



U.F.R. Science et techniques - Nantes

X2BI060 : Travail Encadré d'Étude et de Recherche

**Lapin automate : analyse temporelle de données
physiologiques issues de différents TP
d'expérimentations animales**

Auteur : Barbey Cassandra



SOMMAIRE

Introduction	3
Projet Arduino : lapin automate	3
Présentation du stage	3
Matériel et Méthodes	4
Nettoyage et codage	4
Analyses temporelles	5
Observations tendance et saisonnalité	5
Enveloppe de la séquence	6
Traitement de la fréquence diurèse	6
Classification	6
Résultats	7
Artefacts	7
Interprétation des analyses temporelles	9
Tendance et saisonnalité	11
Observation de l'enveloppe	15
Diurèse et fréquence d'apparition	16
Classification DTW	17
Conclusion / Discussion	18
Conclusion du stage	18
Difficultés rencontrées	18
Points à améliorer	19
Conclusion personnelle	19
Bibliographie / Sitographie	19
Annexes	20
Figure A1 : Code de recodage du temps	20
Figure A2 : Représentation d'un cycle diastole-diastole	20
Figure A3 : Code de l'isolement du 1er cycle diastole-diastole	20
Figure A4 : Code de corrélation par acf() et superposition des graphiques	20
Figure A5 : Code d'obtention de l'enveloppe	20
Figure A6 : Code du calcul de la fréquence diurèse	20
Figure A7 : Code utilisé pour la classification par dtw	20
Figure A8 : Méthodes de calcul de la fréquence diurèse	21

Introduction

- **Projet Arduino : lapin automate**


Notre monde actuel est de plus en plus tourné vers la recherche et l'apprentissage. La curiosité humaine nous pousse à nous renseigner et à découvrir des choses qui nous sont encore inconnues. La biologie est un des domaines suscitant le plus d'intérêt, ce qui est notamment dû à la complexité des organismes vivants ainsi qu'à leur physiologie plus qu'adaptée à la survie. Cependant, afin de nous renseigner sur ce sujet, et de former les générations futures à la continuité des recherches, nous sommes parfois obligés de passer par des étapes jugées non-éthiques comme la dissection ou l'expérimentation animale, nécessaires à la compréhension des réactions chimiques et physiologiques de ces organismes. Afin de contrer ce problème, et de réduire l'utilisation des animaux dans la recherche, plusieurs méthodes ont été tentées comme par exemple l'utilisation de programmes reproduisant les fonctions biologiques d'un être vivant, l'augmentation de la difficulté à obtenir une autorisation de projet utilisant des animaux à des fins scientifiques ou encore la limitation du nombre de sujets inclus dans les études. Cependant, lorsqu'il s'agit de réduire cette utilisation lors de la formation de futures chercheurs ou vétérinaires, cela devient vite plus compliqué étant donné la nécessité d'apprentissage concret afin que les personnes formées soient capables de bien réaliser les travaux futurs.

C'est pour cela qu'en 2018, l'école vétérinaire Oniris de Nantes, l'IMT Atlantique et la Faculté des Sciences et Techniques de Nantes se sont associés afin de réaliser un projet commun : la création d'un lapin automate possédant des réponses physiologiques réelles basées sur les données récoltées de 2013 à 2018 des différents TP d'expérimentations animales réalisés à Oniris. Ce lapin robotisé a pour objectif de remplacer les véritables lapins utilisés lors des TP des vétérinaires en formation afin de diminuer l'utilisation des êtres vivants et de contrer le problème éthique tout en maintenant une formation précise des étudiants.


Ce projet est constitué de plusieurs éléments ^[1] : une interface permettant l'interaction entre l'utilisateur et l'automate ainsi que la visualisation de l'évolution des fonctions physiologiques lors de "l'injection" de la molécule choisie; ainsi qu'un automate introduit dans une peluche constitué d'un moteur, représentant le coeur, d'un ballonnet, représentant les poumons, et émettant un signal sonore répétitif représentatif de la diurèse.

