Analyse d'une base de données sur le sommeil

Thibaut Martini

INCI
5, rue Blaise Pascal
67084 Strasbourg
France

Université de Strasbourg
7, rue René Descartes
67084 Strasbourg
France

Rapport de stage

Intitulé :

Analyse d'une base de données sur le sommeil

Réalisé par **Thibaut MARTINI**
Maître de stage : **André MALAN**

Année universitaire 2011/2012

Remerciements

Tout d’abord, je tiens à remercier Monsieur André MALAN, mon responsable de stage, pour son écoute, son aide et la confiance qu’il m’a porté dans l’exécution quotidienne de mes tâches tout au long du stage. Il a pris tout particulièrement soin à effectuer un suivi régulier de mon travail et a toujours été disponible en cas de problèmes lors de la résolution de mes objectifs.

Je remercie également le professeur BOURGIN, pour m’avoir accueilli dans son service, son financement et l’apport de la base de données qui m’a permis d’effectuer mon stage de fin d’études, sans oublier Juliette CHAMBE et Elisabeth RUPPERT pour l’aide et les conseils qu’elles m’ont apporté.

Merci à tous les chercheurs, les doctorants et les stagiaires de l’INCI qui m’ont accompagné durant le stage, pour leurs disponibilités, leurs bonnes humeurs et leurs soutiens durant toute ma période de ce stage.

Enfin je tiens à remercier Yannick NONN, Annie TU, Jérémy MAGNANENSI et Thomas PERRETTI qui ont pris de leur temps pour la lecture et la correction de ce rapport de stage.

Introduction

Ce stage de fin d'études du Master Statistique s'est déroulé au sein de l'INCI (Institut des Neurosciences Cellulaires et Intégratives), laboratoire du CNRS. Cet organisme public français est le plus grand en France et c'est avec fierté que j'ai pu y effectuer mon stage du 6 février 2012 jusqu'au 31 juillet 2012.

J'ai intégré le « Service des Maladies du Sommeil de la Clinique Neurologique » de l'hôpital universitaire de Strasbourg dirigé par le Professeur BOURGIN. Le Professeur BOURGIN anime également une équipe de recherche fondamentale sur le sommeil, au sein de l'INCI. C'est dans ce cadre que s'est développé le projet d'une collaboration entre l'équipe du Professeur BOURGIN et Monsieur André MALAN, directeur de recherche émérite, également à l'INCI.

Cette expérience m'a permis de découvrir le monde de la recherche publique, étant donné que j'avais effectué mes précédents stages dans des entreprises privées. Mon sujet de stage fut enrichissant. En effet, j'ai acquis des connaissances en statistique dont le thème n'a pas été abordé lors de ma formation.

Lors de ce stage, j'ai également pu me rendre au service des troubles du sommeil, dans lequel les patients sont hospitalisés.

Sommaire

1. Contexte et objectif du stage

1.1 Le laboratoire

L'INCI est un laboratoire du CNRS dont les études visent à comprendre le fonctionnement des cellules nerveuses et endocrines (qui libèrent une substance hormonale dans le système nerveux) et des circuits neuronaux. L'approche développée est multidisciplinaire et caractérisée par différents niveaux d'investigation : génomique, protéomique, cellulaire, intégré et comportemental.

Douze équipes de recherche se répartissent en 3 départements :

- Neurobiologie des rythmes
- Neurotransmission et sécrétion neuroendocrine
- Nociception et douleur

Ce stage s'est déroulé dans le service de Neurobiologie des rythmes.

1.2 Le département de neurobiologie des rythmes

Les recherches effectuées dans ce département ont pour objectif de comprendre les mécanismes nerveux et neuroendocrines impliqués dans le contrôle des rythmes biologiques. Ces rythmes permettent à l'organisme de s'adapter aux variations journalières et saisonnières de l'environnement. Le dysfonctionnement causé par une déstructuration des rythmes biologiques entraine des troubles graves. On les retrouve chez les personnes souffrant de dépressions saisonnières, les insomniaques, les aveugles ou chez les personnes âgées.

Au laboratoire, cette thématique est abordée chez différentes espèces de mammifères, comme le hamster, le rat ou la souris. Une des équipes de ce département étudie le sommeil en association avec le service des troubles du sommeil des hôpitaux universitaires de Strasbourg.

Les données utilisées lors du stage sont issues des différents examens effectués sur les patients hospitalisés.

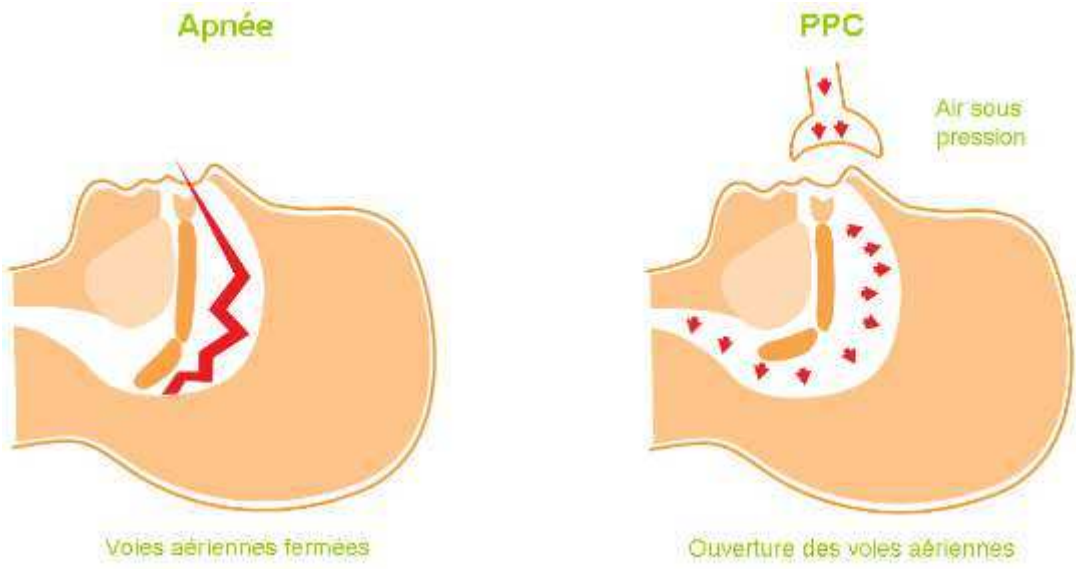
1.3 Les troubles du sommeil

L'apnée du sommeil constitue un problème de santé publique. Ce trouble se traduit par des courts arrêts respiratoires durant la nuit. A long terme, la mauvaise oxygénation de l'organisme entraîne des troubles cardiaques, de l'hypertension artérielle voir même des risques d'infarctus.

Les troubles comprenant le syndrome des apnées hypopnées du sommeil correspondent à un ensemble d'affections respiratoires nocturnes qui se caractérisent essentiellement par l'affaissement partiel ou complet du pharynx (carrefour entre les voies aériennes et les voies digestives) au moment de l'endormissement. L'affaissement complet correspond à l'apnée qui interdit tout passage de l'air de dix secondes à plus d'une minute pouvant se répéter plusieurs fois au cours d'une heure de sommeil. Les apnées peuvent être précédées ou suivies par les affaissements partiels, on parle alors d'hypopnées. Les conséquences biologiques de ces apnées sont une diminution du taux d'oxygène dans le sang et l'apparition d'une somnolence diurne excessive due à un fractionnement du sommeil par de multiples micro-éveils (la plupart du temps invisibles : une distinction est fait entre des micro-éveils supérieurs ou inférieurs à 20 secondes).

Le traitement de référence des apnées consiste à porter toute la nuit un masque relié à un appareil de pression positive continue (PPC). Il s'agit d'un petit compresseur qui envoie de l'air sous pression suffisante pour maintenir les voies aériennes supérieures ouvertes. Ce procédé est le plus souvent efficace et permet de corriger l'obstruction pharyngée et les hypoxies, en rétablissant une respiration nocturne normale.

Figure 1 : Schéma du traitement par PPC



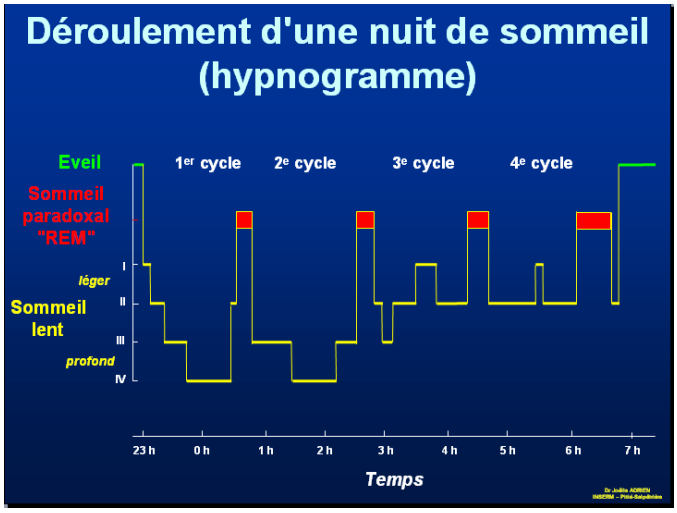
Le sommeil contient 4 stades importants qui constituent un cycle :

- La somnolence (stade1)
- Le sommeil léger (stade2)
- Le sommeil lent profond
- Le sommeil paradoxal

L'électro-encéphalogramme qui enregistre l'activité électrique dans le cerveau permet de spécifier chaque stade du sommeil pendant la nuit d'hospitalisation d'un patient.

Ces cycles se répètent au courant de la nuit. Une bonne nuit est caractérisée par plusieurs cycles réalisés. Les apnées du sommeil empêchent de passer les stades, et donc de compléter les cycles du sommeil.

Figure 2 : Hypnogramme représentant les différents cycles au cours de la nuit



1.4 Contexte et objectif

Le service des troubles du sommeil dispose d'une exceptionnelle base de données clinique sur le sommeil, initiée à l'époque par le Professeur KRIEGER. Cette base comprend maintenant des observations portant sur plusieurs milliers de patients, sur chacun desquels ont été enregistrés plus d'une centaine de données, répétées ou non. Après une assimilation et une compréhension de la base de données et des différentes connaissances biologiques, le projet consiste à utiliser cette base pour rechercher les corrélations entre les variables décrivant le sommeil et ses pathologies, celles-ci pouvant être considérées suivant les cas, soit comme des causes, soit comme des conséquences des anomalies du sommeil. Pour tenir compte des colinéarités entre variables au sein d'un même groupe on utilisera la méthode PLS. Le logiciel utilisé pour le traitement statistique est Statistica.

Le travail a été fait en collaboration avec le Professeur Bourgin, le Docteur Ruppert et le Docteur Chambe pour m'encadrer sur les plans biologiques, et du Docteur André Malan pour les méthodes statistiques.

1.5 Présentation de la base

Depuis 1983, les données des patients hospitalisés dans la clinique des troubles du sommeil sont stockées dans le logiciel de gestion de base de données « Foxprow ». Ces données sont réparties selon divers critères. Les cliniciens effectuent plusieurs examens sur les patients. En Annexe 1 se trouve l'organigramme des différents fichiers représentant les divers examens. Ces fichiers sont :

- « *Question* » : un questionnaire réalisé par les infirmières
- « *Emi* » : regroupe les variables polysomnographiques. Il s'agit des données des patients durant leur nuit d'hospitalisation. Il y a les informations concernant les apnées et les variables respiratoires.
- « *Cardio* » : regroupe différentes variables biologiques, fonctionnelles, cardiaques et respiratoires :
  - facteurs physiques : le tour du cou (Tourcou), le poids (poids), la taille, l'indice de masse corporel (BMI)
  - facteurs pneumologiques : la capacité pulmonaire totale (capa_pulm_tot), le volume résiduel (vol_res)
  - facteurs cardio-vasculaires : la pression artérielle systolique et diastolique couché et debout (PAScouch, PASdebout, PADcouch et PADdebout), la rénine plasmatique (Prenin)
  - facteurs hépatiques : bilurubine totale (bilitot), transaminase alat et asat (transgo, transgp), l'enzyme Gamma GT
  - facteurs métaboliques : la glycémie (glucose), la triglycérine (trigly), le cholestérol (cholest),
  - facteurs hormonaux : des hormones thyroidienne comme la thyréostimuline (Tshus), et la T4libr, la somatomédine (somatom), la testostérone (Testot), la testostérone libre (Testol), la dihydrotestostérone (Dhtesto), FSH, LH
  - autres facteurs : la vitesse de coagulation du sang (Tquick), le fibrinogène (Fibrin) le calcium (Culcium), Phospo, Albumin
- « *MSLT* » : les tests de latence d'endormissement sont présents avec les questionnaires du sommeil d'Epworth et Pichot. Ce sont des questionnaires spécifiques aux patients permettant d'attribuer une note sur l'importance et la gravité des troubles.
- « *Fren* » : regroupe les données des fonctions rénales avant et après traitement du PPC
- « *Emblett* » : il s'agit du contrôle à domicile des patients utilisant le PPC, réalisé par un prestataire.

Chaque fichier correspond à l'année de suivi du patient. Par exemple, *Emi1* représente la première hospitalisation, *Emi5* le suivi du patient à la cinquième année, *EmiD* à la dixième, *EmiQ* à la quinzième et *EmiV* à la vingtième. Cette règle est valable pour tous les fichiers.

2. Gestion et premier traitement de la base de données

Le début du stage m'a permis d'assimiler différentes notions essentielles à la compréhension de la base. Les données peuvent être entrées manuellement ou automatiquement en fonction du type de l'examen. En effet, les variables de sommeil sont issues d'un appareil qui enregistre directement les données dans le logiciel. Les contrôles sanguins par contre sont entrés manuellement dans le logiciel. Ces entrées manuelles sont à l'origine de nombreuses erreurs (valeurs aberrantes, erreurs de saisie). En effet, durant des années différents facteurs ont interféré sur les données. Différentes personnes ont saisi les données, engendrant des erreurs de saisie par période. Des unités de mesures ont pu changer dans le temps. D'autres problèmes ont été rencontrés notamment celui de nombreux doublons et de mauvaises importations des fichiers Excel.

2.1 Premier contact avec la base de données

La première analyse à effectuer fut d'étudier l'effet des différents facteurs biologiques sur les troubles du sommeil. Quels sont ceux qui influent le plus? Lesquels agissent de la même manière ? Par exemple, on sait déjà que le poids et l'âge ont un lien avec les symptômes du sommeil. Quels sont les autres ? Avec quelle intensité ? On étudie la cohérence de la base de données avec certaines connaissances ainsi que d'autres perspectives de résultat.

Pour cela, on a décidé d'utiliser seulement les informations de la première hospitalisation des patients (non prise en compte des patients qui reviennent ou répétés sur 5, 10, 15 et 20 ans). On obtient donc un échantillon indépendant.

Toutes ces informations sont dispersées dans les différents fichiers : il a fallu les regrouper dans un seul et même fichier pour homogénéiser l'analyse sur le logiciel Statistica. Les fichiers utilisés pour regrouper les variables biologiques et polysomnographiques sont :

- *Cardio1* : données biologiques
- *Emi1* : données du sommeil
- *Question* : pour récupérer le sexe des patients (non présent dans les autres fichiers et variable fortement importante)

Les fichiers *Cardio1* et *Emi1* contiennent respectivement 5353 et 4385 observations. L'exportation de « Foxprow » vers Excel n'a pas permis de restituer des fichiers « propres ». Chaque patient présent dans *Cardio1* ne se trouve pas forcément dans le fichier *Emi1* (et inversement). J'ai aussi pu constater des patients qui peuvent être en doublon (voir plus), avec des valeurs présentes seulement chez l'un ou l'autre. Pour éviter de perdre de l'information, je n'ai pas utilisé de procédure supprimant ces doublons pour en garder un seul. J'ai essayé de conserver toutes les informations dans une seule observation.

