Différentes manières de mesurer les interactions entre les santés humaine, sociale et environnementale

2 mai 2024

Table des matières

1 Mesurer les variables influentes **2**
- 1.1 Corrélations . . . 2
- 1.2 Indices HSIC . . . 4
- 1.3 Méthodes de forêts aléatoires . . . 7
- 1.4 Indices de Sobol . . . 10
- 1.5 Comparaison des différentes méthodes . . . 12

2 Visualisation de la structure globale des variables **14**
- 2.1 Matrices d'indices . . . 14
- 2.2 Recherche de communauté . . . 15
- 2.3 Graphe de causalité . . . 17
- 2.4 Lorsque le graphe du modèle graphique est déjà connu . . . 19

3 Application à des variables de santé **20**
- 3.1 Quelques réflexions liés aux variables . . . 20
- 3.2 Détection des variables importantes . . . 21
- 3.3 Arbitrage . . . 21

A Liste des variables utilisées et/ou accessibles **23**

Introduction

Dans le Manifeste pour une santé commune ([2]), la santé globale d'un territoire se décompose selon trois axes inerdépendants : la santé humaine, la santé sociale, et la santé des mileux. On cherche à faire une analyse de l'interconnectivité de ces trois santés en utilisant des méthodes statistiques. On cherche donc à : représenter les interdépendances entre les trois santé, quantitifer ces dépendances, mais aussi construire un outil qui permet de déterminer des leviers d'action, c'est à dire des variables à s'intéresser en priorité, pour améliorer la santé globale.

Pour mesurer l'état de santé d'un territoire, on peut mesurer beacoup de variables/indicateurs de santé (humaine, sociale, environnementale), et ainsi obtenir d'énormes tableaux de comptabilité. Mais comment analyser ces tableaux ? Est-il pertinent de regarder les variables individuellement ? Si l'on cherche à améliorer la valeur d'une de ces variables sur un territoire, comment faire pour que cela ne dégrade pas d'autres santés ? Comment faire en sorte de contrôler quelles autres variables sont influencées, et comment ?

1 Mesurer les variables influentes

Le cadre mathématique est le suivant. On considère des variables $X_0, X_1, X_2, \ldots$, chacune représentant une grandeur ou un indicateur d'une santé mesurable sur un territoire donné. Ces variables peuvent appartenir à l'un des trois santé humaine, sociale ou envirronementale, ou être à cheval entre plusieurs de ces santés. On peut par exemple penser au variables suivantes : Densité de population, espérance de vie à la naissance, taux de chômage, proportion de cadres, proportion de ménages monoparentaux, proportion de patients sous traitement antidépresseur, poids de la biomasse, indice de qualité de l'eau, etc. Un liste d'indicateurs disponibles sur BALISES est donnée en Annexe A.

Dans cette section, on fixe une variable particulière X_0, et on étudie l'influence des autres variables $X_1, X_2, \ldots, X_p$ sur celle ci.

1.1 Corrélations

L'approche la plus simple consiste à regarder les corrélations entre la variable X_0 et les autres variables X_i, $1 \leq i \leq p$. Cette corrélation (de Pearson) est définie par

$$\mathrm{Cor}(X_0, X_i) := \frac{\mathrm{Cov}(X_0, X_i)}{\sigma(X_0)\sigma(X_i)} = \frac{\mathbb{E}[X_0 X_i] - \mathbb{E}[X_0]\mathbb{E}[X_i]}{\sqrt{\mathbb{E}[(X_0 - \mathbb{E}[X_0])^2]\mathbb{E}[X_i^2]}}.$$

Si l'on n'a accès qu'à un échantillon des variables X_0, X_i, on peut approximer cette corrélation en prenant pour éspérance les moyennes empiriques. Si l'on note $X_0^{(1)}, X_i^{(1)}, X_0^{(2)}, X_i^{(2)}, \ldots, X_0^{(n)}, X_i^{(n)}$ les valeurs des variables sur n échantillons, on peut prendre par exemple

$$\widehat{\mathrm{Cor}}(X_0, X_i) := \frac{\frac{1}{n}\sum_{1\leq j\leq n} X_0^{(j)} X_i^{(j)} - (\frac{1}{n}\sum_{1\leq j\leq n} X_0^{(j)})(\frac{1}{n}\sum_{1\leq j\leq n} X_i^{(j)})}{\sqrt{\left(\frac{1}{n}\sum_{1\leq j\leq n}\left(X_0^{(j)} - (\frac{1}{n}\sum_{1\leq k\leq n} X_0^{(k)})\right)^2\right)\left(\frac{1}{n}\sum_{1\leq j\leq n}\left(X_i^{(j)} - (\frac{1}{n}\sum_{1\leq k\leq n} X_i^{(k)})\right)^2\right)}}.$$

(Je ne sais pas si c'est cette formule qui est utilisée par `R`, ou s'il y a des $\frac{1}{n}$ ou $\frac{1}{n-1}$ devant les variances par exemple...)

Remarque 1.1. Les corrélations sont une bonne première manière simple de mesurer la dépendance entre deux variables. Cependant, elles ne mesurent que le dépendances linéaires : On peut avoir $\mathrm{Cor}(X, Y) = 0$ pour deux variables X, Y, sans qu'elles ne soient indépendantes par exemple.

Un avantage des corrélations est également qu'elles fournissent un signe, qui représente si une augmentation de X_i se traduit plutôt par une augmentation ou une diminution de X_0.

On peut aussi regarder les corrélations de Spearman et de Kendall, définies ici uniquement sur des échantillons.

Definition 1.2. Soient $X_0^{(1)}, X_i^{(1)}, X_0^{(2)}, X_i^{(2)}, \ldots, X_0^{(n)}, X_i^{(n)}$ les valeurs des variables X_0 et X_i sur n échantillons. On définit les rangs des échantillons $R_0^{(1)}, R_i^{(1)}, R_0(2), R_i^{(2)}, \ldots, R_0^{(n)}, R_i^{(n)} \in \{1, \ldots, n\}$, avec $X_0^{(j)} < X_0^{(k)} \iff R_0^{(j)} < R_0^{(k)}$ par exemple. La *corrélation de Spearman* entre X_0 et X_i est définie par

$$\widehat{\mathrm{Cor}}_S(X_0, X_i) := \widehat{\mathrm{Cor}}(R_0, R_i)$$

où $\widehat{\mathrm{Cor}}$ est la corrélation classique (de Pearson). La *corrélation de Kendall* entre X_0 et X_i est définie par

$$\widehat{\mathrm{Cor}}_K(X_0, X_i) := \frac{1}{n(n+1)}\left|\left\{(j,k) \in \{1,\ldots,n\}^2 \mid j < k, X_0^{(j)} - X_0^{(k)} \text{ et } X_i^{(j)} - X_i^{(k)} \text{ ont même signe}\right\}\right| - 1.$$

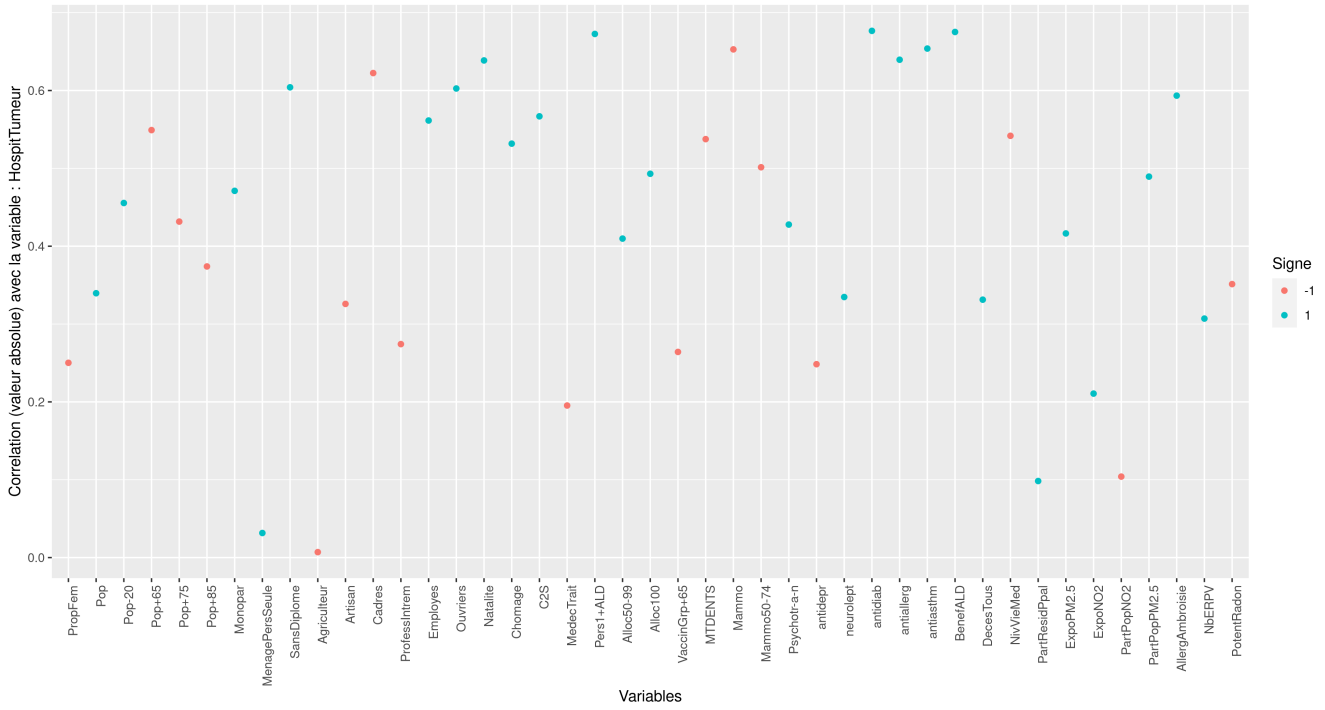


FIGURE 1 – Corrélation entre $X_0 =$ "Taux (standardisé) de patients hospitalisés pour tous motifs" et différentes variables de santé humaine, sociale ou environnementale. A comparer avec les Figures 3, 4 et 5 .

Remarque 1.3. On a $\widehat{\mathrm{Cor}}_S(X_0, X_i) \in [-1,1]$ et $\widehat{\mathrm{Cor}}_K(X_0, X_i) \in [-1,1]$. De plus,

$$\widehat{\mathrm{Cor}}_S(X_0, X_i) = 1 \iff \widehat{\mathrm{Cor}}_K(X_0, X_i) = 1 \iff R_0 = R_i.$$

Les corrélations de Spearman et Kendall ont l'avantage d'être stables par renormalisation croissante : Si $\varphi : \mathbb{R} \to \mathbb{R}$ est strictement croissante ou décroissante, alors

$$\widehat{\mathrm{Cor}}_S(\varphi(X_0), \varphi(X_i)) = \widehat{\mathrm{Cor}}_S(X_0, X_i)$$
$$\widehat{\mathrm{Cor}}_K(\varphi(X_0), \varphi(X_i)) = \widehat{\mathrm{Cor}}_K(X_0, X_i),$$

ce qui n'est pas le cas pour la corrélation de Pearson.

