

Visualisation pour les applications de suivi des activités physiques et du sommeil

Travail d'Études et de Recherche

soutenu le 6 juin 2017

pour l'obtention de la première année du

Master Informatique de l'Université d'Artois

par

Hugues Wattez

Encadrant : Karim TABIA

Remerciements

Je souhaite remercier mon encadrant M. Karim Tabia, pour ses nombreux conseils, pour sa présence tout au long de la période d'étude et pour l'initiation au travail de recherche.

De même, je souhaite remercier mon binôme, Nicolas Ydée, pour son sérieux et son travail en collaboration avec le mien. Ce fut un plaisir de partager ces recherches avec lui.

Ensuite, je veux remercier mes enseignants de l'IUT et de la faculté Jean-Perrin à Lens, pour les connaissances qu'ils m'ont prodiguées.

Enfin, je souhaite remercier mes lecteurs et correcteurs pour leur patience et leur participation à ce rapport.

Table des matières

Table des figures	vi
Introduction générale	1

Partie I Etat de l'art

Chapitre 1 Visualisation	4
1.1 Introduction	4
1.2 Objectifs et qualités d'une visualisation	4
1.3 Types de visualisation	4
1.3.1 En fonction de la nature des données	5
1.3.2 En fonction de l'objectif de la visualisation	9
1.4 Visualisation et Big Data	16
1.5 Technologies pour la visualisation	17
1.5.1 Sans programmation	17
1.5.2 Avec programmation	17
1.6 Conclusion	18
Chapitre 2 Visualisation dans les applications de suivi d'activités physiques et du sommeil	19
2.1 Introduction	19
2.2 Histoire	19
2.3 Données et suivi d'activités physiques et du sommeil	21

2.3.1	Données environnementales et spatiales	21
2.3.2	Données corporelles	22
2.4	Données et visualisation pour les activités physiques et du sommeil	23
2.4.1	Visualisation officielle	23
2.4.2	Visualisation non-officielle	23
2.5	Limites	26
2.5.1	Limites matérielles	26
2.5.2	Limites de l'analyse	27
2.6	Conclusion	28

Partie II Analyse

Chapitre 3 Analyse des données initiales et uniformisation	30	
3.1	Introduction	30
3.2	Analyse de fichiers aux données hétérogènes	30
3.3	Mise en base de données	31
3.4	Enrichissement de la base de données	32
3.5	Conclusion	32
Chapitre 4 Choix d'un indicateur pour la qualité du sommeil	33	
4.1	Introduction	33
4.2	Tester l'efficacité d'un indicateur et ses corrélations	33
4.2.1	Définition d'une corrélation	33
4.2.2	Formule de Pearson	34
4.2.3	Formule de Spearman	35
4.2.4	Formule de Kendall	36
4.3	Tests d'un indicateur "sophistiqué"	37
4.4	Choix d'indicateurs plus basiques	38
4.5	Conclusion	38

Partie III Visualisation et implémentation

Chapitre 5 Technologies utilisées	40
5.1 Introduction	40
5.2 Technologies pour la visualisation	40
5.2.1 Utilisation de R	40
5.2.2 Utilisation de javascript et D3.js	40
5.3 Technologies pour l'application web et mobile	41
5.3.1 Programmation côté client	41
5.3.2 Programmation côté serveur	41
5.4 Conclusion	42
Chapitre 6 Visualisation des données physiques et du sommeil	44
6.1 Introduction	44
6.2 Visualisations basiques	45
6.2.1 Visualisation libre	45
6.2.2 Visualisation de l'influence des types d'activité sur la qualité du sommeil	46
6.2.3 Visualisation de la qualité du sommeil	47
6.3 Visualisations de corrélations	48
6.3.1 Graphique à nuage de points	49
6.3.2 Cercle de corrélations	50
6.4 Conclusion	51

Conclusion générale	52
----------------------------	-----------

Bibliographie

Table des figures

1.1	Diagramme à barres colorées	5
1.2	Diagramme à barres empilées	5
1.3	Diagramme à nuage de points	6
1.4	Diagramme en escaliers	6
1.5	Diagramme à barres empilées avec proportions	7
1.6	Treemap	7
1.7	Diagramme à couches empilées	8
1.8	Vue des Etats-Unis avec et sans agrandissement	8
1.9	Chômage des Etats-Unis de 2004 à 2009	9
1.10	Matrice de diagrammes à nuages de points	10
1.11	Diagramme à tiges	11
1.12	Histogramme de la répartition du taux de natalité à travers le monde de 1960 à nos jours	12
1.13	Carte chaude	13
1.15	Matrice d'étoiles et matrice de diagrammes polaires	13
1.14	Faces de Chernoff	14
1.16	Structure à coordonnées parallèles	14
1.17	Diagramme avec échelonnement multi-dimensionnel	15
1.18	Histogramme et diagramme à surface avec observations aberrantes	15
1.19	Propriété des méthodes de visualisation	17
2.1	Chronologie des objets connectés	21
2.2	Dashboard de l'application web Fitbit	23
2.3	Dashboard amateur des données Fitbit	24
2.4	Visualisation avancée des données Fitbit	24
2.5	Première visualisation artistique sur le nombre de pas parcourus	25
2.6	Deuxième visualisation artistique sur le nombre de pas parcourus	25
2.7	Troisième visualisation artistique sur le nombre de pas parcourus	26
3.1	Schéma de la base de données	31
3.2	Schéma de la table météo	32
4.1	Relation entre X et Y sous forme de nuages de points	34
4.2	Corrélogramme avec la formule de Pearson	35
4.3	Comparaison des coefficients de Spearman et Pearson pour une fonction non-affine et monotone	35
4.4	Corrélogramme avec la formule de Spearman	36
4.5	Corrélogramme avec la formule de Kendall	36
4.6	Corrélogramme (Pearson) des indicateurs basiques et de nos données	38

5.1	Schéma des communications client-serveur	42
6.1	Visualisation libre des trois tables sous forme de courbes	45
6.2	Histogrammes à couches empilées et proportionnelles des activités et courbe de la qualité du sommeil	46
6.3	Calendrier sous forme de carte chaude (classé par semaine) de la qualité du sommeil . . .	47
6.4	Superposition de graphiques représentant la qualité du sommeil en fonction des mois et semaines	48
6.5	Nuage de points sur les paires de données les plus corrélées	49
6.6	Diagramme en cordes des corrélations entre l'ensemble des variables	50

Introduction générale

Ces dernières années, nous avons connu l'essor des technologies, non seulement dans leur sophistication mais aussi dans leur miniaturisation. Cette expansion prend en compte les objets connectés. Ces objets, remplis de capteurs en tout genre, permettent de collecter une quantité phénoménale d'informations sur nous et notre environnement. Ces informations transitent ensuite entre notre smartphone et ce qu'on appelle le "cloud". Tout est enregistré et archivé dans d'immenses serveurs aux quatre coins du monde. Il s'agit en réalité du phénomène de "Big Data". Dans ce TER, nous allons étudier l'ensemble de ces notions, de ces dispositifs et plus spécifiquement nous intéresser aux objets connectés liés à la prise d'informations d'activités physiques et du sommeil.

Le problème, lorsque nous possédons une telle quantité de données, est la représentation que nous tentons d'en faire. En effet, il s'agit d'une prise quotidienne d'une foule d'informations liées aux activités physiques et du sommeil, ce qui en un an peut représenter une quantité de données difficilement appréciable pour une simple vision humaine. Ce qui nous incite donc à nous demander par quels moyens peut-on visualiser un tel volume de données ?

Un autre problème vient se joindre à l'analyse de ces données. Est-on capable de retirer un enseignement de ces données ? Les analyses, liées aux applications actuelles des objets connectés, permettent-elles de produire de l'information supplémentaire (corrélation, chronicité, prédiction, etc) ? Comment peut-on visualiser ce type d'analyse ?

Le but de ce TER est donc d'analyser les données d'activités physiques et du sommeil collectées par les objets connectés. En plus de cette analyse, nous devons rendre compte de ces données, à l'utilisateur, à travers des visualisations intelligentes et dynamiques. Grâce à ces visualisations, il pourra se rendre compte de troubles qu'il peut avoir durant son sommeil et du lien que cela peut avoir avec ses activités physiques ou son quotidien.

Bien que nous survolions la partie analytique durant ce rapport, celle-ci est bien plus complète et détaillée dans les travaux de mon collaborateur.

Première partie

Etat de l'art

Chapitre 1

Visualisation

Sommaire

1.1	Introduction	4
1.2	Objectifs et qualités d'une visualisation	4
1.3	Types de visualisation	4
1.3.1	En fonction de la nature des données	5
1.3.2	En fonction de l'objectif de la visualisation	9
1.4	Visualisation et Big Data	16
1.5	Technologies pour la visualisation	17
1.5.1	Sans programmation	17
1.5.2	Avec programmation	17
1.6	Conclusion	18

1.1 Introduction

La représentation sous forme visuelle constitue l'un des meilleurs moyens pour explorer et comprendre un large ensemble de données. Elle permet de raconter une histoire à travers ces données traduites graphiquement. Il existe différentes manières de présenter les informations. Les supports diffèrent en fonction de l'importance de ces informations. Il s'agit non seulement d'un outil mais aussi d'un support de communication.

1.2 Objectifs et qualités d'une visualisation

Comme nous le disions précédemment, la visualisation de données nous permet de raconter une histoire. Il est nécessaire de toujours viser la vérité et de ne pas tomber dans une visualisation à la conclusion trompeuse. Pour cela, il suffit dans un premier temps de se poser une question appropriée suite à un examen approfondi des données. Ensuite, il faut déterminer le but du graphique. Enfin il ne reste plus qu'à créer le chef-d'œuvre qui permettra à n'importe quelle personne d'interpréter les résultats et d'y associer une histoire à la hauteur.

1.3 Types de visualisation

Afin de raconter une histoire correcte aux lecteurs de nos futurs graphiques, il est nécessaire d'énumérer les principaux graphiques existants, en fonction de leurs données, mais aussi, en fonction de l'objectif

que nous souhaitons donner à ces derniers.

1.3.1 En fonction de la nature des données

Commençons par énumérer les graphiques les plus généralistes tout en essayant d'être un minimum exhaustif. Examinons les différents types de graphiques à travers les prochaines sections.

Séries temporelles

Que ce soit dans les opinions publiques, les migrations de population ou même le développement d'entreprise, les séries temporelles sont omniprésentes. Le temps est la composante commune de toutes ces données. Les variables associées au temps permettent de mesurer les changements. Il existe deux types de données que sont : les données discrètes ou les données continues. A travers le temps, nous cherchons donc à accumuler des données, à créer une vision d'ensemble et ainsi trouver des modèles.

Le graphique à barres (pouvant être utilisé dans d'autres catégories que temporelle), est un des graphiques les plus répandus. En abscisse, nous avons le temps, et en ordonnée, une valeur associée (une variable quelconque). Dans ce type de graphique, il est conseillé de débiter l'axe des ordonnées à 0 afin d'être capable de comparer chacune des barres entre elles. De façon générale, il faut toujours afficher les libellés à chacun des axes du graphique, accompagnés de leurs valeurs et sans oublier l'intitulé. On peut colorer les différentes barres du graphique afin d'ajouter une dimension.

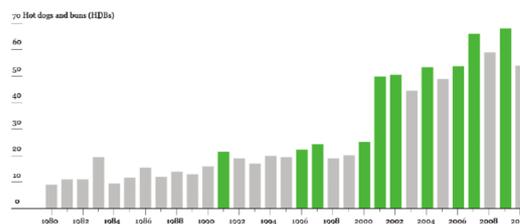


FIGURE 1.1 – Diagramme à barres colorées

Le graphique à barres empilées peut être aussi utilisé dans d'autres catégories. Il est pratique de l'utiliser dans les cas où il existe des sous-catégories de données afin de les comparer aisément.

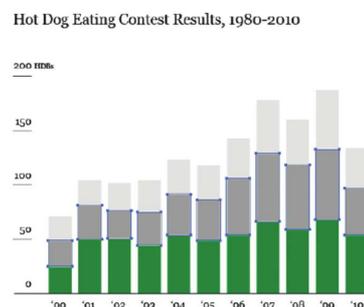


FIGURE 1.2 – Diagramme à barres empilées

Le graphique à nuage de points est pratique dans le cas où nous avons besoin d'une vue plus claire. Il donne une impression de continuité.

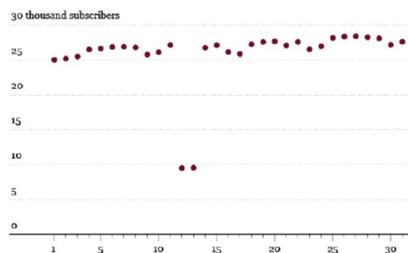


FIGURE 1.3 – Diagramme à nuage de points

Le graphique en escalier donne lui aussi une impression de continuité. Il relie les points discrets entre eux et crée un graphique continu. Dans le cas de l'augmentation du prix des timbres, il permet de constater que la croissance n'est pas continue mais bel et bien ponctuée d'augmentations discrètes.

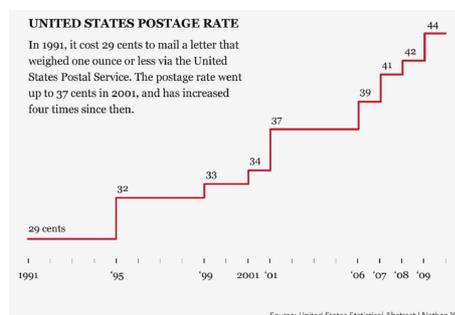


FIGURE 1.4 – Diagramme en escaliers

Lorsque nous possédons énormément de données discrètes ou des courbes très fluctuantes, certaines méthodes nous permettent de les visualiser de façon continue. Tel est le but de la méthode LOESS (Local regrESSion) de William Cleveland et Susan Devlin.

La dimension du temps faisant partie intégrante de nos vies, ces courbes appartiennent aux graphiques les plus intuitifs pour la vision humaine. Les résultats sont donc faciles à interpréter. Elles permettent de rapidement observer les défauts, et ainsi, de se questionner sur leurs origines.

Distribution

Les données d'une distribution sont assemblées en groupes, sous-groupes, catégories, population, etc... A travers une distribution, nous pouvons facilement trouver un maximum et un minimum, et ce, de façon plus rapide quand les données sont classées dans un ordre croissant.

L'un des plus connus du genre est le graphique en camembert de William Playfair créée en 1801. Nous disposons d'un disque coupé en plusieurs portions représentant un tout. Ainsi, la somme des parties représentent 100% du camembert final. L'un des défauts de ces graphiques est le manque de précision à côté des graphiques à barres. En effet, il est compliqué de comparer les portions entre elles, car il est difficile d'évaluer visuellement leurs angles, contrairement à la comparaison de barres voisines. Pour pallier ce problème, nous pouvons nous servir d'un dégradé de couleurs afin de mieux estimer l'importance de chaque proportion.

Le graphique en anneau est son petit cousin avec un trou central supplémentaire. Celui-ci nous oblige à n'utiliser que peu de proportions puisque nous ne pouvons plus les comparer avec leurs angles.

Le graphique à barres empilées permet de comparer plus facilement les proportions. Nous pouvons même ajouter une dimension interactive aux barres en affichant leur proportion au survol.

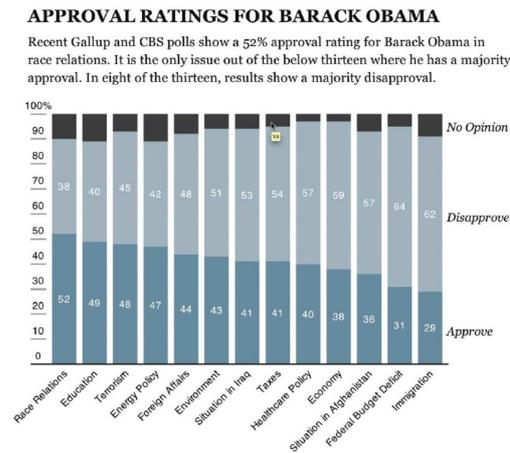


FIGURE 1.5 – Diagramme à barres empilées avec proportions

Un arbre peut être aussi transcrit en un graphique à proportions appelé Treemap. En 1990, Ben Shneiderman souhaite avoir un visuel sur l'occupation de la mémoire de son disque, et par ce biais, pense à une nouvelle visualisation. Le visuel est une surface composée de rectangles représentant une mesure. Ces mêmes rectangles peuvent être aussi composés de sous-rectangles (une sous-mesure du rectangle parent). Ainsi, ce type de graphique est destiné à une structure hiérarchique.

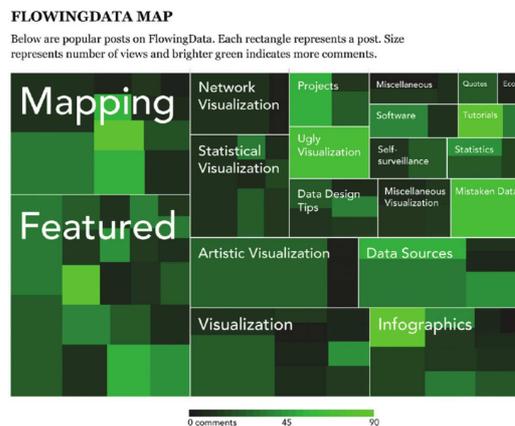


FIGURE 1.6 – Treemap

Le diagramme à couches empilées est un diagramme utilisant le temps comme abscisse. Prenons l'exemple des groupes d'âges en France de 1900 à 2000 et observons la proportion que prennent, à travers le temps, les différentes catégories d'âge. A chaque année donnée, nous obtenons une distribution. Nous pouvons facilement observer le progrès de la médecine, et le vieillissement de la population.

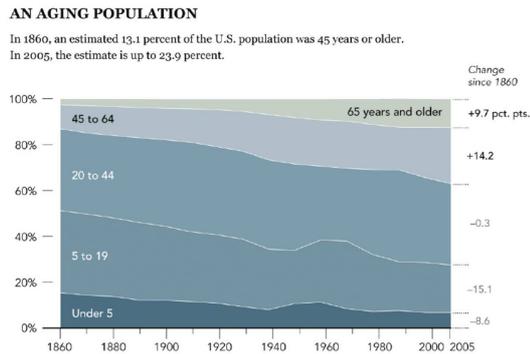


FIGURE 1.7 – Diagramme à couches empilées

Lorsque le nombre de couches devient trop important pour ce graphe, la lecture en devient compliquée. La solution consiste à ajouter de l'interactivité, en proposant l'affichage unique d'une catégorie, en ajustant l'axe des ordonnées pour donner un effet de "zoom".