On recherche tous les éléments qui peuvent biaiser l'analyse tels que les valeurs manquantes et aberrantes, les erreurs de code et les variables redondantes. Une première lecture des fichiers indique beaucoup de variables inutiles ou inexploitables, étant

inintéressante à l'analyse ou possédant aucune valeur malgré la présence dans le fichier. Ces variables sont supprimées de l'analyse.

2.2 Traitement Access et SQL

Le fichier « nettoyé » doit contenir les observations sur une même ligne. Un simple « copier-coller » n'est pas efficace pour effectuer le regroupement, compte tenu du nombre de patients ainsi que de la non-correspondance des patients entre chaque fichier. Ayant la liberté du choix pour la méthode de regroupement des données, j'ai voulu appliquer les connaissances en base de données et le langage SQL. J'avais à disposition le logiciel Access (gestion de base de données de « Office »). Le grand avantage d'appliquer le langage SQL est la rapidité immédiate d'exécution. L'utilisation de macro sur Visual Basic (VBA) avec un nombre conséquent de données prend beaucoup de temps. J'ai utilisé les caractéristiques de ce langage SQL : chaque patient est déterminé de manière unique ou représenté par une clé primaire (langage Base de données) par son nom, son prénom et sa date de naissance.

Les fichiers sont importés dans des tables Access. La relation commune ou clé primaire permet la jointure des tables importées. Par exemple, la concaténation du fichier *Cardio1* des variables biologiques et du fichier *Emi1* des variables du sommeil doit aboutir aux résultats suivants :

Figure 3 : Exemple de regroupement (en bas) entre le fichier Cardio1 (en haut à gauche) et Emi1 (en haut à droite)

NOM	PRENOM	DATE_NAISS	DATE_EX1	TOURCOU1	POIDS1	TAILLE1	...
NOM1	Edmond	03/02/1930	06/08/1992		100.0	167.0	
NOM2	Hatem	16/02/1951	16/11/1998	49.0	150.0	172.0	
NOM3	André	11/02/1948	02/10/1996	44.0	93.0	170.0	
NOM4	Jacques	06/09/1921	03/12/1997	42.0	78.0	168.0	

+

nom	prenom	date_naiss	date_ex1	age1	ahi1	sao2m1	...
NOM1	Edmond	03/02/1930	06/08/1992	62.5	131.7	0	
NOM2	Hatem	16/02/1951	16/11/1998	47.8	99.6	94	
NOM3	André	11/02/1948	02/10/1996	49.2	73.5	96	
NOM4	Jacques	06/09/1921	03/12/1997	76.3	74.9	95	

=

NOM	PRENOM	DATE_NAISS	DATE_EX1	TOURCOU1	POIDS1	TAILLE1	...	age1	ahi1	sao2m1	...
NOM1	Edmond	03/02/1930	06/08/1992		100.0	167.0		62.5	131.7	0	
NOM2	Hatem	16/02/1951	16/11/1998	49.0	150.0	172.0		47.8	99.6	94	
NOM3	André	11/02/1948	02/10/1996	44.0	93.0	170.0		49.2	73.5	96	
NOM4	Jacques	06/09/1921	03/12/1997	42.0	78.0	168.0		76.3	74.9	95	

Néanmoins, des problèmes sont présents pour la concaténation des fichiers. *Cardio1* et *Emi1* ont des problèmes de compatibilité de format pour les dates ou les noms.

Une illustration de ces problèmes est :

- L'incompatibilité avec les dates (de naissance) : Le format des dates dans *Emi1* est « 01-Feb-12 » alors que le fichier *Cardio1* a le format « 01/02/1912 ». Le mois de la variable « DATE_NAISS » d'*Emi1* prend les 3 premières lettres du mois en Anglais qu'Excel ne reconnait pas. Une requête sous Access est créée pour corriger et l'adapter. Ensuite un patient né avant 1925,codé par ces 2 derniers chiffres, par exemple « 12 », est transformé en 2012 au lieu de 1912. Il a fallu créer une autre requête pour corriger cette erreur.
- Le problème de restitution des noms et prénoms : L'importation de certains fichiers inclut une erreur pour toutes les voyelles avec des accents, par exemple un « Ù » à la place du « è ». Chaque type d'erreurs potentielles a dû être repéré, pour les remplacer par leur lettre appropriée sans accent, grâce à une requête Access. Pour homogénéiser tous les fichiers à regrouper, tous les accents ont été supprimés.

La requête Access qui traite ces 2 problèmes est en Annexe 2.

Ces erreurs sont essentielles pour montrer la sensibilité du jeu de données. Le nombre de patients est trop important pour pouvoir les traiter un à un, impliquant surement une perte de quelques observations.

Figure 4 : Fichier Cardio1 réel

nom	prenom	date_naiss	date_ex1	age1	ahi1	sao2m1
NOM1	Edmond	03-Feb-30	05-Aug-92	62.5	131.7	0
NOM2	Hatem	16-Feb-51	16-Nov-98	47.8	99.6	94
NOM3	André	11-Feb-48	26-Mar-97	49.2	73.5	96
NOM4	Jacques	06-Sep-21	03-Dec-97	76.3	74.9	95

Figure 5 : Fichier Emi1 réel

NOM	PRENOM	DATE_NAISS	DATE_EX1	TOURCOU1	POIDS1	TAILLE1
NOM1	Edmond	03/02/1930	06/08/1992		100.0	167.0
NOM2	Hatem	16/02/1951	16/11/1998	49.0	150.0	172.0
NOM3	AndrÚ	11/02/1948	02/10/1996	44.0	93.0	170.0
NOM4	Jacques	06/09/1921	03/12/1997	42.0	78.0	168.0

Chaque fichier contient la date de l'examen effectué. Il arrive que la nuit d'observation ne soit pas effectuée en même temps que les examens biologiques. Une marge de 6 mois d'écart est fixée comme seuil d'acceptabilité. J'ai remarqué par la suite, lors de la réalisation du fichier avec les répétitions, que ces données correspondent à une hospitalisation à la cinquième ou la quinzième année.

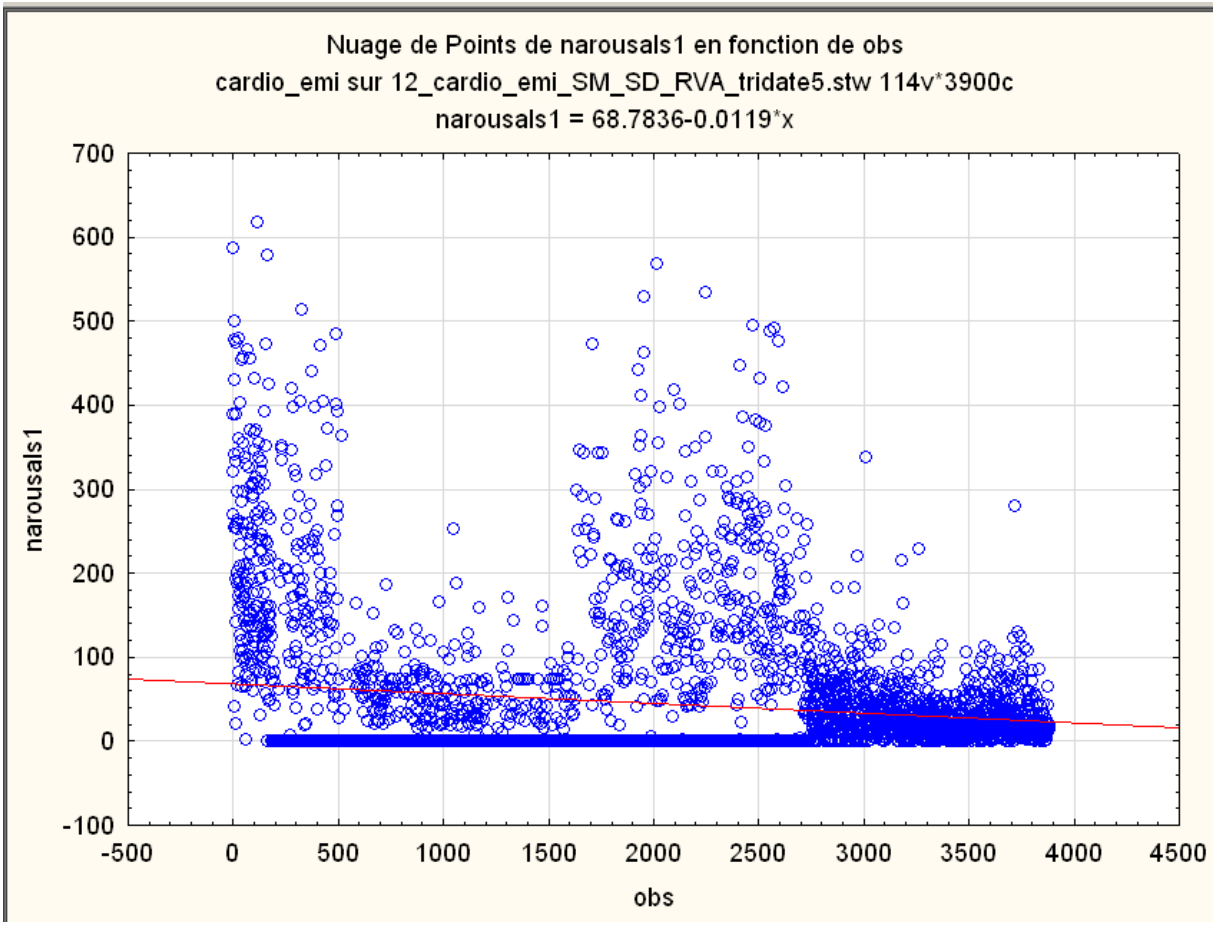
Le fichier final contient 3728 observations après la suppression des nombreux doublons. On peut donc remarquer qu'avec *cardio1* et *emi1* contenant respectivement 5353 et 4345 observations, un certain nombre de données a effectivement été perdu. Celles-ci pouvant être leur non-présence dans un des deux fichiers, des doublons supprimés ou des dates d'examen différentes.

2.3 Analyse univariée et premiers résultats

Avant d'effectuer l'analyse statistique, on réalise une étude simple des variables pour en connaître les principales caractéristiques. On calcule les principaux indicateurs tels que la moyenne, l'écart type, le maximum, le minimum et la boite à moustache. Cette simple étape permet de visualiser une dispersion incohérente du jeu de données. Il a fallu déterminer quelles sont les limites acceptables de chaque variable. Par exemple des pourcentages au-delà de 100%, la taille d'un patient dépassant les 20 mètres ou des taux de glycémie aberrants. Pour cela on a décidé de réaliser un fichier (exemple en Annexe 3) où est indiqué chaque intervalle de conservation des données. Il ne s'agit pas d'éliminer à cette étape des points atypiques, mais de garder les valeurs cohérentes à la réalité.

En effectuant un tri des valeurs par date d'examen (du plus au récent au plus vieux), allant de 1983 à 2009, on peut observer de nombreuses irrégularités sur les variables. La figure suivante montre le résultat pour la variable « Narousals » (le nombre de micro-réveil) en fonction du temps.

Figure 6 : Nombre de micro-réveil en fonction du temps



On peut observer que la variable possède une absence de régularité et de cohérence. Nos données ne doivent pas dépendre du temps. L'observation 0 représente la première observation du patient en 1983 jusqu'au 3728ème patient en 2009. Les premières années, le nombre de micro-éveil est au maximum jusqu'à 620, ensuite le maximum est de 256, pour de nouveau se rapprocher de 600. On peut aussi noter un nombre conséquent de 0. Il peut s'agir d'une vraie valeur ou d'une donnée manquante. En effet, certaines données manquantes sont notées 0 : pour la variable « Narousals », cette valeur empêche de différencier s'il s'agit d'un vrai 0 ou d'une donnée manquante. Cette variable est inutilisable pour l'analyse. Elle sera supprimée du fichier. Cette particularité temporelle sur la variable « Narousals » est présente pour d'autres facteurs. Les raisons peuvent varier : les changements d'appareil de mesure pour la prise d'information des données du sommeil durant la nuit, les changements de personnes qui entrent les données ou les changements d'unité de mesure (mg/l à g/mol) sont autant de facteurs qui peuvent influer sur ces changements.

Après le nettoyage des valeurs aberrantes du fichier, on vérifie les distributions des variables. Certaines données présentent des distributions dissymétriques telles que l'indice de masse corporel (BMI), la glycémie (glucose), la bilurubine totale (bilitot), la Gamma GT, la transaminase ALAT et ASAT, (transgp et transgo), la thyréostimuline (Tshus) et la T4libre. Ces données sont transformées par logarithme. La variable IAH (indice apnée hypopnée ou nombre moyen d'apnée par heure) est transformée par la racine carré. Le but est d'empêcher

une trop forte importance des valeurs éloignées à la moyenne. En Annexe 4 se trouve l'histogramme des données du glucose avant et après transformation logarithme.

Certaines variables sont issues d'un calcul effectué avec d'autres variables comme l'indice de masse corporelle, qui est le poids divisé par la taille au carré. On peut vérifier si les formules sont correctes.

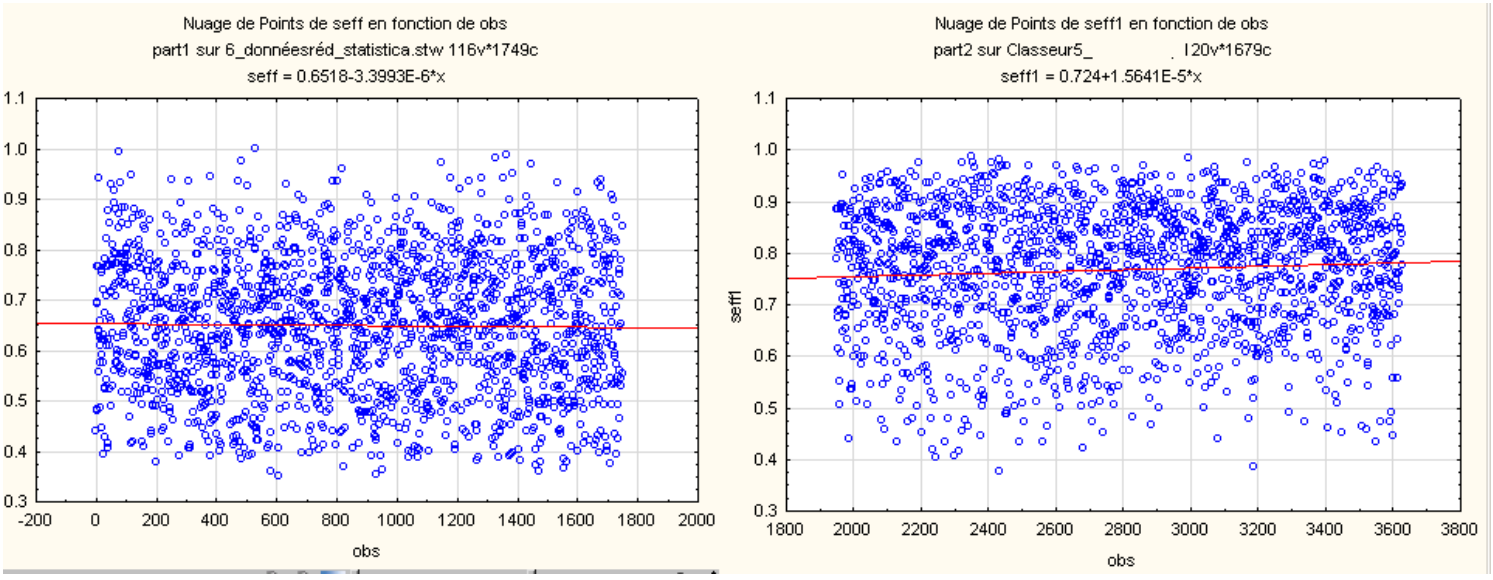
D'autres critères ont été pris en compte pour les patients. Les patients mineurs ont été supprimés, ainsi que ceux ayant dormi moins de 3h durant la nuit d'hospitalisation.