Remarque 1.4. En `R`, on peut accéder facilement à $\widehat{\mathrm{Cor}}(X_0, X_i)$. On suppose les échantillons sont stockés dans une matrices M, où une colonne représente une variable $X_0, X_1, \ldots, X_p$ et une ligne représente un échantillon.

```
# Exemple de matrice, avec p=3 et n=100 echantillons
M = matrix(NA, ncol=4, nrow=100)
M[,1] = 3 * runif(100)
M[,2] = M[,1] + rnorm(100)
M[,3] = - 2 * M[,1] + rnorm(100)
M[,4] = 0.5 * M[,1] + 3 * runif(100)
colnames(M) = c("X0","X1","X2","X3")

# Correlations une par une
Correl12 = cor(M[,1], M[,2])
# Correl12 contient la correlation empirique entre les variables des deux premieres colonnes
```

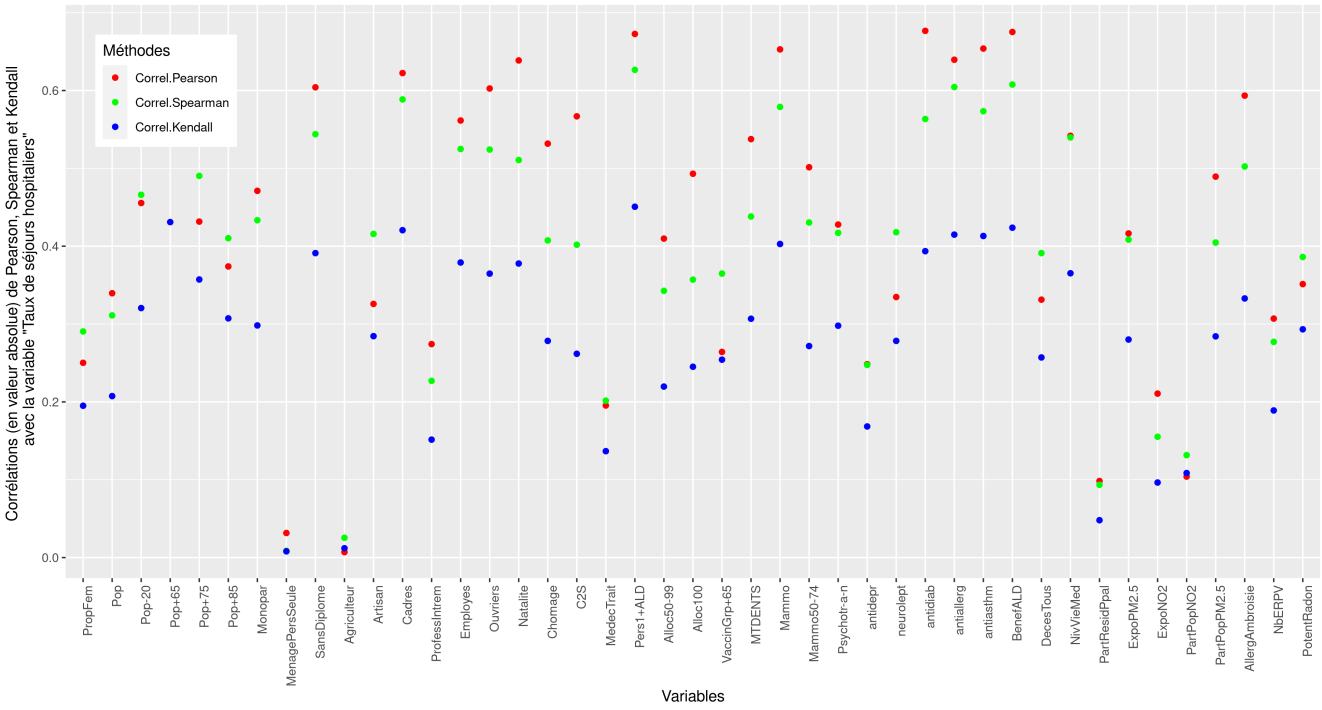


FIGURE 2 – Corrélations de Pearson, Spearman et Kendall entre X_0 = "Taux (standardisé) de patients hospitalisés pour tous motifs" et différentes variables de santé humaine, sociale ou environnementale.

```
# Correlations totales
Correl = cor(M)
# Correl est un tableau (p+1) sur (p+1)
# Correl[i,j] contient la correlation empirique entre la i-eme et la j-eme colonne.
barplot(Correl[1,2:4], names.arg = c("X1","X2","X3"), ylab = "Correlation (de Pearson)")

# Correlation de Spearman
Correl_S = cor(M, method="spearman")
barplot(Correl_S[1,2:4], names.arg = c("X1","X2","X3"), ylab = "Correlation de Spearman")

# Correlation de Kendall
Correl_K = cor(M, method="kendall")
barplot(Correl_K[1,2:4], names.arg = c("X1","X2","X3"), ylab = "Correlation de Kendall")

# Ici, les trois correlations sont similaires, car il n'y a pas vraiment de non-linearite
```

1.2 Indices HSIC

Les espace RKHS (Reproducting Kernel Hilbert Space, voir [6]) permettent de généraliser les indices de corrélation, et présentent l'avantage d'être un peu plus précis, dans le sens où ils ne détectent pas uniquement les dépendance linéaires, ou même monotone.

Definition 1.5. Soit E un ensemble, quelconque. On définit un *noyau* sur E comme une application $k : E \times E \to \mathbb{R}$ symétrique et définie positive, c'est à dire telle que

- (Symétrie) Pour tous $x, y \in E$, on a $k(x,y) = k(y,x)$.
- (Définie positivité) Pour tous $n \geq 1$, $x_1, \ldots, x_n \in E$, $\lambda_i, \ldots, \lambda_n \in \mathbb{R}$, on a $\sum_{1 \leq i,j \leq n} \lambda_i \lambda_j k(x_i, x_j) \geq 0$, avec égalité si et seulement si tous les λ_i sont nuls.

Exemple 1.6. (à vérifier) Voici quelques exemples de noyaux :
- $x, y \mapsto xy$ est un noyau sur $\mathbb{R}$ (dit linéaire),
- $x, y \mapsto \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right)$ est un noyau sur $\mathbb{R}$, pour $\sigma > 0$ fixé (dit gaussien),
- $x, y \mapsto \exp\left(-\frac{|x-y|}{2\sigma^2}\right)$ est un noyau sur $\mathbb{R}$ (dit de Laplace),
- $x, y \mapsto \exp\left(-\frac{d(x,y)^2}{2\sigma^2}\right)$ est un noyau sur un espace métrique (E, d), pour $\sigma > 0$ fixé (dit gaussien).

Definition 1.7. Soit E un ensemble. Un RKHS (Reproducting Kernel Hilbert Space) sur E est un espace de Hilbert $\mathcal{H} \subseteq \mathcal{F}(E, \mathbb{R})$ tel que les applications d'évaluation $f \in \mathcal{H} \to f(x) \in \mathbb{R}$, pour $x \in E$, sont toutes continues.

On va voir que ces deux notions, noyau et RKHS, sont équivalentes.

Definition 1.8. Soit $\mathcal{H} \subseteq \mathcal{F}(E, \mathbb{R})$ un RKHS. Pour $x \in E$, l'application $f \in \mathcal{H} \to f(x) \in \mathbb{R}$ est continue, il existe donc $K_x \in \mathcal{H}$ tel que pour tout $f \in \mathcal{H}$ on ait $f(x) = \langle f, K_x \rangle$. On définit, pour $x, y \in E$, $k_{\mathcal{H}}(x, y) = \langle K_x, K_y \rangle$.

Proposition 1.9. *Il y a une correspondance bijective entre RKHS sur E et noyaux sur E. Cette correpondance est donnée par $\mathcal{H} \mapsto k_{\mathcal{H}}$.*

Remarque 1.10. La correspondance inverse peut être construite comme suit. Soit k un noyau sur E. On définit

$$\mathcal{H}_0 := \mathrm{Vect}(k(x, \cdot), x \in E) \subseteq \mathcal{F}(E, \mathbb{R}),$$

que l'on muni du produit scalaire $\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y)$, prolongé par bilinéarité. On définit $\mathcal{H}_k$ comme la complétion de $\mathcal{H}_0$ pour la norme induite par son produit scalaire. Il est possible de montrer que $\mathcal{H}_k$ est un RKHS, et que $k_{\mathcal{H}_k} = k$ (référence ?)

Une propriété intéressante des certains RKHS est que l'on peut y plonger l'espace des mesures de probabilités sur E.

Definition 1.11. Fixons une tribu $\mathcal{T}$ sur E. Soit k un noyau sur E, et $\mathbb{P}$ une loi de probabilité sur $(E, \mathcal{T})$. On définit $\mu_{\mathbb{P}} := \mathbb{E}_{x \sim \mathbb{P}}[k(x, \cdot)] \in \mathcal{H}_k$. (je ne suis pas sûr de voir pourquoi $\mu_{\mathbb{P}}$ est bien défini, ni pourquoi $\mu_{\mathbb{P}} \in \mathcal{H}_k$ et je n'ai vu nulle part où ce questions sont mentionnées)

On dit que k est *caractéristique* si $\mu : \mathcal{M}_1(E, \mathcal{T}) \to \mathcal{H}_k, \mathbb{P} \mapsto \mu_{\mathbb{P}}$ est injective, où $\mathcal{M}_1(E, \mathcal{T})$ est l'ensemble des mesures de probabilité sur $(E, \mathcal{T})$.

On peut alors construire un test d'indépendance entre deux variables aléatoires.

Definition 1.12. Soit $(E, \mathcal{T})$ un espace mesurable, et X, Y deux variables aléatoires dessus. On considère deux noyau k_X et k_Y sur E, de RKHS respectifs $\mathcal{H}_X$ et $\mathcal{H}_Y$. On définit l'opérateur de covariance croisée $C_{X,Y} : \mathcal{H}_Y \to \mathcal{H}_X$ par, pour tous $f \in \mathcal{H}_X$ et $g \in \mathcal{H}_Y$,

$$\langle f, C_{X,Y}(g) \rangle = \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

Proposition 1.13. *On considère le cadre de la définition précédente, et l'on suppose de plus que k_X et k_Y sont caractéristiques. Alors X et Y sont indépedantes si et seulement si $C_{X,Y} = 0$.*

Definition 1.14. Toujours dans le même cadre, on définit

$$HSIC_{k_X, k_Y}(X, Y) := ||C_{X,Y}||^2_{\mathcal{L}(\mathcal{H}_Y, \mathcal{H}_X)}.$$

Cette quantité pourra être approchée à l'aide d'échantillons, et utilisées pour la création d'un test d'indépendance.

Definition 1.15. On suppose maintenant que l'on dispose de n échantillons de (X,Y), que l'on note $(X^{(1)},Y^{(1)}),\dots,(X^{(n)},Y^{(n)})$. On va approcher la quantité $HSIC_{k_X,k_Y}(X,Y)$ à l'aide de ces échantillons par

$$\widehat{HSIC}_{k_X,k_Y}(X,Y) := \frac{1}{n^2}\operatorname{Tr}(C_X H C_Y H),$$

où C_X, C_Y sont des matrices $n\times n$ ayant pour coordonnées $C_X(i,j) = k_X(X^{(i)},X^{(j)})$, $C_Y(i,j) = k_Y(Y^{(i)},Y^{(j)})$, et $H(i,j) = \delta_{i,j} - \frac{1}{n}$.

Pour revenir à notre problème initial, où l'on a $p+1$ variables $X_0, X_1, \dots, X_p$, on peut considérer l'indice suivant pour estimer l'influence de X_i sur X_0 :

$$I_{HSIC,k_0,k_i}(X_0,X_i) := \frac{\widehat{HSIC}_{k_0,k_i}(X_0,X_i)}{\sqrt{\widehat{HSIC}_{k_0,k_0}(X_0,X_0)\cdot\widehat{HSIC}_{k_i,k_i}(X_i,X_i)}}$$

où l'on a choisit deux noyaux k_0 et k_i sur $\mathbb{R}$ (dans le cas où les X_i sont à valeur dans $\mathbb{R}$). Le dénominateur sert de renormalisation pour comparer les indices calculés avec différents i, et aussi de sorte à ce que $I_{HSIC,k_0,k_i}(X_0,X_i) \in [0,1]$.

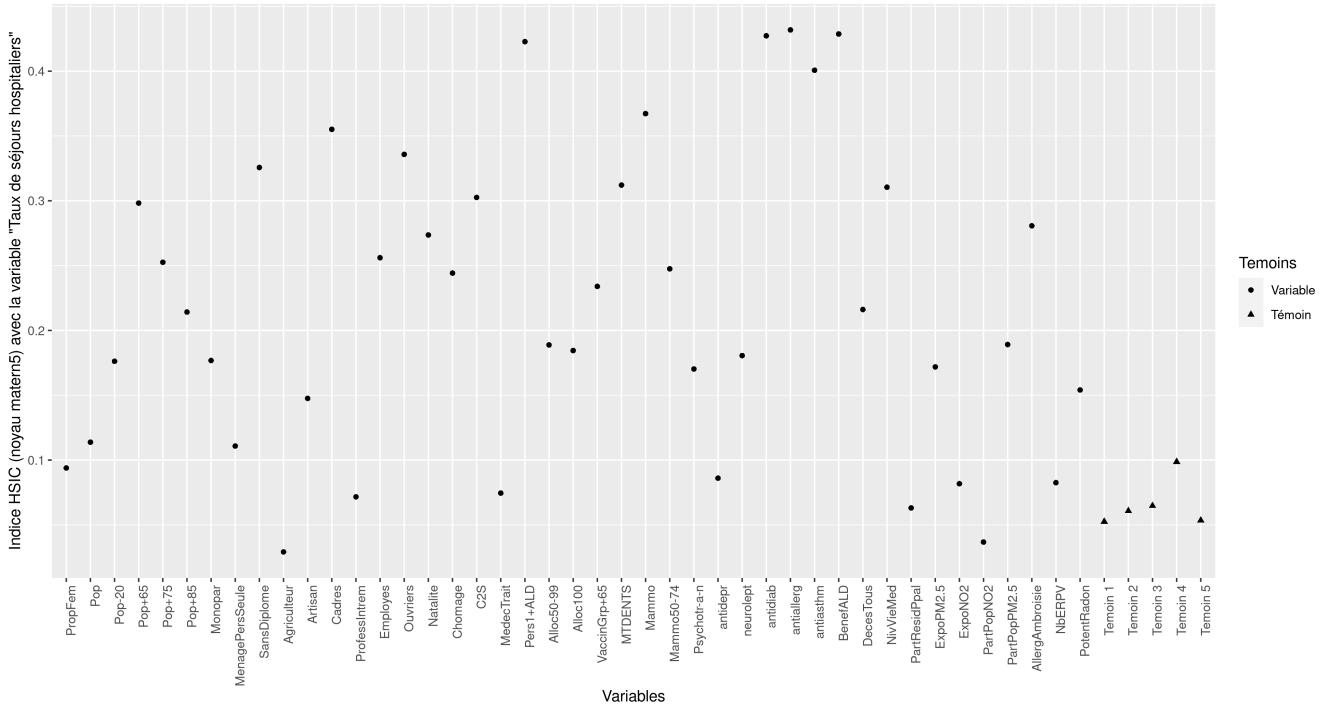


FIGURE 3 – Indices HSIC pour un noyau Matern (`matern5`), entre la variable X_0 = "Taux de séjours hospitaliers pour tous motifs" et différentes variables de santé humaine ou sociale. A comparer avec les Figures 1, 4 et 5.