La différence entre les diagrammes de distribution et les autres est la représentation des parties en un tout. Chaque valeur individuelle a une signification particulière. Il faut retenir, que s'il n'y a besoin d'afficher que quelques catégories, il faut privilégier le diagramme en camembert. S'il y en a plusieurs, il faut alors privilégier un graphique à barres empilées, ou mieux, un graphique en couches pour un effet continu. Ainsi, ce genre de diagramme permet de mieux se rendre compte des proportions. Avant toute conception, il faut toujours se demander ce que nous souhaitons retirer de nos données. Il faut aussi se demander si un graphique statique le permet ou s'il y a besoin d'ajouter de l'interactivité (sans que cela ne devienne complexe).

Données spatiales

Les données spatiales permettent l'utilisation d'un nouveau type de visualisation très intuitive. Il permet de voir les données collectées sur une version réduite du monde réel. A la place de nos habituelles abscisses et ordonnées, nous travaillons avec une latitude et une longitude. Il peut être aussi intéressant d'introduire la dimension du temps sur ce genre de graphique. Pour ce faire, il suffit de le dupliquer en plusieurs exemplaires et d'associer à chacune des copies, un temps précis.

Un premier exemple de carte utile consiste à placer des points uniformes (comme des punaises) à travers l'espace. Ainsi, au premier coup d'œil, on peut se rendre compte de quelques caractéristiques tel qu'un regroupement ou une dispersion des points. Il faut prêter attention à ne jamais trop agrandir une zone de la carte (pour se rendre compte des détails), dans le risque de perdre des informations plus discrètes et hors de la zone où nous nous concentrons.

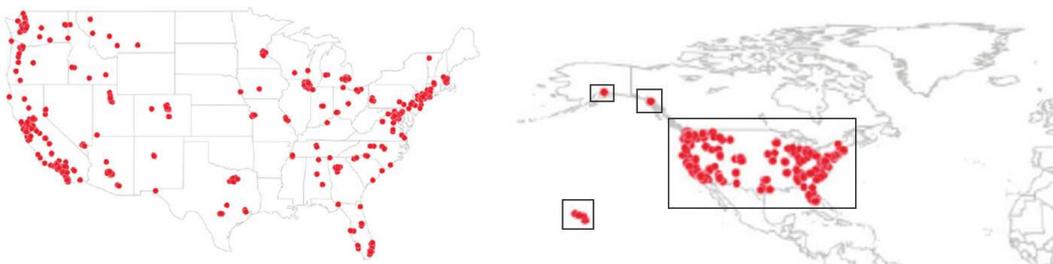


FIGURE 1.8 – Vue des Etats-Unis avec et sans agrandissement

Différentes variantes de cartes existent telles que les cartes avec tracés, représentant des connexions entre points (ce qui peut rappeler des flux de migrations ou encore des lignes aériennes). Une autre variante de la carte à points existe et consiste à donner une épaisseur à chaque point. Il s'agit des cartes à bulles.

Nous pouvons aussi créer des cartes colorées par région. Il s'agit de cartes permettant de représenter divers caractères de chaque région du monde avec une certaine nuance de couleurs. Ces informations sont d'ordre quantitatif ou qualitatif (chômage au niveau d'un pays, la fertilité mondiale...). Toutefois, il faut veiller à ne pas trop nuancer les couleurs afin de ne pas perdre le lecteur. Pour cela, il suffit de limiter le nombre de classes, par exemple, au nombre de quartiles.

Ainsi, à partir de ces cartes colorées, nous pouvons réaliser des matrices de cartes pour observer une évolution au fil du temps. Par exemple, nous pouvons observer le taux variable du chômage d'un pays par rapport à ses différentes régions.

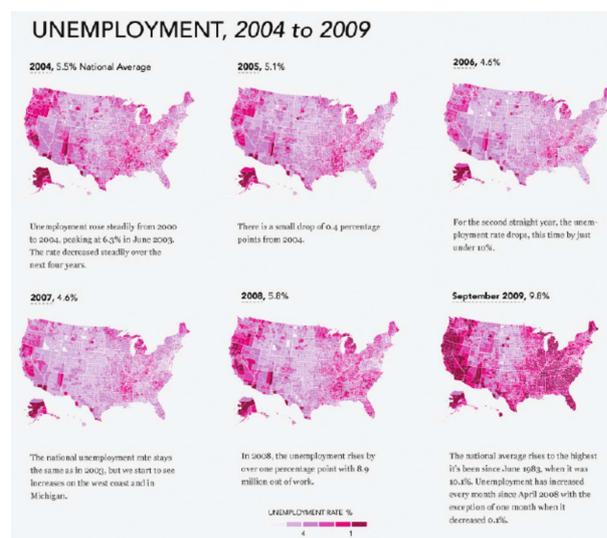


FIGURE 1.9 – Chômage des Etats-Unis de 2004 à 2009

Il n'est pas indispensable de représenter plusieurs graphiques pour montrer l'évolution d'une population. En effet, il suffit d'établir la différence de nuances de couleurs entre deux cartes pour observer une évolution.

Un dernier type de carte vise à ajouter une dimension interactive. Il s'agit d'un moyen judicieux et divertissant d'afficher l'évolution de données. Reprenons l'exemple du chômage, il pourrait être intéressant de n'utiliser qu'une seule carte défilant d'une année à l'autre de façon continue.

L'aspect géographique lié à la visualisation à travers une carte est plus délicate. En revanche, elle peut être très intuitive et gratifiante (dans le sens où elle apporte de nouvelles informations) pour la représentation de données et une exploration approfondie.

1.3.2 En fonction de l'objectif de la visualisation

Jusqu'ici, nos relations temporelles ou proportionnelles nous permettent de trouver aisément le minimum, le maximum et quelques intermédiaires. Le but est maintenant de trouver des relations entre nos différentes variables. On doit donc observer des relations croissantes ou décroissantes, linéaires ou non, entre nos variables. On doit aussi examiner si les données sont espacées, regroupées, ou si des aberrations sont constatées. Nous allons observer qu'il existe des moyens pour comparer plusieurs distributions

à travers une vue d'ensemble.

Visualiser des relations

En général, les statistiques s'efforcent de trouver des relations qui unissent les données, et ainsi, de détecter des corrélations. Par exemple, si la taille d'une population augmente, il en est très certainement de même pour le poids moyen. Il s'agit d'une corrélation simple. Plus il y a un grand nombre de facteurs, plus la recherche de corrélations est complexe. La recherche de corrélations ne donne pas systématiquement des modèles linéaires.

Une corrélation signifie, dans la définition d'un graphe, qu'une chose a tendance à fluctuer en fonction d'une autre. Par exemple, le prix du lait a tendance à augmenter lorsque le prix de l'essence augmente. Il s'agit donc ici d'une corrélation et non d'une causalité. En effet, si le prix du lait augmente, y a-t-il réellement un lien direct avec le prix du transport ou alors un tout autre facteur ? C'est pourquoi, il est difficile de prouver une causalité à travers un visuel.

Tout d'abord, voyons un nouveau type de graphique. Il s'agit du graphique à nuages de points. A chaque lien entre deux variables, il faut poser un point sur le graphique. Ainsi, nous obtenons un "nuage de points" qui reste assez peu évocateur. C'est pourquoi nous devons tracer une courbe de LOESS afin d'y voir plus clair sur la relation entre les deux variables.

La partie intéressante se situe au moment où nous comparons l'ensemble de nos variables entre elles. Pour ce faire, nous allons utiliser une matrice de graphiques à nuages de points. Elle permet de comparer toutes les paires de variables en un "clin d'œil" et de clarifier les données collectées.

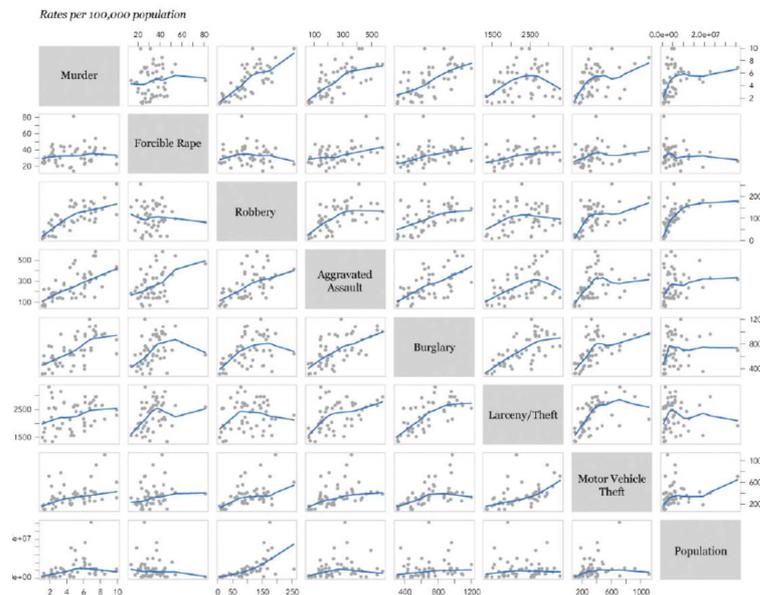


FIGURE 1.10 – Matrice de diagrammes à nuages de points

Un dérivé de ce graphe utilise un graphique à bulles. A l'aide de ses disques (bulles) dont l'aire est proportionnelle à la valeur d'une troisième variable, nous réussissons donc à comparer une composante supplémentaire.

Introduisons maintenant le diagramme à tiges et à feuilles. Il s'agit de l'ancêtre de l'histogramme. Il permet d'avoir un rapide aperçu sur la répartition de nos données. Par exemple, la répartition du taux de naissances à travers le monde avec des ensembles allant de 2 en 2 (8-10%, 10-12%, ..., 50-52%). A

chaque donnée récupérée, nous plaçons sa première décimale à droite des ensembles.

```

8 | 2371334468999
10 | 01223455566999001222334555777889
12 | 00011111356789993789
14 | 0034566788991237
16 | 227779123677889
18 | 00233677888900448
20 | 0024445688912455679
22 | 0057834579
24 | 11456677771347
26 | 31335667
28 | 014999
30 | 124234
32 | 1449069
34 | 556049
36 | 8890
38 | 023455823468
40 | 23125
42 | 699
44 | 17
46 | 252
48 |
50 |
52 | 5

```

FIGURE 1.11 – Diagramme à tiges

Par exemple, s'il y a un taux de naissance de 8,2% en France, nous allons ajouter 2 à droite de l'ensemble tronqué à 8. Si le Japon possède un taux de naissance de 11,0%, alors nous ajoutons 0 à droite de l'ensemble tronqué à 10. Ainsi de suite, nous obtenons notre diagramme rapidement, et nous remarquons un histogramme basculé à 45 degré. Nous observons la répartition des naissances à partir des différentes classes d'ensemble.

Après définition du diagramme à tiges, il est facile de comprendre le principe d'un histogramme ou diagramme de distribution. Il s'agit du diagramme à barres traditionnel avec, à la place du temps, une distribution dans différents intervalles. Comme dans le précédent diagramme, la hauteur (la largeur précédemment) permet de se rendre compte de la fréquence de chaque classe. La précision de la médiane peut être une information supplémentaire et utile. On peut varier le nombre de classes du diagramme en fonction de nos besoins : s'il faut plus ou moins de clarté ou de précisions.

Pour accompagner un histogramme, nous pouvons tracer un diagramme de densité. Il s'agit d'une courbe linéaire qui à chaque point en abscisse correspond à une probabilité en ordonnée. L'aire sous sa courbe doit être égale à 1. Ainsi, l'association du tracé de densité et des barres de distribution permet de comparer aisément le maximum, le minimum et la médiane.

Détaillons maintenant le comportement d'une matrice d'histogrammes. Elle permet d'avoir un rapide aperçu d'une certaine répartition de variables, par exemple, à travers le temps. Ci-dessous, nous avons la répartition du taux de natalité à travers le monde de 1960 à nos jours.

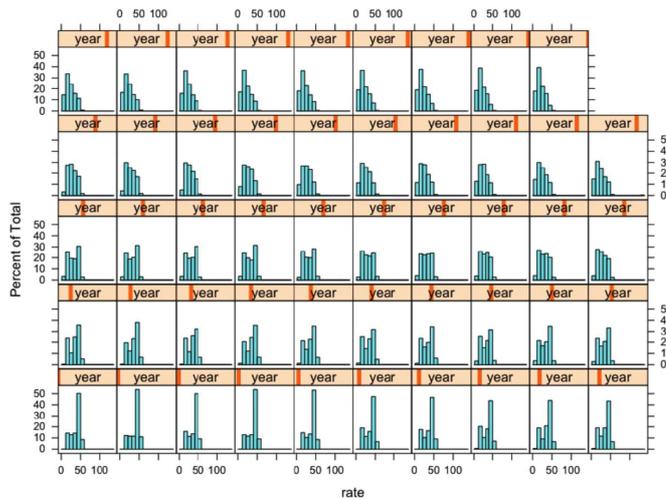


FIGURE 1.12 – Histogramme de la répartition du taux de natalité à travers le monde de 1960 à nos jours

Pour finir, la recherche de corrélations dans une diversité de données demande une plus grande réflexion et de nouveaux diagrammes nous rendant capable de les analyser. Elle se révèle être gratifiante du fait qu'elle nous apporte beaucoup d'informations. Nous pouvons chercher des modèles à travers ces distributions, et ainsi, donner des explications probables en rapport au contexte.

Identifier des différences

Jusque là, la comparaison de quelques variables est assez aisée. Qu'en est-il s'il faut en comparer une centaine ? Et comment remarquer la moindre différence, similitude ou encore relation ?

Notre première solution consistera à utiliser des cartes chaudes. A travers ces cartes, nous montrerons les données des différentes variables dans leur totalité. A la place de valeurs numériques, nous aurons des couleurs. Ainsi, plus la valeur est élevée, plus la case correspondante sera foncée. Une astuce supplémentaire concernant la visualisation est de pré-trier la carte en fonction de l'une des variables (une variable clé), afin d'avoir une meilleure vision d'ensemble.

Les visages de Chernoff permettent, grâce à l'œil humain, d'identifier rapidement les têtes les plus "attrayantes". En effet, chaque visage composant la palette de faces de Chernoff, associe plusieurs variables modifiant leurs traits. Cela permet d'associer facilement les têtes les plus communes, et remarquer les têtes sortant de l'ordinaire.

Les graphiques en étoile remplacent les visages multi-proportionnés de Chernoff par des étoiles auxquelles chaque branche est associée à une variable. L'essentiel est de ne pas surcharger l'étoile afin qu'elle reste lisible. Une petite variante existe et correspond au graphique de Nighingale (ou diagramme polaire). A la place d'une étoile, nous utilisons des arcs de cercle composant un cercle entier, avec des rayons et couleurs différents pour chaque variable. Les deux peuvent être utilisés sous forme de matrices dans le but de représenter une population.

Top 50 scorers during the 2008-2009 season

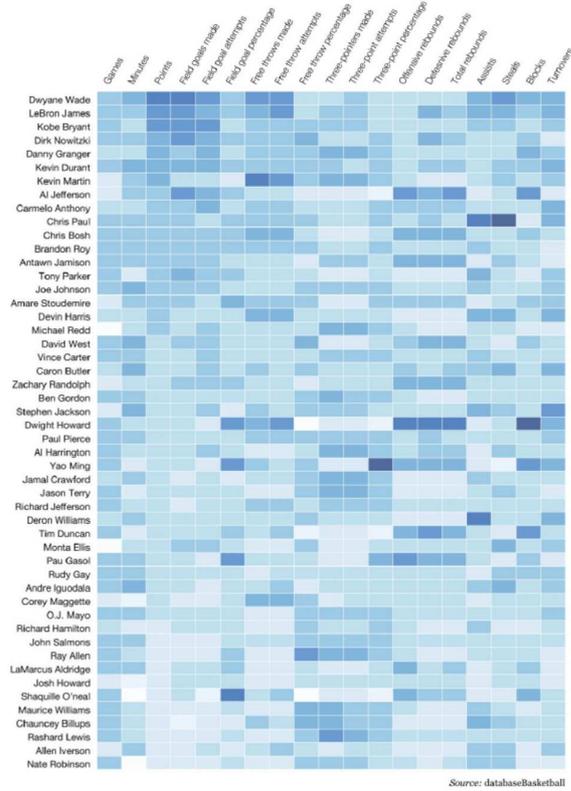


FIGURE 1.13 – Carte chaude

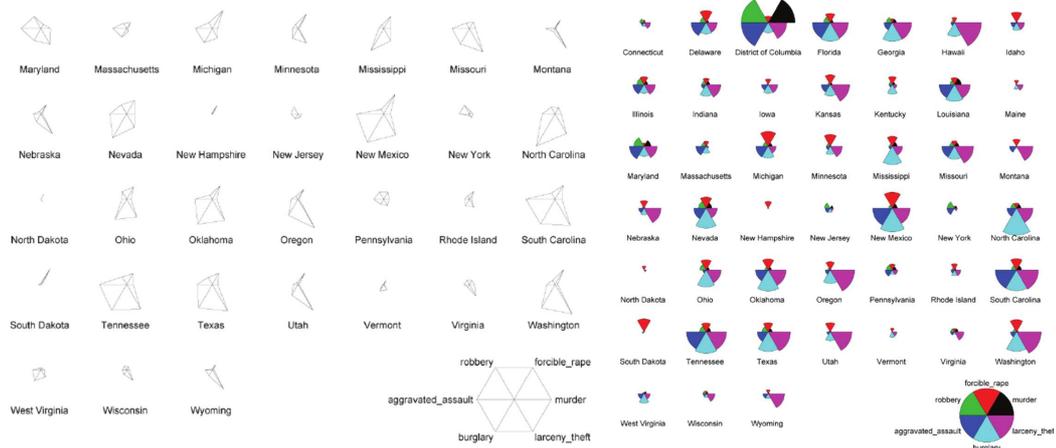


FIGURE 1.15 – Matrice d'étoiles et matrice de diagrammes polaires

La structure à coordonnées parallèles permet elle aussi de comparer plusieurs variables. On place plusieurs axes parallèlement et chacun d'entre eux représente une variable. Le haut de l'axe représente le maximum d'une variable et le bas son minimum.