- **Présentation du stage**

Dans le cadre de ce projet, j'ai été amené à travailler avec des données, issues de séries temporelles d'une population de lapins (*Oryctolagus cuniculus*), représentatives de différentes fonctions physiologiques (pression artérielle, fréquence cardiaque, diurèse) à l'état basal et après injection de molécules (acétylcholine, adrénaline, angiotensine II, desmopressine ...). Mon objectif était d'analyser ces données et de permettre leur classification en plusieurs groupes d'individus afin de voir les variations des réactions possibles.

Mon stage s'est étendu sur une durée de 2 mois (du 23/03/2021 au 15/05/2021) et a été tutoré par 2 enseignants-chercheurs : Mme HERVE Julie (Service de Physiologie/Laboratoire IECM à l'Ecole Vétérinaire Oniris) et Mr PRUD'HOMME Charles (Equipe TASC, DAPI, IMT Atlantique). 

Matériel et Méthodes

Pour réaliser mon stage, j'ai eu accès aux données récoltées lors de la réalisation des TP de physiologie cardio-respiratoire et cardio-rénale de 2013 à 2018. Au vu du nombre de données présentes et de la durée du stage, je n'ai pu travailler qu'avec les données cardio-rénales. Ces dernières ont été obtenues à l'aide de 2 logiciels de mesures continues, différents : le logiciel Biopac  le plus ancien, sur lequel ont été réalisés les récoltes de données de 2013 à 2016 et le logiciel LabChart, le plus récent, sur lequel ont été réalisés les récoltes de mesures de 2017 à 2018.

La première étape que j'ai dû effectuer lorsque l'on m'a permis d'accéder aux données, a été de trier les séquences utilisables et non-utilisables. En effet, lors de l'enregistrement des mesures, certains fichiers ont été dupliqués lors de leur création provoquant ainsi l'apparition de fichiers vides. Suite à cela, j'ai ensuite exporté les fichiers, présents sous forme de tracé, en fichier texte .txt afin de pouvoir facilement les manipuler et les analyser.

- **Nettoyage et codage**

Une fois les données récupérées et importées sur mon ordinateur personnel, et sur mon environnement Rstudio, j'ai effectué un travail de nettoyage et de recodage de ces dernières. En effet, les séries de données importées possédaient une durée différente, et une forme différente selon leur logiciel de mesure initial, rendant ainsi la comparaison et l'analyse compliquée. Pour toutes les données, j'ai donc recodé la variable temps et fait commencer les séries à 0.005 secondes (cf. **Figure Annexe 1**) car les valeurs ont été mesurées toutes les 5 millisecondes. J'ai également renommé, et supprimé, certaines colonnes représentant les commentaires faits sur le logiciel lors du TP, comme par exemple "injection d'acétylcholine", n'étant pas utiles pour réaliser mon objectif.

Après ce travail de remise au propre, j'ai voulu observer la représentation des différentes variables des séries pour vérifier la présence ou non d'artefacts, provoqués par le déplacement des sondes lors des TP, empêchant la bonne exploitation des données. Une fois cela fait, il me fallait trouver une base qui me permettrait d'effectuer mes comparaisons et mes observations de façon précise. Mes données possédant chacune 3 variables différentes : pression artérielle, fréquence cardiaque et diurèse; je me suis donc basée sur la pression artérielle (PA) et plus particulièrement sur sa cyclicité. Pour cela, j'ai cherché à identifier le 1er cycle présent dans ma séquence de données en me basant sur un cycle diastole-diastole, c'est-à-dire sur les plus petites valeurs (cf. **Figure Annexe 2**). Afin de le déterminer, j'ai créé plusieurs "fonctions" (cf. **Figure Annexe 3**) dépendantes les unes des autres. La première "fonction" m'a permis de déterminer la 1ère valeur la plus faible, représentative de la pression diastolique, et d'enregistrer sa position dans une variable appelée *num*; à partir de cette dernière, j'ai pu appliquer ma deuxième "fonction" qui avait pour objectif de déterminer la 1ère valeur la plus forte, représentative de la pression systolique, et d'enregistrer sa position dans la variable *num2*. Puis, à partir de *num2*, j'ai pu appliquer ma troisième "fonction" ayant pour objectif de déterminer et d'enregistrer la position de la 2ème pression diastolique dans la variable *num3*. Une fois les 2 extrémités trouvées (*num* et *num3*), j'ai calculé la différence entre les 2 afin d'obtenir la durée et le nombre de données d'un cycle.