Remarque 1.16. Une des forces des RKHS est que ces indices sont très généraux. Par exemple, si les variables X_i ne sont pas à valeurs réelles, alors il suffit de prendre des noyaux sur des espaces différents. De plus, X_0 et X_i ne sont pas obligatoirement à valeur dans le même espace, puisque les noyaux k_0 et k_i sont séparés.

Le risque est alors de choisir le mauvais noyau...

Remarque 1.17. Les indices HSIC sont implémentés en `R` dans le package `sensitivity`. La librairie dispose aussi d'un test d'indépendance basé sur ces indices HSIC. L'hypothèse nulle est "X_i est indépendante de X_0", et est équivalente à $I_{HSIC,k_0,k_i}(X_0,X_i) = 0$ si les noyaux k_0 et k_i sont caractéristiques.

```r
library(sensitivity)
# Exemple de matrice, avec p=4 et n=100 echantillons
M = matrix(NA, ncol=5, nrow=100)
M[,1] = 3 * runif(100)
M[,2] = M[,1] + rnorm(100)
M[,3] = - 2 * M[,1] + rnorm(100)
M[,4] = 0.5 * M[,1] + 3 * runif(100)
M[,5] = runif(100)
colnames(M) = c("X0","X1","X2","X3","X4")

library('sensitivity')
# Calcul des indices; "rbf" est le noyau gaussien
sensi = sensiHSIC(X = M[,2:5], kernelX = "rbf", kernelY = "rbf")
tell(sensi, M[,1])
# On recupere les indices comme vecteur de taille p avec sensi$S$original
# Un posibilite pour visualiser les indices
barplot(sensi$S$original, names.arg = c("X1","X2","X3","X4"), ylab = "Indice HSIC")

# Calcul de la p-valeur du test d'independance
test.HSIC <- testHSIC(sensi)
plot(1:4, test.HSIC$pval$original, xaxt = 'n',
        xlab = "Variables", ylab = "p-valeur du test d'independance")
axis(1, 1:4, c("X1","X2","X3","X4"))
```

Remarque 1.18. Voici un exemple (choisi exprès pour qu'il marche bien...) où les indices HSIC détectent une dépendance que les corrélations ne détectent pas.

```r
X = matrix(1:314/100, ncol=1)
Y = sin(X)
plot(X,Y)
# Indice HSIC
sensi = sensiHSIC(X = X, kernelX = "rbf", kernelY = "rbf")
tell(sensi, Y)
print(sensi$S$original) # = 0.342 >> 0
# Correlations
print(cor(X,Y, method='pearson')) # = -0.009
print(cor(X,Y, method='spearman')) # = -0.005
print(cor(X,Y, method='kendall')) # = -0.003
```

1.3 Méthodes de forêts aléatoires

Les forêts aléatoires sont un outil d'apprentissage, c'est à dire un ensemble de méthode à qui on donne un ensemble de données "entrées", un ensemble correspondant de données "sorties", et qui essaie d'estimer la fonction "entrée $\mapsto$ sortie" sous-jacente. La plupart du temps, l'apprentissage est effectué dans un objectif de prédiction : Une fois la fonction sous-jacente estimée, on récupère de nouvelle données "entrée", pour lesquelles on ne connaît pas les données "sorties" correspondantes, et on utilise l'estimation de notre fonction pour prédire ces données "sorties".

Cepedant, cela n'est pas ce qui nous intéresse ici. On considère quand même des données "entrées", qui correspond aux variables $X_1,\dots,X_n$, et des données "sorties", qui correspondent à la variable X_0. Cepedant, à partir d'une estimation de fonction, on peut déduire une estimation de l'importance de chaque variable X_i dans la prédiction. C'est cette importance que nous utiliserons comme un indice d'influence de chacune des variables X_i sur X_0.

Definition 1.19. On considère des échantillons $(X_0^{(1)},\dots,X_p^{(1)}),(X_0^{(2)},\dots,X_p^{(2)}),\dots,(X_0^{(n)},\dots,X_p^{(n)})$ des $p+1$ variables. Les méthode des forêts aléatoires prend ces données en argument (ainsi que d'autres paramètres, peu importants ici, jouant un rôle similaire aux noyaux pour les indices HSIC), et renvoie une fonction $f_{RF}:\mathbb{R}^p\to\mathbb{R}$. Cette fonction est construite pour approcher la fonction $X_1,\dots,X_p\mapsto X_0$. Nous ne détaillons pas le fonctionnement de l'algorithme des forêts aléatoires ici.

Noter que la fonction f_{RF} n'est pas la même à chaque appel de l'algorithme, même si les données sont les mêmes. C'est le côté "aléatoire" de celui ci.

Definition 1.20. On suppose avoir séparé nos échantillons en deux catégories, de tailles n_1 et $n_2=n-n_1$. A partir des n_1 premiers échantillons, on applique l'algorithme RF pour obtenir une fonction f_{RF}. A partir des n_2 échantillons restant, on peut tester la qualité de f_{RF} en définissant

$$err := \frac{1}{n_2}\sum_{k=1}^{n_2}\left|f_{RF}(X_1^{(k)},\dots,X_p^{(k)})-X_0^{(k)}\right|^2,$$

où la somme porte sur les échantillons non utilisés dans l'apprentissage.

Ensuite, pour estimer l'importance d'une variable spécifique X_i, on peut procéder comme suit. On tire au hasard une permutation π à n_1 éléments. On définit un deuxième jeu de variable entrées $(X')_j^{(k)}, 0\le j\le p, 1\le k\le n_1$ par $(X')_j^{(k)}=X_j^{(k)}$ si $j\neq i$, et $(X')_i^{(k)}=X_i^{(\pi(k))}$. Autrement dit, on permute aléatoirement les échantillons de la variable X_i en question, mais pas les autres. L'idée est que dans ce jeu d'échantillons, la variable X_i n'a presquz plus aucun pouvoir prédictif. À partir de ces échantillons $(X')_j^{(k)}$, on applique à nouveau l'algorithme RF pour obtenir f'_{RF}, et on peut estimer une nouvelle erreur

$$err'_i := \frac{1}{n_2}\sum_{k=1}^{n_2}\left|f'_{RF}(X_1^{(k)},\dots,X_p^{(k)})-X_0^{(k)}\right|^2.$$

Remarquons que l'on peut calculer cette erreur avec les échantillons originaux $X_j^{(k)}$, non permutés. On définit finalement l'indice d'importance de la variable X_i comme étant

$$I_{RF}(X_i,X_0) := \frac{err'_i-err}{err}.$$

Cet indice mesure à quel point l'erreur quadratique moyenne augmente lorsque la variable X_j ne plus plus être utilisée efficacement pour la prédiction. Remarquons que cet indice peut être négatif : si par exemple une variable n'est pas influente, alors on peut avoir $err'_i<err$ uniquement par le bruit. Cependant, il est très probable que dans ce cas $|err'_i-err|\ll err$, c'est à dire $I_{RF}(X_i,X_0)\simeq 0$.

Remarque 1.21. L'indice d'influence $I_{RF}(X_i,X_0)$ dépend des valeurs des autres variables X_j. En effet, la fonction f_{RF} est entraînée sur l'ensemble des variables. C'est une différence assez fondamentale entre cette méthode d'estimation d'indices et les méthodes par corrélation ou HSIC, pour lesquelles les autres variables X_j n'ont pas d'effet. Une autre différence, liée à la précédente, est que l'influence de X_0 sur X_i n'est pas la même que celle de X_i sur X_0.

De plus, l'indice d'influence $I_{RF}(X_i,X_0)$ est un indice de prédiction : il décrit à quel point une variable peut être utile pour en décrire une autre.

Enfin, cet indice n'est calculé qu'empiriquement. Ce n'est pas l'estimation empirique d'un indice "parfait" qui proviendrait de la connaissance exacte de toute la loi jointe totale, contrairement aux indices de corrélation (de Pearson au moins) et HSIC (et à Sobol, voir section 1.4).

Remarque 1.22. La méthode de calcul de err, err'_i, puis I_{RF} est généralisable à n'importe quelle méthode d'apprentissage, il suffit de remplacer f_{RF} et f'_{RF} par d'autres estimations de la fonction "entrée $\to$ sortie" sous-jacente.

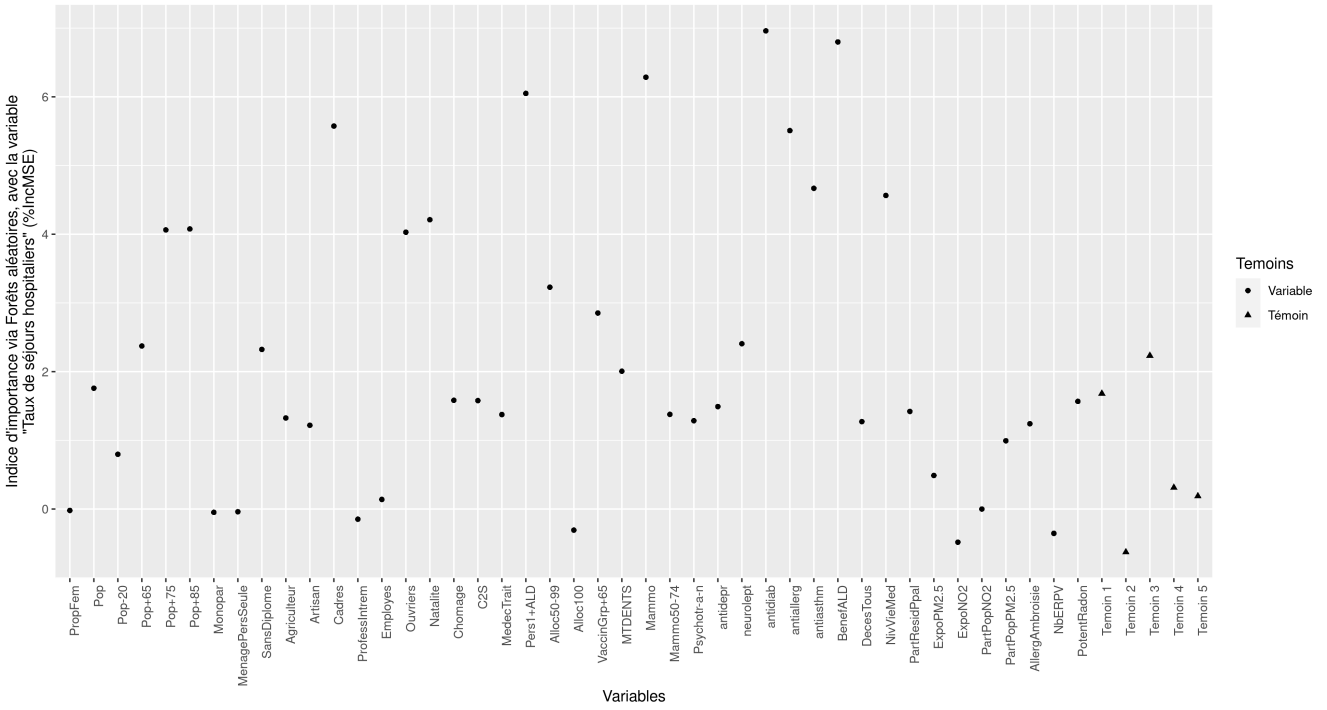


FIGURE 4 – Indices "%IincMSE" via Forêts aléatoires, entre la variable X_0 = "Taux de séjours hospitaliers pour tous motifs" et différentes variables de santé humaine ou sociale. A comparer avec les Figures 1, 3 et 5.