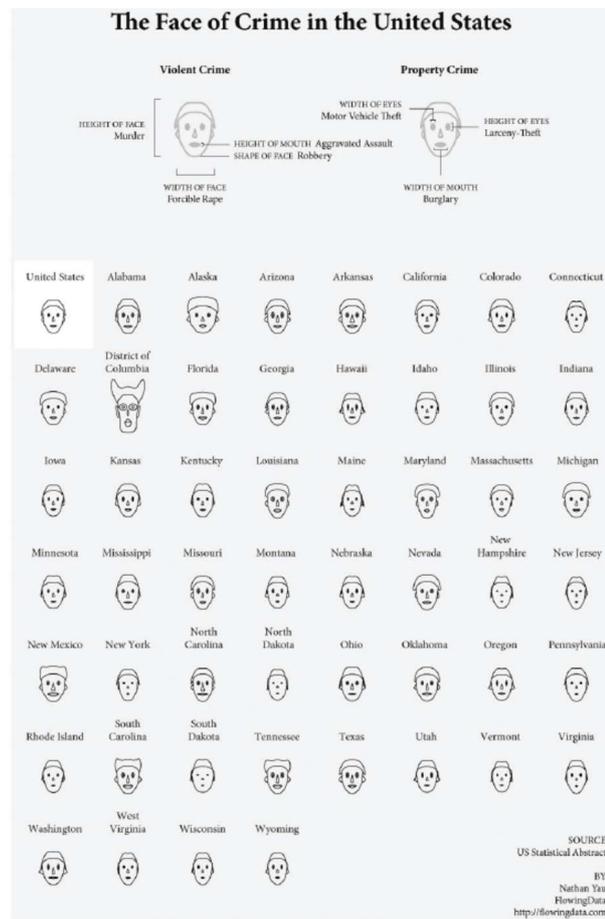


FIGURE 1.14 – Faces de Chernoff

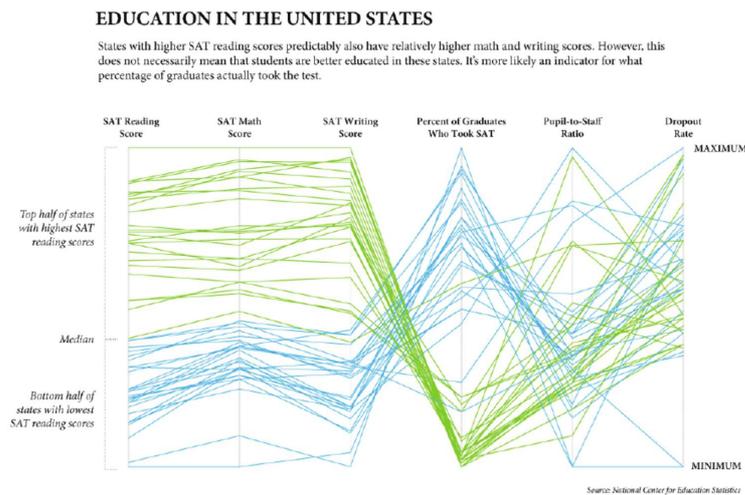


FIGURE 1.16 – Structure à coordonnées parallèles

Il existe une méthode permettant de réduire le nombre de dimensions : MDS (Multi-Dimensional-Scaling). Cette méthode permet, par exemple, de placer sur un segment, différents points avec des va-

riables fusionnées. Par exemple deux pays : l'un ayant une forte population et l'autre avec une faible population. Les deux pays vont donc se trouver opposés l'un à l'autre dans ce segment. Ainsi, si un troisième pays apparaît avec une population moyenne, il se retrouverait entre ces deux premiers pays. Ajoutons une troisième variable représentant le PIB. Le pays faiblement peuplé possède un petit PIB et un fort PIB pour le grand pays. Il se trouve que le troisième pays possède un assez fort PIB. Ainsi, ce troisième pays va être représenté beaucoup plus proche, dans le segment, du pays au fort PIB et à la forte population.

Selon une certaine fonction que l'on se fixe, nous pouvons ainsi ajouter autant de variables que nous le souhaitons, et ainsi réordonner toutes les capitales mondiales sur une carte en deux dimensions (comme nous le montre l'exemple ci-dessous).

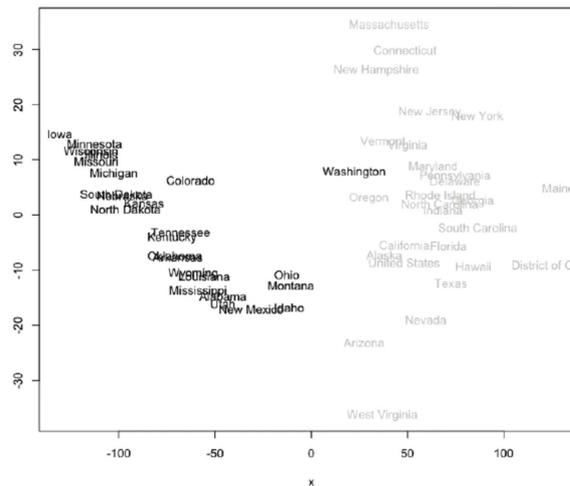


FIGURE 1.17 – Diagramme avec échelonnement multi-dimensionnel

Certains graphiques sont pratiques pour observer des données aberrantes. Ils nous permettent de trouver des points qui diffèrent du reste de la population. S'il ne s'agit pas d'erreurs de données, ces points peuvent être très intéressants à étudier. A l'aide des graphiques déjà présentés, nous pouvons déjà remarquer ce genre d'anomalies. Par exemple, cet histogramme représente une grande population dans l'une de ses classes, et à l'autre bout une petite population isolée, avec entre deux, le néant.

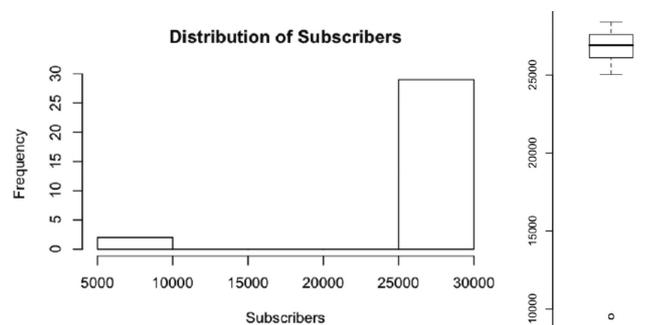


FIGURE 1.18 – Histogramme et diagramme à surface avec observations aberrantes

Ce genre de diagrammes représente une aide précieuse à la formulation de la question initiale. Ils permettent de mettre en évidence, soit des regroupements, des dispersions ou des aberrations permettant d'étudier avec une plus grande précision nos données.

1.4 Visualisation et Big Data

Le Big Data est comme son nom l'indique, une quantité massive de données. Les données sont si importantes qu'il en devient difficile de les travailler avec des outils classiques. Le support visuel peut être un outil permettant de les analyser et d'en avoir un aperçu général.

Il existe deux grands challenges dans la visualisation de Big Data. Il s'agit d'une part de la "scalabilité" (traduction directe de l'anglophone scalability), c'est à dire de l'adaptation au changement d'échelle et au maintien de ses fonctionnalités et, d'autre part, du dynamisme. Jusque là, nous travaillions avec une petite (voire moyenne) quantité de données statiques. Il s'agit maintenant de gérer de grande quantité d'informations et le tout de façon dynamique. En effet, ces données, en plus d'être conséquentes, peuvent être plus ou moins diversifiées et hétérogènes (pas forcément de structure fixe).

L'un des principaux caractères attendus dans ce domaine est la vitesse de réponse à une requête telle qu'un visuel sur l'ensemble de données. L'utilisation d'index efficaces est donc nécessaire pour les bases de données comportant de telles données mais peut s'avérer être une tâche complexe. Ensuite, nous avons besoin d'utiliser le cloud computing afin d'accéder à une quantité suffisante de puissance de calcul parallèle. Ainsi, les différents problèmes liés au Big Data sont départagés en tâches indépendantes et pouvant être exécutées en concurrence. L'utilisation de méthodes réduisant le nombre de dimensions peut être aussi un moyen d'accélérer le processus. Mais qui dit une réduction de dimensions, dit aussi une perte d'informations et peut être la perte d'un modèle, d'une corrélation ou d'une aberration intéressante.

Du côté visuel, quelques problèmes demeurent. En effet, l'utilisateur peut éprouver des difficultés à distinguer l'ensemble des objets présentés dans le graphique. En plus des capacités limitées de la perception humaine, nous pouvons aussi avoir une limite au niveau de la résolution de l'écran. De même, pour contrer ce problème, nous pouvons réduire l'information, au risque de perdre des informations importantes. En somme, il est compliqué d'afficher communément, beaucoup d'informations et beaucoup de dimensions.

Quelques solutions pour pallier ces problèmes existent. En premier lieu, nous pouvons augmenter considérablement la puissance matérielle par l'utilisation de la parallélisation que nous propose le cloud computing. L'analytique peut être aussi un outil indéniable. Grâce à l'analyse préliminaire des données que l'on manipule, on peut en réduire considérablement la manipulation. Il faut aussi s'assurer de la qualité des données. On peut afficher plusieurs petites quantités de données au lieu d'un ensemble incompréhensible. Enfin, nous pouvons supprimer les données isolées et les placer dans des graphiques à part.

Quelques progrès ont eu lieu dans ce domaine. La visualisation doit avant toute chose, donner une pré-visualisation pour ensuite agrandir, filtrer et afficher les détails à la demande. Elle doit facilement permettre l'apparition de modèles et de tendances. Le développement de plateformes telles que Hadoop est un réel progrès dans le cloud computing et la gestion du big data. En effet, elle permet la distribution parallèle de processus et la visualisation en temps réel de données issues du big data.

Il existe différentes visualisations optimisées pour l'affichage de données issues de ce large domaine. Certaines ont déjà été présentées :

- Treemap ;
- Circle Packing est équivalente à Treemap mais est représentée par des cercles hiérarchiques ;
- Sunburst est elle aussi sous forme de cercles pour des données hiérarchiques. Le centre du cercle représente la racine et plus on s'éloigne, plus nous nous enfonçons dans la hiérarchie. On ne parle plus de longueurs et de largeurs, mais d'angles et de rayons ;
- Parallel Coordinates ;
- StreamGraph est un équivalent du diagramme à couches empilées avec un axe central supplémentaire ;
- Circular Network Diagram où les données sont placées autour d'un cercle et sont liées par des

tracés incurvés. Ces tracés sont plus ou moins foncés et épais afin d'exprimer la force de leur relation.

Method name	Large data volume	Data variety	Data dynamics	Method name	Big data class
Treemap	+	-	-	Treemap	Can be applied only to hierarchical data
Circle packing	+	-	-	Circle packing	Can be applied only to hierarchical data
Sunburst	+	-	+	Sunburst	Volume + Velocity
Parallel coordinates	+	+	+	Parallel coordinates	Volume + Velocity + Variety
Streamgraph	+	-	+	Streamgraph	Volume + Velocity
Circular network diagram	+	+	-	Circular network diagram	Volume + Variety

FIGURE 1.19 – Propriété des méthodes de visualisation

Les outils traditionnels pour visualiser le big data ne sont pas adaptés. Parfois, ce dernier peut représenter jusqu'à des zetta-octets de données (un milliard de terra-octets). En définitive, pour une meilleure approche de ce monde rempli de données, nous avons l'utilisation du cloud et de sa programmation parallèle. Nous avons aussi l'utilisation de design évolutif avec une approche par réduction. Avec ceci, nous avons des visuels interactifs et intuitifs permettant la découverte aisée de modèles. La visualisation en direct semble impossible sans un minimum d'analyse préliminaire des données. Le domaine de l'analytique possède un rôle important dans la démystification de ce genre de données.

Un récent domaine permettrait même d'immerger les sens humains de part sa technologie. Il s'agit de la réalité virtuelle. En effet, cet aspect, encore peu exploré, permettrait une visualisation multi-dimensionnelle et une immersion totale de l'individu. Il permettrait la visite de dimensions graphiques encore peu étudiées. Cela engendrerait une meilleure et bien plus intuitive perception des données que le traditionnel bureau d'ordinateur.

1.5 Technologies pour la visualisation

Il existe une grande diversité de moyens pour afficher correctement des données sous la forme des graphes. Ces moyens sont, au choix, avec ou sans programmation. Notons qu'un minimum de pratique en programmation est nécessaire pour parvenir à des visualisations de données plus fines et intuitives.

1.5.1 Sans programmation

Le plus connu et le plus accessible de tous est incontestablement Excel. Il permet, entre autres, avec peu de moyens, de créer un premier graphique. Il génère un rapide aperçu, mais celui-ci ne peut pas être personnalisé à souhait et n'affiche pas toujours la précision d'information souhaitée. Dans le même rang, il existe aussi Google Sheet. Il s'agit d'une version en ligne d'Excel avec, en plus des fonctions de base, une fonction de partage. Il permet aussi de créer quelques graphes animés.

Toujours dans le même ordre d'idée, il existe Many Eyes de chez IBM. De même, il permet une rapide pré-visualisation. Comme son nom le suggère, il permet de créer des graphes visibles par "plusieurs yeux". En effet, il permet d'analyser collectivement les graphes mis à disposition publiquement (toute information chargée sur l'outil appartient au domaine publique), afin de détecter des caractéristiques particulières, et ce, plus rapidement qu'avec une simple paire d'yeux. Mais comme tout outil sans programmation, il ne permet pas une grande flexibilité graphique.

1.5.2 Avec programmation

Il existe différents langages de programmation pour construire un graphique. Tout d'abord, commençons par Python. Grâce aux bibliothèques NumPy et SciPy, ce langage est capable de créer quelques

graphiques. Ces derniers ne sont pas très raffinés ni flexibles à souhait. Dans le même genre, nous avons PHP, qui permet d'être facilement couplé à une base de données SQL. Cela implique qu'il n'y ait plus besoin de manipuler de CSV.

Processing est un langage de programmation open-source orienté graphique et visualisation. Il permet de faire des graphiques flexibles et plutôt sophistiqués. De plus, celui-ci est disponible en java et en JS. Un autre et très connu logiciel de programmation statistique est R. En effet, il s'agit de l'un des logiciels les plus utilisés par les statisticiens. Il possède beaucoup de packages composés d'innombrables graphiques. En revanche, le rendu des graphiques doit souvent être retouché à l'aide de logiciels annexes tel qu'Illustrator, afin d'embellir quelques aspects vieillissants du logiciel. Il ne permet pas non plus de créer des graphiques interactifs et n'est pas adapté au web dynamique.

Nous arrivons maintenant à des langages permettant de dynamiser nos graphiques. Il s'agit, dans un premier temps, de Flash et Action Script. Ils sont très puissants et permettent de modéliser quasiment à volonté ce que l'on souhaite. Mais aujourd'hui, ces langages tendent à être remplacés par la récente révolution de l'HTML en sa version 5, CSS3 et un langage qui lie le tout : le JavaScript. De plus, ces langages sont bien mieux intégrés au web d'aujourd'hui. A l'aide de la bibliothèque D3, gratuite et en open source, il est possible d'accéder à de nombreux modèles pré-existants. Elle peut aussi créer ses propres modèles graphiques. Ainsi, tout graphique est interprété côté client, permettant de ne pas surcharger le serveur. Il existe quelques autres bibliothèques telles que jQuery Sparkline, JavaScript InfoVis Toolkit, Google Charts API, etc.

1.6 Conclusion

Les visualisations peuvent être statiques ou dynamiques. Les visualisations interactives conduisent souvent à un meilleur travail de découverte que les outils statiques. Ces outils interactifs peuvent aussi aider à une excellente prise de connaissance du Big Data. Pour en arriver à ces résultats, il convient d'utiliser les bons outils. Le Web semble faciliter le processus scientifique. En effet, le Web permet d'obtenir des données dynamiques en un temps opportun et de maintenir les visualisations à jour. Les méthodes pour visualiser le Big Data sont en constantes évolutions. La réalité virtuelle semble être une issue convenable à la recherche de motifs particuliers. Il ne faut surtout pas perdre de vue qu'un graphique doit, avant toute chose, raconter une histoire et ne surtout pas induire en erreur.

Ainsi, la visualisation possède de nombreux domaines d'application compte tenu de son éventail de possibilités graphiques. Les données liées aux objets connectés font parties des domaines qui concentrent énormément d'informations. Serait-il possible d'associer ce domaine à la visualisation afin de mieux interpréter cette quantité astronomique de données ? Pourrait-on trouver plus aisément des relations entre chacune des composantes qui nous entourent ?

Chapitre 2

Visualisation dans les applications de suivi d'activités physiques et du sommeil

Sommaire

2.1	Introduction	19
2.2	Histoire	19
2.3	Données et suivi d'activités physiques et du sommeil	21
2.3.1	Données environnementales et spatiales	21
2.3.2	Données corporelles	22
2.4	Données et visualisation pour les activités physiques et du sommeil	23
2.4.1	Visualisation officielle	23
2.4.2	Visualisation non-officielle	23
2.5	Limites	26
2.5.1	Limites matérielles	26
2.5.2	Limites de l'analyse	27
2.6	Conclusion	28

2.1 Introduction

Nous vivons dans un monde rempli de capteurs. De notre smartphone, en passant par les détecteurs composants nos maisons (sensibles à la fumée et certains gaz), mais aussi dans tout nouvel appareil high-tech. Leur capacité s'accroît et se diversifie constamment. A travers eux, nous possédons comme des "super-pouvoirs" capables de ressentir ce que nos cinq sens sont incapables de maîtriser. Ils nous permettent de mieux appréhender notre environnement, mais aussi d'agréments notre connaissance de soi. Cet agrément se traduit par une quantité phénoménale d'informations. N'étant doté que d'une simple vision humaine, existe-t-il un moyen d'afficher les informations liées au suivi d'activités physiques et du sommeil et d'en comprendre l'ensemble? Peut-on corrélérer ces informations? Avant toute chose, commençons par un bref historique des objets connectés.

2.2 Histoire

Depuis très longtemps, l'Homme tente de réduire les objets les plus complexes en un format le plus minime possible. C'est le cas aujourd'hui avec les objets connectés (appelés aussi "wearables")

et les capteurs qu'ils embarquent. Afin de se rendre compte du processus de miniaturisation, voyons d'où sommes-nous partis pour en arriver à de telles avancées aujourd'hui. Ci-après, quelques exemples d'ancêtres de nos objets connectés.

Nous remontons, tout d'abord, 5 siècles auparavant et atteignons la période de l'Œuf de Nuremberg, créé par Peter Henlein. Nous sommes donc au XVI^{ème} siècle où nous trouvons un premier objet connecté, tout simplement, au temps. Il s'agit, en effet, de l'une des caractéristiques les plus importantes de ces objets. Cette horloge portable se porte autour du cou comme un collier. Elle manque légèrement de précision mais devient très populaire en 1580.

La première montre-bracelet apparaît en 1812. Elle est offerte par Abraham Louis Breguet (horloger et physicien français) à Caroline Bonaparte, la plus jeune sœur de Napoléon I^{er}. Dans un premier temps, seules les femmes appartenant à la noblesse peuvent l'acquérir. Puis, vient le tour des aviateurs et artilleurs d'équiper ce genre de montre à bracelet, pour finalement se répandre au grand public.

Nous arrivons maintenant dans le courant de la Première Guerre Mondiale. L'avancée en photographie, et l'allègement des appareils photos, permettent de créer un premier ancêtre organique des drones actuels. Il s'agit de la caméra pigeon. Le dispositif comprend une caméra attachée autour du cou d'un pigeon dressé dans le but d'espionner les lignes ennemies. Dans un temps où les avions, bâtis en bois, sont loin d'être furtifs et discrets, la caméra pigeon est la seule alternative à une surveillance au-delà du front.