Cette durée va me servir par la suite pour décomposer la série afin d'avoir des représentations plus précises au vu du nombre d'observations présentes par séries.

- **Analyses temporelles**

Dans le cadre de mon stage, les données manipulées sont des données temporelles issues de séries chronologiques, les analyses effectuées sont donc légèrement différentes des analyses courantes. Pour commencer, j'ai tout d'abord réalisé une double représentation graphique de mes données : d'une part, j'ai effectué une représentation classique affichant chaque variable en fonction du temps, et de l'autre, j'ai effectué un histogramme de ces dernières me permettant de visualiser la distribution des données de chaque variable au cours d'un processus donné, ici le TP d'expérimentations animales avec ou sans injection de molécules.

Suite à cela, j'ai ensuite appliqué à mes séries le test de Ljung-box ^[2] permettant de vérifier la présence d'autocorrélation entre les données d'une même séquence chronologique. Ce test, particulièrement efficace lors de la présence de grandes séquences de données, se base sur des hypothèses concernant la distribution des résidus : l'hypothèse nulle (H0) est que les résidus sont distribués de façon indépendante tandis que l'hypothèse alternative (H1) est que les résidus ne sont pas distribués indépendamment, mais qu'ils présentent une corrélation en série. J'ai donc utilisé ce test sur chaque couple séquence molécule/variable en utilisant un risque alpha à 5% et par extension, une p-value limite de 0,05.

- **Observations tendance et saisonnalité**

Après avoir effectué les analyses temporelles, j'ai cherché à observer la tendance des séries de données. "La tendance est l'orientation générale d'une série d'observations à la hausse ou à la baisse sur une période assez longue" ^[3]. Au vu de la quantité de données présentes dans une série (30 000 à 130 000 observations), j'ai décidé de les moyenner en me servant de la durée de cycle précédemment calculée. Pour ce faire, j'ai calculé la moyenne des données présentes sur chaque cycle afin de réduire le nombre d'observations et de mieux visualiser la tendance qui se dégage de la représentation graphique de ces moyennes en fonction du temps.

Suite à cela, je me suis penchée sur la saisonnalité des séries de données. La saisonnalité peut-être vue comme une similarité entre le comportement des données à différents temps ^[4]. Ici, elle représente la cyclicité de notre série sur une période de temps en millisecondes. Pour obtenir la saisonnalité des mes séquences de données, je suis passée par une fonction d'autocorrélation, acf, effectuant des mesures d'association entre les valeurs "actuelles" et "futures" d'une même série avec un décalage de k mesures. Afin d'avoir un résultat et une représentation concluante, j'ai effectué la fonction acf sur 6 cycles avec une visualisation tous les 3 cycles afin de bien observer la corrélation entre les mesures, et ce, de manière distinctive sans chevauchement de données (cf. **Figure Annexe 4**).

Pour appuyer cette saisonnalité, j'ai également comparé deux à deux chaque cycle et comparé le premier et le dernier cycle, par le biais de représentations graphiques, afin de voir leur superposition et donc la cyclicité de mes séries de données (cf. **Figure Annexe 4**).