Séparation de l'échantillon en 2 sous-échantillons indépendants

Le fichier de données est séparé en 2 parties. La première partie contient les patients jusqu'en décembre 1999, la deuxième de mars 2001 à 2009. On a respectivement 1680 et 1750 observations. La cause de cette séparation est le fruit de divers changements effectués lors de la fin de l'année 1999. La mesure de certaines variables, comme la testorérone, n'est plus effectuée à partir de l'année 1999, tandis que d'autres variables changent d'unité de mesure au cours de cette même période. Le nombre d'observation reste conséquent pour chaque sous-échantillon et l'analyse statistique permettra de confronter les résultats.

La figure suivante montre un exemple d'évolution survenue aux alentours de l'observation 2000 représentant l'année 1999. Avant cette année, la tendance est constante autour de 0.65, alors qu'après celle-ci, elle est constante autour 0.75.

Figure 7 : Efficacité du sommeil en fonction du temps



2.4 Analyse multivariée

On s'intéresse à la corrélation entres les variables biologiques et polysomnographiques. Lors d'un modèle de régression, cette étape est utile pour ôter des variables prédictives fortement liées, puisque les résultats seraient très instables. La régression PLS (expliquée au chapitre suivant) attribue un poids en fonction de la covariance de chaque couple de variable. La matrice de corrélation permet d'avoir un premier regard sur ces poids.

Pour rappel le coefficient de corrélation linéaire (de Pearson) entre une variable X1 et X2 est :

$$cor(X1, X2) = \frac{cov(X1, X2)}{\sqrt{var(X1)var(X2)}}$$

Ayant 2 échantillons indépendants (partie1 et partie2), on teste si la corrélation c_1 entre X1 et Y1 du premier échantillon est identique à la corrélation c_2 entre X2 et Y2 du deuxième échantillon. On teste l'hypothèse :

$$H_0 : c_1 = c_2$$

$$H_1 : c_1 \neq c_2$$

Pour cela on effectue la transformation de Fischer sur c_1 et c_2 :

$$z_i = \frac{1}{2}\ln(\frac{1 + c_i}{1 - c_i}) \ , \quad \text{i} = 1,2$$

En notant n_1 la taille de l'échantillon 1 et n_2 celle de l'échantillon 2, on pose la statistique :

$$U = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

U suit asymptotiquement une loi normale centrée réduite. En prenant la valeur absolue de U on obtient un rejet de H_0 au risque α si :

$$|U| \geq u_{1-\frac{\alpha}{2}}$$

Pour un risque α = 0.05, on a $u_{1-\frac{\alpha}{2}} = 1.96$ quantile de la loi normale centrée réduite et région critique du test.

En annexe 5 se trouve les résultats concernant la corrélation entre les variables biologiques et du sommeil pour les 2 parties, ainsi que le test d'égalité des covariances. Les corrélations ne sont pas toujours très fortes, mais les tests concluent à une non-corrélation non significative pour de nombreux couples. Le test d'égalité des covariances n'est pas significativement différent entre les 2 échantillons. On déduit une stabilité et une représentativité de la matrice de covariance entre les 2 parties.

De plus, la matrice de corrélation des variables du sommeil indique une forte corrélation entre celles-ci. C'est pour la raison pour laquelle chaque variable ne sera pas utilisé séparément, mais sera contenu dans une matrice Y pour une régression PLS2.

Le TAU de KENDALL

Le coefficient de corrélation permet d'établir la relation linéaire et n'est pas forcément adapté à notre jeu de données. Le TAU de KENDALL permet d'étudier une relation monotone d'une variable. Il s'agit d'une méthode non paramétrique. De plus nos données ne suivent pas forcément une loi normale.

Le coefficient de Kendall est interprété à l'aide de paires concordantes ou discordantes sur un couple de variable (X, Y) avec n observations :

- Les paires d'observation est concordante si $x_i > x_j$ alors si $y_i > y_j$
- Les paires d'observation est discordante si $x_i > x_j$ alors si $y_i < y_j$

On définit C le nombre de paire concordante, et D le nombre de paires discordantes. On définit le Tau de Kendall :

$$m = \frac{C - D}{\frac{1}{2}(n-1)n}$$

La comparaison entre les matrices de corrélation avec le coefficient de Pearson et le tau de Kendall n'indique pas une grande différence.

2.5 Conclusion

La compréhension et l'étude de la base de données a été le facteur le plus important du stage. Cette partie peut paraitre sommaire, mais elle a nécessité le travail tout le long du stage. J'ai voulu réaliser des fichiers en perdant le moins d'informations disponibles. De plus, cette étape est cruciale à l'analyse statistique pour avoir les meilleurs résultats.

La gestion de cette base a beaucoup ralenti la progression des analyses statistiques, et même rendu non informatif (ou partiellement informatif) celles-ci. La robustesse des modèles, notamment pour les données répétées, a été compromise pour ces raisons.

3. Méthodologie et premières analyses

3.1 Description méthode

3.1.1 Problème de la régression linéaire multiple

L'objectif est de régresser un bloc de variable X (données biologiques) qui contient p variables explicatives sur un bloc Y (données polysomnographiques) contenant r variables dépendantes sur un échantillon de taille n. Il est possible de régresser chaque variable Y_i i=1..r de la matrice Y séparément mais la forte corrélation entre les données du sommeil conseille de garder le bloc en entier. Le bloc X est aussi fortement corrélé. Le modèle de régression linéaire multiple s'écrit :

$$Y = XB + E$$

avec B le vecteur des coefficients et E le terme d'erreurs

Une régression linéaire multiple constitue une solution simple pour établir un modèle linéaire cherchant à expliquer les différentes variables réponses à l'aide des p variables explicatives. Par contre elle empêche de prendre en compte les données manquantes, ce qui conduit souvent au rejet de beaucoup d'observations incomplètes et pouvant contenir de l'information utile. Notre jeu de données peut être considéré comme ayant des valeurs manquantes réparties aléatoirement. Ceci implique une forte élimination d'observations : la partie 1 contient 342 patients, parmi les 1680, qui contiennent des données sur toutes les variables prédictives et dépendantes, c'est à dire qu'on a 80% des observations supprimées.

Un autre défaut de la régression linéaire multiple est la sensibilité de la méthode à la colinéarité entre les variables entrainant une variance infinie de l'estimateur. En effet, on sait que l'erreur quadratique moyenne de l'estimateur vaut :

$$EQM = \sigma^2 trace((X'X)^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$$

Avec λ_j la j-ième valeur propre de la matrice $X'X$. En cas de multicolinéarité, cette valeur propre est proche de 0, donc l'erreur moyenne quadratique tend vers un (trop) grand nombre, impliquant une interprétation impossible des coefficients du modèle.

3.1.2 La régression PLS

La régression PLS est une technique permettant de contourner tous ces obstacles. C ette méthode combine les caractéristiques de l'analyse en composantes principales et de la régression multiple. La particularité de cette régression est d'utiliser également l'information contenue dans les covariables Y_i. Elle réduit la dimension d'un ensemble de variables en

minimisant la variance de X et Y, mais en maximisant leur covariance. On utilisera la régression PLS2 ou PLS multivariée puisque le nombre de variables prédictives est supérieur à 1. Elle va créer une nouvelle matrice (appelée « matrice des scores ») à partir de la matrice d'origine, dont les H colonnes seront les composantes principales (ou variables latentes), avec H < p. Une régression de la variable réponse sur les composantes trouvées est effectuée à chaque étape de la méthode. Un autre avantage de cette méthode est le peu d'hypothèses probabilistes à vérifier, notamment sur les résidus. Des nouvelles coordonnées du sous-espace sont calculées pour chaque individu.

L'algorithme NIPALS est utilisé pour effectuer cette régression PLS. En effet, celui-ci permet d'effectuer une réduction de dimension sur une matrice de variables et d'observations et permet notamment de traiter les données manquantes. Il est adapté à la régression PLS pour résoudre les problèmes de colinéarité et de données manquantes. Il n'y a aucune suppression des données manquantes, ni d'estimations lors du déroulement de la régression PLS mais une estimation finale peut être établie.

3.2 Théorie PLS

La régression PLS mesure le lien entre en ensemble de variables X et Y. D'abord l'objectif est de trouver H composantes principales combinaisons linéaire de X et expliquant au mieux X et Y. Chaque composante, notée $t_h , h = 1..H$ est ensuite régressée sur chaque variable de Y, puis sur les variables X d'origine.

Pré-requis : Un modèle linéaire gaussien $Y = XB + E$ avec $E \approx N(0, \sigma^2 I_n)$ et les hypothèses habituelles avec X'X inversible, permet de trouver l'estimateur de Gauss Markov :

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

PLS réalise un enchainement de régressions multiples de ce type, un utilisant un vecteur à la place de la matrice X.

Notation :

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (x_1 \dots x_j \dots x_p)$$

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1r} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nr} \end{pmatrix} = (y_1 \dots y_j \dots y_r)$$

Les tailles des matrices X et Y sont respectivement {n*p} et {n*r}.

Pour une matrice $X = (\ x_1 \ldots\ x_j \ldots\ x_p)$, on note sa transposée $X^{'} = \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_p \end{pmatrix}$

On initialise $X_0 = X$ et $Y_0 = Y$ la première colonne de Y.

Algorithme :

Les différentes étapes de la régression PLS avec l'algorithme NIPALS sont décrites ci-dessous :

Etape 1 :

Pour chaque étape h = 1….H, soit u_h la première colonne de Y_{h-1}.

Etape 2: Tant que l'algorithme n'a pas convergé, on réalise les sous-étapes suivantes :

Etape 2.1 : calcul du poids de *X* en faisant la régression de u'_h sur X'_{h-1} :

$$w_h = \frac{X'_{h-1}\, u_h}{u'_h u_h}$$

Etape 2.2 : On norme le vecteur w_h :

$$w_h = \frac{w_h}{||w_h||} = \begin{pmatrix} w_{h,1} \\ \vdots \\ w_{h,j} \\ \vdots \\ w_{h,p} \end{pmatrix}$$

w_h est un vecteur colonne {1*p}. Chaque $w_{h,j}$ représente le coefficient de régression de u_h dans la régression de la variable $x_{h-1,j}$.

$w_h u'_h$ est une matrice {p*n} qui estime X'_{h-1}

Etape 2.3 : calcul de la composante principale h en faisant la régression de w'_h sur X_{h-1} :

$$t_h = \frac{X_{h-1}\, w_h}{w'_h w_h}$$

t_h est un vecteur colonne {1*n}. Chaque t_{hj} représente le coefficient de régression de w_h dans la régression de la variable $x_{h-1,j}$.

$t_h w'_h$ est une matrice {n*p] qui estime X_{h-1}

Etape 2.4 : calcul du poids de Y en faisant la régression de t'_h sur Y'_{h-1} :

$$c_h = \frac{Y'_{h-1}\, t_h}{t'_h t_h}$$

c_h est un vecteur colonne {1*r}. Chaque $c_{h,j}$ représente le coefficient de régression de u_h dans la régression de la variable $x_{h-1,k}$.

$c_h t'_h$ est une matrice {r*n} qui estime Y'_{h-1}

Etape 2.5: calcul de la nouvelle valeur de u_h en faisant la régression de c_h sur Y_{h-1} :

$$u_h = \frac{Y_{h-1}\, c_h}{c'_h c_h}$$

u_h est un vecteur colonne {1*n}. Chaque $u_{h,i}$ représente le coefficient de régression de $u_h$dans la régression de la variable $x_{h-1,j}$.

$u_h c'_h$ est une matrice {n*r} qui estime Y_{h-1}

Etape 2.6 : On teste la convergence de u_h : si $||u_{new} - u_{old}|| < \varepsilon$, avec ε un seuil à déterminer, alors on considére que notre algorithme a convergé, sinon on reprend à **l'étape 2.1** avec la nouvelle valeur de u_h.

Etape 3 : Une fois la convergence effectuée, calcul de la nouvelle valeur de poids sur X en utilisant la composante principale t_h finale, appelée « poids factoriels », en faisant la régression de t'_h sur X'_h :

$$p_h = \frac{X'_{h-1}\, t_h}{t'_h t_h}$$

Cela permet d'ôter la contribution des composantes obtenue sur X et Y

Etape 4 : Le résidu des matrices X et Y est utilisé pour conditionner la matrice à l'étape h suivante :

$$X_h = X_{h-1} - t_h p'_h$$

$$Y_h = Y_{h-1} - t_h c'_h$$

Données manquantes :

Ces différentes régressions sont sans problème lors de la présence de données complètes. L'intérêt réside dans le cas de données manquantes, ainsi pour chaque étape, en prenant l'exemple à **l'étape 2.1** :

$$w_h = \frac{X'_{h-1}\, u_h}{u'_h u_h}$$

En notant la composante u_h avec les données manquantes incluses :

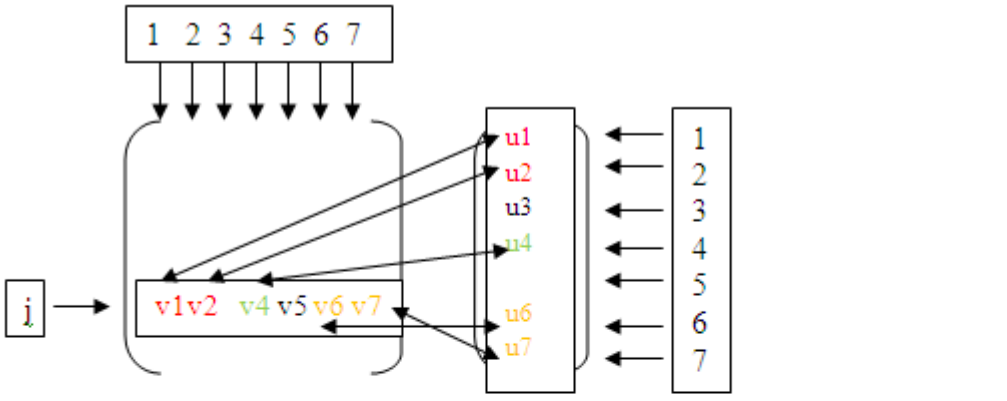
$$u_h = \begin{pmatrix} u_{h,1} \\ : \\ u_{h,i} \\ : \\ u_{h,n} \end{pmatrix}$$

On calcule les coefficients du vecteur w_h de la maniére suivante

$$w_{h,j} = \frac{\sum_{\{i\,:\,si\ x_{ij}\ et\ u_{hi}\ existent\}} x_{ij} u_{hi}}{\sum_{\{i\,:\,si\ u_{hi}\ existe\}} {u_{hi}}^2} \quad , \ \mathrm{j} = 1..\mathrm{p}$$

Le schéma suivant indique un exemple de calcul de $w_{h,j}$ pour la j-ième variable avec 7 observations :

Figure 8 : Schéma du calcul des coefficients avec données manquantes



X'_{h-1} $\qquad u_h = v1u1+v2u2+v4u4+v6u6+v7u7$

$$||u_h|| = \sqrt{u1^2 + u2^2 + u3^2 + u4^2 + u6^2 + u7^2}$$

$X'_{h-1,j3}$ est manquant ainsi que $u_{h,5}$

Donc $w_{h,j} = X'_{h-1,j1} u_{h,1} + X'_{h-1,j2} u_{h,2} + X'_{h-1,j4} u_{h,4} + X'_{h-1,j6} u_{h,6} + X'_{h-1,j7} u_{h,7}$

$||u_h||$ peut se calculer avec u3, car seul u5 est manquant

Le procédé est le même pour chaque matrice $X_h$et Y_h (et leur transposée).