Les premiers vêtements connectés apparaissent en 1961. Créée par Edward Thorpe et Claude Shannon, il s'agit d'une chaussure embarquant un dispositif permettant de tricher à la roulette (jeu de casino). Il permet de calculer des probabilités de tirage avec une précision de 44%. Les calculs sont ensuite transmis, par le biais d'un émetteur radio, à l'oreillette du fraudeur.

En 1972, Hamilton crée la première montre digitale, la Pulsar P1. Celle-ci n'est destinée qu'aux riches amateurs. Elle est constituée d'or et vendue à un tarif équivalent à 9800 euros actuels. Trois ans plus tard, cette même firme commercialise la Pulsar Calculator Watch, en ajoutant une fonction calculatrice à sa montre.

La Seiko UC 2000, produite en 1984, est la première montre "programmable" dans le sens où nous pouvons y enregistrer du texte. Pour ce faire, il suffit de connecter la montre à un module annexe permettant d'inscrire du texte par induction. Le module était accompagné d'une imprimante thermique.

Peu de temps auparavant, le constructeur japonais produit une montre embarquant un écran LCD de 1,25 pouces permettant de régler les fréquences hertziennes d'une télévision en s'y connectant.

Nous faisons maintenant un bond en avant d'une vingtaine d'année (2006) pour arriver aux chaussures connectées de Nike. Celles-ci, associées à l'Ipod de chez Apple, permettent d'accéder à diverses performances telles que la vitesse et distance d'entraînement.

Nous arrivons en 2008, avec la sortie du bracelet connecté Fitbit. Il est parmi les premiers à proposer un traqueur d'activités sans fil. Il calcule, lui aussi, un certain nombre de performances tels que les pas, la fréquence cardiaque ou même la qualité de sommeil. Ce dernier n'a pas tardé à se faire concurrencer, avec l'arrivée, par exemple, de Pebble et de son support sur Android Wear.

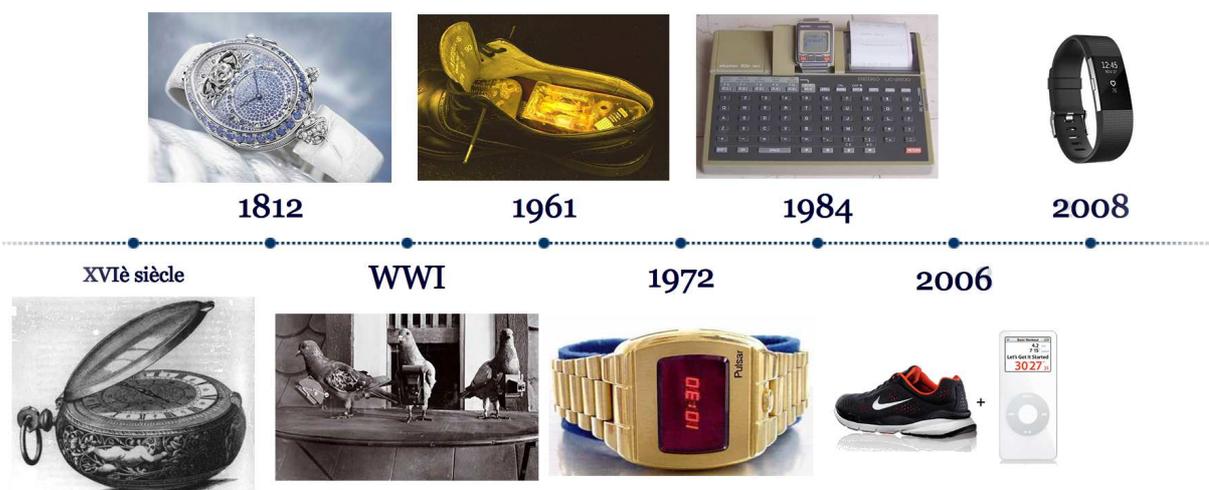


FIGURE 2.1 – Chronologie des objets connectés

Après toutes ces années d'évolution, nous pouvons nous demander : de quoi sont composés les objets connectés actuels ? Comment peuvent-ils obtenir tous ces renseignements sur nous et sur notre environnement ?

2.3 Données et suivi d'activités physiques et du sommeil

Comme nous l'avons déjà vu précédemment, il existe une multitude de capteurs. Ils sont de tout genre et permettent de prélever des informations appartenant à notre environnement ou même à notre état corporel. Nous allons observer, à travers certains objets connectés du marché, quels sont les différents types de données pouvant être collectés. Nous énumérerons uniquement les données utiles au suivi d'activités physiques et au suivi du sommeil.

2.3.1 Données environnementales et spatiales

Notre environnement est riche d'informations et peut être caractérisé par une multitude de mesures. A travers les capteurs composant les objets connectés, nous pouvons amasser abondamment ces données.

Voyons tout d'abord les capacités de stations météo connectées comme la station Hector de chez French Tech. La plupart des stations, dont Hector, sont composées des capteurs suivant :

- un thermomètre mesurant des températures oscillant souvent de -30 à 60 degré Celsius ;
- un baromètre captant la pression atmosphérique et pouvant prédire les tendances météo à venir ;
- un hygromètre permettant de capturer l'humidité ambiante.

Il existe aussi d'autres objets connectés que sont les capteurs et purificateurs d'air ou de pollution. Les exemples sont nombreux. Nous pouvons notamment citer Koto, Footbot, Clarity ou encore Lapka PEM, composé pour la plupart de capteurs de (en omettant ceux précédemment cités) :

- COV évaluant la présence de composés organiques volatiles ;
- luminosité ou d'ultra-violet ;
- monoxyde et dioxyde de carbone ;
- particules fines présentes dans l'air ;
- radiation ou plus communément appelé compteur Geiger, captant les particules radio-actives ;
- matières organiques détectant des quantités importantes de nitrate dans nos aliments (dues à l'utilisation d'engrais synthétiques) ;

- champs électromagnétiques pouvant être causés par des appareils électroniques, émetteurs sans fil, des lignes électriques, etc ;
- ammoniac.

Les smartphones, bracelets connectés et autres caméras connectées (comme la Fisheye) sont aussi composés d'une multitude de capteurs tels que :

- capteur photo-sensible qu'est composé une caméra, qui dans l'exemple de Fisheye, permet de capter toute intrusion dans un domicile ;
- détecteur de proximité qui désactive l'écran d'un smartphone pendant un appel ;
- magnétomètre à trois axes servant de guide GPS, ou même de boussole ;
- gyromètre qui capte les mouvements ;
- accéléromètre permettant de capter une accélération linéaire ;
- altimètre qui est intimement lié au baromètre puisque tout deux se basent sur la pression atmosphérique.

2.3.2 Données corporelles

D'autres capteurs sont plutôt destinés à donner des informations liées aux caractéristiques corporelles.

Commençons par les objets les plus connus et les plus populaires du moment : les bracelets et montres connectés. Décrivons une montre connectée destinée aux seniors : la LifePlus. Celle-ci permet de prévenir un centre de télé-assistance en cas de signes vitaux irréguliers par le biais du réseau LORA. Elle est composée de :

- un capteur prélevant le rythme cardiaque ;
- un oxymètre permettant la mesure de la concentration d'oxygène dans les cellules.

Les moniteurs de sommeil sont aussi très en vogue. La plupart d'entre eux utilisent des capteurs environnementaux, tels que les capteurs de luminosité, de sons, de température, d'humidité, etc. Mais certains, comme Zeo ou Neuroon ont quelques particularités plus professionnelles comme :

- un capteur de type EEG (électroencéphalogramme), afin de suivre l'activité du cerveau pendant le sommeil
- un capteur de type EOG (électro-oculogramme) pour mesurer les mouvements des yeux pendant la nuit

Pour conclure cette partie, voici un exemple des capacités d'une application mobile, ApneaApp. L'application a été créée par un groupe de chercheurs d'une université à Washington. Le but de l'application est de capter la respiration de l'utilisateur afin de remarquer s'il ne fait pas d'apnée du sommeil. Cette anomalie est difficilement décelable et nécessite de séjourner à l'hôpital afin de procéder à la pose de capteurs spécifiques (EEG, capteur de flux aérien, capteur du rythme cardiaque...). La solution de l'université américaine a été de subtiliser le haut parleur, afin d'émettre des ultra-sons qui rebondissent sur le corps du patient et sont captés par le microphone du téléphone. Ainsi, tout mouvement respiratoire de l'individu est enregistré. L'application a fait l'objet d'une étude clinique prouvant son efficacité dans 95 à 99 % des cas.

A travers ces différents exemples, nous remarquons qu'il existe une multitude de capteurs permettant de collecter de très diverses données. Ces données nous permettent de quantifier les multiples qualités de notre environnement et nos signes vitaux. Ces données sont très utiles puisqu'elles nous permettent de mieux nous comprendre à travers notre environnement. Plus précisément, dans le cadre de cette étude, elles nous permettront de mieux cerner, par la visualisation, la corrélation entre nos activités physiques et notre sommeil.

2.4 Données et visualisation pour les activités physiques et du sommeil

Dans cette partie, nous allons surtout nous concentrer autour du bracelet Fitbit. Etant le leader du marché, il est celui le plus abouti et avec lequel le public a le plus développé les façons de visualiser les données.

2.4.1 Visualisation officielle

Dans un premier temps, voyons ce que propose le support officiel du bracelet Fitbit. La figure 2.2 correspond au tableau de bord de l'application. Le "dashboard" est composé de plusieurs visualisations que nous avons déjà pu rencontrer dans le premier chapitre.



FIGURE 2.2 – Dashboard de l'application web Fitbit

Tout d'abord, nous apercevons un diagramme à barres empilées. Les différentes classes représentent la même donnée, il s'agit juste de nuances d'intensité. Ce graphique peut informer l'utilisateur sur son nombre de pas, le nombre d'étages parcourus et la quantité de calories brûlées.

Ensuite, on aperçoit plusieurs graphiques représentés par des cercles partiels (équivalent au graphique en anneau) nous informant sur la complétion d'activités journalières. Ces activités sont diverses :

- le nombre de pas parcourus ;
- le distance parcourues ;
- le nombre d'étages montés ;
- le nombre de calories brûlées ;
- les moments d'activités et de sédentarités.

Nous voyons aussi des barres de progression nous informant sur la quantité de temps en activités intenses, moyennes et légères. Un dernier graphique représente une courbe temporelle en rapport avec la fluctuation du poids de la personne.

2.4.2 Visualisation non-officielle

Quelques utilisateurs de Fitbit se sont amusés à créer, à partir de leurs données exportées ou alors de l'API Fitbit, à créer de nouveaux graphiques plus informatifs et originaux.

Voyons une première personnalisation des diagrammes provenant des données de l'objet connecté.



FIGURE 2.3 – Dashboard amateur des données Fitbit

Nous apercevons à nouveau les demi-cercles représentant l'accomplissement du nombre de calories brûlées ainsi que les étages montés quotidiennement.

Du côté des nouveautés, nous avons l'apparition du camembert de Playfair. Celui-ci est représenté par trois classes liées par le nombre de minutes actives (forte, moyenne et faible activité).

Enfin, nous avons la visualisation des pas parcourus, des étages montés et des moments d'activités sous forme de diagrammes à barres et au fil des jours du mois. Le troisième diagramme à barres, plus précisément, empilées, représente le nombre de minutes actives en trois classes (comme le camembert) au fil des jours du mois. Ce diagramme est le plus informatif de ce tableau de bord.

Nous allons maintenant voir la figure 2.4 nous proposant une analyse plus fine des données Fitbit recueillies.



FIGURE 2.4 – Visualisation avancée des données Fitbit

Chacune des colonnes représentent, respectivement, une comparaison du nombre d'étages montés, du nombre de pas parcourus et des minutes d'activités. Ces colonnes sont en fonction des lignes, respectivement, pendant et hors des vacances, pendant les jours de semaine et du week-end, et enfin, durant les différents jours de la semaine. Ainsi, à travers ce tableau de graphes, nous pouvons, par exemple, déterminer que cet utilisateur parcourt plus de pas et monte plus d'étages pendant la semaine que le week-end. Chacun des graphes nous apporte une information complémentaire assez gratifiante pour l'utilisateur.

Il existe même des représentations visuelles plus créatives, plus originales et détachées de tout modèles présentés jusqu'ici.

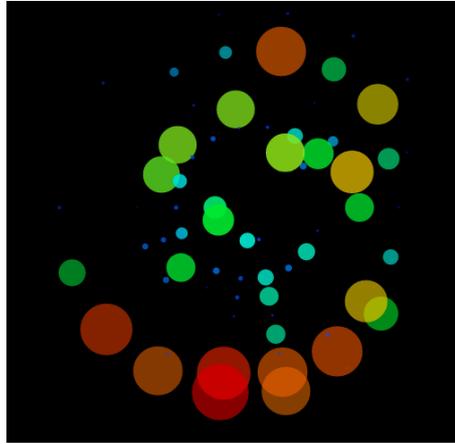


FIGURE 2.5 – Première visualisation artistique sur le nombre de pas parcourus

Il s'agit d'un visuel dynamique qui, toutes les 5 minutes, ajoute une bulle au milieu qui représente le nombre de pas parcourus durant ce laps de temps. Chaque nouvelle bulle pousse les anciennes vers l'extérieur. Les pas du début de journée apparaissent donc sur les extrémités de la figure. Si on ne note aucune activité durant cinq minutes, aucune bulle n'est affichée. Le rayon et la couleur de la bulle varie en fonction du nombre de pas.

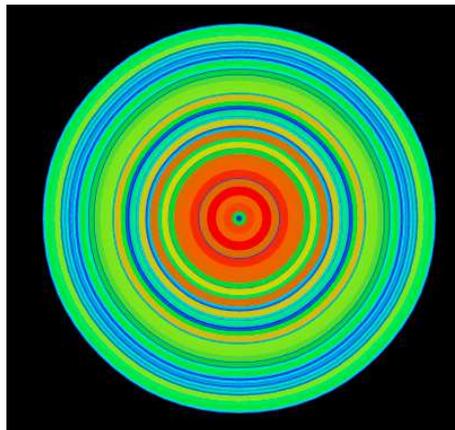


FIGURE 2.6 – Deuxième visualisation artistique sur le nombre de pas parcourus

A la même fréquence que le précédent graphique, de nouveaux cercles apparaissent autour du disque représentant le nombre de pas. Du bleu vers le rouge et une largeur de trait plus importante représente le nombre de pas pendant la même durée de cinq minutes. La taille totale du disque représente l'ensemble des pas de la journée.

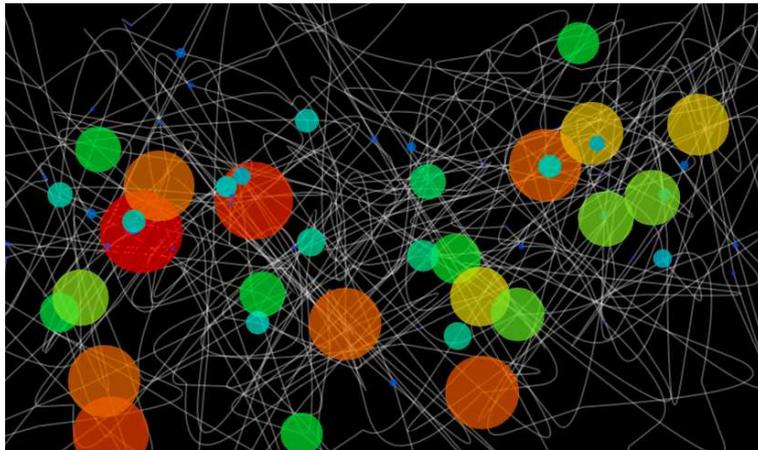


FIGURE 2.7 – Troisième visualisation artistique sur le nombre de pas parcourus

Ce dernier graphique représente toujours le même modèle à la même fréquence. La trait blanc représente la ligne du temps et est tracée aléatoirement. Toutes les cinq minutes, une bulle colorée apparaît avec les caractéristiques précédemment citées.

2.5 Limites

2.5.1 Limites matérielles

Dans cette partie, nous observons la qualité de précision des capteurs actuels, notamment ceux utilisés dans le domaine du grand public. Il est à savoir qu'il existe toute une panoplie de technologies de diverses qualités et précisions. A travers quelques exemples, nous allons estimer l'état de précision actuelle des capteurs nous environnant.

Commençons par l'étude de capteurs permettant l'appréhension de distance par laser. Pour sa voiture intelligente, Google a équipé son véhicule d'un capteur Lidar (permet d'émettre et de recevoir des ondes électromagnétiques) pour un prix avoisinant les 80 000 dollars. Il permet une grande autonomie du véhicule équipé. Moins onéreux et plus imprécis, il existe le Cruise RP-1 s'adaptant à n'importe quelle voiture, permet à celle-ci pour un prix huit fois moindre, de la rendre autonome en ligne droite sur autoroute. Un autre capteur, appelé Lidar-Lite, à un prix grand public, permet lui, une certaine autonomie de petits appareillages tels que les drones. Un dernier dispositif, lié au projet Tango de Google, permet à un smartphone d'analyser son environnement pour ensuite le retranscrire en un environnement 3D. La précision de ce dernier est loin derrière celle de la Google Car. Ceci est dû à la faible capacité de calcul d'un smartphone à côté des ordinateurs embarqués de la voiture autonome (recevant une quantité impressionnante de signaux à traiter). Ce que nous pouvons tirer de cette première analyse est, qu'il existe des capteurs tel que Lidar, capable d'une grande précision et applicable dans des domaines importants de la recherche, mais, qu'il existe aussi d'autres capteurs beaucoup moins onéreux mais perdant en précision. Ainsi, la miniaturisation et le rabais des coûts vont de pair avec l'efficacité des dispositifs.

Prenons maintenant l'exemple de Chemisense. Il s'agit d'un capteur chimique portable mis au point pour détecter une douzaine de produits toxiques. La startup a voulu la rendre optimale mais n'a pas réussi à atteindre la qualité d'analyse des stations de surveillance. En effet, s'il est contraint à analyser plusieurs produits chimiques simultanément, cela crée une dysfonction du capteur. La remarque est identique au précédent exemple ; la miniaturisation nuit, pour le moment, à la capacité de renseigner correctement l'état de l'environnement.