Remarque 1.23. En `R`, on peut calculer les indices d'influence grâce au package **`randomForest`**. Celui-ci permet en fait de calculer deux indices d'influence. Le premier s'appelle "%IncMSE" (% Increase in Mean Square Error) et correspond à I_{RF}. Le second est nommé "IncNodePurity" et n'est pas décrit dans ces notes (voir fin du paragraphe 4.1 de Genuer et Poggi [5] pour une brève explication, ou sur Wikipédia).

```
# Exemple de matrice, avec p=3 et n=100 echantillons
M = matrix(NA, ncol=4, nrow=100)
M[,1] = 3 * runif(100)
M[,2] = M[,1] + rnorm(100)
M[,3] = - 2 * M[,1] + rnorm(100)
M[,4] = 0.5 * M[,1] + 3 * runif(100)
colnames(M) = c("X0","X1","X2","X3")

library('randomForest')
RF = randomForest(x = M[,2:4], y = M[,1], importance = TRUE)
imp = importance(RF)
par(mfrow = c(1,2))
barplot(imp[,1], names.arg = c("X1","X2","X3"),
        xlab = "Variables", ylab = "Indice d'importance par erreur quadratique moyenne")
barplot(imp[,2], names.arg = c("X1","X2","X3"),
        xlab = "Variables", ylab = "Indice d'importance par purete du noeud")
```

1.4 Indices de Sobol

Definition 1.24. Soit $X_1, \dots, X_p \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ des variables aléatoires, et $f : \mathbb{R}^p \to \mathbb{R}$ und fonction quelconque. On définit les indices de sensibilité de Sobol

$$S_i := \frac{\mathrm{Var}\,\mathbb{E}[f(X_1, \dots, X_p) \mid X_i]}{\mathrm{Var}(f(X_1, \dots, X_p))}.$$

Dans la définition précédente, il faut penser que $X_0 = f(X_1, \dots, X_p)$, et on a alors $S_i = \dfrac{\mathrm{Var}\,\mathbb{E}[X_0 \mid X_i]}{\mathrm{Var}(X_0)}$. Cet indice de Sobol n'est **pertinent que lorsque les variables X_i sont indépendantes** (voir [3], la raison principale est que la décomposition de Hoeffding n'est orthogonale que dans le cas indépendant). Cette définition est théorique, et pour l'appliquer en pratique, il faut en construire un estimateur à partir d'échantillons. Le Lemme suivant va permettre ceci en donnant une expression alternative de S_i faisant disparaître les espérances conditionnelles.

Lemme 1.25. *("Pick-freeze") On suppose que les variables X_j, $1 \le j \le p$ sont indépendantes. On définit des copies indépendantes de ces variables $X'_j, 1 \le j \le p$, où X'_j à la même loi que X_j. On définit $X_0 = f(X_1, \dots, X_p)$ et $X_0^i = f(X'_0, \dots, X'_{i-1}, X_i, X'_{i+1}, \dots, X'_p)$. Alors on a*

$$S_i = \frac{\mathrm{Cov}(X_0, X_0^i)}{\mathrm{Var}(X_0)}.$$

Noter que $\mathrm{Var}(X_0) = \mathrm{Var}(X_0^i)$*, car X_0 et X_0^i ont même loi. De plus,*

$$\mathrm{Cov}(X_0, X_0^i) = \mathbb{E}[X_0 X_0^i] - \mathbb{E}[X_0]^2$$

L'expression donnée par le Lemme est maintenant facilement estimable par des moments.

Definition 1.26. On suppose que les variables X_j, $1 \le j \le p$ sont indépendantes. On considère des échantillons indépendants $X_j^{(k)}$ et $(X')_j^{(k)}$, avec $1 \le j \le p$ et $1 \le k \le n$. On définit $X_0^{(k)} = f(X_1^{(k)}, \dots, X_p^{(k)})$ et $X_0^{i(k)} = f((X'_0)^{(k)}, \dots, (X'_{i-1})^{(k)}, X_i^{(k)}, (X'_{i+1})^{(k)}, \dots, (X'_p)^{(k)})$ On définit

$$\widehat{S}_i := \frac{\widehat{\mathrm{Cov}}(X_0, X_0^i)}{\widehat{\mathrm{Var}}(X_0)},$$

avec

$$\widehat{\mathrm{Var}}(X_0) = \frac{1}{n}\sum_{k=1}^{n} (X_0^{(k)})^2 - \left(\frac{1}{n}\sum_{k=1}^{n} X_0^{(k)}\right)^2$$
$$\widehat{\mathrm{Cov}}(X_0, X_0^i) = \frac{1}{n}\sum_{k=1}^{n} X_0^{(k)}(X_0^{i(k)} - X_0^{(k)}).$$

Remarque 1.27. Il existe des estimateurs alternatifs à $\widehat{\mathrm{Cov}}(X_0, X_0^i)$, par exemple

$$\frac{1}{n}\sum_{k=1}^{n} X_0^{(k)} X_0^{i(k)} - \left(\frac{1}{n}\sum_{k=1}^{n} X_0^{(k)}\right)\left(\frac{1}{n}\sum_{k=1}^{n} X_0^{i(k)}\right)$$

ou

$$\frac{1}{n}\sum_{k=1}^{n} X_0^{(k)} X_0^{i(k)} - \left(\frac{1}{2n}\sum_{k=1}^{n} (X_0^{(k)} + X_0^{i(k)})\right)^2.$$

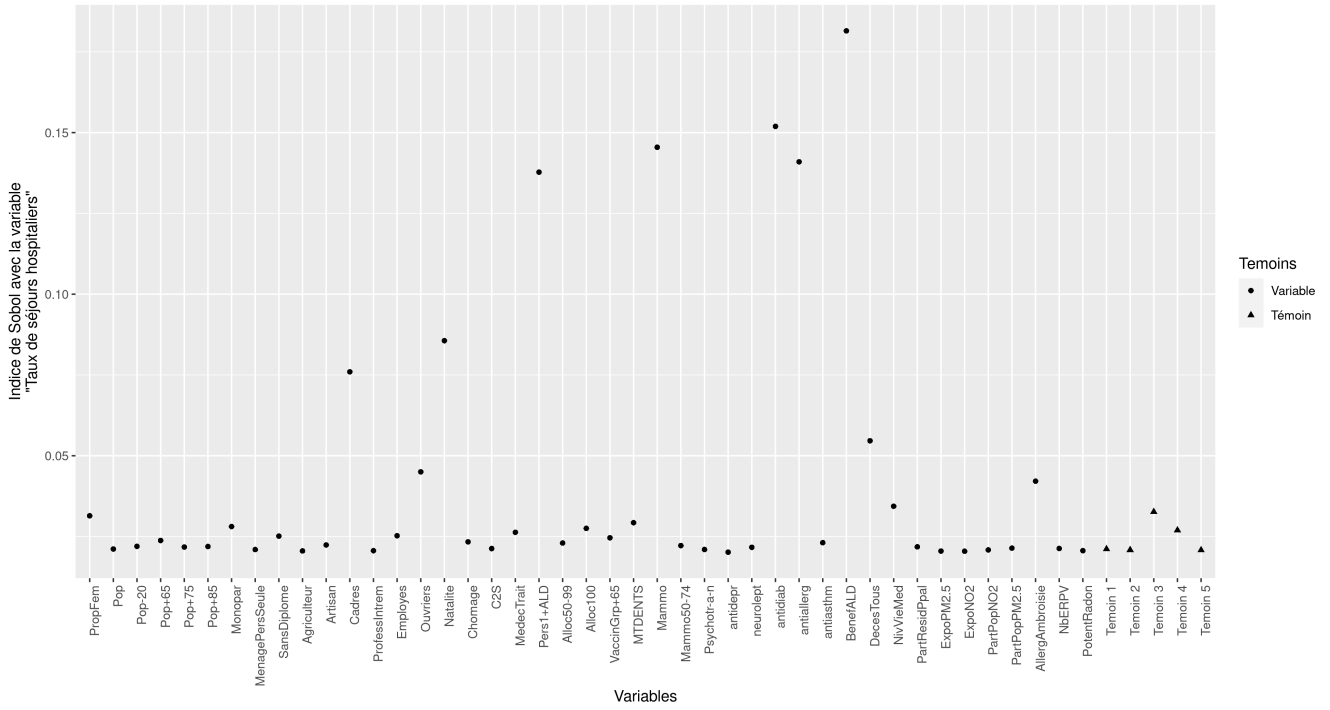


FIGURE 5 – Indices de Sobol entre la variable X_0 = "Taux de séjours hospitaliers pour tous motifs" et différentes variables de santé humaine ou sociale. L'apprentissage a été fait par Fôrets aléatoires. A comparer avec les Figures 1, 3 et 4.

Remarque 1.28. L'indépendance des variables $X_1, \ldots, X_p$ est une hypothèse cruciale pour le calcul des indices de Sobol. Pour cette raison, on utilise de l'apprentissage (ici, par forêts aléatoires), pour pouvoir choisir les valeurs des variables $X_1, \ldots, X_p$ indépendamment. Cependant, les données générées peuvent ne pas correspondre à une commune réaliste. Par exemple, il est possible qu'une commune fictive générée ait comme valeur (en éxagérant) "Propotion de cadres = 50%" et "Proportion d'ouvriers = 60%", ce qui est bien sûr impossible.

Remarque 1.29. En `R`, le package **`sensitivity`** permet de calculer les indices de Sobol.

```
# Exemple de matrice, avec p=5 et n=1000
M = matrix(NA, ncol=3, nrow=1000)
M[,1] = rnorm(1000)
M[,2] = runif(1000)
M[,3] = rnorm(1000)
colnames(M) = c("X1", "X2", "X3")
M1 = M[1:500,]
M2 = M[501:1000,]
# Les echantillons a mettre dans la machine de Sobol sont tres contraignants :
# il faut que les variables soient independantes, et il faut deux jeux de
# donnees independantes (ici M1 et M2)

Fonc = function(X) {
# La fonction (qui est censee etre une boite noire) que l'on etudie est :
# (x1, x2, x3) -> x1 + 2*x2 + 0.5*x3^2 + bruit
# Ici on l'applique a chaque ligne d'une matrice
        n = nrow(X)
        return(X[,1] + 2*X[,2] + 0.5*X[,3]^2 + 0.1 * rnorm(n))
}
```

```r
library('sensitivity')
Sobol = sobolEff(model = NULL, X1 = M1, X2 = M2)
Y = Fonc(Sobol$X)
tell(Sobol, y = Y)
barplot(Sobol$S$original, names.arg = c("X1","X2","X3"),
        xlab = "Variables", ylab = "Indice de Sobol")
# Noter que l'indice de X2 est plus petit que celui de X1 meme si son coefficient
# est plus grand dans la fonction Fonc. Cela est du au fait que les valeurs de X2 ont
# une variance plus petite (uniforme sur [0,1] au lieu de gaussienne de variance 1)
```

1.5 Comparaison des différentes méthodes

Rappelons les différents indices construits jusqu'ici.

- Les corrélations $|\widehat{\mathrm{Cor}}(X_0, X_i)|$, à valeur dans $[0,1]$ (ou $[-1,1]$ sans les valeurs absolues)
- Les indices HSIC, $I_{HSIC,k_0,k_i}(X_0, X_i)$, à valeur dans $[0,1]$.
- Les indices d'importance $I_{RF}(X_0, X_i)$, à valeur dans $\mathbb{R}$, mais jamais vraiment négatifs (on peut considérer qu'un indice négatif est zéro).
- Les indices de Sobol $\widehat{S}_i$, à valeur dans $[0,1]$.

Pour comparer ces différents indices, il est nécessaire d'avoir une échelle commune. La plupart des indices sont à valeur dans $[0,1]$, mais sans véritable comparaison directe possible. Par exemple, un certain indice HSIC de 0.1 pourra être considéré comme élevé, alors qu'une corrélation de 0.1 sera plutôt faible.

Une manière de remédier à ce problème est de regarder le rang des indices au lieu de leur valeur. Cependant, cette solution ne permet pas de différencier de fossé entre les variable plutôt influentes, et les variables plutôt indépendantes. Par exemple, si deux variables explicative ont des indices beaucoup plus grands que les autres, alors on ne verra pas qu'elles sont deux, et pas trois.

Un autre manière de remédier au problème est de normaliser chaque indice, de manière à ce sa moyenne soit 0 et sa variance 1.

En plus de comparer les méthodes entre elles, on peut aussi comparer les variables X_0 entres elles : Si on a deux variables X_0, X_0', on peut mesurer l'influence sur elles des variables $X_1, \ldots, X_p$, et comparer les deux graphes obtenus.