Le logiciel Statistica permet de choisir un nombre minimum de variables et d'observations manquantes.

Remarque : Les variables étant centrées réduites, les coefficients $w_{h,j}$ peuvent s'écrire en fonction de la covariance lors de la première étape de l'algorithme pour la composante h :

En notant :

$$w_h = \begin{pmatrix} w_{h,1} \\ \vdots \\ w_{h,j} \\ \vdots \\ w_{h,p} \end{pmatrix} \; et \; ||w_h|| = \sqrt{\sum_{j=1}^{\mathrm{p}} w^2{}_{h,j}}$$

On a :

$$w_{h,j} = \frac{cov(x_{h-1,j}\,, u_h)}{\sqrt{\sum_{m=1}^{p} cov^2(x_{h-1,m}\,, u_h)}}$$

Avec $x_{h-1,j}$ la j-ième colonne de la matrice X'_{h-1}

Pour la première composante, les données sont centrées réduites donc l'espérance vaut 0. X'_{h-1} avec h > 1 est le résidu issu de la régression linéaire gaussienne donc son espérance vaut également 0.

On a:

$$* \; cov\big(x_{h-1,j}\,, u_h\big) = E\big(x_{h-1,j}\, u_h\big) = \frac{\sum_{i=1}^{n} (x_{h-1,ij}\, u_{h,i}\,)}{n}$$

$$* \; \sqrt{\sum_{m=1}^{p} cov^2\big(x_{h-1,m}\,, u_h\big)} = \sqrt{\sum_{m=1}^{p} \frac{\big(\sum_{i=1}^{n} (x_{h-1,im}\, u_{h,i})\big)^2}{n^2}}$$

$$= \frac{\left(\sqrt{\sum_{m=1}^{p} \big(\sum_{i=1}^{n} (x_{h-1,im}\, u_{h-1,i})\big)^2}\right)}{n}$$

Donc

$$\frac{cov(x_{h-1,j}\,, u_h)}{\sqrt{\sum_{m=1}^{p} cov^2(x_{h-1,m}\,, u_h)}} = \frac{\sum_{i=1}^{n} (x_{h-1,ij}\, u_{h,i}\,)}{\left(\sqrt{\sum_{m=1}^{p} \big(\sum_{i=1}^{n} (x_{h-1,im}\, u_{h,i})\big)^2}\right)} \qquad (1)$$

De plus

$$w_h = \frac{X'_{h-1}\, u_h}{u'_h u_h} = \begin{pmatrix} \sum_{i=1}^{n} x_{h-1,i1}\, u_{h,i} \\ \vdots \\ \sum_{i=1}^{n} x_{h-1,ij}\, u_{h,i} \\ \vdots \\ \sum_{i=1}^{n} x_{h-1,ip}\, u_{h,i} \end{pmatrix} / \sum_{i=1}^{n} u^2{}_{h,i}$$

Et

$$||w_h|| = \left(\sqrt{\sum_{j=1}^{p} \left(\sum_{i=1}^{n} x_{h-1,ij} u_{h,i} \right)^2} \right) / \sum_{i=1}^{p} u^2_{h,i}$$

Après la normalisation de w_h

$$\frac{w_{h,j}}{||w_h||} = \frac{\sum_{i=1}^{n} (x_{h-1,ij}\, u_{h,i})}{\left(\sqrt{\sum_{m=1}^{p} \left(\sum_{i=1}^{n} (x_{h-1,im}\, u_{h,i}) \right)^2} \right)} \qquad (2)$$

Finalement, on obtient l'égalité entre (1) et (2) :

$$w_{h,j} = \frac{cov(x_{h-1,j}, u_h)}{\sqrt{\sum_{m=1}^{p} cov^2(x_{h-1,m}, u_h)}}$$

3.3 Le choix du nombre de composantes

L'objectif étant la réduction de la dimension de la matrice, si l'algorithme ne s'arrête pas, le nombre de composante H sera égal au nombre de variables explicatives p. Le but est de savoir combien de composantes sont nécessaires et significatives au modèle pour empêcher un sur-ajustement du modèle. Ainsi il faut arrêter la construction de nouvelles composantes lorsqu'elles ne sont plus significatives.

La validation croisée est une technique robuste pour définir le nombre de variables latentes. Elle utilise les critères du PRESS (PRediction Error Sum of Square) et RESS (REsidual Sum of Squares). Le PRESS permet de calculer la capacité prédictive du modèle avec un nombre de composante déterminé. Pour la h-ième composante et la variable y_k , $k = 1..r$, on calcule les prévisions $\hat{y}_{h,ik}$ et $\hat{y}_{h,(-i)k}$; cette dernière est une prévision calculée sans l'observation i. On a donc :

$$\mathrm{RESS}_{kh} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (y_{ik} - \hat{y}_{h,ik})$$

$$\mathrm{PRESS}_{kh} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (y_{ik} - \hat{y}_{h,(-i)\mathrm{k}})$$

On calcule l'indice $\mathrm{Q}^2_{\mathrm{hk}}$ suivant de la k-ième variable et de la h-ième composante :

$$\mathrm{Q}^2_{\mathrm{hk}} = 1 - \frac{\mathrm{PRESS}_{kh}}{\mathrm{RESS}_{k(h-1)}}$$

On calcule aussi l'indice Q_h^2 sur l'ensemble des variables Y

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{r} PRESS_{kh}}{\sum_{k=1}^{r} RESS_{k(h-1)}}$$

Ces indices permettent de mesurer l'apport de chaque composante au pouvoir prédictif du modèle. Ils permettent également de décider de l'apport significatif d'une composante principale :

- Si $Q_h^2 \geq (1 - 0.95^2)$ alors la composante est significative : apport global des variables dépendantes
- Si $Q_{hk}^2 \geq (1 - 0.95^2)$ alors la composante est significative : apport global de la variable dépendante k.

3.4 Importance des Variables dans la Projection (VIP)

Un indicateur du pouvoir de modélisation d'une covariable prédictive est la VIP (importance des variables dans la projection).

Soit $w_{h,j}$ issu du vecteur w_h, SC_h la somme des carrés expliquée de la composante principale h et SC_{tot} la somme des carrés totale expliquée par la régression. On obtient l'importance de la variable j :

$$VIP_j = \sqrt{\frac{\sum_{h=1}^{H} w^2_{h,j} SC_h}{SC_{tot}} * p} \quad , \qquad j = 1..p$$

3.5 Application et résultats

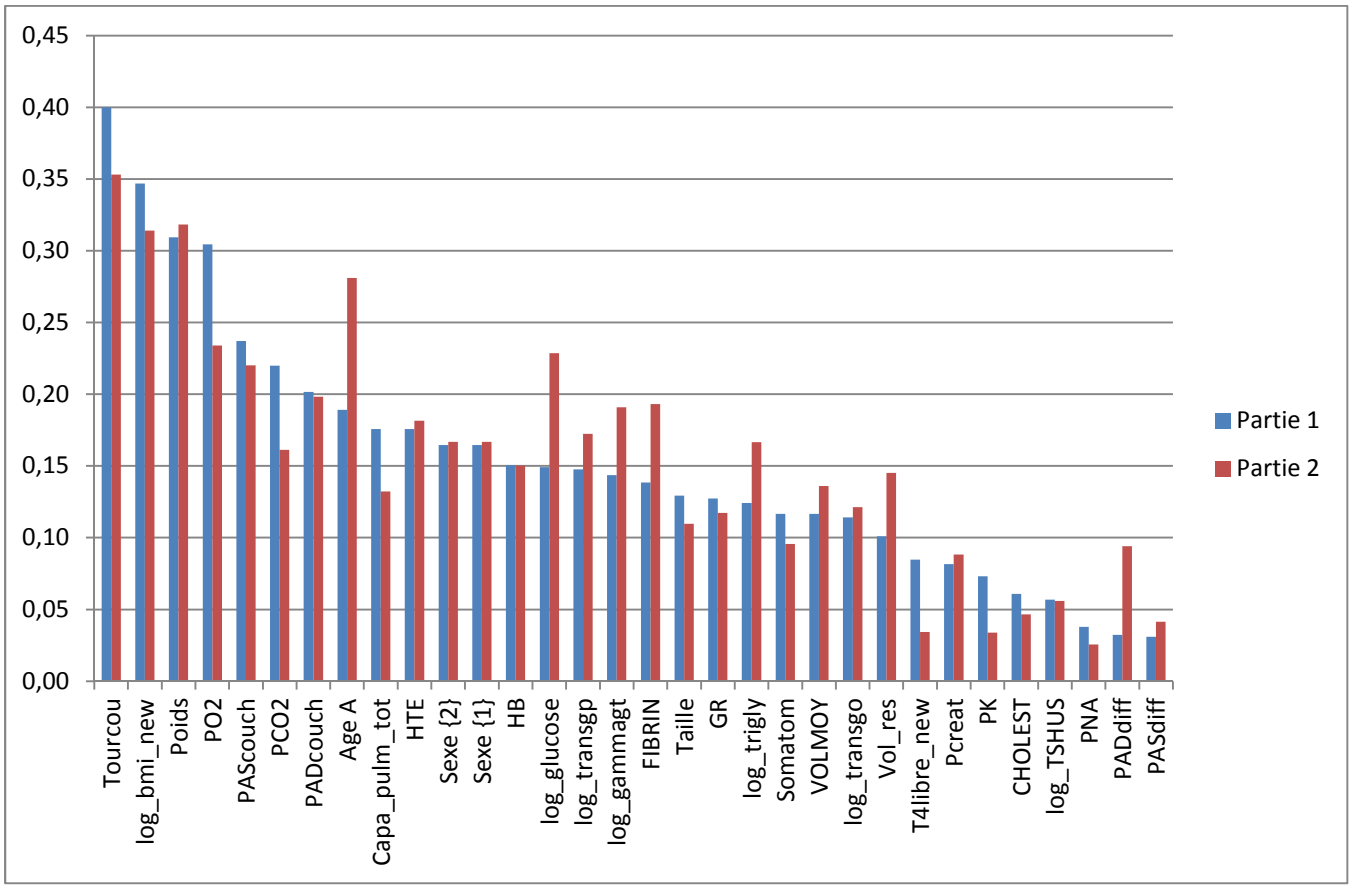
Le premier objectif de ce stage est de décrire des relations entre les variables du sommeil et les données biologiques pour les patients qui ont eu une première hospitalisation. (C'est à dire ceux dont on possède des informations sur le sommeil). Le jeu de données que j'ai réussi à constituer pour cette analyse comprend :

10 variables polysomnographiques : IAH, HYI, Tcpinf90, Tst1+2_pc, Tstlp_pc, Tstp_pc, seff, sao2mi, sao2m, nchang avec transformation des variables IAH et nchang

31 variables biologiques : Tourcou, poids, taille, bmi, capacité pulmonaire totale, volume résiduel, po2, pco2, pascouché, padcouché, pasdiff, paddiff, fibrin, cholest, t4libre, somatom, pcrat, pna, pk, gr, hb, hte, volmoy, gammagt, transgp, transgo,trigly, tshus, glucose et le sexe (variable qualitative mais pris en compte dans l'analyse avec l'indicatrice)

Chaque sous-échantillon peut être considéré indépendant. Ainsi, les résultats des 2 analyses permettent de confronter les résultats et ainsi de vérifier leur cohérence.

Figure 9 : Comparaison VIP entre les 2 parties



On peut remarquer une similarité du VIP entre les 2 parties. Les variables les plus influentes sont le tour du cou, le BMI et le poids pour chaque partie.

Lors de cette régression, on possède l'information du R^2 sur chaque composante principale pour les variables dépendantes et explicatives. De plus, la régression PLS permet d'obtenir 3 composantes principales significatives pour les 2 parties.

Figure 10 : Comparaison R^2 des variables prédictives entre les 2 parties

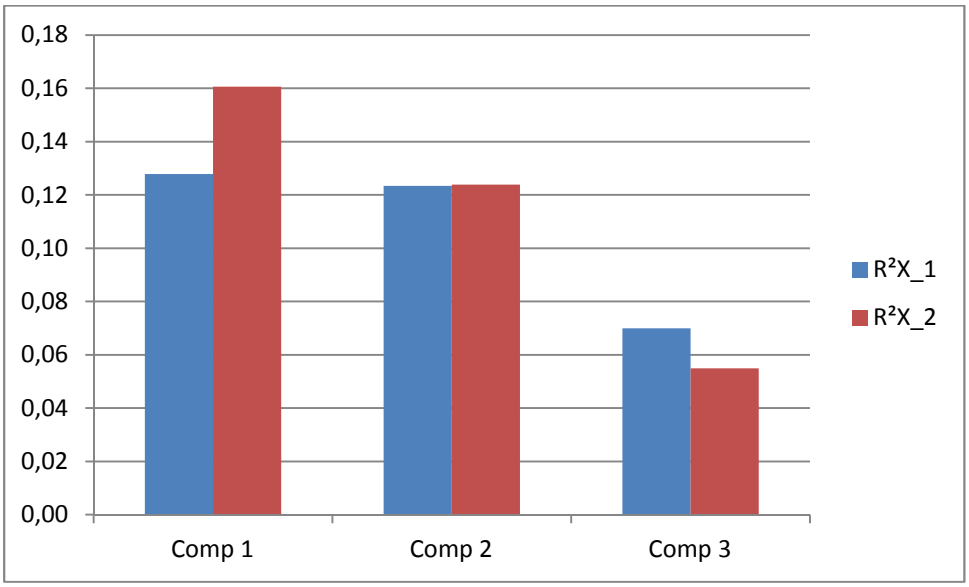
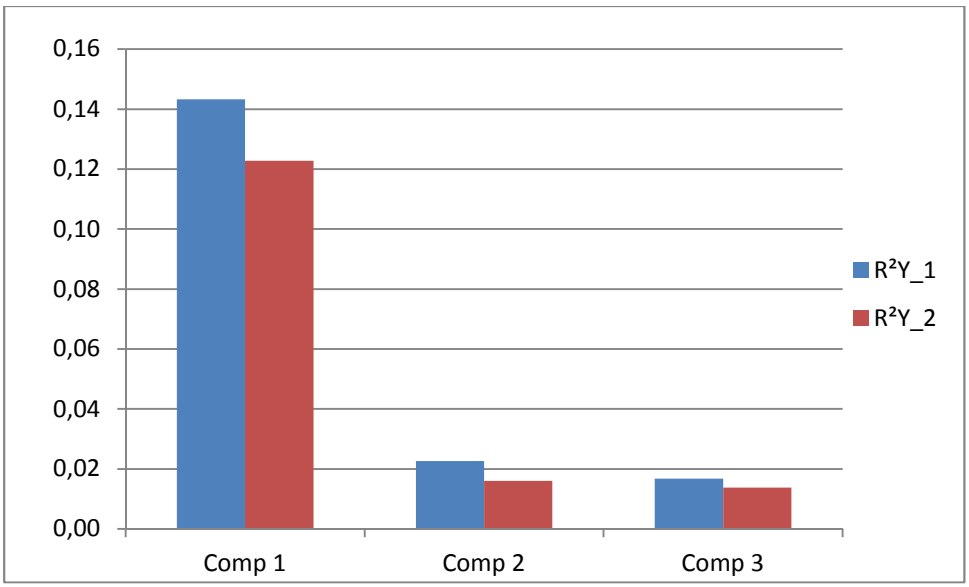


Figure 11 : Comparaison R^2 des variables dépendantes entre les 2 parties



Les 3 premiers axes factoriels expliquent 31% (respectivement 33%) des variables explicatives et 16% (respectivement 15%) des variables dépendantes de la partie 1 (respectivement partie 2). Ce résultat est assez faible, mais il y a une forte similarité entre les 2 parties. L'apport de 2% est significatif, ceci est dû à la règle de Q^2_{hk} qui impose seulement à au moins une des variables prédictives d'ajouter de l'information.