Venons-en, enfin, à l'exemple concernant les capteurs cardio-fréquence-mètre. Ceux-ci font partie intégrante des bracelets connectés et représentent un atout indéniable. Il a été testé cinq bracelets connectés de marques différentes et comparé à un électrocardiographe médical. Les résultats de cette étude ont été mitigés. Il apparaît quelques différences entre les prises les plus respectives possibles de l'appareil professionnel et celles des bracelets. En effet, l'un utilise la technologie optique (nécessite un sujet immobile) et l'autre est composé de capteurs sensibles aux impulsions électriques liés aux battements cardiaques. Ainsi les professionnels de santé préconisent ce genre d'appareillage uniquement dans le cadre privé et non médical. Dans le cas contraire, on se doit d'utiliser une sangle au niveau de la poitrine, bien plus sensible aux battements ou la traditionnelle prise de pouls au niveau du poignet.

2.5.2 Limites de l'analyse

Dans cette section, nous allons voir les limites que nous pouvons rencontrer actuellement avec le traitement des données. Notamment, nous pouvons nous rendre compte qu'au delà de relevés concernant le sommeil, nous n'avons pas les moyens de quantifier la qualité du sommeil. Nous devons être capable de créer des mesures objectives de qualité de sommeil à partir de ces données subjectives.

Principalement à travers les capteurs de mouvements, nous relevons diverses informations très intéressantes concernant l'activité du sommeil. Nous pouvons repérer les différents cycles et ainsi relever les minutes de sommeil, d'éveil et le nombre de réveils. Bien que ces données nous en apprennent beaucoup sur le sommeil, elles ne permettent pas de quantifier clairement la qualité du sommeil. Il existe différents moyens de rendre quantifiable la qualité du sommeil.

Il existe des questionnaires du type PSQI (indicateur de qualité du sommeil de Pittsburgh), permettant une interprétation assez fine des données. Sa qualité provient du questionnement de l'utilisateur sur son impression de qualité de sommeil. Par exemple, il peut demander si l'utilisateur a eu froid ou chaud pendant la nuit, s'il a cauchemardé, s'il a des douleurs en plus des données que nous avons précédemment citées, mais aussi des questions relatives au dernier mois de sommeil. De plus, il prend aussi en compte l'ingestion de certains médicaments et les problèmes liés à l'insomnie. Ce genre de questionnaire demande un effort supplémentaire à l'utilisateur de l'objet connecté.

D'autres indicateurs n'utilisent que des données numériques. Il s'agit notamment de l'indicateur PSG (polysomnographique). Afin de déterminer son indicateur, il utilise différentes variables :

- TST (total sleep time) temps total d'endormissement ;
- TWT (total wake time) temps total d'éveil ;
- TMT (total movement time) temps total d'agitation ;
- TRT (total recording time) temps total d'enregistrement :

$$TRT = TST + TWT + TMT$$

- SE (sleep efficiency) efficacité du sommeil :

$$SE = \frac{TST}{TRT}$$

La formule générale de PSG prend aussi en compte certains autres indicateurs comme :

- WASO (wake after sleep onset) réveil après endormissement ;
- Les différents stades de sommeil (phase 1, phase 2, sommeil paradoxal, etc)

Or ces indices ne sont pas recensés par les objets connectés actuels. Il est à ajouter que ces objets ne prennent pas en compte l'état de santé du sujet, comme son humeur ou son anxiété. Ou encore, les capteurs à l'état actuel, ne différencient pas un sommeil paradoxal (état de sommeil agité), d'un réveil du sujet.

Nous en concluons que les capteurs environnementaux "grands publics" sont à utiliser avec parcimonie et en connaissance de cause, du fait de l'imprécision des données collectées. A cette imprécision, vient s'ajouter un manque d'information nous empêchant d'utiliser un indice existant. Dans le cadre de

notre étude, le manque de précision des appareils n'a pas une grande incidence tant que ceux-ci donnent des résultats proportionnels (dans le sens où deux mesures d'un même pouls ne donnent pas des résultats différents). L'avenir de la miniaturisation des capteurs sensoriels a tout de même de belles années devant elle avec la diversification des dispositifs et la recherche d'informations de plus en plus fidèles. Comme la loi de Moore le conjecture, dans sa version populaire, la puissance de ces dispositifs grandira constamment au même titre que leur taille diminuera au fil du temps.

2.6 Conclusion

A travers ces différentes parties, nous avons vu que les objets connectés, liés aux activités physiques et du sommeil, peuvent recueillir énormément d'informations à travers les différents capteurs qu'ils embarquent. A travers ces données certains s'efforcent de trouver des visualisations capables de retracer le plus fidèlement possible les événements environnementaux et physiques. D'autres essaient d'extraire des informations supplémentaires, dérivées de premières analyses, et ce parfois même, de façon artistique. Ce qui nous amène donc à se demander : est-il possible de créer des visualisations capables de dévoiler des modèles issus des données extraites de ces capteurs ? Peut-on révéler des corrélations à travers toutes ces informations ?

Deuxième partie

Analyse

Chapitre 3

Analyse des données initiales et uniformisation

Sommaire

3.1	Introduction	30
3.2	Analyse de fichiers aux données hétérogènes	30
3.3	Mise en base de données	31
3.4	Enrichissement de la base de données	32
3.5	Conclusion	32

3.1 Introduction

Pour ce chapitre, nous allons essayer de mettre en oeuvre l'apprentissage précédent. Nous disposons de données d'activités physiques, corporelles et du sommeil. Il s'agit d'enregistrements retraçant plusieurs mois d'activités d'un sujet. Ces activités sont retranscrites à travers plusieurs fichiers CSV (ou XLS) que nous devons analyser.

3.2 Analyse de fichiers aux données hétérogènes

Nous disposons de données très hétérogènes que nous devons dans un premier temps "parser" et uniformiser. Ces données ont été recueillies à travers un dispositif Fitbit. L'analyse préliminaire de ces données consiste à donner un aspect homogène à cet ensemble de données afin de faciliter les analyses plus approfondies qui suivront.

Dans un premier temps, nous avons converti l'ensemble des fichiers en un seul et même format : CSV. Le format CSV (Comma-separated values) représente les données sous forme de tableau où chacune d'entre elles est séparée par un caractère et chaque ligne du tableau est représentée par une ligne du fichier CSV. Un autre problème d'hétérogénéité est lié à ce format. En effet, il n'existe pas de spécification formelle quant au choix du caractère de séparation. Ainsi, il nous faut aussi uniformiser les fichiers CSV entre eux. Notre choix s'est tourné vers le standard RFC 4180 et ainsi à l'utilisation du caractère ";" pour séparer les données. Nous avons utilisé des "régex" afin de remplacer les différents motifs. Une fois les fichiers CSV uniformisés, il nous a fallu séparer les grands types de données en : activités, sommeil et corps.

3.3 Mise en base de données

La mise en base de données est une étape importante puisqu'elle permettra par la suite d'accéder facilement à n'importe quelle donnée, présente initialement dans les fichiers CSV, à travers une simple requête SQL.

Nous avons choisi un SGBD (Système de Gestion de Base de Données) relationnel afin de manipuler efficacement ces données. Contrairement au paradigme NoSQL, les SGBD relationnels mettent à disposition des outils permettant de facilement manipuler nos informations. Par exemple, nous pouvons plus aisément les regrouper par date clé à l'aide d'un "group by" ou plus généralement, manipuler les dates plus facilement.

Notre choix initial s'est tourné vers MySQL. Il s'agissait d'une question de simplicité quant à l'import des fichiers CSV suite à la création des tables : activités, sommeil, corps et utilisateur.

L'utilisation d'un SGBD nous a permis de facilement mettre de côté les tuples non valides. Un tuple non valide correspond, la plupart du temps, à un défaut d'utilisation donnant des résultats nuls. Par exemple, nous avons considéré qu'un tuple de la table "activity" était invalide lorsque les minutes sédentaires étaient égales à 1440. En effet, cela correspondrait à une journée entière sans aucun mouvement de l'utilisateur. De même pour la table "sleep", les tuples comportant zéro minute de sommeil ont été invalidés.

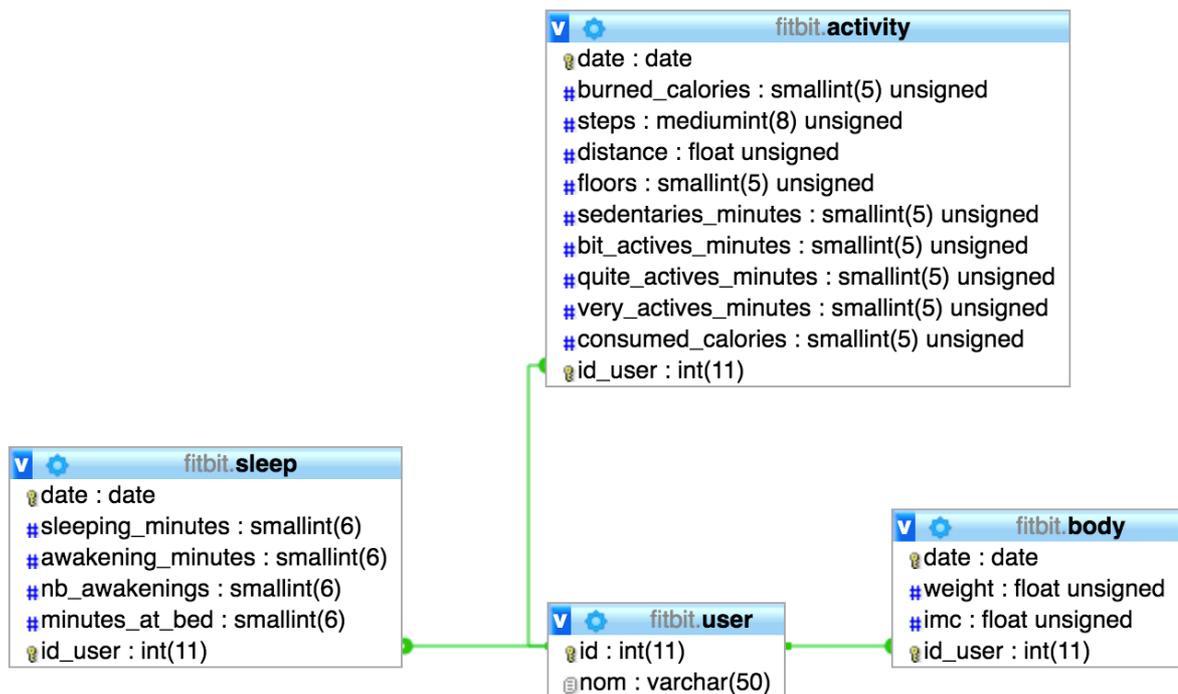
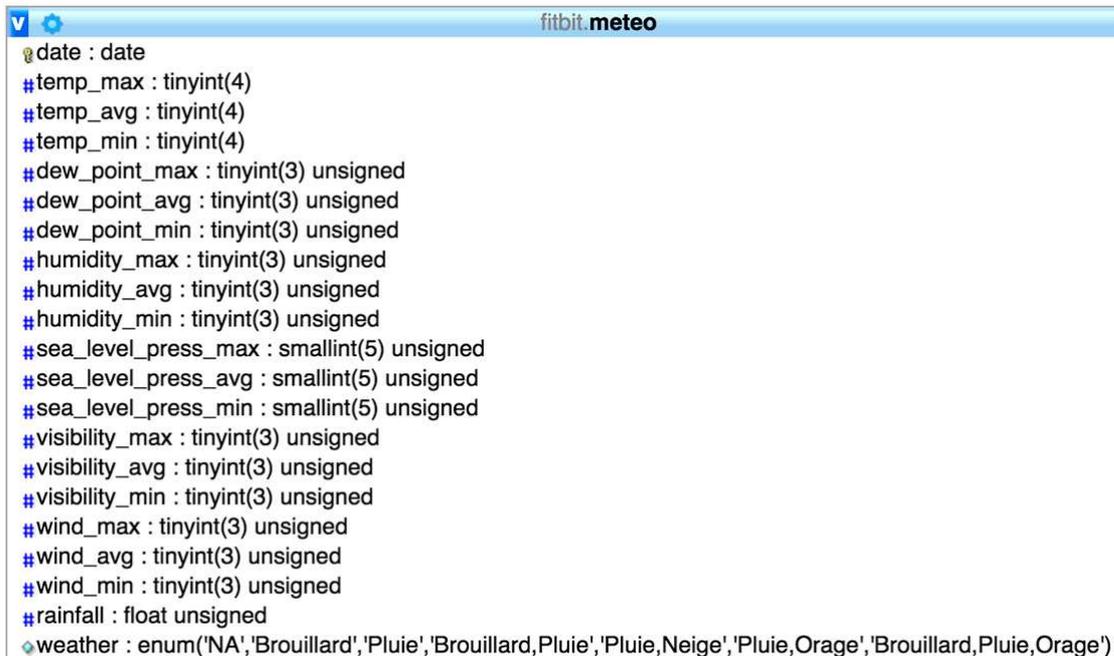


FIGURE 3.1 – Schéma de la base de données

Suite à ce choix et à la complétion de la base, nous nous sommes finalement redirigés vers une base PostgreSQL. Ce choix est dû à une utilisation particulière des triggers que nous n'avons pas pu appliquer en MySQL. En effet, par la suite nous utiliserons des indicateurs pour la qualité du sommeil. Un indicateur prend en compte un tuple mais aussi, l'ensemble des tuples existants. Ainsi, à chaque ajout de données de la part de l'utilisateur, le SGBD doit mettre à jour l'ensemble de ses tuples (une fois l'ensemble des tuples insérés).

3.4 Enrichissement de la base de données

Une des dernières étapes de la base de données fut l'ajout de données externes pouvant aider à l'obtention de corrélations par rapport au sommeil de l'utilisateur. Nous avons ajouté des données telles que la météorologie pouvant influencer directement avec le sommeil, mais aussi des données telles que le taux d'allergènes ou de polluants atmosphériques. Nous avons pu récupérer toutes ces données à l'aide d'API appartenant à "météociel" ou encore la pollution atmosphérique et les allergènes par le biais de <http://www.atmo-hdf.fr> et <http://ias.openhealth.fr>. Une API (Application Programming Interface) est une interface applicative permettant, à n'importe quel acteur autorisé, d'accéder aux données d'un site externe.



```
fitbit meteo
@date : date
#temp_max : tinyint(4)
#temp_avg : tinyint(4)
#temp_min : tinyint(4)
#dew_point_max : tinyint(3) unsigned
#dew_point_avg : tinyint(3) unsigned
#dew_point_min : tinyint(3) unsigned
#humidity_max : tinyint(3) unsigned
#humidity_avg : tinyint(3) unsigned
#humidity_min : tinyint(3) unsigned
#sea_level_press_max : smallint(5) unsigned
#sea_level_press_avg : smallint(5) unsigned
#sea_level_press_min : smallint(5) unsigned
#visibility_max : tinyint(3) unsigned
#visibility_avg : tinyint(3) unsigned
#visibility_min : tinyint(3) unsigned
#wind_max : tinyint(3) unsigned
#wind_avg : tinyint(3) unsigned
#wind_min : tinyint(3) unsigned
#rainfall : float unsigned
weather : enum('NA','Brouillard','Pluie','Brouillard,Pluie','Pluie,Neige','Pluie,Orage','Brouillard,Pluie,Orage')
```

FIGURE 3.2 – Schéma de la table météo

3.5 Conclusion

Suite à la transcription des fichiers initiaux en données homogènes dans une base de données cohérente, nous pouvons désormais les étudier avec sérénité. Cette étude sera accompagnée d'éléments externes. En effet, avec l'ajout d'une table retraçant les événements météorologiques, nous allons pouvoir étendre nos possibilités de recherche. Mais avant toutes recherches concernant nos données, il nous faut encore agrémenter notre table liée au sommeil. Jusque là, nous possédons les diverses caractéristiques de sommeil de l'utilisateur, mais comment comparer cet ensemble de caractéristiques aux autres variables que nous possédons ? Comment transformer ces caractéristiques afin d'obtenir une information pouvant quantifier la qualité du sommeil ?

Chapitre 4

Choix d'un indicateur pour la qualité du sommeil

Sommaire

4.1	Introduction	33
4.2	Tester l'efficacité d'un indicateur et ses corrélations	33
4.2.1	Définition d'une corrélation	33
4.2.2	Formule de Pearson	34
4.2.3	Formule de Spearman	35
4.2.4	Formule de Kendall	36
4.3	Tests d'un indicateur "sophistiqué"	37
4.4	Choix d'indicateurs plus basiques	38
4.5	Conclusion	38

4.1 Introduction

Le choix d'un indicateur est une étape cruciale dans la recherche de corrélations avec nos données initiales. C'est pourquoi, nous nous sommes concentrés sur la recherche d'un indicateur de qualité du sommeil pouvant réellement la quantifier.

4.2 Tester l'efficacité d'un indicateur et ses corrélations

Avant de nous aventurer dans le choix d'un indicateur nous permettant de quantifier la qualité du sommeil, nous avons dû trouver un moyen de vérifier si un indicateur joue bien son rôle. L'un de nos principaux outils est R. Il nous permet de comparer facilement un ensemble de variables entre elles afin de vérifier leur corrélation. Dans un premier temps, nous allons définir ce qu'est une corrélation et ensuite nous verrons trois formules nous permettant d'estimer les corrélations.

4.2.1 Définition d'une corrélation

Dans notre cas, l'étude d'une corrélation entre les variables X et Y, correspond à l'intensité de relation qui les unie. Pour ce faire, nous avons une formule qui nous permet de calculer un coefficient appartenant à $[-1;1]$. Ainsi, plus le coefficient approche les extrémités de l'ensemble, plus la corrélation est forte.

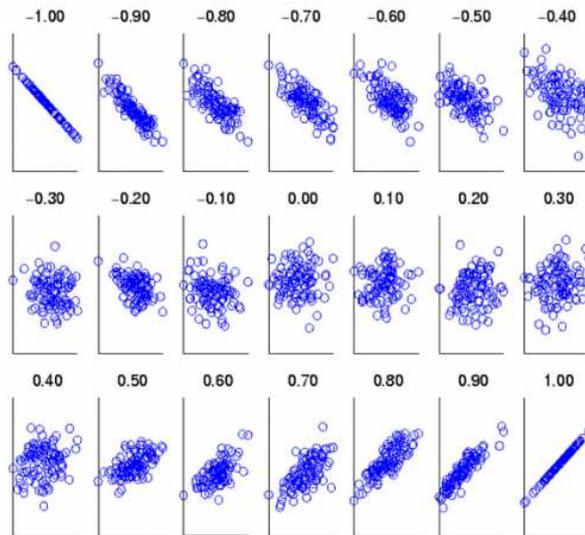


FIGURE 4.1 – Relation entre X et Y sous forme de nuages de points

Imaginons, à travers l'exemple ci-dessus, que X est en abscisse et Y en ordonnée. Nous observons donc que :

- plus X est grand et Y devient petit, plus le coefficient se rapproche de -1 ;
- plus X est grand et Y devient grand, plus le coefficient se rapproche de 1.