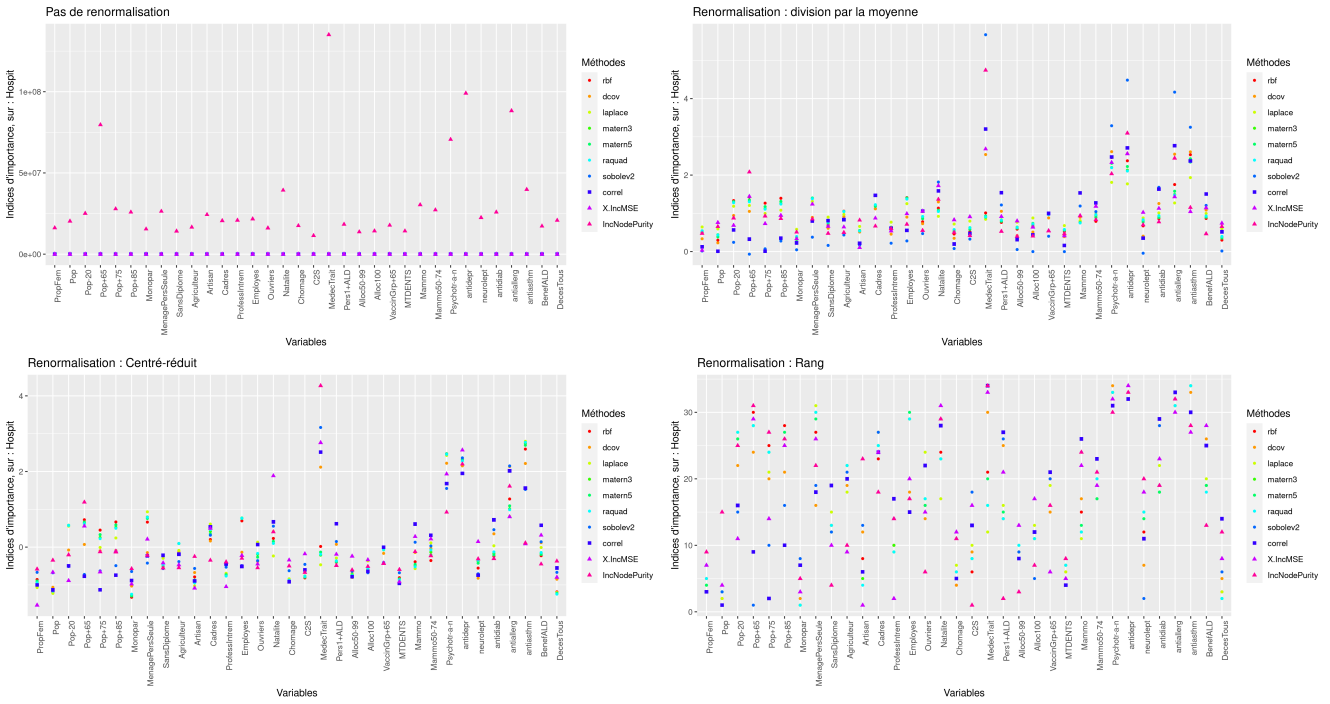


FIGURE 6 – (La figure est un peu petite... mais il n'y a pas grand chose d'intéressant, ce que je veux montrer c'est qu'on ne voit rien sur ces graphiques...

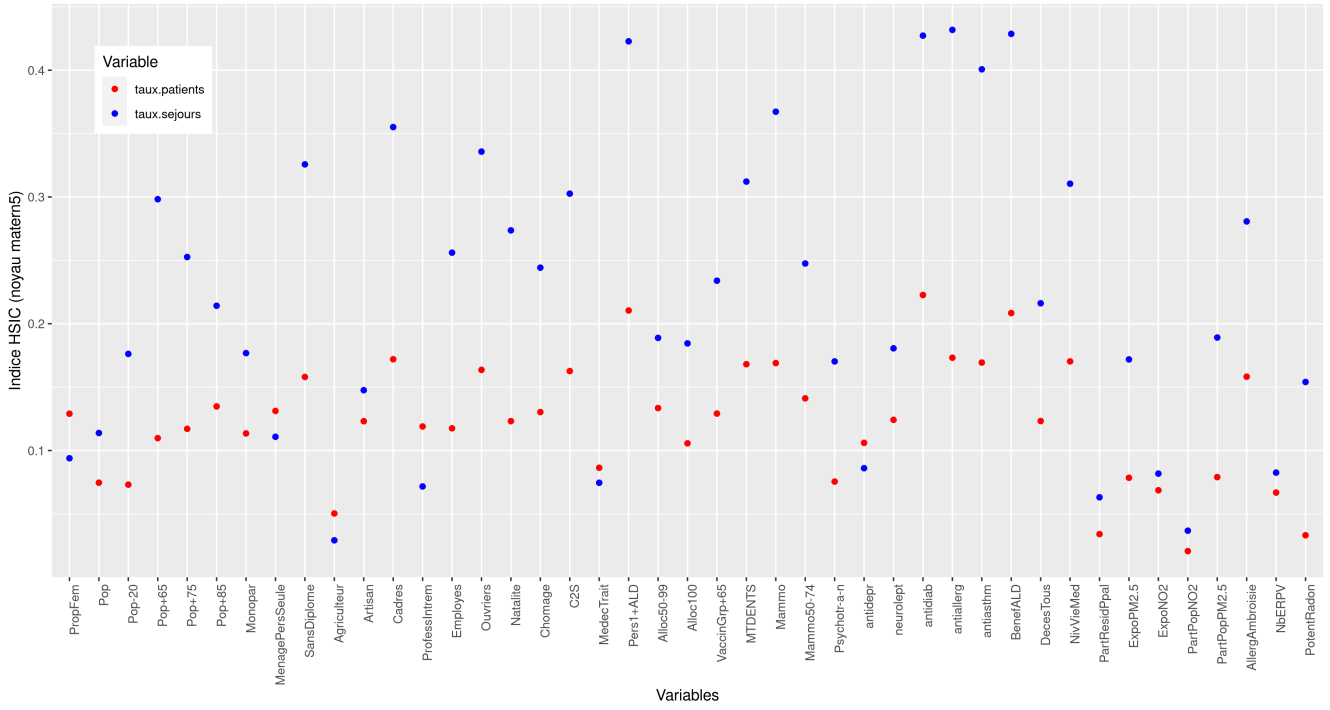


FIGURE 7 – Indices HSIC entre d'une part les variables "Taux de séjours hospitalier toutes causes" et "Taux de patients hospitalisés toutes causes", et d'autre part différentes variables de santé. On peut voir qu'en général, les indices pour les séjours hospitaliers sont plus grands : Cela signifie (pour moi) que les séjours hospitaliers sont un meilleur indicateur à prendre.

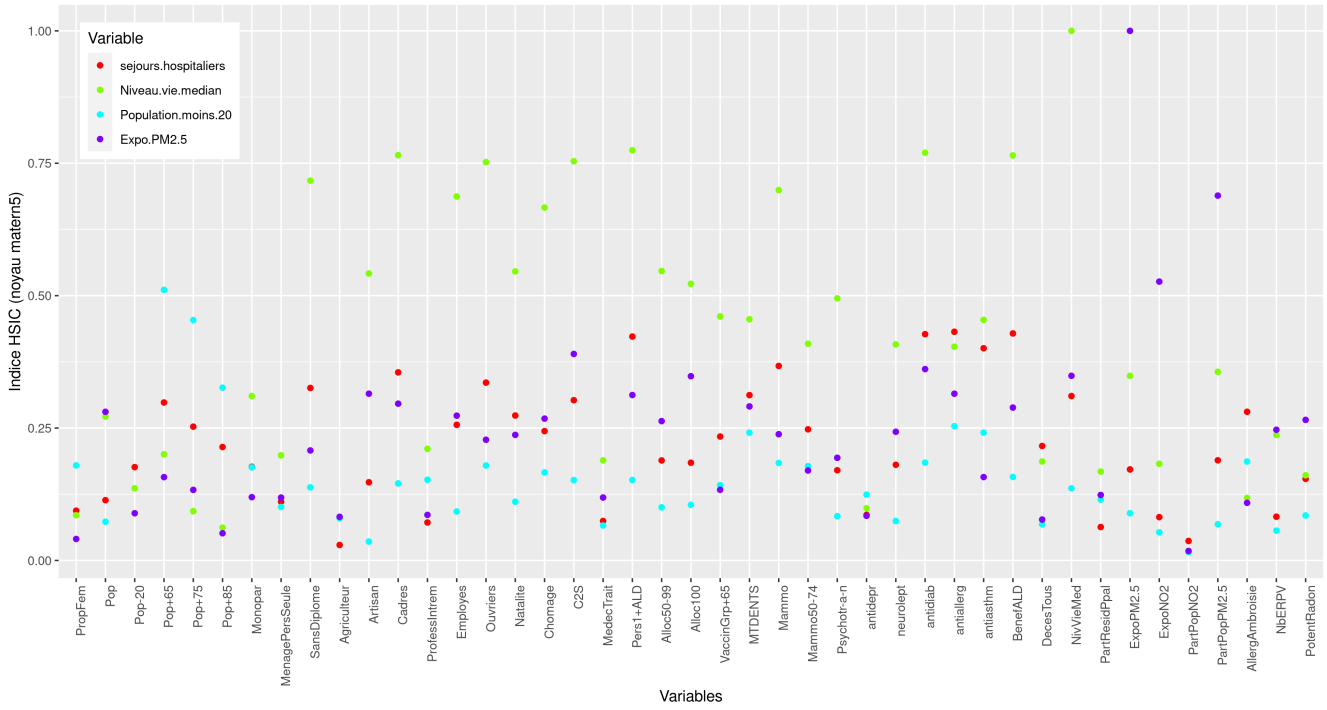


FIGURE 8 – Même graphique que la Figure 7, mais avec 4 variables en ordonnée : "Taux de séjours hospitaliers toutes causes", "Niveau de vie médian", "Proportion de la population de moins de 20 ans", "Exposition de la population aux PM2.5". On voit que la variable "Niveaude vie médian" est celle qui est le plus reliée à la plupart des autres variables, ce qui est logique car la plupart de ces variables sont sociales ou medico-sociales. On voit aussi des indices plus hauts pour "Population de moins de 20 ans" dans les variables démographiques, et des indices plus hauts pour les PM2.5 dans les variables environnementales.

2 Visualisation de la structure globale des variables

Un désavantage des méthode vues jusqu'ici est qu'elle se concentrent seulement sur une variable à expliquer X_0. Pour avoir une vue plus globale des intéractions entre les variables, on peut permuter les variables pour faire jouer le rôle de X_0 à chacune des variables. En d'autres termes, pour chaque variable X_j, on va chercher quelles sont les variables qui ont le plus d'influence sur elle. Noter que le terme "influence" n'est pas à prendre ici dans le sens causal, mais seulement dans le sens où "X_i est influente sur X_j" si "Une grande variation de mesure sur X_i implique une grande variation de mesure sur X_j".

2.1 Matrices d'indices

On considère $X_1,\dots,X_p$, et $Y_1,\dots,Y_q$ des variables, on cherche à étudier l'influence de chacune des variables X_i sur chacune des variables Y_j. Selon la notation de la section 1, cela revient à prendre à tour de rôle $X_0 = Y_j$. Pour une méthode donnée, on obtient pq indices, qui sont des indicateurs de l'influence de chacun des X_i sur chacun des Y_j. Il est possible de comparer ces indices par colonnes, c'est à dire à Y_j fixé, cela revient aux études faites en section 1. Il est aussi possible de comparer ces indices par lignes, c'est à dire à X_i fixé (ce qui n'apporte rien si les indices sont symétriques). Dans ce cas, on fixe une variable explicative, et on regarde sur quelles autres variables elle a de l'influence. Il est également possible de comparer les colonnes/lignes entre elles. Par exemple, un colonne qui aura en moyenne des indices plus grand que les autres est une variable influente sur beaucoup d'autres variables.

Il est possible de choisir $p = q$ et $Y_i = X_i$. Dans ce cas, une autre analyse des résultats est possible, qui s'approche du "Stochastic Block Model" (SBM). On peux essayer de rassembler les variables en

plusieurs groupes qui s'influencent beaucoup entre elles, mais peu entre les groupes. Cet objectif s'appelle la recherche de communautés, et fait l'objet de la section 2.2.

Remarque 2.1. Dans le cas $p = q$ et $Y_i = X_i$, et pour les indices de corrélation et HSIC, la matrice de indices sera symétrique, et l'analyse par lignes devient la même que l'analyse par colonnes. Cela vient du fait que les indices en question sont symétriques : par exemple, $\mathrm{Cor}(X_i, X_j) = \mathrm{Cor}(X_j, X_i)$ et $I_{HSIC,k_i,k_j}(X_i, X_j) = I_{HSIC,k_j,k_i}(X_j, X_i)$.