La comparaison des poids factoriels (figure 12) sur chaque nouvelle composante principale permet de comparer l'information contenue sur cette première composante. Malgré l'apport du R^2 faible, on s'attend à ce que les poids soient identiques.

Figure 12 : Poids factoriels des X sur la première composante

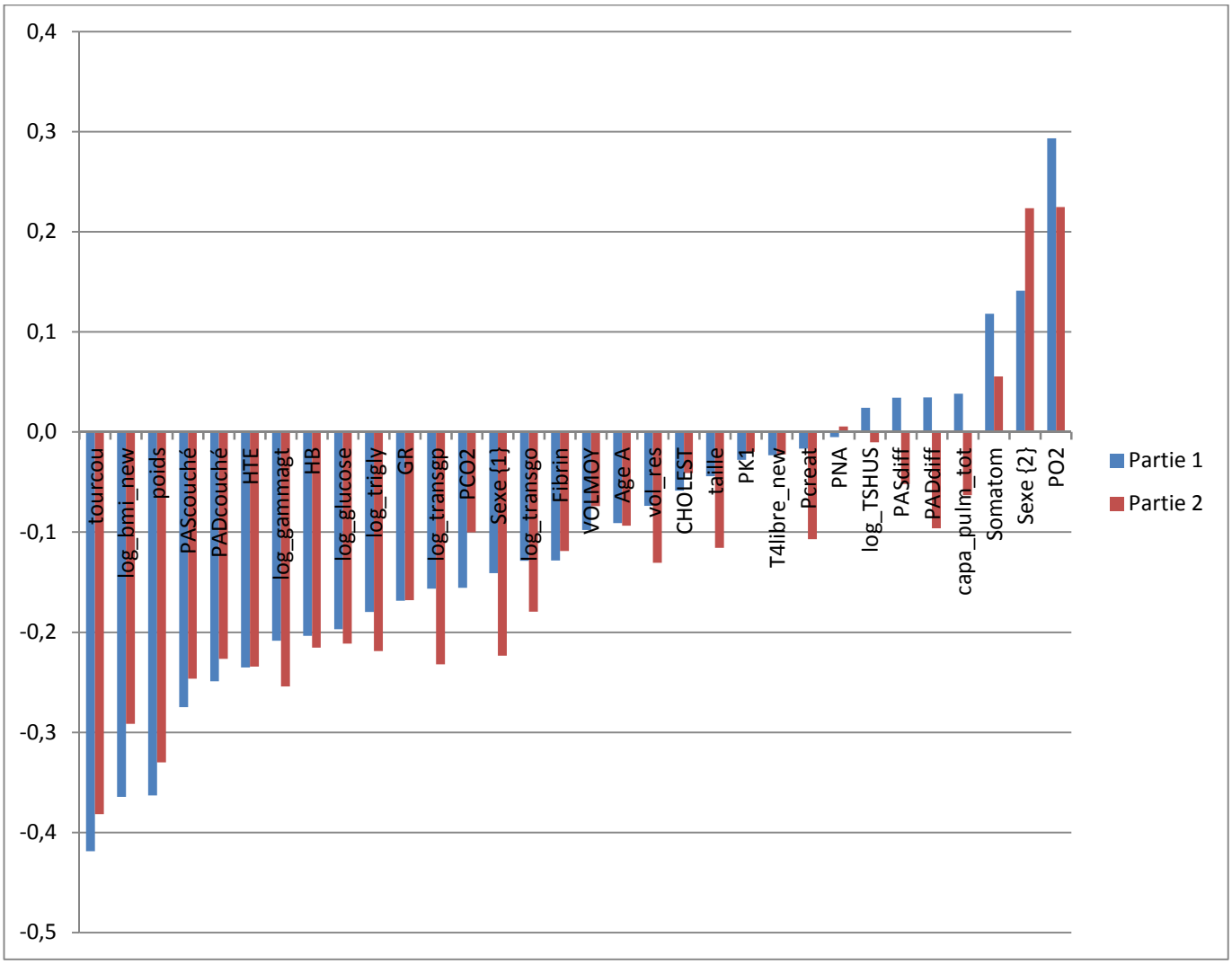
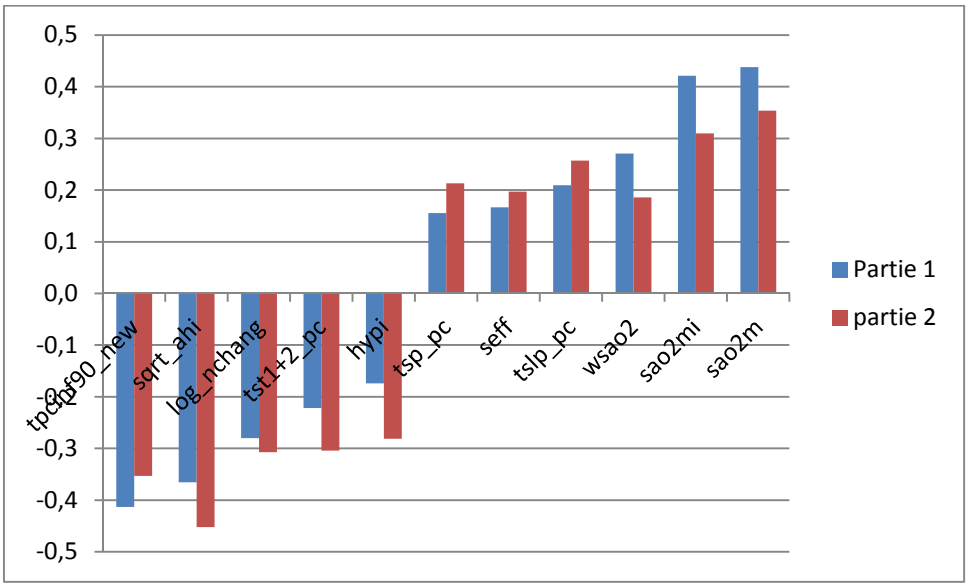


Figure 13 : Poids des Y sur la première composante



Comme pour le VIP, les poids sont similaires entre les 2 parties.

3.6 Conclusion

On remarque une faible information contenue dans les composantes principales, malgré tout l'apport de la régression PLS. D'autres analyses PLS ont été effectuées, notamment en séparant les variables du sommeil en 2 parties : les données respiratoires et les données observées ou en réduisant le nombre de variables prédictives, mais elle n'excède pas les 30% de variance expliquée.

L'objectif était de mieux comprendre les relations entre les variables du sommeil et les variables prédictives. L'étude réalisée a permis d'établir une hiérarchie sur l'importance de ces liens, grâce à la similarité de la régression PLS des 2 sous échantillons indépendants. La forte dispersion de nos données (et peut être la faible linéarité) empêche d'établir un modèle prédictif. La part expliquée par les résidus est trop importante, impliquant une faible valeur du R^2.

4. Données longitudinales

4.1 Présentation

L'intérêt principal est la répétition des données. On s'intéresse à l'évolution des variables suivantes : la glycémie, la T4libre, le tshus, la trigly, la transgp, la transgo, la pression artérielle et le cholestérol en recherchant si les facteurs de troubles du sommeil peuvent jouer un rôle et avoir un effet.

Un nouveau fichier doit être créé pour effectuer l'analyse. Le nombre restreint de données sur les patients 15 ans et 20 ans incite à utiliser seulement les données jusqu'à 10 ans. Tous les patients répétés sont présents dans la partie 1, ainsi il n'y a plus de séparation pour cette étude. Le principe de regroupement est le même que dans la première partie. Le fichier final possède 85 observations.

Une nouvelle variable s'ajoute au jeu de donnée : l'observance. Il s'agit du temps moyen en heures d'utilisation du traitement par nuit. Il s'agit d'une variable continue et elle permet d'ajouter un facteur d'influence sur l'évolution des troubles du sommeil.

Un exemple de la fragilité de la base de données

Lors de ce regroupement, j'ai voulu intégrer une variable diabète qui est présente dans le fichier. Cette variable est codée 1 si la personne est diabétique, 0 sinon. Le diabète a pour conséquence une augmentation de la glycémie (glucose). Après traitement, ce taux de glycémie redevient normal, c'est-à-dire aux alentours de 5 mmol/l. Un patient ayant un taux de glycémie supérieur à 7.5 mmol/l est considéré diabétique. Les données sont établies sans protocole expérimental, des patients peuvent venir effectuer les analyses cliniques en ayant pris ou non leur traitement, donc leur taux de glycémie pouvait être très haut (s'il ne l'avait pas pris) ou dans les normes (s'il l'avait pris). Ceci engendre un biais important sur l'analyse sachant qu'une des questions principales est l'évolution du diabète pour les patients souffrant de maladie du sommeil. De plus, une analyse approfondie montre des patients pouvant avoir des taux de glycémie très élevés, donc le patient est diabétique, sans pour autant avoir la variable diabète codée 1. Ce cas de figure est corrigeable, mais à l'inverse, si un patient a pris son traitement du diabète, et que la variable diabète n'a pas été indiquée, il est impossible à savoir si cette personne est diabétique. Cette notion montre toute la fragilité et la difficulté d'interprétation de la base de données.

On s'attend à une analyse peu robuste, comme il a été le cas pour la régression PLS.

4.2 Le modèle

Pour étudier l'évolution des variables, on a décidé d'utiliser un modèle de régression multiple. Le principe est d'utiliser cette régression multiple pour trouver les effets des facteurs polysomnographiques sur la différence du facteur dépendant entre le temps t1 (la 10ème année) et t0 (la 1ère année).

Le modèle de régression est :

$$Y = XB + E$$

Plus précisément :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_{ij} \quad , i = 1..n$$

Où Y_i est la différence entre la mesure d'une variable au temps t1 et t0. Le paramètre β_0 est l'effet moyen de la différence.

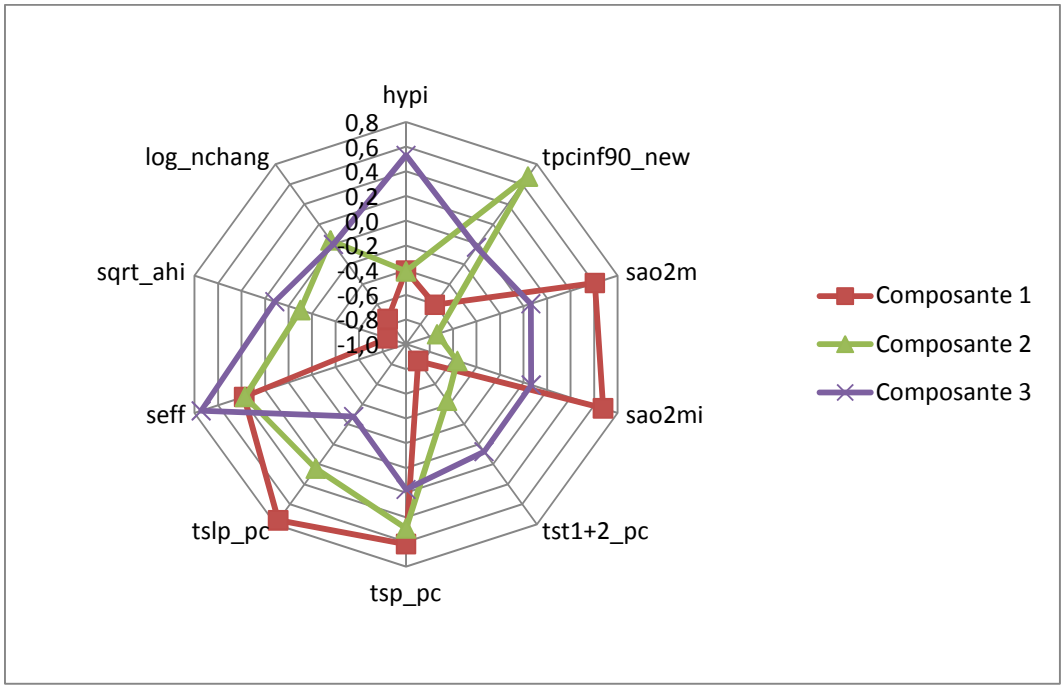
Les hypothèses à vérifier lors de cette régression sont :

- L'indépendance des erreurs : les patients sont supposés indépendants et ne sont donc pas liés
- L'homogénéité des résidus : on tracera les valeurs prévues en fonction des résidus
- La normalité des résidus : on utilise la droite de Henry et le test de Shapiro-Wilk

Le modèle a un problème de colinéarité entre les régresseurs avec l'utilisation des données brutes. Pour cela, on utilise les composantes principales pour y remédier. On s'est interrogé si les composantes trouvées dans l'analyse PLS peuvent être utilisées. PLS est un modèle de régression, donc l'utilisation des composantes principales peut déstructurer l'information contenue. En effet, les composantes principales trouvées sont calculées à l'aide des covariances entre toutes les variables, incluant une information supplémentaire dans le calcul des composantes principales. On a préféré regrouper l'information des variables polysomnographiques dans des composantes principales réalisées grâce à l'algorithme NIPALS sur un seul bloc de variable (comme une ACP). L'avantage est la prise en compte des valeurs manquantes, ainsi chaque patient possède une coordonnée. Ces composantes principales sont calculées dans le jeu de données de la partie 1, puisque tous les patients répétés sont présents dans cette partie. Leur score (ou poids sur la composante) est ensuite intégré dans le fichier contenant les 85 patients avec l'utilisation d'Access.