Les valeurs intermédiaires dépendent de la disparité des points dans le graphique. Maintenant, il nous faut calculer de façon efficace le coefficient de corrélation selon le type de données dont nous disposons.

Les graphiques présentés par la suite, pour donner un exemple d'application de nos formules, sont des corrélogrammes. Les nuages de points ne sont plus tracés et seul le coefficient de corrélation apparaît implicitement à travers une pastille de couleur. Plus la pastille est épaisse, plus la corrélation est forte. Le coefficient tend vers 1 quand la pastille devient bleu et vers -1 si celle-ci s'approche du rouge. Comme nous l'expliquions dans l'état de l'art, il s'agit d'une visualisation de données permettant un rapide aperçu pour l'œil humain.

4.2.2 Formule de Pearson

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

où

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N ((X_i - \bar{X}) \cdot (Y_i - \bar{Y}))$$

Cette formule permet de détecter la présence ou l'absence de relation linéaire entre les variables X et Y. Il existe une certaine limite au coefficient de Pearson. Si la relation que nous étudions possède une forme exponentielle, Pearson ne la considérera pas corrélée.

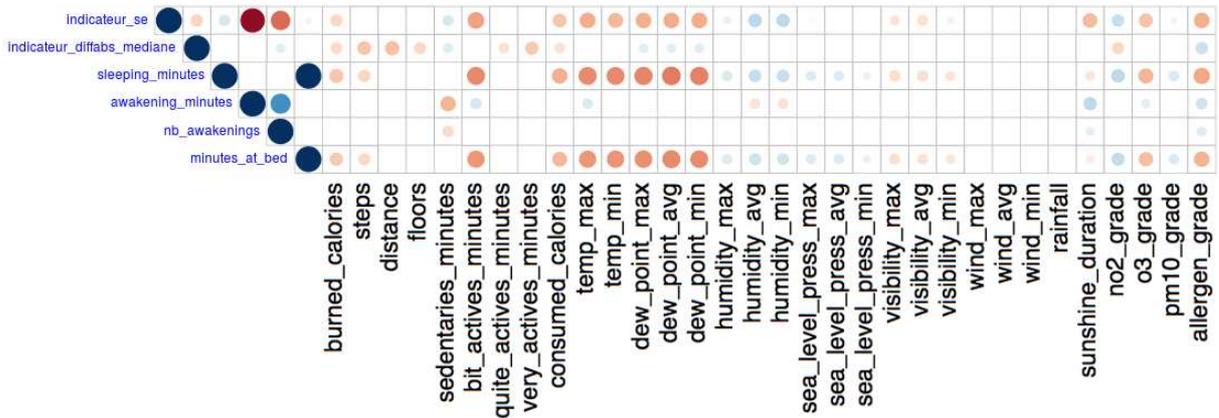


FIGURE 4.2 – Corrélogramme avec la formule de Pearson

Nous pouvons constater, à travers ce premier corrélogramme, l'apparition de corrélations entre les différents attributs de notre base de données.

4.2.3 Formule de Spearman

$$\rho(X, Y) = 1 - \frac{(6 \cdot \sum_{i=1}^N [r(X_i) - r(Y_i)]^2)}{N^3 - N}$$

où

$r(X_i)$ est le rang de la distribution $r(X_1), \dots, r(X_n)$

$r(Y_i)$ est le rang de la distribution $r(Y_1), \dots, r(Y_n)$

La corrélation de Spearman est utilisée lorsque nos variables X et Y semblent corrélées sans que la relation entre les deux variables soit de type affine. Pour obtenir une corrélation de Spearman parfaite (+1 ou -1), il faut que l'une des variables soit une fonction monotone parfaite de l'autre.

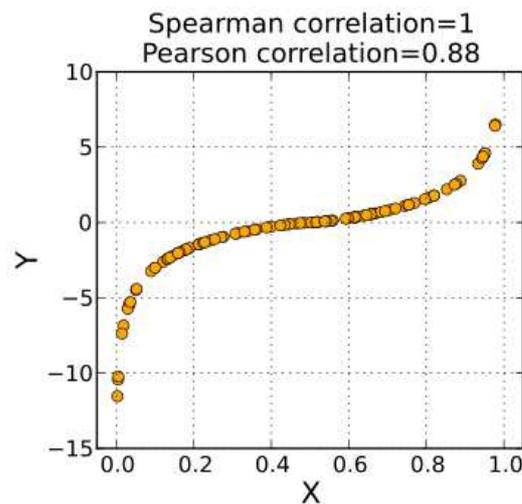


FIGURE 4.3 – Comparaison des coefficients de Spearman et Pearson pour une fonction non-affine et monotone

Ainsi, nous voyons que Spearman donne une relation parfaite pour cette relation non-linéaire alors que Pearson donne un résultat inférieur.

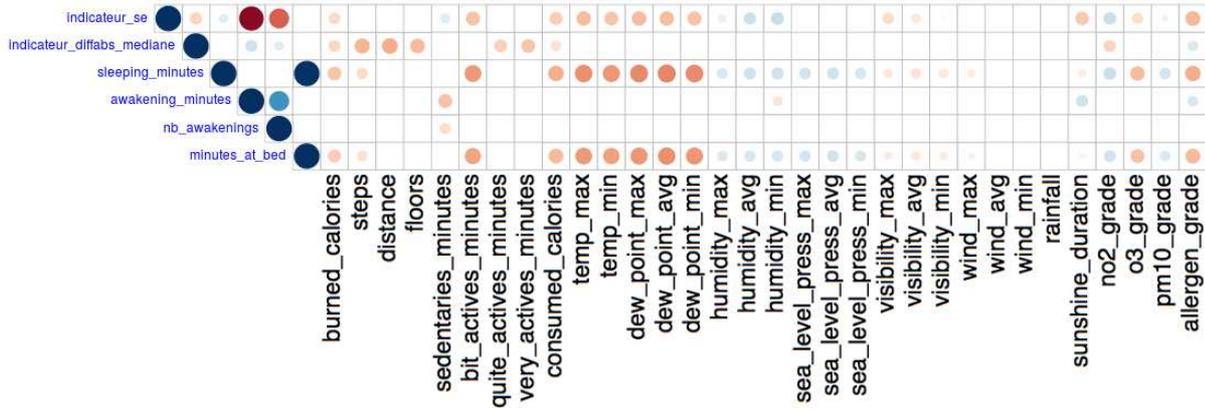


FIGURE 4.4 – Corrélogramme avec la formule de Spearman

A travers ce nouveau corrélogramme, nous remarquons que Spearman met en valeur les mêmes corrélations que le précédent graphique.

4.2.4 Formule de Kendall

Avant de donner la formule permettant de calculer le tau de Kendall, on doit définir deux notions utilisées dans la définition de la formule.

Soit $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ un ensemble d'observations des variables jointes X et Y tel que les valeurs des (x_i, y_i) sont uniques. Les paires d'observations (x_i, y_i) et (x_j, y_j) sont dites concordantes si $x_i \leq x_j$ et $y_i \leq y_j$ ou si $x_i \geq x_j$ et $y_i \geq y_j$. Les paires sont discordantes si $x_i \leq x_j$ et $y_i \geq y_j$ ou si $x_i \geq x_j$ et $y_i \leq y_j$. Dans le cas où $x_i = x_j$ ou $y_i = y_j$, la paire n'est ni concordante ni discordante.

Le tau de Kendall est défini de la manière suivante :

$$\tau = \frac{(\text{nombre de paires concordantes}) - (\text{nombre de paires discordantes})}{\frac{1}{2} \cdot n \cdot (n-1)}$$

Le tau de Kendall ne prend en compte, lui aussi, que des relations monotones.

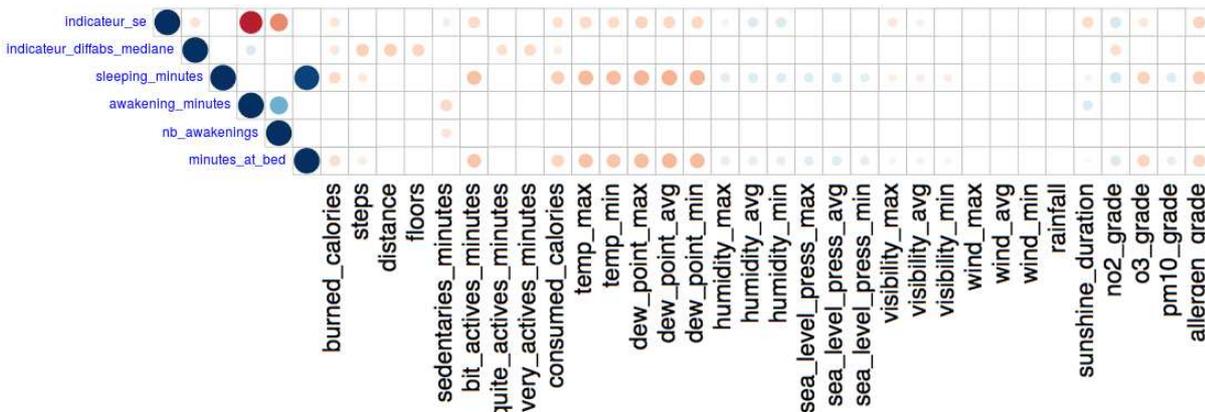


FIGURE 4.5 – Corrélogramme avec la formule de Kendall

Nous remarquons à nouveau que les corrélations mises en évidence par Pearson et Spearman sont

toujours présentes mais de façon moins intenses.

4.3 Tests d'un indicateur "sophistiqué"

Notre choix s'est tout d'abord orienté vers une formule reprenant l'ensemble des données liées au sommeil. Cette formule est scindée en deux parties :

évaluation du sommeil seul + *évaluation du sommeil par rapport à l'historique*

L'évaluation du sommeil seul se calcule en prenant en compte l'efficacité du sommeil :

$$\frac{\text{nombre de minutes endormi}}{\text{nombre de minutes passées au lit}}$$

Cette formule donne un résultat dans [0;1]. A ce résultat nous soustrayons ce qui peut être néfaste pour un sommeil tel que le nombre de réveils :

$$\frac{1}{10} * \text{nombre de réveils néfastes}$$

Où le nombre de réveils néfastes se calcul comme suis :

$$\text{nombre de réveils néfastes} = \max(0, \text{nombre de réveils} - 1)$$

Ainsi chaque réveil supplémentaire à 1, dit néfaste, retire $\frac{1}{10}$ du coefficient d'efficacité du sommeil. Le tout donne la première partie de notre formule, soit l'évaluation du sommeil en lui-même :

$$\max\left(0, \frac{\text{nombre de minutes endormis}}{\text{nombre de minutes passées au lit}} - \frac{1}{10} * \text{nombre de réveils néfastes}\right)$$

Nous arrivons maintenant au calcul de la partie prenant en compte les habitudes de l'utilisateur. Pour ce faire, nous allons calculer la médiane des minutes de sommeil de l'utilisateur. Ce calcul ne prend en compte que les sommeils de bonne qualité. Un sommeil de bonne qualité comprend au maximum un seul réveil et moins de 20 minutes d'éveil au total (selon la partie analytique de ce TER).

Le côté historique de la formule donne donc :

$$1 + \frac{1}{10} * \frac{\min(0, 60 - |\text{médiane}(\text{minutes de sommeil}) - \text{minutes de sommeil}|)}{30}$$

Ainsi, si l'utilisateur a un écart de plus d'une heure sur son sommeil habituel, celui-ci aura un sommeil de moins bonne qualité de l'ordre de $-\frac{1}{10}$ par demi-heure d'écart.

La formule générale donne donc :

$$\max\left(0, \frac{\text{ndme}}{\text{ndmpal}} - \frac{1}{10} * \text{ndrd}\right) + 1 + \frac{1}{10} * \frac{\min(0, 60 - |\text{med}(\text{mds}) - \text{mds}|)}{30}$$

Les analyses sur R n'ont pas donné de résultat très corrélés entre cet indicateur et les autres attributs liés à l'activité et à la météorologie. C'est pourquoi nous nous sommes redirigés par la suite sur des formules plus simples et plus corrélées.

4.4 Choix d'indicateurs plus basiques

Suite à ces premières observations, nous nous sommes restreints à des indicateurs aux formules plus simplistes. Nous avons utilisé un premier indicateur reprenant la formule de notre état de l'art :

$$\text{indicateur se} = \frac{\text{nombre de minutes endormis}}{\text{nombre de minutes passées au lit}}$$

Il s'agit de la formule d'efficacité du sommeil. Notre deuxième formule reprend l'historique de l'utilisateur :

$$\text{indicateur diffabs mediane} = |\text{médiane}(\text{minutes de sommeil}) - \text{minutes de sommeil}|$$

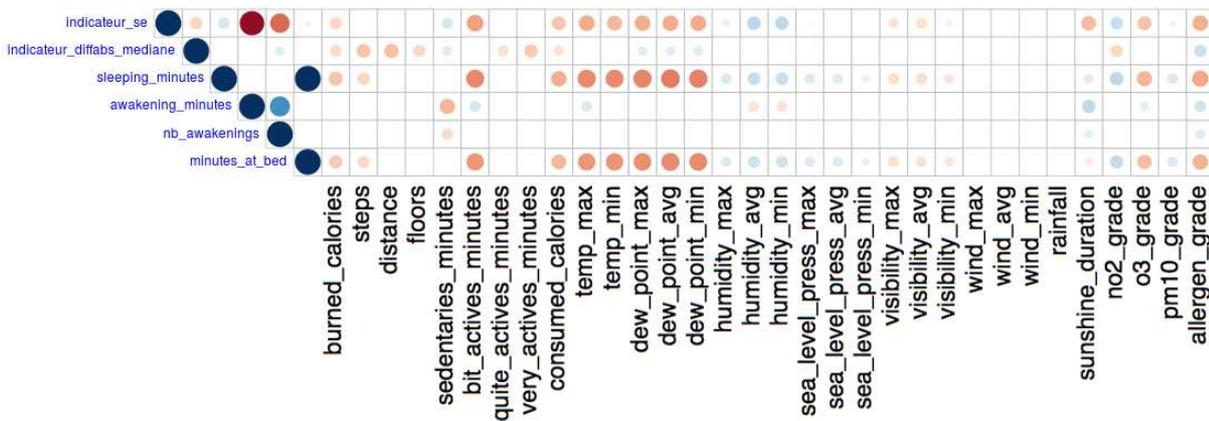


FIGURE 4.6 – Corrélogramme (Pearson) des indicateurs basiques et de nos données

Nous remarquons à travers ce corrélogramme, que nos indicateurs d'efficacité du sommeil (indicateur se) et celui en relation avec l'historique (indicateur diffabs mediane) sont corrélés avec les données d'activités physiques et météorologique. Par exemple, nous pouvons déterminer que, moins l'utilisateur passe de minutes de sédentarité dans la journée, plus son sommeil est efficace. Nous remarquons aussi que, moins l'utilisateur marche dans la journée, plus ses bonnes habitudes de sommeil changent.

4.5 Conclusion

Suite à la définition d'une corrélation et à l'étude de différentes formules permettant de la calculer, nous nous sommes attardés sur le choix d'un indicateur quantifiant le plus fidèlement possible la qualité du sommeil. En confrontant ce qui existe et ce que la partie analytique de ce TER a conclu d'un bon sommeil, nous avons produit trois indicateurs. Notre premier indicateur, le plus complexe, possède une formule reprenant la qualité du sommeil en lui même et son évaluation par rapport à l'historique de la personne. Cet indicateur, n'étant que très peu corrélé avec les données de sommeil et les autres données, nous avons décidé de le scinder en deux. Ainsi, nous nous sommes retrouvés avec l'indicateur de qualité du sommeil (sleep efficiency) et l'indicateur de la distance à la médiane séparés. Nous avons tout de même gardé le premier indicateur comme référence pour les tests de la partie suivante (bien qu'il soit peu corrélé, il possède tout de même un lien avec l'ensemble des variables).

Troisième partie

Visualisation et implémentation

Chapitre 5

Technologies utilisées

Sommaire

5.1	Introduction	40
5.2	Technologies pour la visualisation	40
5.2.1	Utilisation de R	40
5.2.2	Utilisation de javascript et D3.js	40
5.3	Technologies pour l'application web et mobile	41
5.3.1	Programmation côté client	41
5.3.2	Programmation côté serveur	41
5.4	Conclusion	42

5.1 Introduction

Afin de mener à bien la création de nos visualisations ainsi que leur implémentation dans un environnement applicatif, nous avons à choisir les plateformes, langages et technologies les plus appropriés.

5.2 Technologies pour la visualisation

5.2.1 Utilisation de R

Dans un premier temps, il nous faut prendre connaissance des données dont nous disposons. Afin d'allier un rapide aperçu et l'utilisation de notre base de données, il nous faut un langage de programmation capable de communiquer avec. En effet, nous aurions pu obtenir un rapide aperçu de nos données à l'aide d'un tableur, mais nous aurions été contraint d'utiliser des formats de données intermédiaires.

C'est pourquoi, nous nous sommes tournés vers le langage de programmation R. R permet de communiquer aisément avec notre SGBD. Il permet aussi de créer de rapides aperçus et parfois même de compléter notre base de données avec des informations supplémentaires. Ces informations supplémentaires sont diverses. Il s'agit, par exemple, de données issues d'une étude de corrélations. Ainsi, pour chaque utilisateur, nous contrôlons les données fortement corrélées entre elles.

5.2.2 Utilisation de javascript et D3.js

Tout ne peut pas être synthétisé avec un langage statique tel que R. En effet, R permet de grandes choses, il permet de créer des aperçus efficaces de nos données mais ne permet pas de les dynamiser.

L'utilisateur, peut parfois, souhaiter interagir avec les données. Cela peut se traduire par la modification d'un paramètre, au survol ou au clique permettant d'accéder à d'autres données. Les issues sont multiples et seule l'imagination fait barrage. L'une des meilleures manières d'explorer ces issues est l'utilisation d'un langage permettant une interaction simple de l'utilisateur dans un environnement qui lui est familier. Le tout devra communiquer avec notre base de données.

Notre choix s'est tourné vers D3.js. Il s'agit d'une bibliothèque graphique Javascript datant de 2011. Il permet de créer des graphiques en SVG à partir de données numériques. Le format SVG (Scalable Vector Graphics) est un format de données pour décrire des éléments graphiques vectoriels. Ce format permet donc des agrandissements et rétrécissements d'image infinis contrairement aux formats statiques tels que jpg et png. Son environnement de programmation, le javascript, permet une interaction aisée avec l'utilisateur.

5.3 Technologies pour l'application web et mobile

5.3.1 Programmation côté client

Ainsi, l'environnement de programmation côté client imposé par D3.js est le Javascript. Il s'agit d'un langage de programmation de script. Outre le fait qu'il serve de support à D3.js, il est aussi très utile afin de capter les interactions de l'utilisateur avec le DOM HTML. Il permet, par exemple, de capter le survol d'une division du DOM ou encore de l'un des éléments graphiques appartenant à un objet SVG. Nous pouvons aussi capter des événement tel que le clique (depuis une souris ou une surface tactile) ou encore le défilement d'une page.