- **Enveloppe de la séquence**

J'ai ensuite cherché à obtenir l'enveloppe des données de chaque variable afin d'observer la variation d'amplitude des mesures. Pour cela, j'ai isolé les maximums et les minimums de chaque cycle, à l'aide d'une boucle while, et les ai enregistrés dans un nouveau dataframe créé en amont. J'ai ensuite affiché les tracés des données isolées par superposition des plots (cf. **Figure Annexe 5**).

- **Traitement de la fréquence diurèse**

Contrairement à la pression artérielle et à la fréquence cardiaque, la variable diurèse est beaucoup plus complexe à analyser, car ce n'est pas l'amplitude qui nous intéresse. En effet, mon objectif pour cette dernière était d'isoler la fréquence du nombre de pics pour chaque séquence afin d'observer si l'injection de molécules modifiait ou non la production d'urine par les reins. Pour répondre à cette question, j'ai tout d'abord effectué une représentation graphique de la diurèse des 6 premiers cycles en fonction du temps afin d'observer la valeur maximale d'un pic de diurèse ainsi que le nombre de pics présents sur 6 cycles. Suite à cela, j'ai ensuite procédé à une extraction de tous les maximums de la variable diurèse, à l'aide d'une boucle while, et je les ai enregistrés dans un dataframe précédemment créé. Étant donné que chaque extraction débute à partir de la dernière valeur de l'extraction précédente, il va y avoir la création de doublons lorsque la valeur maximale se trouve au niveau de cette jonction. Afin de contrer cela, j'ai créé une fonction me permettant de recoder les doublons successifs en NA tout en gardant le 1er doublon intact. Après cela, j'ai effectué un recodage de mon tableau en me servant de la valeur maximale du pic précédemment observée afin d'identifier les pics de diurèse parmi les autres valeurs maximales. Enfin, j'ai calculé le nombre de pics présents sur 6 cycles et en ai déduit la fréquence d'apparition à l'aide de deux boucles jointes : une boucle while et une boucle for (cf. **Figure Annexe 6**).

- **Classification**

La dernière étape de mon stage consistait à effectuer une classification des différentes séries et variables afin de créer des clusters représentatifs des différentes réactions et fonctions biologiques selon la molécule injectée. Pour atteindre mon objectif final, il me fallait regrouper tous les individus selon leur injection et selon une variable. J'ai donc créé de multiples tableaux dans lesquels j'ai enregistré les données d'une variable pour chaque individu. Pour avoir une classification optimale et égale pour toutes les données, j'ai identifié la séquence de données avec le moins d'observations afin de m'en servir pour imposer une taille aux tableaux par sélection.

Je me suis ensuite occupée de la partie classification en utilisant une classification par DTW (Dynamic Time Warping). "Le principal avantage de DTW est la possibilité de regrouper des séries chronologiques selon leurs modèles ou formes même si ces modèles ne sont pas synchronisés (décalage)" ^[5]. Pour cela, j'ai tout d'abord importé et installé les différentes bibliothèques utilisées pour réaliser mes futures lignes de commandes : "TSclust", "dtwclust", "ggdendro", "dplyr", "ggplot2", "gridExtra" et "dendextend". Puis j'ai ensuite effectué la classification en la décomposant en plusieurs étapes : calcul de la distance entre chaque série chronologique, sélection du nombre de cluster voulu, création du dendrogramme et des plots servant à la classification, enregistrement des représentations et obtention de la représentation de la classification.