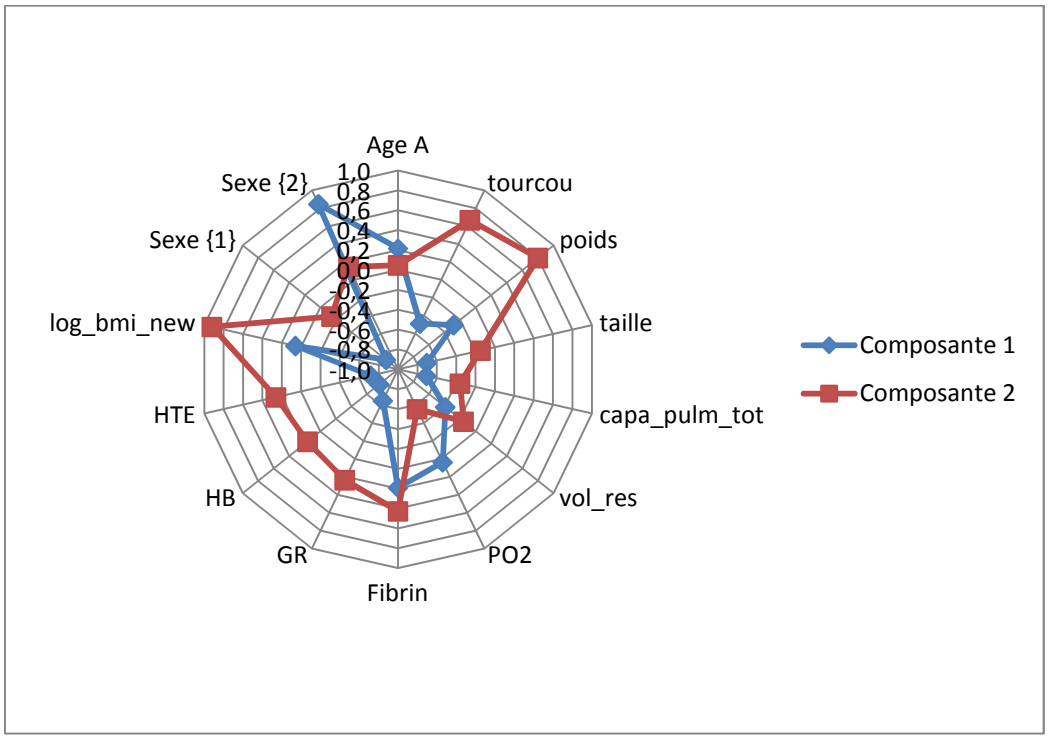
Les données polysomnographiques permettent d'avoir 75% de la variance expliquée pour les 3 composantes. En Annexe 6 se trouve la sortie Statistica qui calcule la contribution sur chaque composante.

Figure 14 : Contribution des variables polysomnographiques sur les 3 composantes



La même méthode est utilisée pour les données biologiques. On s'est servi des variables ayant le plus d'influence dans la première sans intégrer les données dont l'évolution sera étudiée. L'ACP permet d'obtenir 60% de la variance expliquée pour les 2 composantes principales trouvées.

Figure 15 : Contribution des variables biologiques sur les 2 composantes



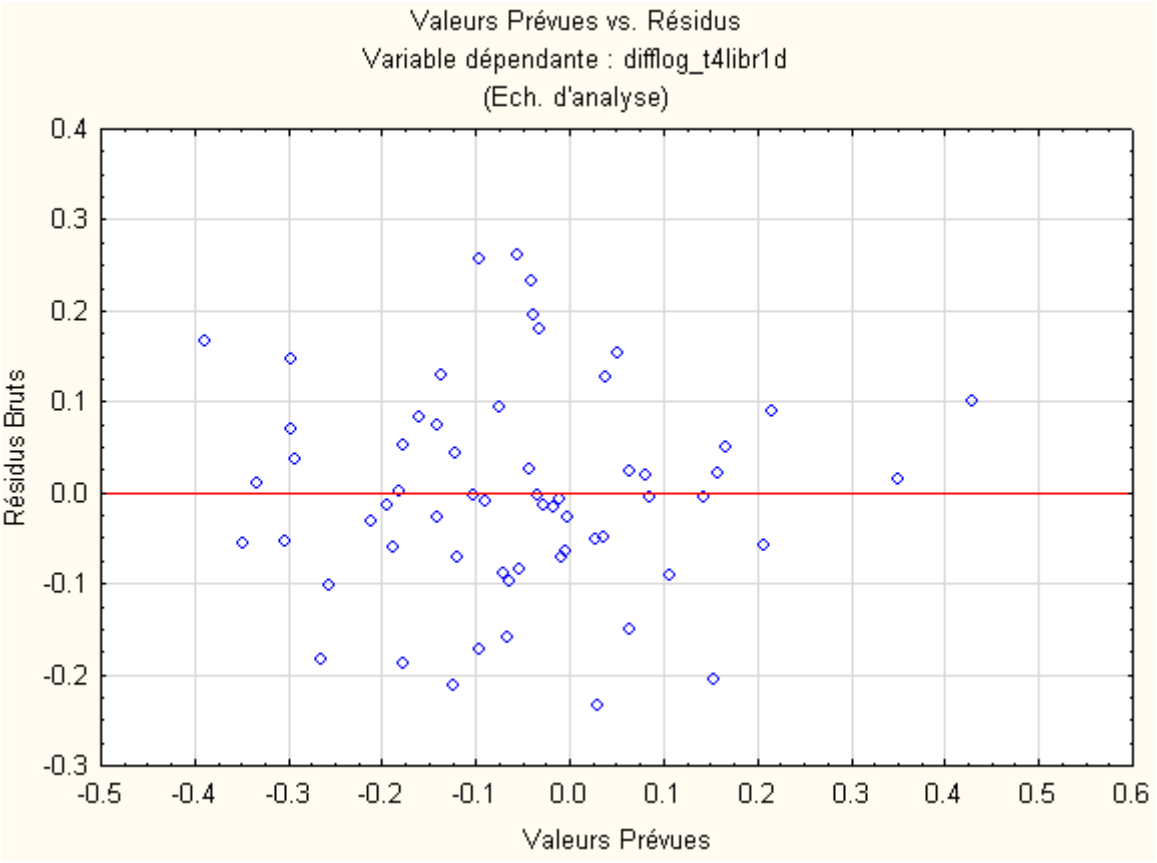
On remarque une forte contribution du sexe sur la première composante, ainsi que du BMI, du tour du cou et du poids sur la deuxième composante.

4.3 Application

Cet exemple traite de la régression sur la différence entre la T4libr au temps t1 et au temps t0. Les valeurs manquantes sur une des 2 variables permettent de retenir 61 observations. On réalise donc la régression avec la valeur de Y au temps t0, les coordonnées (ou scores) des composantes principales pour chaque observation, plus l'utilisation de la covariable « observance ».

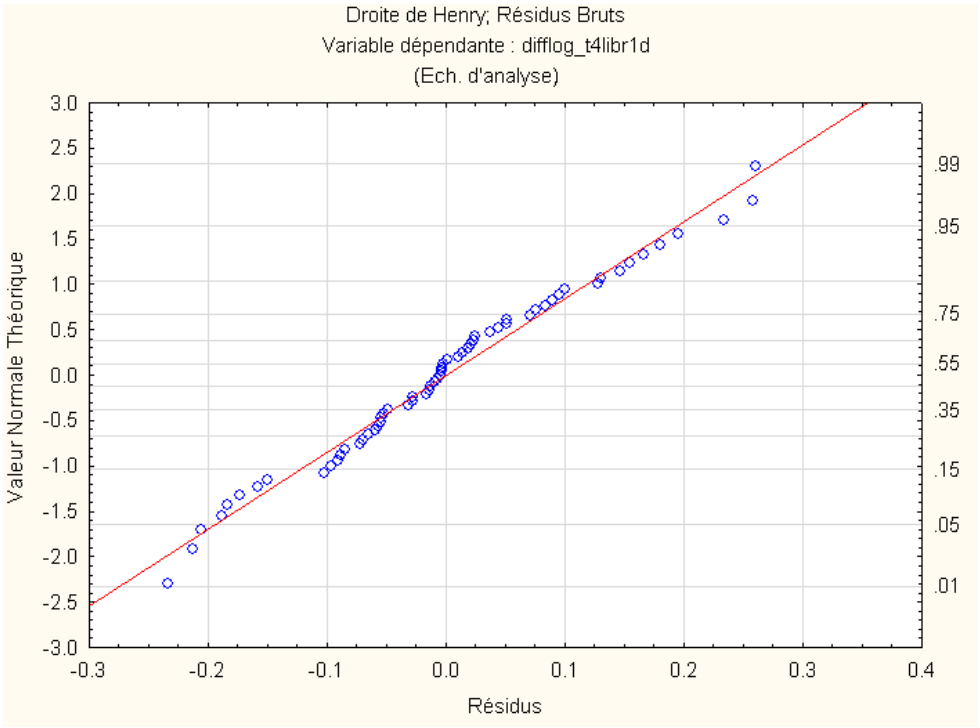
On vérifie l'homogénéité des résidus en traçant le nuage de points des résidus en fonction des valeurs prévues et on utilise la droite de Henry pour vérifier l'hypothèse de normalité des erreurs :

Figure 16



On constate que les résidus ont une espérance nulle et sont répartis autour de la valeur nulle, on peut donc conclure à l'homogénéité des variances.

Figure 17 : Droite de Henry des résidus



Les résidus sont alignés sur la droite. Le test de Shapiro-Wilk trouve une valeur de la statistique valant 0.9803 représentant une p_valeur de 0.4290, ainsi l'hypothèse de normalité des résidus est acceptée.

On vérifie si le modèle est intéressant. Pour cela, on calcule les caractéristiques du modèle en prenant compte du calcul de la variance expliquée (R^2), ainsi que le test :

$$H_0 : \beta_1 = \cdots = \beta_p = 0\,,\ \ p = 7$$

Contre l'hypothèse alternative :

$$H_1 : Au\ moins\ un\ \beta_j \neq 0\,,\ \ j = 1..7$$

La figure suivante montre les résultats obtenus :

Figure 18

Test de SC Modèle Complet vs. SC Résidus

	Multiple - R	Multiple - R^2	Ajusté - R^2	SC - Modèle	dl - Modèle	MC - Modèle	F	p
difflog_t4libr1d	0.820481	0.673190	0.630026	1.624277	7	0.232040	15.596	0.000000

On remarque que pour une p-valeur associée au risque 0.05, on a un rejet du test H_0. Le R^2 vaut 67%, indiquant un bon modèle, mais avec un tiers de la variance non-expliquée.

On calcule l'effet significatif de chaque variable à l'aide du test de Fischer suivant :

$$H_0 : \beta_j = 0\,,\ \ j = 1..7$$

Contre l'hypothèse alternative :

$$H_1 : \beta_j \neq 0\,,\ \ j = 1..7$$

La figure suivante montre les résultats obtenus :

Figure 19

Tests Univariés de Significativité de difflog_t4libr1d

	SC	Degré de - Liberté	MC	F	p
Ord.Orig.	0.948443	1	0.948443	63.74834	0.000000
observance	0.006702	1	0.006702	0.45046	0.505033
Comp1_acpbio	0.088143	1	0.088143	5.92445	0.018335
Comp2_acpbio	0.009844	1	0.009844	0.66163	0.419625
Comp1_acpsomno	0.083391	1	0.083391	5.60501	0.021592
Comp2_acpsomno	0.000213	1	0.000213	0.01431	0.905245
Comp3_acpsomno	0.003643	1	0.003643	0.24485	0.622773
log_t4libr1	1.044280	1	1.044280	70.18989	0.000000
Erreur	0.788530	53	0.014878		

Les premières composantes principales des données du sommeil et biologiques ont un effet au risque 0.05 sur l'évolution de la T4_libr. Les variables polysomnographiques de cette composante ayant une forte contribution sont l'efficacité du sommeil (seff), la pression en oxygène minimum et moyenne (po2m et po2mi), le pourcentage de sommeil paradoxal et lent-profond. La variable biologique ayant le plus d'importance sur la première composante est le sexe. De plus la T4libr (log_t4libr après transformation logarithme) au temps t0 à un effet sur son évolution.

Le modèle explique 67% de la variance, qui représente donc un bon modèle.

En effectuant le modèle de régression et en vérifiant les hypothèses sur les autres évolutions entre t0 et t1, on obtient :

Variable dépendante	**Variable(s) explicative(s) significative(s)**	**R^2**	**Nombre d'obs. utilisé**
Diff_t4libr	log_t4libre1, comp1_somno, comp1_acpbio	67%	61
Diff_tshus	Aucun effet	17%	
Diff_gluc	Aucun effet	15%	65
Diff_transgo	log_transgo1	60%	58
Diff_transgp	log_transgp1, comp1_acpsomno	61%	56
Diff_trigly	log_trigly1, observance	24%	58
Diff_cholest	cholest1	33%	59
Diff_ahi	sqrt_ahi1	30%	75
Diff_Nchang	log_chang1	31%	68
Diffpascouch	pascouch1	50%	55

Diffpadcouch	padcouch1	63%	53
Diffpasdebout	observance, pasdebout1	66%	54
Diffpaddebout	paddebout1	80%	55

4.4 Conclusion

Certains modèles possèdent une bonne part de la variance expliquée, aux alentours de 65%, d'autres sont plus restrictifs et n'expliquent qu'entre 10% et 30%. De plus, les variables au temps t0 ont toujours un effet significatif à son évolution. Des modèles fiables sont difficiles à dégager des analyses. Les raisons possibles sont :

- le traitement des patients concernant le diabète ou l'hypertension artérielle qui n'est pas contrôlé, biaisant ainsi l'analyse.
- les facteurs ne sont simplement pas significatifs. Les simples composantes ne suffisent pas à l'explication du modèle. En essayant de régresser avec les variables d'origine et sans les composantes principales, le nombre d'observations est très faible à cause des valeurs manquantes. Seulement 3 observations contiennent des informations sur toutes les variables.

On a débattu sur l'évolution de cette base de données, qui a sûrement demandée une attente trop longue avant d'utiliser l'information. La récolte et le nettoyage des données sont des étapes primordiales à une bonne analyse statistique lors des études cliniques. On doit savoir à l'avance comment les données vont être analysées avant de créer la base de données. Une mise à jour de la base de données a été entreprise par une clinicienne, qui a finalement quitté le service, laissant son travail en suspense.

Cette partie a été effectuée à la fin du stage, le manque de temps a empêché la mise en place d'un modèle peut être plus adapté et robuste.

5. Perspective

Le traitement sur les données longitudinales n'a pas pu aboutir aux résultats souhaités et la régression PLS n'explique qu'une faible partie de l'information.

Tout d'abord, une amélioration de la base de données est nécessaire, ou du moins un suivi plus régulier à travers les années. Le logiciel « Foxprow » n'est peut être pas adapté à la quantité d'informations et à l'exportation des fichiers. Par ailleurs une méthode plus rigoureuse serait envisagée quant à l'incorporation des données dans la base, pour minimiser le nombre d'erreurs. Un protocole précis pourrait profiter à l'analyse statistique des données répétées. La mise en place d'un protocole est complexe à instaurer. Cela demande un coût élevé que ne peut pas se permettre le service.

Sur le plan statistique, un travail plus approfondi sur la distribution des variables pourrait améliorer l'analyse. La dispersion et la non-normalité peuvent empêcher le modèle de régression d'utiliser le maximum d'information. D'autres méthodes PLS plus spécifiques peuvent être appliquées, comme la multiblock PLS traitant de la non-linéarité entre les facteurs ou la PLS avec l'utilisation du Bootstrap. Le logiciel Statistica n'est pas adapté pour ces techniques, il conviendrait d'utiliser R.

Bilan

Ce stage au sein de l'INCI a été très enrichissant professionnellement et m'a permis de découvrir le monde de la recherche. J'ai pu effectuer un travail dont les résultats intéressaient fortement les cliniciens. Ma mission a été effectuée en autonomie, avec les conseils toujours judicieux de mon maître de stage, le Docteur MALAN. Cela m'a permis d'avoir une grande liberté sur les techniques employées, ce qui n'est pas souvent le cas lors d'une étude clinique.

L'interaction entre le clinicien et le statisticien m'a permis de comprendre l'importance de celle-ci. C'est lors de ces entrevues que le clinicien expose ses attentes et que le statisticien réfléchit aux différentes méthodes qu'il peut employer. Chaque analyse change en fonction du résultat souhaité et des variables choisies. Durant le stage, on s'est souvent demandé si certaines variables quantitatives devaient être transformées en variable qualitative. Quelques analyses spécifiques et simples m'ont été demandées sur la base de données, comme l'évolution d'apnées en fonction du sexe, mais sans analyse statistique importante.