HTML, CSS et Bootstrap

Comme il a été dit précédemment, nous nous sommes penchés vers la programmation Web pour développer un environnement simple et intuitif pour l'utilisateur. Il est très pratique de développer des applications Web car celles-ci sont lisibles sur une grande diversité de supports. Ces supports vont du plus commun des ordinateurs à tout objet mobile doté d'un navigateur. Il s'agit de supports allant du téléphone portable (à la petite résolution) jusqu'au tablette (à la résolution quasi égale à celle de nos écrans d'ordinateur), en passant par des résolutions intermédiaires très diverses. L'utilisation du Web peut être un choix judicieux et à la fois contraignant de part cet aspect.

Il ne suffit donc pas seulement d'utiliser de l'HTML et CSS classique. Il nous faut un moyen de convenir à tout type de résolution. C'est pourquoi nous avons opté pour l'utilisation de Bootstrap. Il s'agit d'un framework HTML et CSS. En somme, il s'agit de classes CSS pré-faites, utilisant les Media Queries en CSS3, afin de subvenir à quatre grandes catégories de résolution (smartphone et tablette en portrait et paysage et les écrans d'ordinateur).

5.3.2 Programmation côté serveur

Du côté du serveur, nous avons dû choisir un langage permettant l'interaction entre les requêtes de l'utilisateur et la base de données. Pour cela, il existe une panoplie de langages permettant tous de créer une interaction fluide. Nous avons choisi le langage PHP accompagné d'un MVC (Model View Controller) afin de gérer séparément ces différents aspects de la requête.

5.4 Conclusion

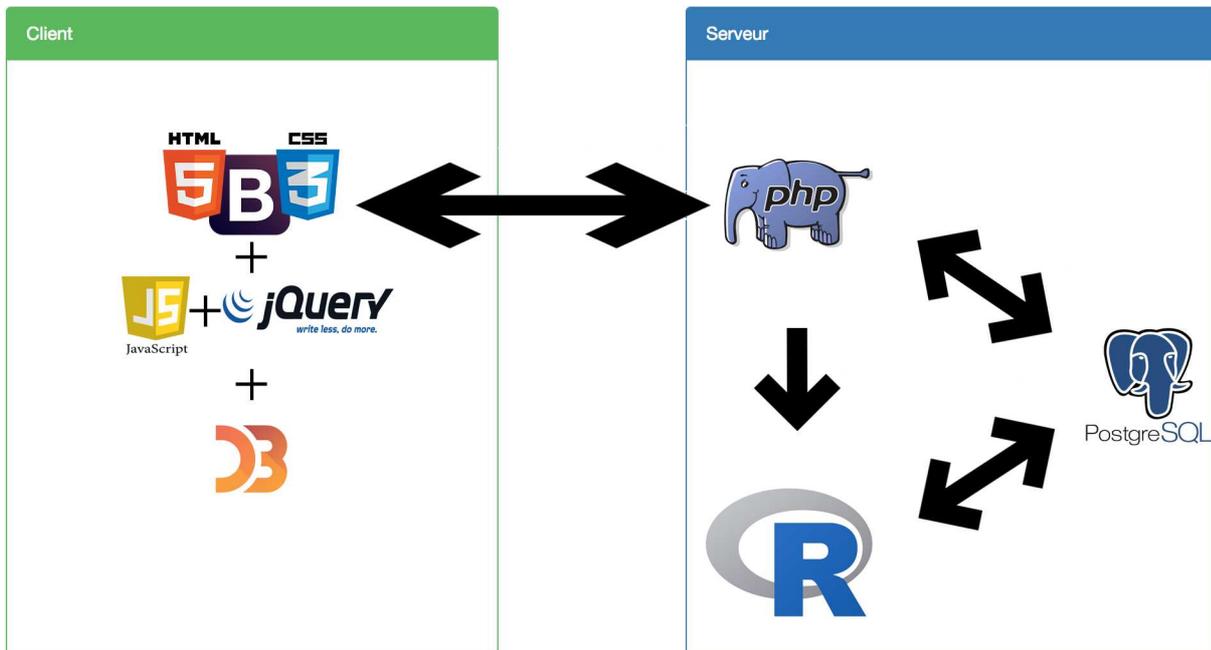


FIGURE 5.1 – Schéma des communications client-serveur

Comme le montre ce schéma, l'ensemble des technologies s'harmonise de cette façon :

- côté client :
 - les traditionnels HTML et CSS à la base de tout site, renforcé par l'utilisation de Bootstrap et son comportement à l'épreuve de toute résolution ;
 - le Javascript comme langage de programmation renforcé par :
 - jQuery pour l'utilisation simplifiée des sélecteurs d'élément du DOM et leur modification ;
 - D3.js pour la création simplifiée de graphique en SVG.
- côté serveur :
 - PHP se charge de réceptionner les requêtes clientes :
 - lors de la création d'un compte utilisateur, il se charge de parser les fichiers CSV et de les envoyer en base (fonctionne partiellement) ;
 - il envoie une requête à R afin qu'il accomplisse ses calculs de corrélation (fonctionne partiellement) ;
 - et se charge de toute requête liée à la demande de données à visualiser.
 - PostgreSQL communique avec PHP tout comme avec R :
 - pour toute requête de sélection ;
 - pour des requêtes d'insertion :
 - à la création d'un utilisateur ;
 - à l'ajout d'informations de la part de R.
 - R reçoit une requête de la part de PHP avec l'id de l'utilisateur à analyser et modifie les données en base.

Les tâches au fonctionnement partiel nécessitent une action humaine pour fonctionner totalement. Il s'agit ici d'un défaut de temps par rapport à l'implémentation. Ces tâches sont réalisables mais auraient

nécessité un délai supplémentaire pour en venir à bout. Cette partie de l'implémentation est secondaire et c'est pourquoi nous ne nous sommes pas attardé avec. Bien-sûr, cela n'a pas compromis la réalisation finale.

Chapitre 6

Visualisation des données physiques et du sommeil

Sommaire

6.1	Introduction	44
6.2	Visualisations basiques	45
6.2.1	Visualisation libre	45
6.2.2	Visualisation de l'influence des types d'activité sur la qualité du sommeil .	46
6.2.3	Visualisation de la qualité du sommeil	47
6.3	Visualisations de corrélations	48
6.3.1	Graphique à nuage de points	49
6.3.2	Cercle de corrélations	50
6.4	Conclusion	51

6.1 Introduction

Dans cette dernière partie, nous allons enfin mettre en pratique l'apprentissage et les analyses précédentes. Dans un premier temps, nous allons tester les possibilités de D3.js puis faire quelques comparaisons intéressantes entre la qualité du sommeil seule mais aussi avec les activités physiques. Puis dans un second temps, nous observerons les corrélations existantes à travers des graphiques de plus en plus élaborés.

6.2 Visualisations basiques

6.2.1 Visualisation libre

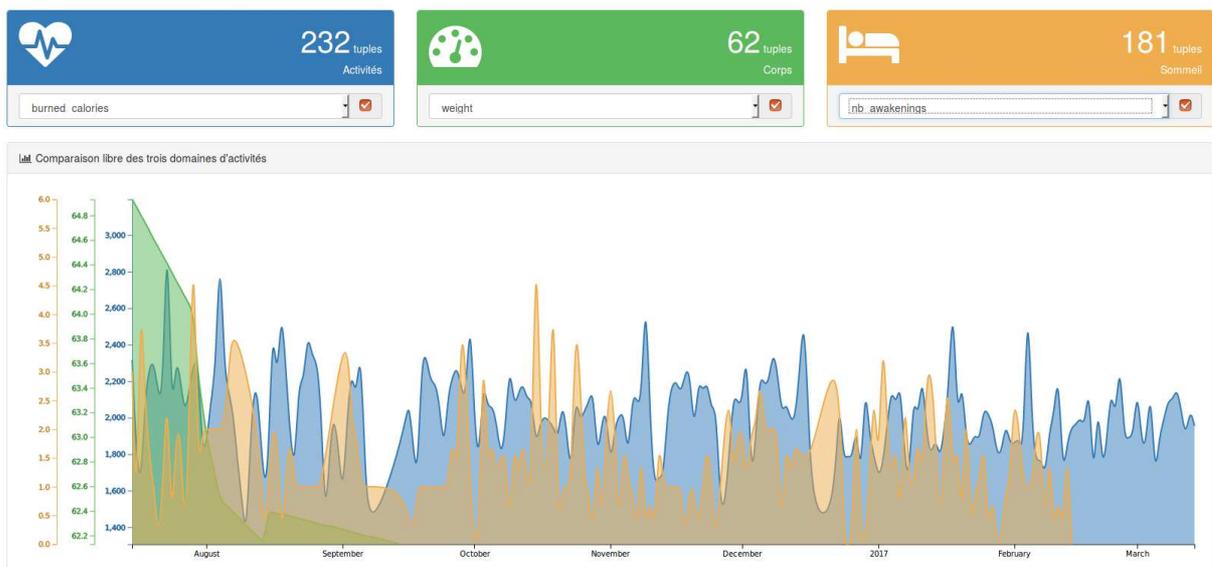


FIGURE 6.1 – Visualisation libre des trois tables sous forme de courbes

Ce premier graphique nous a permis d'expérimenter les capacités de D3.js et des manipulations possibles à l'aide des "listeners" en javascript ou encore d'association de classes CSS permettant d'afficher ou masquer des éléments du DOM.

Ce graphique représente trois courbes. Ces courbes correspondent à nos trois grands types de données : activités physiques, données corporelles et activités du sommeil. Ainsi, pour chaque type de données, nous pouvons choisir entre l'une de leurs colonnes. L'utilisateur choisit alors quelles colonnes afficher ou s'il souhaite masquer certaines courbes à l'aide de leur checkbox. Une dernière interaction permet à l'utilisateur de zoomer sur les courbes, élargissant ainsi l'axe du temps. Cela se fait par le biais de la roulette de la souris ou au double-clic (plus adapté pour l'environnement mobile).

Le graphique dans son ensemble permet d'avoir un premier aperçu des données et une première comparaison entre elles. Les limites de ce graphique se font sentir par une quantité de données trop importantes. Bien que nous puissions regarder les données en détails, il nous manque, pour le moment, le moyen d'amener des informations pertinentes à l'utilisateur.

6.2.2 Visualisation de l'influence des types d'activité sur la qualité du sommeil

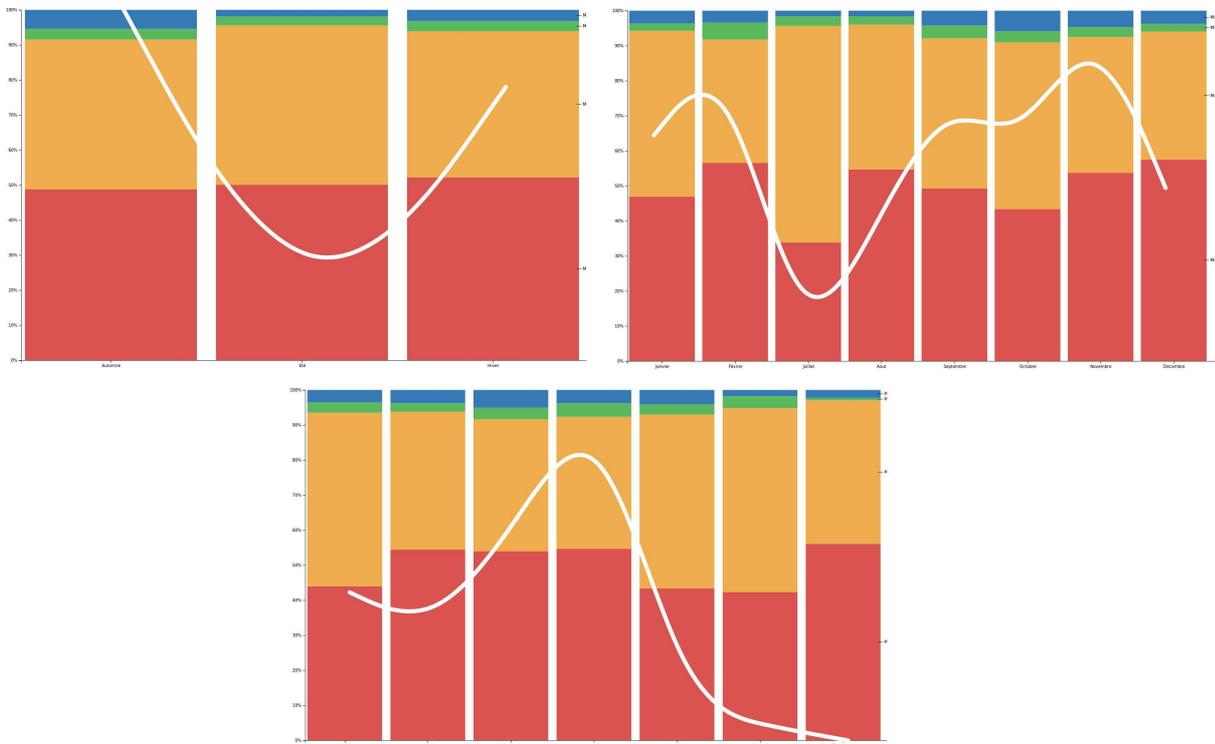


FIGURE 6.2 – Histogrammes à couches empilées et proportionnelles des activités et courbe de la qualité du sommeil

Pour ce deuxième graphique, nous avons représenté nos données d'activités physiques sous forme d'histogrammes à couches empilées. Chacune des couches représentées en rouges, jaunes, vertes et bleues correspondent respectivement à la quantité d'activité sédentaire, peu actives, moyennement active et très active. En plus de ces informations, nous avons incrusté une courbe représentant la qualité du sommeil. Les trois graphiques sont en fonction du temps correspondants aux saisons, aux mois de l'année et enfin aux jours de la semaine.

Pour une meilleure clarté des informations, nous avons soustrait les minutes de sommeil aux minutes sédentaires. En effet, nous avons remarqué que celles-ci étaient incluses dans les données d'activités physiques.

A travers cette figure, nous remarquons que l'utilisateur possède un bien meilleur sommeil durant l'automne, un moins bon en été et enfin, un sommeil moyen en hiver. Ces résultats peuvent s'expliquer par le fait que la qualité de sommeil diffère en fonction de la chaleur et de la quantité de fortes activités.

Voyons maintenant le diagramme traitant les mois. Tout comme celui traitant les saisons, nous remarquons un bien meilleur sommeil durant les mois les plus frais. En général, nous remarquons que l'utilisateur possède un meilleur sommeil lorsque ses activités physiques sont plus intenses.

Enfin, nos précédentes remarques sont une fois de plus confirmées par le diagramme traitant les jours de la semaine. Une remarque supplémentaire peut être considérée. Nous voyons que trois jours consécutifs (mardi, mercredi et jeudi) possèdent la même quantité d'activité sédentaire. Durant ces trois jours, nous remarquons un accroissement de la qualité du sommeil. L'utilisateur peut conclure que la qualité de son sommeil devient correcte lorsqu'il garde une certaine régularité dans ses activités.

Suite à ces observations, nous avons pu dégager quelques informations intéressantes sur le lien entre

la qualité du sommeil et les données physiques. Jusqu'ici nous avons comparé une moyenne de la qualité du sommeil en fonctions des jours de la semaine et des semaines. Il pourrait être maintenant intéressant de trouver un moyen de comparer l'ensemble des semaines (et mois) séparément, et voir s'il est possible de dégager quelques chronicités.

6.2.3 Visualisation de la qualité du sommeil

Semaine	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Moyenne
Semaine 7 de l'année 2017	98.95 %	87.95 %	98.95 %				88.84 %	83.52 %
Semaine 6 de l'année 2017	94.13 %	87.92 %	88.21 %	99.34 %	88.7 %	96.78 %	99 %	93.44 %
Semaine 5 de l'année 2017	99.5 %	98.8 %	98.42 %	76.16 %	86.82 %*	84.43 %	89.52 %	90.52 %
Semaine 4 de l'année 2017	94.99 %	98.59 %	98.14 %	88.05 %	98.75 %	98.35 %	67.85 %	92.1 %
Semaine 3 de l'année 2017	81.77 %	88.78 %	77.57 %	95.31 %	82.5 %	98.99 %	76.91 %	85.98 %
Semaine 2 de l'année 2017	86.96 %	99.06 %	88.53 %	93.51 %	73.88 %	68.76 %	72.29 %	79 %
Semaine 1 de l'année 2017	97.13 %	46.92 %	87.93 %	99.04 %	89.11 %	99.25 %		74.2 %
Semaine 53 de l'année 2016					99.22 %	80.07 %	54.13 %	83.94 %
Semaine 52 de l'année 2016	98.93 %	99.12 %	58.26 %	99.32 %	86.82 %*	65.29 %	93.19 %	85.85 %
Semaine 51 de l'année 2016	86.82 %*	86.82 %*	86.82 %*	86.82 %*	86.82 %*	99.15 %	86.82 %*	88.58 %
Semaine 50 de l'année 2016	84.35 %	97.63 %	88.11 %	86.82 %*	88.1 %	87.87 %	99.51 %	90.34 %
Semaine 49 de l'année 2016	67.9 %	78.14 %	85 %	88.44 %	87.99 %	92.73 %	58.05 %	79.75 %
Semaine 48 de l'année 2016	86.82 %*	77.7 %	98.48 %	88.74 %	99.22 %	99.26 %	97.66 %	92.55 %
Semaine 47 de l'année 2016	83.24 %	98.44 %	86.82 %*	88.07 %	97.85 %	98.09 %	98.98 %	93.07 %
Semaine 46 de l'année 2016	95.91 %	98.53 %	98.44 %	98.49 %	98.84 %	89.43 %	89.06 %	95.53 %
Semaine 45 de l'année 2016	98.8 %	98.68 %	86.98 %	92.33 %	99.52 %	88.31 %	99.22 %	94.83 %
Semaine 44 de l'année 2016	87.07 %	88.12 %	60.43 %	88.15 %	98.12 %	99.23 %	83.56 %	86.38 %
Semaine 43 de l'année 2016	73.03 %	67.08 %	77.82 %	97.8 %	99.49 %	98.25 %	95.66 %	87.02 %
Semaine 42 de l'année 2016	71.58 %	98.32 %	87.66 %	54.99 %	97.88 %	98.78 %	37.32 %	78.08 %
Semaine 41 de l'année 2016	98.22 %	88.82 %	96.39 %	86.07 %	97.72 %	86.82 %*	77.25 %	90.18 %
Semaine 40 de l'année 2016	86.3 %	67.57 %	98.49 %	86.9 %	77.43 %	98.62 %	97.64 %	87.66 %
Semaine 39 de l'année 2016	98.8 %	88.88 %	98.29 %	66.56 %	97.86 %	98.91 %	98.87 %	92.6 %
Semaine 38 de l'année 2016	98.76 %	98.21 %	98.01 %	98.8 %	99.04 %	86.82 %*	99.81 %	97.07 %
Semaine 37 de l'année 2016	86.82 %*	86.82 %*	86.82 %*	86.82 %*	86.82 %*	86.82 %*	86.82 %*	86.82 %

FIGURE 6.3 – Calendrier sous forme de carte chaude (classé par semaine) de la qualité du sommeil

Afin de remarquer les chronicités entre semaine, nous avons emprunté le principe de la carte chaude appliqué à un calendrier superposant les semaines. Ainsi, nous pouvons comparer les nuances de couleurs de semaine en semaine. Les nuances de couleurs vont du vert au rouge et donc d'un sommeil de bonne qualité à une qualité moindre. Cette visualisation n'utilise pas D3.js et est entièrement construite par des éléments HTML. Nous pouvons réordonner chacune des semaines en fonction des colonnes du tableau par ordre croissant ou décroissant.