- **Artefacts**

Comme je l'ai expliqué précédemment, j'ai tout d'abord procédé à un recodage de la variable temps ainsi qu'à une suppression de la variable commentaire, puis après cela, j'ai vérifié la présence d'artefacts pour tous les couples variables/molécules de tous les individus. On peut observer ci-dessous deux représentations graphiques du couple pression artérielle/sans injection pour l'individu 1 (cf. **Figure 1**) et pour l'individu 2 (cf. **Figure 2**). La représentation graphique de l'individu 2 montre clairement la présence d'artefacts de 0 à 1200 millisecondes, se traduisant par une non-stabilité de la courbe et la présence de plusieurs pics de variations. On remarque cependant qu'après 1200 millisecondes, il y a une stabilisation et un retour à la "normale" si l'on se fie à l'intervalle de variation de la pression artérielle de l'individu 1 (aux alentours de 40 et 60). Nous ne pouvons pas utiliser directement la séquence de l'individu 2 pour nos analyses, il a donc fallu isoler la partie exploitable tout en vérifiant que la quantité de données restantes soit suffisante, ce qui est le cas si l'on observe les **Figures 3** et **4** représentant le nombre de données restantes après sélection sur l'individu 2 sans injection.

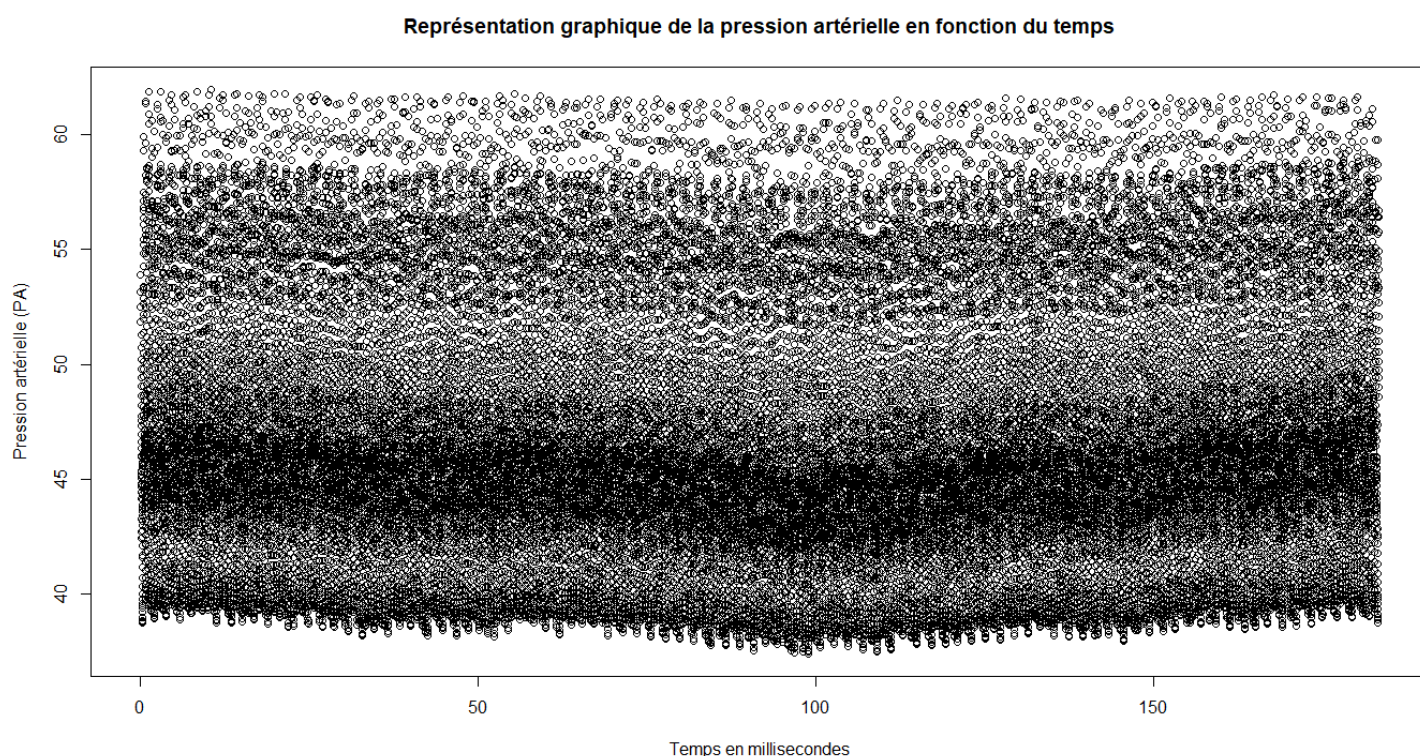


Figure 1 : Représentation graphique du couple pression artérielle/sans injection pour l'individu 1 (absence d'artefacts)