J'ai pu appliquer quelques méthodes statistiques vues en cours à des données réelles, et ainsi comprendre leur utilité tout en approfondissant mes connaissances. Enfin, j'ai surtout pu approfondir mes connaissances sur d'autres méthodes statistiques, tels que la régression PLS, qui m'était inconnue. Pour cela, j'ai dû comprendre la manière d'interpréter les résultats et les limites de validité de la méthode.

Bibliographie

Livres :

- M. Tenenhaus, (2010) La régression PLS, Théorie et pratique, *TECHNIP*

- Stéphane Tufféry, (2010, troisième édition actualisée et augmentée), DataMining et statistique décisionnelle, l'intelligence des données, *TECHNIP*

Sites internet :

- Cours de statistiques [en ligne] : http://wikistat.fr/ : site complet sur les modèles statistiques et les traitements à effectuer

- http://jml85.pagesperso-orange.fr/Pages/cours_VBA.htm : premier pas avec Access

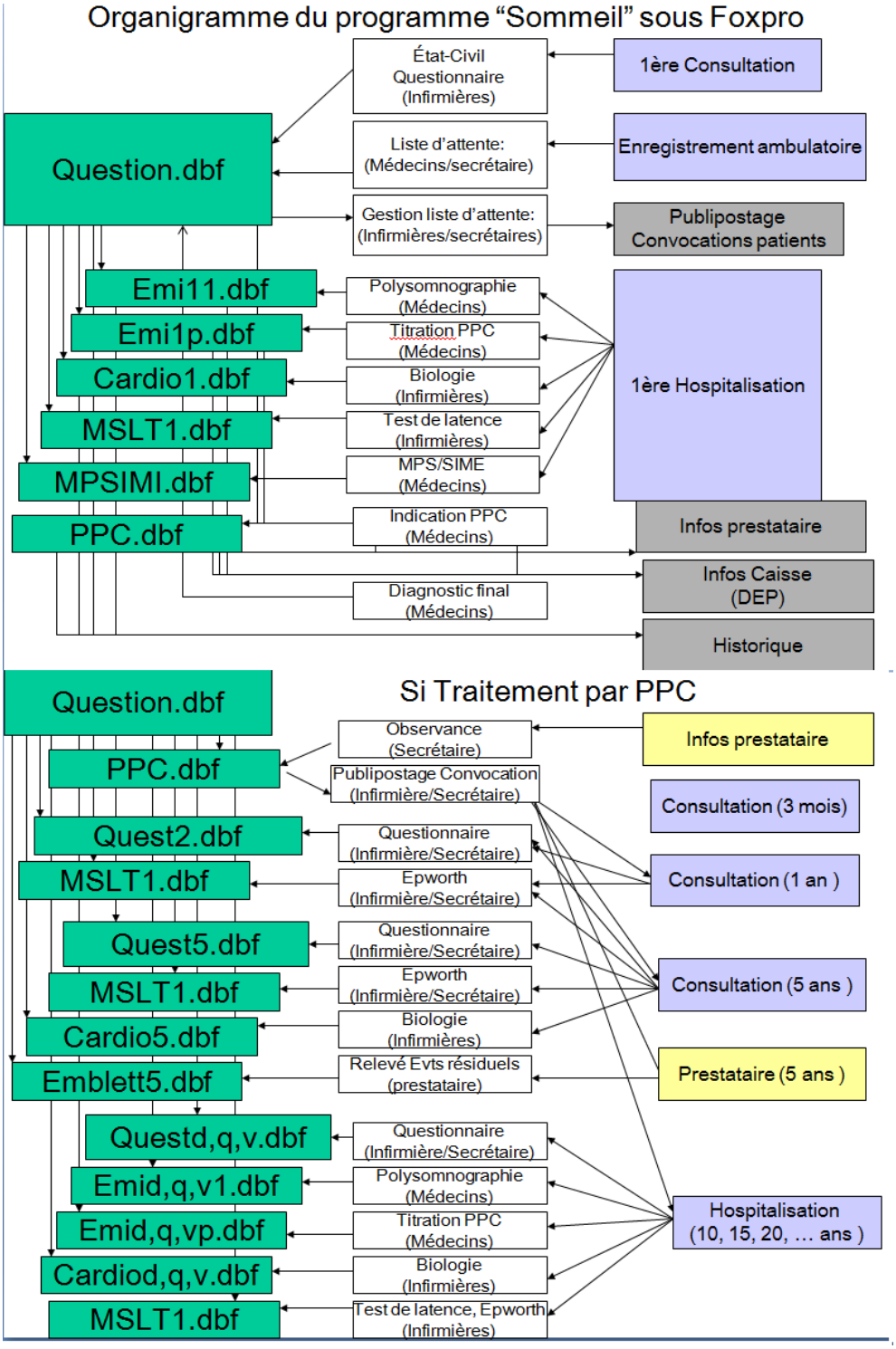
- http://cerig.efpg.inpg.fr/tutoriel/bases-de-donnees/ : approfondissement des connaissances Access

- http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf : cours sur les corrélations et les tests

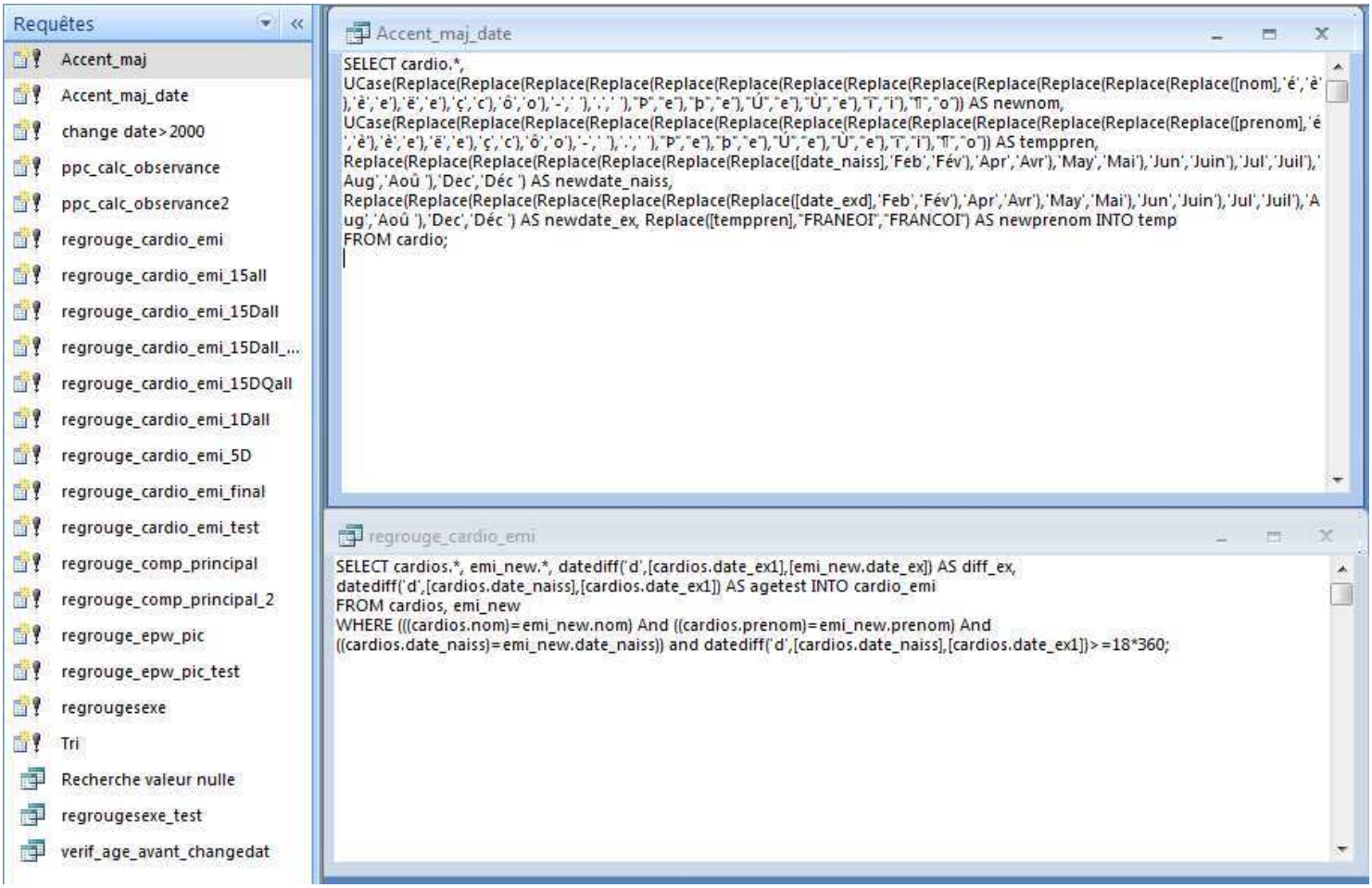
- http://www.statsoft.com/textbook/partial-least-squares/ : aide Statistica en ligne

Annexe

Annexe 1 : Organigramme



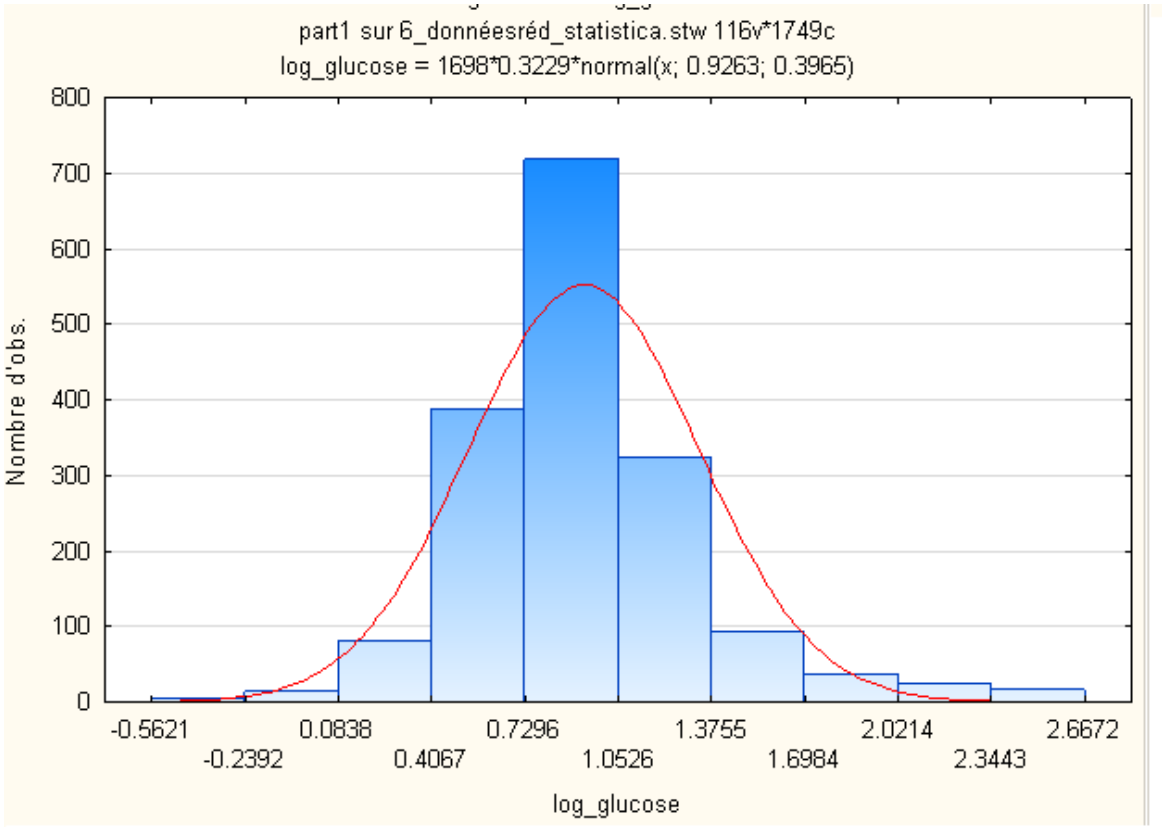
Annexe 2 : Requête Access

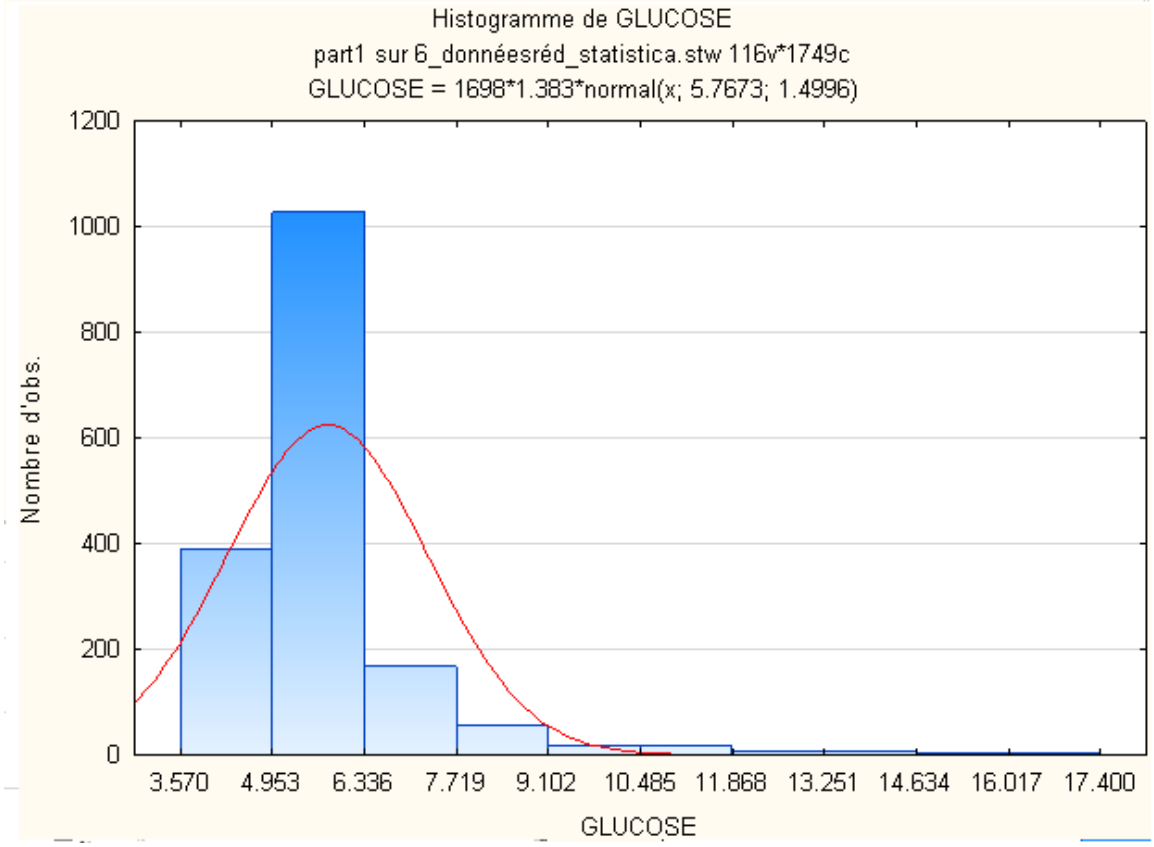


Annexe 3 : Intervalle de conservation des données

Variable	Min	Max	Description
Sexe	variable	catégorielle	Sexe de la personne
tourcour	29	63	Tour du cou de la personne
poids	37	205	poids
taille	140	220	taille
bmi	10	65	Indice de masse corporel
PO2	40	140	Pression o2
PCO2	20	76	Pression co2
PASCOUCH	50	220	Pression art. Sys. Couch.
PADCOUCH	45	220	Pression art. Dias. Couch.
PASDEBOU	50	220	Pression art. Sys. Debout
PADDEBOU	45	220	Pression art. Dias. Debout
Glucose	2	20	Tx de glucose en mmol/l
Calcium	2	3	Mmol/l norm 2.25 à 2.60 donc OK 2 à 3 (OK pour 1.92 et 1.56)
Transgo	2	80	transaminase ALAT ou TGO en U/l norm 5 à 40, mais OK pour les valeurs max 316
Transgp	1	210	Transaminase ASAT ou TGP en U/l, norm 5 à 45, mais ok pour max 210
Gammagt	1	300	Foie en U/l, norm 5 à 35, mais OK pour extrêmes
…	…	…	…