Mise à part une vue d'ensemble assez claire grâce aux nuances de couleur, il est difficile de discerner des chronicités dans ce graphique. C'est pourquoi nous avons essayé un autre visuel à base de courbes, que nous espérons, plus parlant.

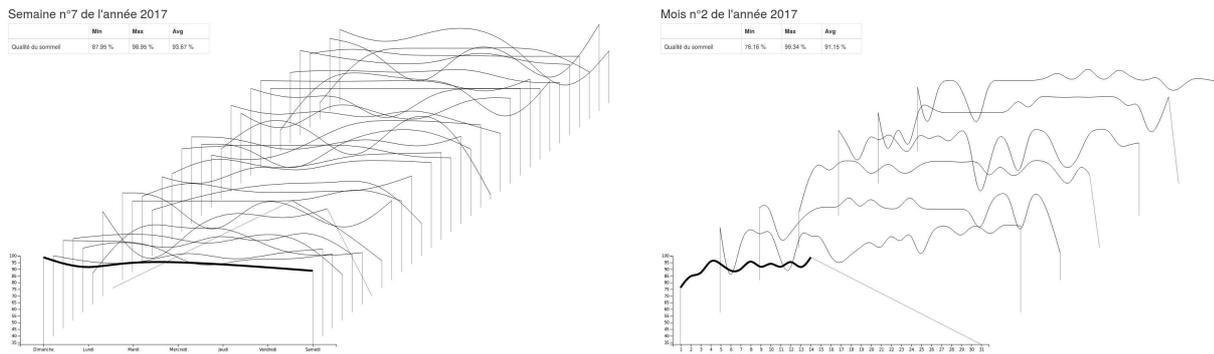


FIGURE 6.4 – Superposition de graphiques représentant la qualité du sommeil en fonction des mois et semaines

A travers ces deux graphiques nous avons voulu représenter les différentes semaines et différents mois de sommeil de l'utilisateur à l'aide de différentes courbes en perspectives. Cette perspective permet d'apercevoir, s'il y a, une certaine répétition entre chaque semaine ou mois.

Nous avons voulu apporter une nouvelle approche à l'utilisateur en lui permettant d'interagir avec la perspective. Ainsi, en appuyant sur "+" ou "-" avec son clavier, l'utilisateur peut déplacer le repère et passer d'une courbe à l'autre.

Une fois de plus, il est compliqué d'observer de nouvelles relations. A travers les données de notre utilisateur, nous n'arrivons pas à discerner des régularités entre les semaines ni même les mois. Il ne s'agit certainement pas ici d'une généralité; d'autres utilisateurs pourraient retrouver, à travers ce graphique, une certaine chronicité de la qualité de leur sommeil en fonction des semaines et mois.

6.3 Visualisations de corrélations

Dans cette nouvelle partie, nous avons de nouvelles informations grâce au langage R. Tout comme nous avons fait avec les tests de nos indicateurs, nous avons analysé chacune des paires de variables dont nous disposons afin de remarquer des corrélations entre elles. Il faut maintenant trouver un moyen de rendre compte de ces corrélations à l'utilisateur.

6.3.1 Graphique à nuage de points

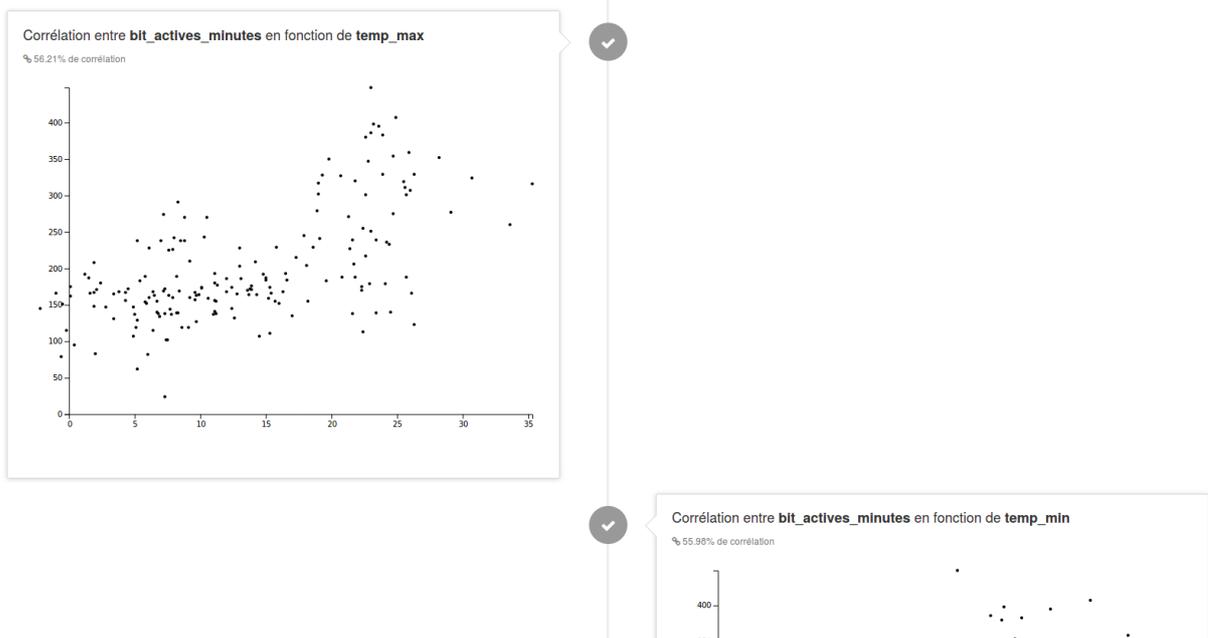


FIGURE 6.5 – Nuage de points sur les paires de données les plus corrélées

Dans un premier temps, nous avons opté pour l'utilisation d'une "timeline" pour afficher l'ensemble des corrélations. Nous avons subtilisé la ligne du temps de ce genre de graphique en la remplaçant par une ligne de corrélations décroissantes. Ainsi, nous voyons apparaître en premières places nos variables les plus corrélées avec leur pourcentage de corrélation. De plus, nous affichons le graphique à nuage de points correspondant à la relation entre la paire de variables.

Ce graphique permet d'avoir un premier aperçu des relations existantes entre nos variables. Étant donné que nous possédons quelques dizaines de variables à comparer, le nombre de paires en est donc conséquent. Il nous faut trouver un autre moyen d'afficher nos corrélations de façon plus compacte.

6.3.2 Cercle de corrélations

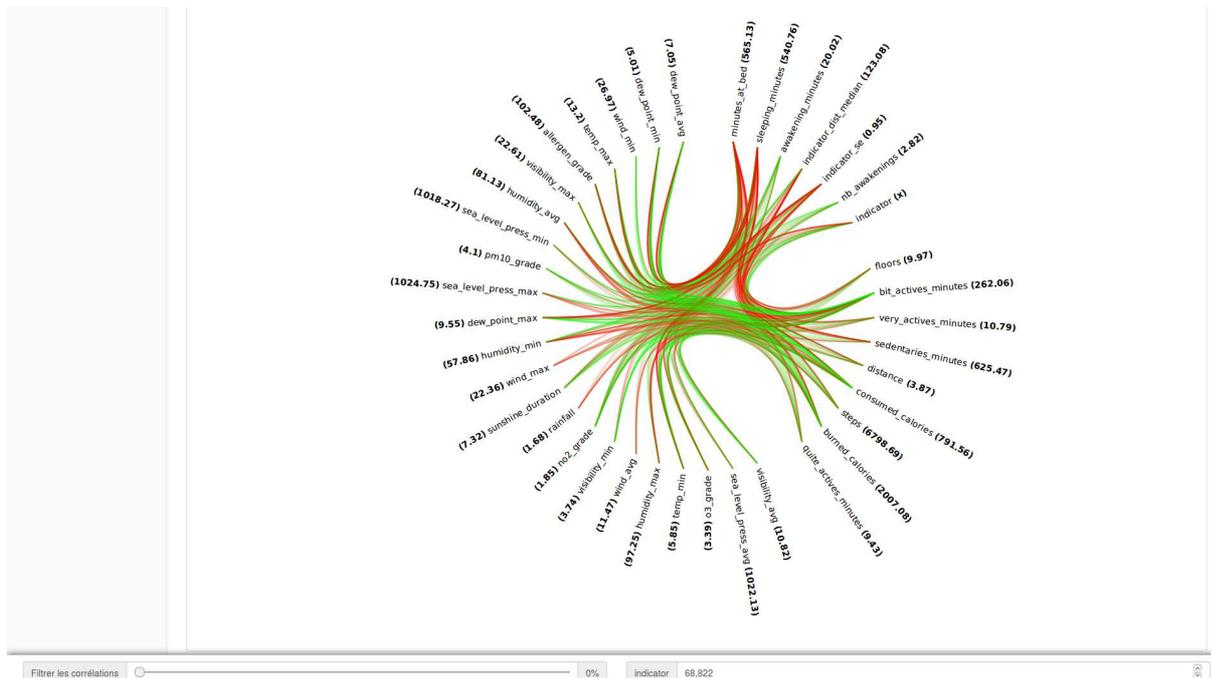


FIGURE 6.6 – Diagramme en cordes des corrélations entre l'ensemble des variables

Ce dernier graphique essaye de remédier aux précédentes remarques en mettant en scène une nouvelle manière de visualiser les données. Il s'agit ici du principe d'un diagramme en cordes. C'est une représentation qui nous permet de projeter nos tables de données et les colonnes qu'elles comportent en variables parcourant le diamètre du cercle. Chacune des paires de variables est reliée par une "corde" plus ou moins épaisse en fonction de leur force de corrélation et du rouge vers le vert en fonction d'une corrélation positive ou négative.

L'utilisateur peut interagir avec le diagramme en survolant les variables avec sa souris et n'afficher que les corrélations correspondantes. En plus de cette interaction, il peut définir un seuil minimal de corrélation à afficher sur le diagramme à l'aide d'un curseur qu'il peut incrémenter ou décrémenter.

A l'aide de R, nous avons, à nouveau, agrémenter notre base de données. En effet, nous avons pu tracer les courbes de régression liées à nos nuages de points initialement introduits dans le graphique précédent. Ainsi, nous avons subtilisé ces courbes afin d'introduire une troisième interaction avec l'utilisateur. Il peut désormais cliquer sur l'une des variables présentées autour du cercle et décider de modifier sa valeur. Au clic, l'ensemble des courbes de régression apparaissent avec un marqueur représentant la valeur avec laquelle l'utilisateur souhaite interagir. En plus du déplacement du marqueur sur la courbe, nous pouvons observer la valeur des variables entourant notre diagramme se modifier dynamiquement.

Grâce à ce graphique, l'utilisateur peut aisément observer les corrélations qui existent entre ses différentes activités, son sommeil (dont la qualité du sommeil) et la météorologie. En plus de cette observation, il peut interagir avec la valeur de ses variables pour observer des augmentations et diminutions parmi les autres valeurs. Par exemple, lorsque nous augmentons la quantité de minutes peu actives dans une journée, nous constatons que nos indicateurs de qualité du sommeil diminuent progressivement.

Ce genre de graphique permet à l'utilisateur d'augmenter sa connaissance de façon drastique. En effet, il lui permet de voir rapidement quels sont ses caractères les plus corrélés et en plus de ça, pouvoir interagir avec eux en modifiant leur valeur et voir les répercussions sur l'ensemble de ses caractères.

6.4 Conclusion

Bien que nos données soient très impressionnantes, il est possible à travers la visualisation, d'apporter énormément d'informations. En effet, le simple fait de tracer une courbe temporelle permettant de retracer une information permet une bien meilleur approche que des chiffres pour la vision humaine. Si à cela nous ajoutons des informations supplémentaires que l'utilisateur peut modifier, sélectionner, survoler ou classer à souhait, la quantité d'informations et de résultats obtenus devient alors tout aussi impressionnante que la quantité de nos données initiales.

C'est ainsi qu'agissent les graphiques précédemment présentés. A travers le croisement d'informations, la mise en valeur de chronicité et l'analyse des corrélations, nous avons pu indiquer des informations aux-quelles l'utilisateur n'aurait jamais pu avoir accès avec les données brutes initiales.

Conclusion générale

Lors de ce TER, nous nous sommes tout d'abord intéressés à l'existant. Nous nous sommes rendus compte qu'il existait énormément de moyens afin de capturer les différentes caractéristiques corporelles décrivant les activités physiques et du sommeil. Ce genre de capture fournit une quantité phénoménale d'informations. L'une des manières les plus optimales pour appréhender ce flot de données concerne le domaine de la data-visualisation. Ainsi, nous nous sommes aperçus qu'il existait une diversité de graphiques cherchant à représenter au mieux ces informations. Bien qu'ils soient nombreux, nous n'avons pas rencontré de graphiques capables de nous apporter une information supplémentaire. En effet, la majeure partie de l'existant s'efforce juste à afficher les données le plus respectivement possible.

Le travail de ce TER a donc été de pousser la partie analytique afin d'acquérir de nouvelles informations telles que des corrélations, des chronicités ou des prédictions. Ainsi, nous avons pu produire de nouvelles visualisations capables de retranscrire les informations de l'analytique mais surtout d'apporter une connaissance nouvelle à l'utilisateur. Cette information lui permet de mieux appréhender les relations qui existent entre ses activités physiques, la météorologie et la qualité de son sommeil.

Toutefois, nos résultats auraient pu être plus concluants en fonction de certains critères. En effet, quelques doutes subsistent quant à la significativité des mesures capturées. Par exemple, le nombre de réveils durant la nuit (qui est un critère d'une grande importance concernant la qualité du sommeil), ne semble pas être très significatif. Le capteur, faisant correspondre les mouvements de l'utilisateur à des réveils, semble parfois nous transmettre des données erronées (ou alors une incompréhension logique de ces données). Ces erreurs peuvent donc induire un indicateur, les utilisant, à devenir inexact. Il aurait été intéressant d'agrémenter le dispositif de suivi, de capteurs pouvant collecter des données telles que la température ou encore la luminosité ambiante. En effet, bien que nous ayons lié la position de l'utilisateur à une météo correspondante, nous ne pouvons pas garantir qu'il s'agissait des températures qu'il subissait. En effet, il peut se situer dans une zone chauffée, climatisée ou encore être en voyage.

Pour de nouvelles recherches à ce sujet, il pourrait être intéressant d'obtenir une grande base d'utilisateurs. Ainsi, nous aurions des utilisateurs ayant testés des solutions afin de remédier à certains problèmes types. Le but serait de répertorier ces problèmes types, afin qu'ils aient une utilité future pour des utilisateurs subissant les mêmes troubles. La correspondance inter-utilisateurs se ferait par leurs caractéristiques (taille, poids, âge, etc), mais aussi par les données d'activités du sommeil (correspondance sur le nombre de réveils, le temps d'éveil, le temps de sommeil, les indicateurs, etc), les corrélations aux activités physiques et la météorologie.

Ce TER m'a beaucoup appris sur la méthodologie de recherche. La démarche empruntée m'a captivé de part un apprentissage intense au tout début de l'étude. Le fait de se documenter et d'apprendre par soi-même est un réel engouement à mes yeux. La partie analytique et la recherche d'indicateurs m'a aussi beaucoup enjoué. Enfin, la partie développement fut pour moi la plus intense de part l'application de ces nouvelles connaissances et de mon attrait à la programmation.

Si je devais choisir à nouveau un sujet de recherche, j'opterais pour un sujet plus ancré à l'informatique. Ce sujet fut très passionnant, enrichissant et bien adapté pour la période à laquelle nous avons

eu droit. Il m'a appris beaucoup sur le fascinant métier du data-scientist. Mais pour de plus longues recherches, je choisirai un sujet plus propre à l'informatique : ma passion originelle.

Bibliographie

- [AUVRAY & GUYOT 2017] Elisabeth AUVRAY et Vivien GUYOT. « Capteurs Environnementaux », 2013 - 2017.
- [BASTIEN 2016] BASTIEN. « Wearables : les objets connectés en 500 ans d’Histoire », 2016.
- [BOSTOCK 2017] Mike BOSTOCK. « Hierarchical Edge Bundling », 2017.
- [COOK 2013] Peter COOK. « UK Temperature History », 2013.
- [D. KRYSYAL & D. EDINGER 2008] Andrew D. KRYSYAL et Jack D. EDINGER. « Measuring sleep quality ». 2008.
- [GRASLAND 2015] Claude GRASLAND. « La corrélation », 2015.
- [J. PILCHER *et al.* 1996] June J. PILCHER, Douglas R. GINTER, et Brigitte SADOWSKI. « Sleep quality versus sleep quantity ». 1996.
- [KELLEHER 2017] Curran KELLEHER. « Introduction to D3 », 2017.
- [LI 2012] Ian LI. « Spark : Visualizing Physical Activity Using Abstract Ambient Art », 2012.
- [OPENCLASSROOM 2013] OPENCLASSROOM. « Introduction au SVG », 2013.
- [RAD 2015] Reza RAD. « Be Fitbit BI Developer in Few Steps : Step 3 Visualization », 2015.
- [SIMONEAU] Jacqueline SIMONEAU. « Quand la météo rend malade ».
- [WANG *et al.* 2015] Lidong WANG, Guanghui WANG, et Cheryl Ann ALEXANDER. « Big Data and Visualization : Methods, Challenges and Technology Progress », 2015.
- [YAU 2011] Nathan YAU. *Data visualisation*. Eyrolles, 2011.
- [YDÉE 2017] Nicolas YDÉE. « Analytique pour les applications de suivi des activités physiques et du sommeil ». 2017.
- [ZAFFAGNI 2015] Marc ZAFFAGNI. « Une application mobile qui détecte l’apnée du sommeil », 2015.