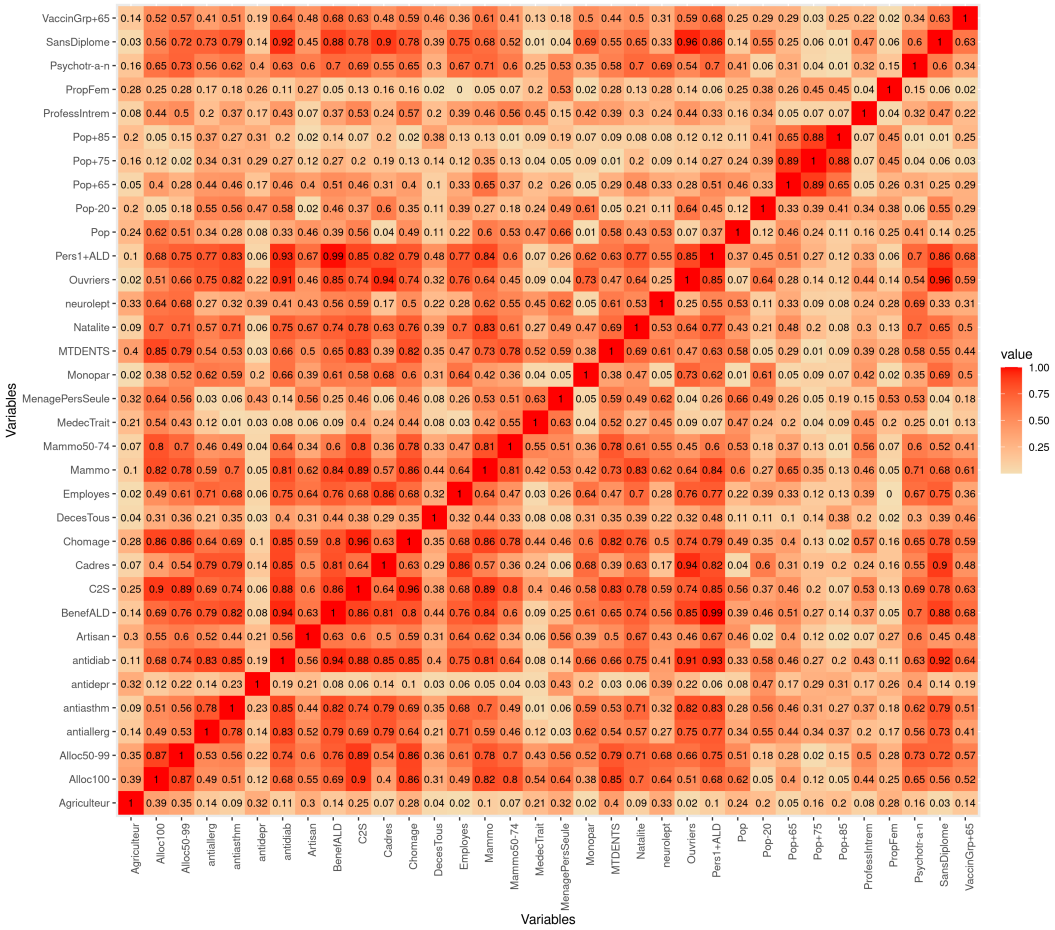


FIGURE 9 – Les valeurs absolues de corrélations entre différentes variables

2.2 Recherche de communauté

Lorsqu'on a beacoup de variables (i.e. $p \gg 1$), alors il peut être difficle d'interprêter une grande matrice d'indice, ou un grand graphe. Pour faire un peu de ménage, on va chercher des groupes de variables qui sont très reliées entre elles. On appelera ces groupes des communautés.

Definition 2.2. Un *graphe pondéré* est (ici) un ensemble de sommets G muni d'une application $\omega : G^2 \to \mathbb{R}$, telle que $\omega(u, v) = \omega(v, u)$. On obtient un graphe (non-dirigé) sous-jacent sur G en prenant pour arêtes $E = \{\{u, v\} \mid \omega(u, v) \neq 0\}$. Inversement, si l'on dispose d'un graphe non dirigé (G, E), on peut obtenir un graphe pondéré en posant $\omega(u, v) = \mathbb{1}_{\{u,v\} \in E}$.

Si $G = \{1, \ldots, p\}$, un graphe pondéré peut être vu de manière équivalente comme une matrice $\Omega \in \mathcal{M}_p(\mathbb{R})$.

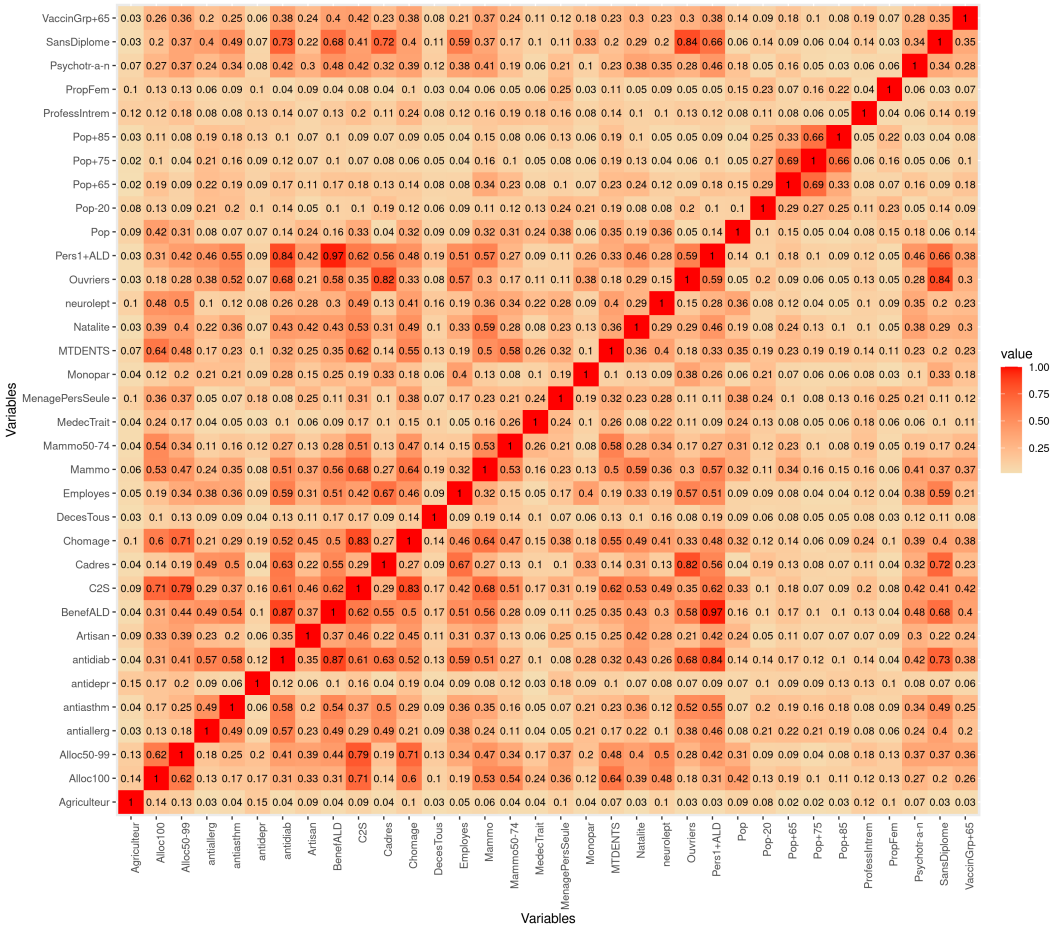


FIGURE 10 – Les indices HSIC entre différentes variables (Voir le code, on peut aussi mettre des 0 sur la diagonale pour faire plus ressortir les autres valeurs)

Definition 2.3. Soit (G,ω) un graphe pondéré. Pour $A,B \subseteq G$, on définit le flux observé

$$I_{A,B} := \sum_{\substack{u\in A\\ v\in B}} \omega(u,v).$$

Pour $A \subseteq G$, on note $\overline{A} := G\backslash A$. Étant donné $A_1,\dots,A_k \subseteq G$ une partition des sommets de G, on définit la *modularité* de cette partition par

$$m(A_1,\dots,A_k) = \frac{1}{(I_{G,G})^2}\sum_{i=1}^{k}\left(I_{A_i,A_i}I_{\overline{A_i},\overline{A_i}} \;-\; I_{\overline{A_i},A_i}I_{A_i,\overline{A_i}}\right).$$

Remarque 2.4. Quelque soit la partition, on a toujours $m(A_1,\dots,A_k) \in [-1,1]$. Il faut interpréter une grande modularité comme un bon découpage en communautés : Les termes positifs $I_{A_i,A_i}I_{\overline{A_i},\overline{A_i}}$ sont ceux compatibles avec la partition, et les termes négatifs $I_{\overline{A_i},A_i}I_{A_i,\overline{A_i}}$ sont ceux qui lui sont incompatibles.

Remarque 2.5. Dans le cas $k=1$, on a toujours $m(A_1)=0$. Il est possible que pour tous $A_1,\dots,A_k$ avec $k\geq 2$ on ait $m(A_1,\dots,A_k)<0$, c'est le cas où "pas de découpage est le meilleur découpage". Selon mes calculs, la borne supérieure $m\leq 1$ est un peu large. On a en fait toujours $m \leq \frac{k-1}{k}$ (si mes calculs sont justes), donc par exemple si $k=2$ on ne peut pas dépasser $m=\frac{1}{2}$.

Remarque 2.6. En `R`, le package `modMax` comporte différents algorithmes de détection de communauté par maximisation de la modularité. On pourra aussi regarder le package `leiden`, qui implémente l'algorithme

Leiden (qui ne fonctionne pas par maximisation de la modularité, mais qui est aussi un algorithme de recherche de communautés. Je n'ai pas approfondi...). Dans ce package, j'ai utilisé les méthodes dite d'optimisation spectrale pour construire des communautés. La première est `spectralOptimization`, elle ne prend que le graphe en entrée. Les trois suivantes sont `multiWay`, `spectral1` et `spectral2`, qui trouvent des communautés à partir du graphe, et on peut leur spécifier un nombre maximal de communautés à trouver. Dans tous les cas, ces fonctions prennent le graphe sous le format d'une matrice d'adjacence.

```
adj = matrix(runif(400), nrow=20, ncol=20)
adj = (adj + t(adj))/2 # Pour obtenir une matrice d'adjecence symetrique
library('modMax')
spectralOptimization(adj)
multiWay(adj, maxComm = 3)
spectral1(adj, maxComm = 3)
spectral2(adj, maxComm = 3)
```

2.3 Graphe de causalité

On se réferre à Murphy [8], chapitre 10 pour la définition d'un réseau bayésien (i.e. modèle graphique dirigé). On pourra regarder le chapitre 19 (Champs aléatoires de Markov, i.e. modèle graphique non dirigé), et la chapitre 26 (Apprentissage d'un modèle graphique).

(Voir aussi mes notes scannées)

Definition 2.7. On considère une graphe dirigé (G,E), où G est une ensemble (fini) de sommets, et $E \subseteq G^2$ est un ensemble d'arêtes (dirigées). On considère que $(u,v) \in E$ correspond à une arête $u \to v$. On note abusivement G pour (G,E). Un cycle (dirigé) est une suite $v_1,\dots,v_n$ de sommets tels que pour tout $0 \leq i \leq n$, $(v_i, v_{i+1}) \in E$, avec $v_{n+1} := v_1$. On dit que G est acyclique, ou que G est un *DAG* (Directed Acyclic Graph) si il n'existe pas de tel cycle dans G.

Soit G un DAG. Un *ordre topologique* sur G est un ordre sur ses sommets tel que pour toute arête $(v_1,v_2) \in E$, on ait $v_1 < v_2$. Un tel ordre existe toujours, mais n'est presque jamais unique. Pour $v \in G$, on note $pa(v) = \{u \in G \mid (u,v) \in E\}$ l'ensemble des parents de v. On définit $pred(v) = \{u \in G \mid$ il existe un chemin dirigé de u à $v\}$ les prédecesseurs de v.

Si $(X_v)_{v\in G}$ est une famille indexée par les sommets G et $A \subseteq G$, on notera $X_A = (X_u)_{u\in A}$. Donc par exemple, $X_G = (X_v)_{v\in G}$ et $X_{pa(v)} = (X_u)_{u\in pa(v)}$.

