
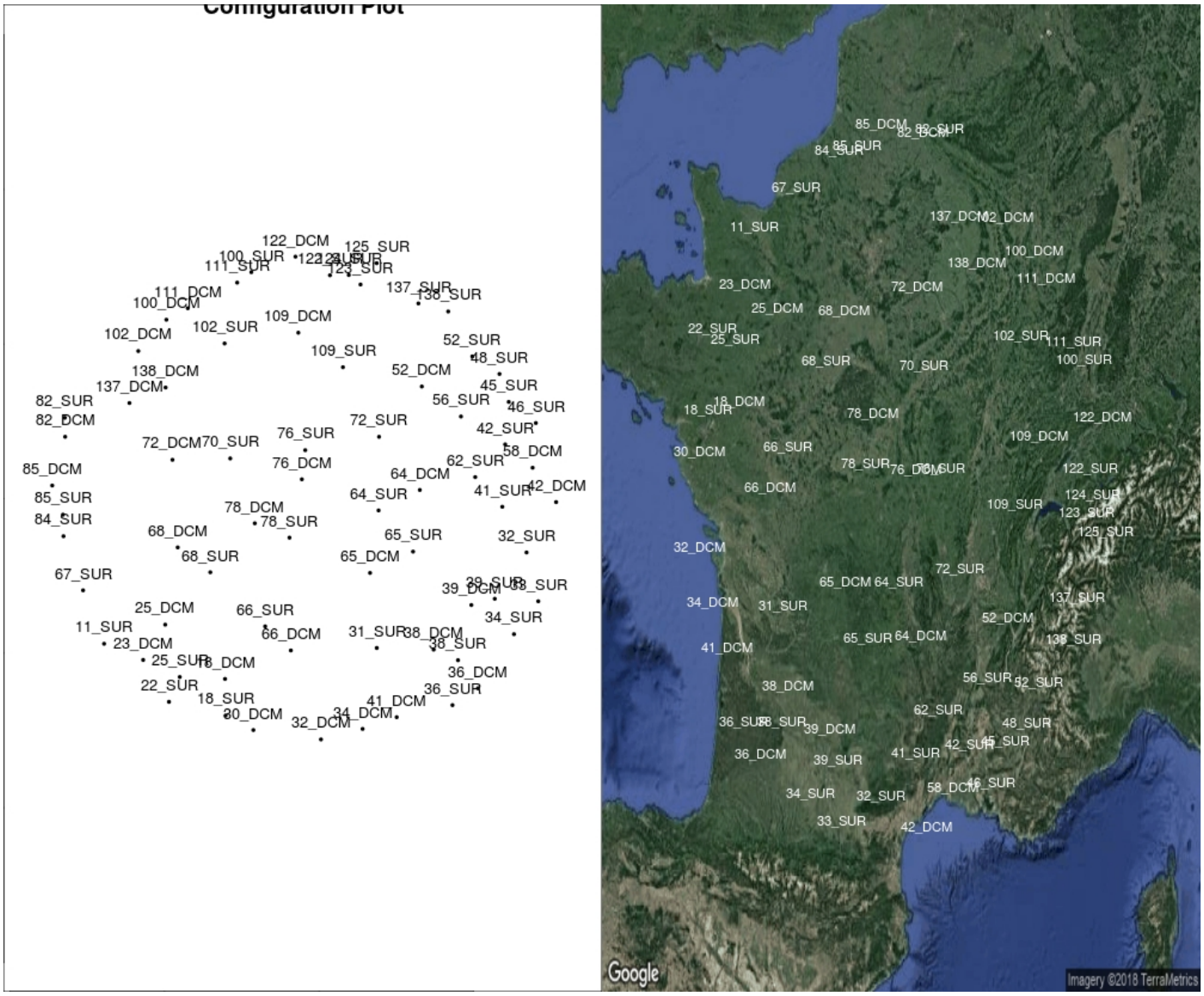


FIGURE 4 – "Scree Plot"

Dendogramme : hclust

6.3 MDS Solution et MDS solution sur la carte de la France


Configuration Plot
Google
Imagery ©2018 TerraMetrics