Annexe 4 : Histogramme de la distribution de la variable glucose avant et après transformation





Annexe 5 : Corrélation entre variables biologiques et polysomnographiques de chaque partie + test d'égalité des covariances

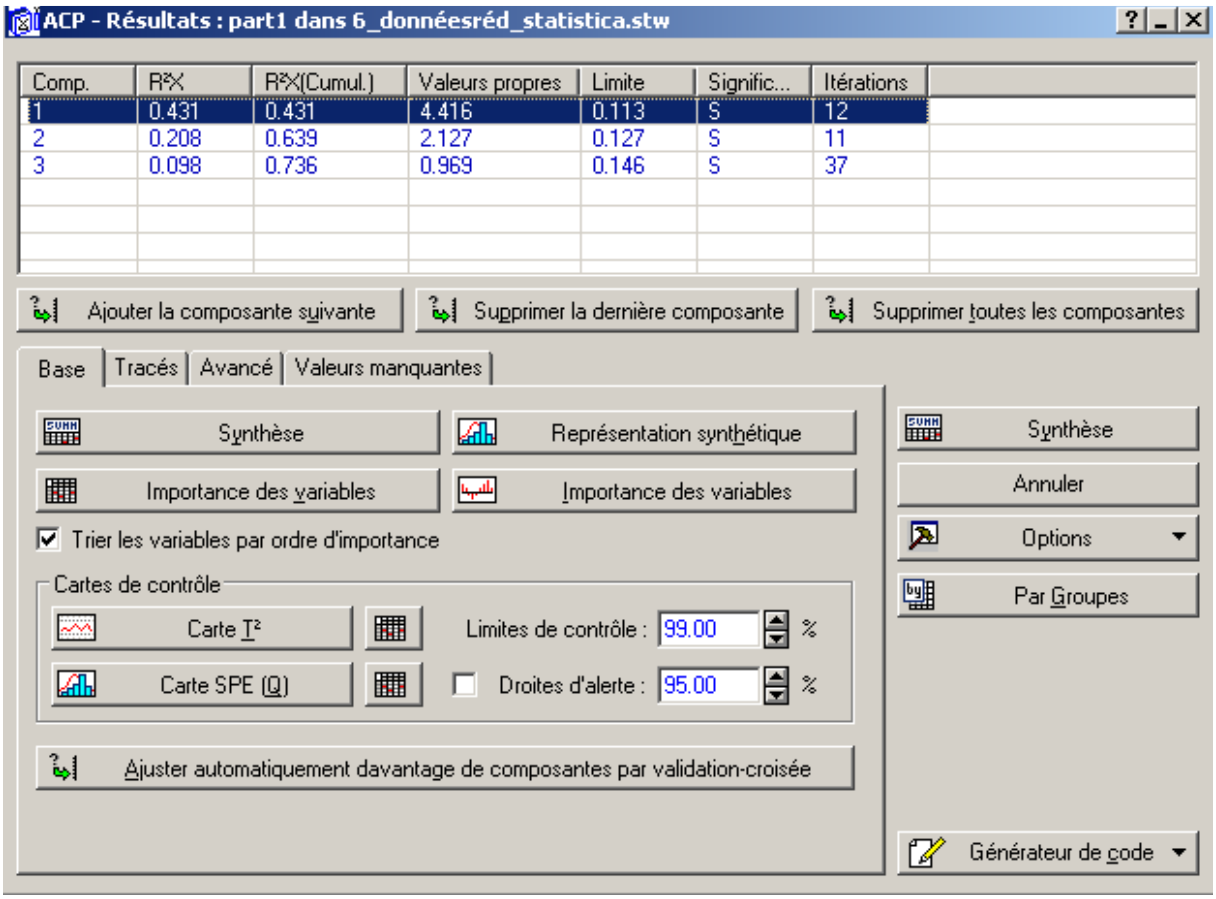
partie 1	tpcinf90_new	sao2m	sao2mi	tst1+2_pc	tslp_pc	tsp_pc	seff	sqrt_ahi	log_nchang
Age A	0.067	-0.109	-0.132	0.052	-0.100	0.017	-0.302	0.101	0.120
tourcou	0.355	-0.367	-0.369	0.245	-0.219	-0.191	-0.200	0.389	0.221
poids	0.347	-0.376	-0.299	0.150	-0.068	-0.188	-0.073	0.262	0.135
taille	-0.040	0.042	0.113	0.032	-0.056	0.005	0.029	0.025	0.024
capa_pulm_tot	-0.074	0.083	0.154	-0.006	-0.039	0.053	0.021	0.012	0.015
vol_res	0.097	-0.123	-0.024	0.020	-0.070	0.041	-0.131	0.060	0.084
PO2	-0.276	0.393	0.254	-0.077	0.055	0.076	0.079	-0.144	-0.113
PCO2	0.203	-0.185	-0.189	0.038	-0.049	-0.013	0.018	0.050	0.031
PAScouché	0.144	-0.178	-0.171	0.061	-0.027	-0.078	-0.119	0.122	0.143
PADcouché	0.097	-0.104	-0.099	0.078	-0.019	-0.117	-0.080	0.134	0.116
PASdiff	0.024	-0.009	0.008	-0.001	-0.003	0.005	-0.013	0.065	0.073
PADdiff	0.010	-0.027	-0.055	-0.014	-0.020	0.047	0.037	0.022	0.040
Fibrin	0.196	-0.218	-0.099	0.106	-0.069	-0.111	-0.028	0.104	0.110
CHOLEST	0.001	-0.006	-0.039	0.056	-0.038	-0.057	-0.080	0.088	0.107
T4libre_new	-0.005	-0.040	-0.085	0.011	-0.042	0.025	0.015	0.105	0.000
Somatom	-0.113	0.115	0.136	-0.010	0.040	-0.027	-0.011	-0.080	-0.101
Pcreat	-0.015	0.005	0.031	-0.006	-0.039	0.053	-0.097	0.053	0.090
PNA	0.004	-0.030	0.011	-0.066	0.079	0.029	0.045	-0.057	-0.026
PK	0.098	-0.074	-0.008	-0.034	0.026	0.030	0.094	0.031	0.053
GR	0.123	-0.098	-0.063	0.007	-0.023	0.012	0.025	0.057	-0.007
HB	0.159	-0.117	-0.039	0.034	-0.053	-0.002	-0.029	0.089	0.058
HTE	0.215	-0.193	-0.091	0.030	-0.041	-0.007	-0.036	0.098	0.046
VOLMOY	0.112	-0.138	-0.026	0.052	-0.043	-0.044	-0.072	0.044	0.068
log_gammagt	0.114	-0.131	-0.174	0.117	-0.109	-0.087	-0.031	0.166	0.106
log_transgp	0.058	-0.071	-0.122	0.076	-0.096	-0.029	0.039	0.146	0.023
log_trigly	0.085	-0.109	-0.122	0.112	-0.025	-0.169	-0.097	0.130	0.071
log_TSHUS	-0.112	0.069	0.028	0.020	0.028	-0.065	-0.004	-0.055	-0.070
log_transgo	0.020	-0.048	-0.105	0.034	-0.075	0.022	0.020	0.084	0.000
log_glucose	0.092	-0.088	-0.176	0.093	-0.091	-0.064	-0.095	0.137	0.065
log_bmi_new	0.385	-0.417	-0.393	0.143	-0.052	-0.195	-0.105	0.265	0.129
wsao2	-0.342	0.476	0.224	0.056	-0.030	-0.067	0.022	-0.009	-0.092

partie 2	tpcinf90_new	sao2m	sao2mi	tst1+2_pc	tslp_pc	tsp_pc	seff1	sqrt_ahi	log_nchang
Age A	0.101	-0.145	-0.083	0.148	-0.111	-0.128	-0.362	0.160	0.161
Tourcou	0.335	-0.348	-0.313	0.274	-0.238	-0.184	-0.194	0.434	0.292
Poids	0.327	-0.330	-0.297	0.232	-0.178	-0.196	-0.067	0.349	0.215
Taille	0.023	-0.044	0.019	-0.017	-0.009	0.052	-0.027	0.067	-0.007
Capa_pulm_to	-0.054	0.027	0.087	-0.066	0.022	0.105	-0.019	-0.013	-0.040
Vol_res	0.089	-0.109	-0.035	0.079	-0.081	-0.031	-0.207	0.119	0.102
PO2	-0.255	0.311	0.204	-0.107	0.096	0.066	0.154	-0.196	-0.138
PCO2	0.072	-0.077	-0.114	0.097	-0.082	-0.070	0.070	0.083	0.086
PAScouch	0.176	-0.154	-0.133	0.189	-0.133	-0.178	-0.191	0.254	0.200
PADcouch	0.098	-0.109	-0.105	0.176	-0.146	-0.128	-0.124	0.215	0.156
PASdiff	0.085	-0.101	-0.046	0.068	-0.056	-0.049	-0.036	0.080	0.105
PADdiff	0.083	-0.104	-0.107	0.090	-0.068	-0.077	-0.045	0.095	0.084
FIBRIN	0.183	-0.162	-0.127	0.156	-0.128	-0.118	-0.061	0.139	0.100
CHOLEST	-0.025	0.004	0.039	0.030	-0.040	0.004	0.020	0.013	0.017
T4libre_new	-0.014	0.003	0.006	-0.006	-0.020	0.048	-0.064	-0.011	-0.022
Somatom	-0.133	0.123	0.154	0.002	-0.047	0.076	0.037	-0.115	-0.078
PCREAT1	0.007	-0.023	-0.037	0.072	-0.062	-0.049	-0.099	0.099	0.074
PNA	-0.004	-0.038	-0.070	-0.080	0.077	0.042	-0.050	0.006	-0.060
PK	0.018	0.002	-0.022	0.009	0.001	-0.022	0.034	-0.019	-0.001
GR	0.036	-0.049	-0.002	0.086	-0.082	-0.046	-0.020	0.121	0.061
HB	0.098	-0.139	-0.017	0.082	-0.068	-0.059	-0.096	0.158	0.098
HTE	0.127	-0.158	-0.044	0.145	-0.130	-0.089	-0.106	0.199	0.140
VOLMOY	0.140	-0.170	-0.071	0.082	-0.065	-0.064	-0.133	0.110	0.121
log_gammagt	0.133	-0.147	-0.101	0.150	-0.134	-0.093	-0.106	0.179	0.117
log_transgp	0.102	-0.132	-0.129	0.157	-0.119	-0.135	-0.031	0.216	0.133
log_trigly	0.104	-0.121	-0.098	0.163	-0.138	-0.115	-0.056	0.214	0.165
log_TSHUS	0.012	-0.021	-0.014	0.050	-0.011	-0.090	-0.020	0.048	0.004
log_transgo	0.076	-0.110	-0.130	0.034	0.003	-0.078	-0.083	0.114	0.092
log_glucose	0.157	-0.175	-0.203	0.170	-0.123	-0.156	-0.117	0.233	0.135
log_bmi_new	0.307	-0.313	-0.304	0.246	-0.179	-0.224	-0.059	0.320	0.225
wsao2	-0.382	0.596	0.170	0.019	-0.055	0.053	0.060	-0.033	-0.048

Le test d'égalité des corrélations indiquent seulement au seuil de 5%, un rejet entre la covariance de wsao2 et de sao2m.

partie 1	tpcinf90_new	sao2m	sao2mi	tst1+2_pc	tslp_pc	tsp_pc	seff	sqrt_ahi	log_nchang
Age A	0.448	0.484	0.657	1.275	0.150	1.928	0.900	0.791	0.559
tourcou	0.299	0.290	0.841	0.413	0.256	0.103	0.084	0.719	1.012
poids	0.297	0.705	0.033	1.137	1.478	0.103	0.082	1.274	1.101
taille	0.831	1.136	1.252	0.644	0.621	0.620	0.742	0.554	0.400
capa_pulm_tot	0.272	0.749	0.894	0.795	0.803	0.699	0.533	0.331	0.733
vol_res	0.110	0.190	0.148	0.777	0.142	0.954	1.046	0.786	0.242
PO2	0.302	1.242	0.698	0.402	0.549	0.123	1.001	0.705	0.328
PCO2	1.772	1.468	1.010	0.791	0.438	0.747	0.690	0.448	0.719
PAScouché	0.427	0.328	0.515	1.716	1.416	1.354	0.970	1.824	0.773
PADcouché	0.010	0.060	0.083	1.309	1.698	0.148	0.583	1.104	0.547
PASdiff	0.803	1.221	0.722	0.910	0.712	0.713	0.313	0.199	0.433
PADdiff	0.969	1.024	0.694	1.375	0.634	1.636	1.087	0.968	0.576
Fibrin	0.178	0.766	0.369	0.675	0.789	0.085	0.437	0.476	0.123
CHOLEST	0.339	0.127	1.042	0.354	0.016	0.809	1.325	0.987	1.187
T4libre_new	0.117	0.572	1.207	0.235	0.289	0.299	1.044	1.538	0.291
Somatom	0.261	0.101	0.254	0.162	1.157	1.359	0.622	0.471	0.298
Pcreat	0.293	0.366	0.899	1.029	0.294	1.358	0.024	0.610	0.203
PNA	0.102	0.110	1.083	0.195	0.036	0.171	1.262	0.830	0.453
PK	1.060	1.003	0.188	0.566	0.334	0.689	0.798	0.651	0.721
GR	1.168	0.650	0.808	1.045	0.773	0.773	0.590	0.854	0.902
HB	0.819	0.297	0.287	0.634	0.209	0.751	0.892	0.926	0.534
HTE	1.200	0.486	0.616	1.532	1.179	1.088	0.935	1.376	1.251
VOLMOY	0.375	0.435	0.602	0.397	0.296	0.267	0.807	0.891	0.700
log_gammagt	0.249	0.221	0.983	0.438	0.331	0.087	0.992	0.182	0.144
log_transgp	0.582	0.811	0.094	1.088	0.306	1.416	0.924	0.957	1.469
log_trigly	0.262	0.169	0.324	0.681	1.500	0.738	0.550	1.150	1.264
log_TSHUS	1.646	1.194	0.561	0.397	0.508	0.327	0.219	1.359	0.986
log_transgo	0.750	0.826	0.331	0.004	1.040	1.334	1.367	0.403	1.224
log_glucose	0.876	1.173	0.367	1.038	0.432	1.226	0.301	1.315	0.937
log_bmi_new	1.182	1.596	1.340	1.414	1.709	0.398	0.603	0.801	1.321
wsao2	0.609	2.232	0.731	0.493	0.331	1.582	0.509	0.312	0.591

Annexe 6 : Sortie STATISTICA des résultats de l'ACP avec l'algorithme NIPALS sur les données du sommeil



Annexe 7 : Sortie STATISTICA des poids factoriels des premières et deuxièmes composantes principales sur les variables explicatives et dépendantes