Definition 2.8. Soit G un DAG, et $X = (X_v)_{v\in G}$ un ensemble de variables aléatoires (définies sur un même espace). On dit que le graphe G décrit la structure de X (terminologie inventée par moi-même, donc qu'on ne retrouve nulle part à part ici) si la loi jointe de X s'écrit

$$d\mathbb{P}_X(x_G) = \prod_{v\in G} d\mathbb{P}_{X_v|X_{pa(v)}=x_{pa(v)}}(x_v).$$

Remarque 2.9. On note $G = \{v_1,\dots,v_p\}$. Si X a pour densité $d\mathbb{P}_X(x_G) = f_X(x_G)d\mu_{v_1}(x_{v_1})\cdots d\mu_{v_p}(x_{v_p})$ et les lois conditionnelles $X_v|X_{pa(v)}$ ont pour densité $d\mathbb{P}_{X_v|X_{pa(v)}} = f_{X_v|X_{pa(v)}}(x_v, X_{pa(v)})d\mu_v(x_v)$, alors le fait que G décrive X se réécrit

$$f_X(x_G) = \prod_{v\in G} f_{X_v|X_{pa(v)}}(x_v, x_{pa(v)}).$$

C'est d'ailleurs la définition donnée dans Murphy [8], où toutes les lois ont une densité (voire sont discrètes).

Remarque 2.10. Étant donné $X = (X_1,\dots,X_p)$, il peut y avoir plusieurs graphes G décrivant la structure de X (et ayant donc pour sommets $\{1,\dots,p\}$). En fait, quelque soit X, le graphe "dirigé complet" G avec pour sommets $\{(i,j) \in \{1,\dots,p\}^2 \mid i<j\}$ décrit toujours la structure de X. Plus généralement, si G décrit la structure de X, alors tout DAG G' contenant G décrit encore la structure de X. Le plus intéressant

est de prendre "le plus petit" possible (même s'il n'en existe pas un plus petit en général, on verra en définition 2.13 une manière de résoudre cette ambiguité).

Proposition 2.11. *Soit G un DAG et $X = X_G$ un ensemble de variables aléatoires dont la structure est décrite par G. Alors pour tout sommet $v \in G$, on indépendance conditionnelle entre un sommet et ses prédecesseurs non direct sachant ses parents :*

$$X_v \perp\!\!\!\perp X_{pred(v)\setminus pa(v)} \mid X_{pa(v)}.$$

Remarque 2.12. De manière plus générale, il existe une notion de d-séparation (voir section 2.1 de [7], ou section 10.5.1 de Murphy [8], moins claire) entre deux ensembles disjoints de sommets A et B par un troisième ensemble E (la notion n'est pas très dure à expliquer mais je ne fais pas ici).

Ainsi, pour trois ensembles disjoints de sommets $A, B, E \subseteq G$, on a "$X_A \perp\!\!\!\perp X_B \mid X_E$ pour tout X décrit par G" si et seulement si A et B sont d-séparés par E.

On peut alors formaliser la notion de "plus petit DAG" décrivant la structure de X.

Definition 2.13. On dit que G décrit *fidèlement* (faithfully en anglais) la structure de X si : pour tous ensembles disjoints d sommets $A, B, E \subseteq G$, on a $X_A \perp\!\!\!\perp X_B \mid X_E$ si et seulement si A et B sont d-séparés par E.

Remarque 2.14. Il n'est pas clair a priori qu'étant donné X, il existe un DAG décrivant fidèlement la structure de X. C'est quand même le cas, et l'algorithme PC va en donner un.

Recherche d'un DAG fidèle à un ensemble de variables aléatoires

(On suit Kalisch-Bühlmann [7])

On suppose disposer de variables aléatoires $X = (X_1, \dots, X_p)$. On veut chercher un DAG G qui décrit fidèlement X. C'est le but de l'algorithme PC (Peter-Clark). Celui-ci prend en entrée des informations d'indépendance conditionnelle sur les variables, et donne en sortie un DAG G représentant fidèlement la structure de X.

Si l'on dispose seulement d'échantillons des variables $X_1, \dots, X_p$, alors on a besoin d'un test d'indépendance conditionnelle (de plus, on va appliquer ce test un certain nombre de fois, donc il faut faire attention aux problématiques de tests multiples). Un cadre dans lequel cela est réalisé est lorsque X est un vecteur gaussien. Dans ce cas, on peut mesurer relativement simplement l'indépendance conditionnelle grace aux covariances.

Remarque 2.15. En `R`, le package `pcalg` permet d'appliquer l'algorithme PC à un ensemble d'échantillons. Il faut également fournir un test d'indépendance conditionnelle, et un test dans le cas gaussien integré dans le package (`gaussCItest`).

```
# Exemple de matrice, avec p=5 et n=100 echantillons
M = matrix(NA, ncol=5, nrow=100)
M[,1] = 3 * rnorm(100)
M[,2] = M[,1] + 0.1*rnorm(100)
M[,3] = - 2 * M[,1] + 0.2*rnorm(100)
M[,4] = -3 * M[,2] + 2 * M[,3] + 0.3*rnorm(100)
M[,5] = - M[,2] + 0.1*rnorm(100)
colnames(M) = c("X1", "X2", "X3", "X4", "X5")

library('pcalg')
library('igraph')
SuffStat = list(C = cor(M), n = nrow(M))
alpha = 0.05 # A choisir sciemment...
# Application de l'algorithme PC, avec un test d'independance gaussien "gaussCItest"
G_PC = pc(SuffStat, indepTest = gaussCItest, alpha = alpha, p = ncol(M))
```

```
Amat = as(G_PC, "amat") # Matrice d'adjecence
G = graph_from_adjacency_matrix(Amat)
tkplot(G, vertex.label = colnames(M), canvas.height=800, canvas.width=1000)
# Ne fontionne pas tres bien sur mon exemple...
```

A noter que la constante α utilisée dans les tests doit être choisie prudemment, notamment à cause de la problématique des tests multiples. Dans Kalisch-Bühlmann [7], il est montré qu'on a consistance en prenant $\alpha_n = 2(1-\Phi(\frac{\sqrt{n}c_n}{2}))$, où n est le nombre d'échantillons, et c_n est une borne inférieure à toutes les valeur absolues de corrélations conditionnelles non-nulles (soit ces corrélations sont nulles, soit elles sont éloignée de 0 d'au moins c_n).

Remarque 2.16. Ces graphes sont à prendre avec beaucoup de pincettes. Il ne faut absolument pas interpreter une flèche comme un lien causal montré par les données (même si c'est l'origine théorique), mais lutôt comme une indication que ces variables sont reliées ente elles.

2.4 Lorsque le graphe du modèle graphique est déjà connu

Dans cette section, on se donne un jeu de variables $X = X_1, \cdots, X_p$. On suppose que l'on connait un DAG G ayant pour sommets $\{1,\dots,p\}$ et décrivant la structure de X (par exemple en ayant demandé à des experts, en essayant de le deviener nous-mêmes, ou en ayant déjà appris le graphe par un algorithme quelconque...).

La distribution des variables s'écrit alors, en intégrant des paramètres :

$$f_X(x_G) = \prod_{v \in G} f_{X_v|X_{pa(v)}}(x_v, x_{pa(v)}, \theta_v).$$

Ensuite, il y a deux cas de figure :

— Si l'on peut observer toutes les variables X_i, $i \in G$. Dans ce cas, à partir de ces observations, on peut estimer les paramètres θ_v à partir des observations pour affiner notre comprehension de la structure globale des variables.

— Si l'on ne peut observer qu'un partie des variables X_i, $i \in E \subsetneq G$. Dans ce cas, il est plus difficle d'estimer les paramètres θ_v, mais cela est encore possible, par exemple avec l'algorithme EM. Dans ce cas, le but peut également de prédire la valeur de variables X_i, $i \in H$ avec $H \subseteq G \backslash E$.

Remarque 2.17. Thibault a conseillé les packages `R` suivants pour manipuler les réseaux bayésiens : `rjags`, `rstan`, (et `rbugs`, pas trouvé). J'ai surtout exploré `rstan` de mon côté.

Quelques remarques dans le cas gaussien.

On fait l'hypothèse que X_G est un vecteur gaussien, et on écrit $G = \{1,\dots,p\}$ avec un ordre topologique. On peut alors écrire

$$X_k | X_{pa(k)} \sim \mathcal{N}(\mu_k + \sum_{l \in pa(k)} w_{l,k} X_l, \sigma_i).$$

On note $W = (w_{l,k})_{1 \le l,k \le p)}$, qui représente la structure du DAG G : il y a une arête $l \to k$ si et seulement si $w_{l,k} \neq 0$. On peut remarquer que $X - WX \sim \mathcal{N}(\mu, \Sigma)$, avec $\mu = (\mu_1,\dots,\mu_p)$ et $\Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_p)$. Or comme G est un DAG, W est triangulaire inférieure stricte et $I - W$ est inversible. En notant $U = (I-W)^{-1}$, on a alors $X \sim \mathcal{N}(U\mu, U\Sigma U^T)$.

Noter que $U\Sigma U^T$ est la décomposition (unique) de Cholesky de la matrice de covariance de X_G. Il est donc possible de retrouver U (et donc W, et donc la structure du graphe G), à partir de la matrice de covariance de X_G.

3 Application à des variables de santé

Cette section vise à expliquer mon avis sur l'utilisation de toutes les méthodes présentées précédemment. Je me suis un peu laisser aller, j'ai mis beaucoup d'idée pas très claires, et je ne sais pas à quel point elles sont pertinentes.

3.1 Quelques réflexions liés aux variables

Les données se trouvent dans une matrice avec p colonnes (les variables), et n lignes (les échantillons). Idéalement, les lignes sont des échantillons indépendant d'une même loi. En pratique, on peut se demander comment on s'approche le plus possible de cette hypothèse.

Dans notre cas, le lignes sont des communes de la métropole de Lyon. Comme ces communes s'influencent entre elles, elles ne sont pas indépendantes. (par exemple deux communes voisines, ou avec le même parti politique à leur tête, sont corrélées).

Je pense qu'il est trompeur de regarder directement les variables, sans mettre en contexte leurs valeurs. Je vais donner trois exemples, en exagérant les chiffres.

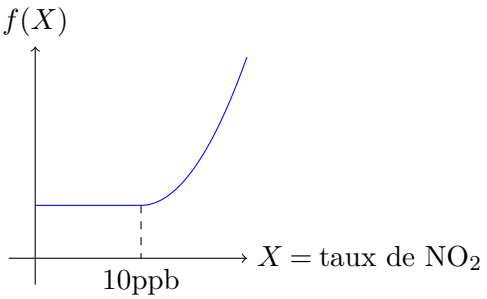
Premier exemple : Il est dans mon sens bien préférable que le niveau de vie médian passe de 10000€/an (pauvre) à 20000€/an (classe moyenne), plutôt qu'il passe de 30000€/an (assez riche) à 60000€/an (très riche).

Autre exemple : Supposons que la quantité pré-industrielle de NO_2 dans l'atmosphère est 10ppb (je ne connais pas la vraie valeur). A cause des non-linéarités, il peut être préférable de passer de 30ppm à 25ppm plutôt que de passer de 15ppm à 10ppm.

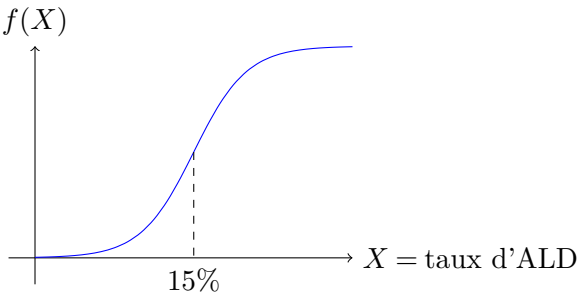
Un dernier exemple (peut être plus discutable) : Faire baisser le taux d'ALD de 35% à 30% est-il plus utile que de le faire baisser de 10% à 5% ? La raison serait la suivante : Que le taux d'ALD soit 10% ou 5% (deux valeurs en dessous de la moyenne nationale actuelle), chaque patient reçoit suffisamment d'aide pour faire face à sa condition. Alors qu'à 30% ou 35% (valeurs très élevées), le système d'aide est saturé, et la quantité d'aide disponible ne change pas, donc diminuer le nombre de patients en ALD augmente par la même occasion la quantité d'aide que chaque personne ayant encore une ALD reçoit.

Un manière de penser le problème est de dire que les indicateurs sont mal choisis. Par exemple, au lieu de regarder le niveau de vie médian, on peut regarder le taux de personnes en situation de pauvreté. Ou encore, au lieu de regarder la quantité de NO_2 en ppm, on peut regarder le risque de développer une maladie à cause de la pollution, ou le risque de dégrader la biodiversité environnante. Cependant, une variable peut aussi être choisie parce qu'elle est facile à mesurer, alors que leurs alternatives ne le sont pas.

Je propose une autre solution possible à ce problème : A une variable X, on applique une fonction $f(X)$ qui reflète la qualité de cette variable. Par exemple, on peut mettre un seuil : si X est le niveau de vie médian, on peut regarder $f(X) = \mathbb{1}_{X \geq 15000}$. On peut aussi mettre plusieurs seuils pour faire une fonction "en escalier". Par exemple, pour la quantité de NO_2, je verrai bien un fonction qui ressemble à :



Pour le taux d'ALD, on peut imaginer qu'avoir 15% est un bon objectif. On peut alors considérer comme renormalisation :



Bien sûr, cette renormalisaiton dépend du contexte (le choix de 15%), et pourrait changer au cours du temps (donc c'est pas un "indicateur universel").

Il y a un dernier problème qui me dérange, qui est assez lié au précédent. Imaginons une variable X qui soit le taux de tel cancer. Dans la plupart des méthodes (corrélations, certains HSIC(?), Random Forest, Sobol...), remplacer X par une de ses transformations affines $aX + b$ ne change rien. Autrement dit, on peut supposer toutes les variables centrées réduites. Mais alors les outils ne sauront pas si X varie plutôt entre 0.1% et 0.2%, ou si X varie plutôt entre 10% et 20%. Dans le premier cas, les outils diront "X a beaucoup évoluée" si X passe de 0.19% à 0.11%. Dans le deuxième cas, les outils diront "X n'a pas beaucoup évoluée" si X passe de 19% à 18%. Alors que dans le premier cas cela concerne 0.08% de la population, et 1% dans le deuxième cas.

J'essaie de dire mes idée encore très floues encore d'une autre manière. On considère la variable $X =$ "taux de telle maladie". On va essayer de comparer 2 scénarios. Dans le premier cas, on arrive à faire passer X de 2% à 0.2%. Dans le deuxième cas, on arrive à faire passer X de 10% à 5%.

Cas 1 : On a divisé le nombre de malades par 10, mais on a guérit 1.8% de la population.

Cas 2 : On a divisé le nombre de malades seulement par 2, mais on a guérit 5% de la population.

Quel est le plus souhaitable ? Est-il possible de dire que l'un est plus souhaitable que l'autre ? Ce que l'on fait actuellement avec nos outils, c'est aucun des deux. Ce que l'on fait, c'est comparer la réduction (1.8% ou 5%) à l'écart-type de la distribution de X parmi les communes choisies.

3.2 Détection des variables importantes

Les données de santé que l'on manipule comportent un grand nombre de variables. Les outils que l'on a développés jusqu'ici pourraient aider à selectionner un plus petit jeu de variables, plus facile à regarder. A destination de politiques qui n'ont pas le temps ou l'envie de lire un rapport de 200 pages par exemple ? On peut alors chercher des variables qui sont fortement corrélées à beaucoup d'autres variables. Pour ce faire, on peut examiner les matrices d'indices, et choisir les variables dont les colonnes sont les plus élevées (les plus rouges).

On peut aussi utiliser l'outil "recherche de communautés" pour cette tâche. On peut par exemple essayer de garder une variable par groupe de variable interconnectées.

On peut aussi détecter des liens entre variables que l'on ne soupconnait pas auparavant ? Par exemple le taux de recours aux dentistes est un bon indicateur du niveau de vie médian.

3.3 Arbitrage

Je vois deux possibilités d'usage des outils pour de l'arbitrage :

— Un projet est proposé. On regarder sur quelles variables le projet agit directement, et on essaie de prédire quel sera les conséquences sur l'ensemble des variables.

— Comme le premier point, mais en amont du projet : On étudie quelle(s) variable(s) devrai(en)t être changée(s) pour avoir un impact positif sur la santé globale. Ensuite, on cherche des projets qui agissent sur ces variables.

Dans les deux cas, la modélisation mathématique est essentiellement la même, c'est un peu l'inverse de l'analyse de sensibilité : On ne cherche pas à voir quelles variables vont influencer une variables finale, mais quelles variables sont influencées par une variables initiale. Avec les indices symétriques (corrélation et HSIC), c'est la même chose. On peut interpréter les indices d'influence dans un sens mais aussi dans l'autre (i.e. on regarde les colonnes au lieu des lignes, mais ce sont les mêmes). En fait, même pour Random Forest et Sobol, on peut aussi faire ca : on génère toute la matrice des indices (ligne par ligne), puis on regarde les colonnes correspondant aux variables impactées directement par le projet.

Dans les faits, on peut utiliser ces matrices dans le but de voir les variables qui vont être le plus influencées, et se demander "est-ce qu'on a pensé à évaluer l'impact sur telle variable ?"

Outre l'utilisation des matrices d'indices, je pense qu'il est pertient d'utiliser des réseaux bayésiens ici, ou au moins d'essayer. On se débrouille pour construire un réseau bayésien qui modélise l'intéraction de nos variables. Ensuite, on donne la valeur de certaines variables, et on simule le reste des variables pour voir comment elles vont évoluer.

A Liste des variables utilisées et/ou accessibles

Récupéré sur BALISES.

L'indice correspond au numéro de la ligne dans le fichier contenant toutes les variables.

Indice	Abbréviation	Nom complet
3	PropFem	Proportion.de.femmes
4	Pop	Population
5	Pop-20	Population.des.moins.de.20.ans (en %)
6	Pop+65	Population.des.65.ans.et.plus (en %)
7	Pop+75	Population.des.75.ans.et.plus (en %)
8	Pop+85	Population.des.85.ans.et.plus (en %)
9	Monopar	Ménage.dont.la.famille.principale.est.monoparentale (en % des ménages)
10	MenagePersSeule	Ménage.de.personne.vivant.seule (en % des ménages)
11	SansDiplome	Population.de.15.ans.et.plus.non.scolarisée.sans.diplôme
12	Agriculteur	Population.d.agriculteurs.exploitants.actifs.de.15.à.64.ans
13	Artisan	Population.d.artisans..commerçants..chefs.d.entreprise.actifs.de.15.à.64.ans
14	Cadres	Population.de.cadres.et.professions.intellectuelles.supérieures.actifs.de.15.à.64.ans
15	ProfessIntrem	Population.de.professions.intermédiaires.actifs.de.15.à.64.ans
16	Employes	Population.d.employés.actifs.de.15.à.64.ans
17	Ouvriers	Population.d.ouvriers.actifs.de.15.à.64.ans
18	Natalite	Taux.de.natalité
19	Chomage	Chômage..au.sens.du.recensement..des.15.64.ans
21	C2S	Bénéficiaires.de.la.complémentaire.santé.solidaire..C2S.
22	MedecTrait	Bénéficiaires.de.15.ans.et.plus.ayant.déclaré.un.médecin.traitant
23	Pers1+ALD	Bénéficiaires.ayant.au.moins.une.ALD
29	Alloc50-99	Allocataires.dont.le.poids.des.prestations.représente.50...à.99...du.revenu.disponible
30	Alloc100	Allocataires.dont.le.poids.des.prestations.représente.100..du.revenu.disponible
63	VaccinGrp+65	Personnes.de.65.ans.et.plus.bénéficiaires.du.vaccin.contre.la.grippe
64	MTDENTS	Jeunes.de.3..6..9..12..15..18..21.et.24.ans.ayant.bénéficié.d.un.examen.bucco.dentaire.gratuit..M.T.dents.
65	Mammo	Femmes.de.tous.âges.ayant.réalisé.une.mammographie
66	Mammo50-74	Femmes.de.50.à.74.ans.ayant.réalisé.une.mammographie.organisée
69	Psychotr-a-n	Patients.sous.traitement.psychotrope.hors.antidépresseur.et.hors.neuroleptique
70	antidepr	Patients.sous.traitement.antidépresseur
71	neurolept	Patients.sous.traitement.neuroleptique
72	antidiab	Patients.sous.traitement.antidiabétique..y.compris.insuline.
73	antiallerg	Patients.sous.traitement.anti.allergique
74	antiasthm	Patients.sous.traitement.antiasthmatique
76	Hospit	Patients.hospitalisés.toutes.causes
77	SejoursHospit	Séjours.hospitaliers.toutes.causes
78	HospitTumeur	Patients.hospitalisés.pour.tumeurs
79	HospitCardioVasc	Patients.hospitalisés.pour.maladies.cardio.vasculaires
80	HospitRespir	Patients.hospitalisés.pour.maladies.respiratoires
84	HospitDiabete	Patients.hospitalisés.pour.diabète
85	HospitTrauma65+	Patients.de.65.ans.et.plus.hospitalisés.pour.traumatisme
87	Accouch15-49	Séjours.pour.accouchement.chez.les.femmes.de.15.à.49.ans
103	BenefALD	Bénéficiaires.d.ALD
104	BenefALDTumeurs	Bénéficiaires.d.une.ALD.pour.tumeurs
105	BenefALDCardioVasc	Bénéficiaires.d.une.ALD.pour.maladies.cardio.vasculaires
106	BenefALDPsy	Bénéficiaires.d.une.ALD.pour.maladies.psychiatriques
107	BenefALDrespir	Bénéficiaires.d.une.ALD.pour.maladies.respiratoires
108	BenefALDdiabete	Bénéficiaires.d.une.ALD.pour.diabète
109	BenefALDAlzhei	Bénéficiaires.d.une.ALD.pour.maladie.d.Alzheimer.et.autres.démences.chez.les.75.ans.et.plus
111	DecesTous	Décès.toutes.causes
126	ExpoPM2.5	AIR...Exposition.moyenne.de.la.population.aux.PM2.5
127	ExpoNO2	AIR...Exposition.moyenne.de.la.population.au.NO2
128	PartPopNO2	AIR...Part.de.la.population.exposée.à.des.niveaux.de.NO2.supérieurs.aux.seuils.de.l.OMS
129	PartPopPM2.5	AIR...Part.de.la.population.exposée.à.des.niveaux.de.PM2.5.supérieurs.à.l.ancien.seuil.de.l.OMS
134	AllergAmbroisie	POLLENS...Part.de.la.population.potentiellement.allergique.à.l.ambroisie
139	NbERPV	BATIMENT.LOGEMENT...Nombre.d.établissements.recevant.des.populations.vulnérables..ERPV.
140	PotentRadon	BATIMENT.LOGEMENT...Radon...Potentiel.d.émission.par.le.sol
123	NivVieMed	Niveau de vie médian

Pour plus de détails sur les données de BALISES, voir ce fichier : `https://www.balises-auvergne-rhone-alpes.org/connexion_stat/compteur_telechargements.php?atg=METADONNEES_20230201_1_11`.

La variable "Niveau de vie médian" n'est pas présente sur BALISES, je l'ai récupérée directement sur

le site de l'INSEE.

Références

[1] Observatoire de santé régional (ORS) Auvergne-Rhône-Alpes. *BALISES*. URL : https://www.balises-auvergne-rhone-alpes.org/index.php.

[2] F. Collart-Dutilleul, O. Hamant, I. Negretiu et F. Riem. *Manifeste pour une santé commune*. Editions Utopia, 2023.

[3] Sébastien Da Veiga, Fabrice Gamboa, Bertrand Iooss et Clémentine Prieur. *Basics and Trends in Sensitivity Analysis*. Philadelphia, PA : Society for Industrial et Applied Mathematics, 2021. DOI : 10.1137/1.9781611976694. URL : https://epubs.siam.org/doi/abs/10.1137/1.9781611976694.

[4] Noé Fellmann, Mathis Pasquier, Christophette Blanchet-Scalliet, Céline Helbert, Adrien Spagnol et Delphine Sinoquet. « Sensitivity analysis for sets : application to pollutant concentration maps ». In : (2023). URL : https://arxiv.org/abs/2311.16795.

[5] Robin Genuer et Jean-Michel Poggi. *Les forêts aléatoires avec R*. Presses Universitaires de Rennes, 2019. URL : https://lesforetsaleatoiresavecr.robin.genuer.fr/.

[6] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray et Mark Crowley. « Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nyström Method, and Use of Kernels in Machine Learning : Tutorial and Survey ». In : (2021). arXiv : 2106.08443. URL : https://arxiv.org/abs/2106.08443.

[7] Markus Kalisch et Peter Bühlmann. « Robustification of the PC-Algorithm for Directed Acyclic Graphs ». In : *Journal of Computational and Graphical Statistics* 17.4 (2008), p. 773-789. DOI : 10.1198/106186008X381927. eprint : https://doi.org/10.1198/106186008X381927. URL : https://doi.org/10.1198/106186008X381927.

[8] Kevin P Murphy. *Machine learning : a probabilistic perspective*. Cambridge, MA, 2012.