Représentation graphique de la pression artérielle en fonction du temps

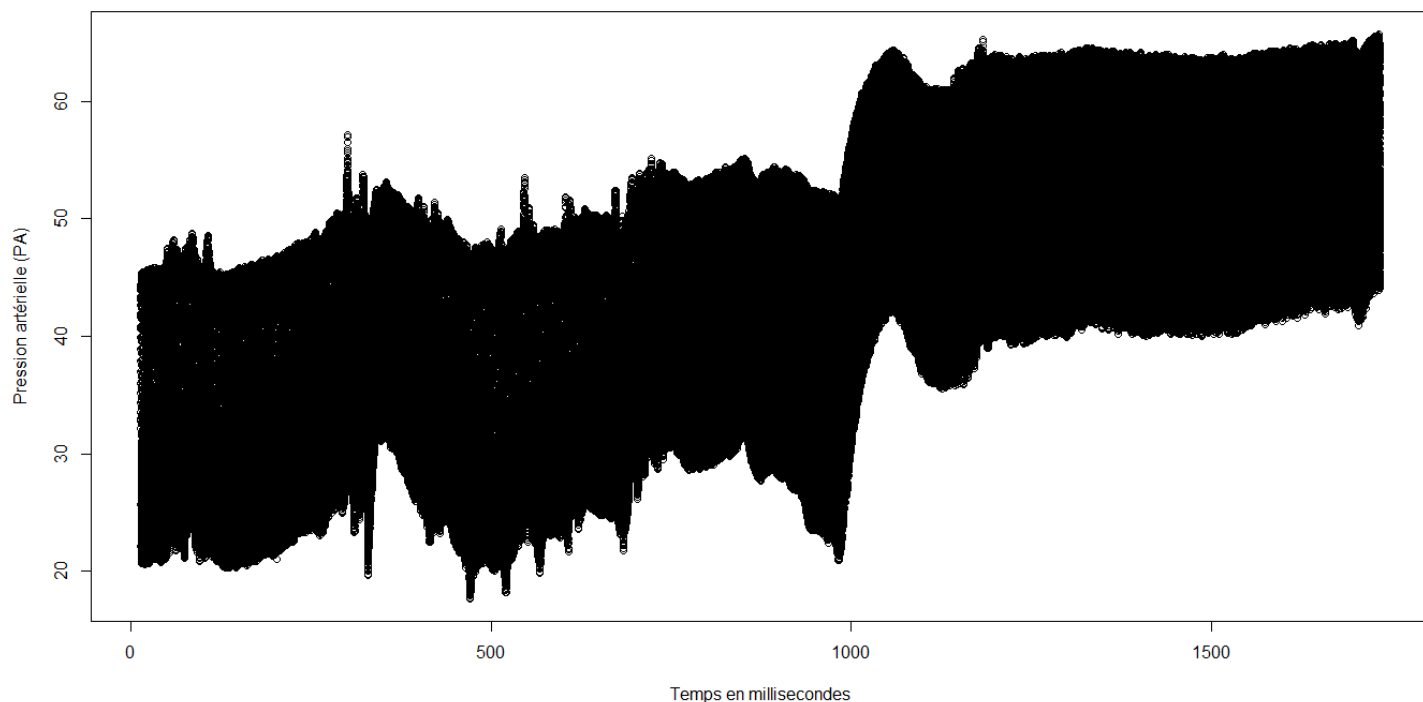


Figure 2 : Représentation graphique du couple pression artérielle/sans injection pour l'individu 2 (présence d'artefacts)

Représentation graphique de la pression artérielle en fonction du temps

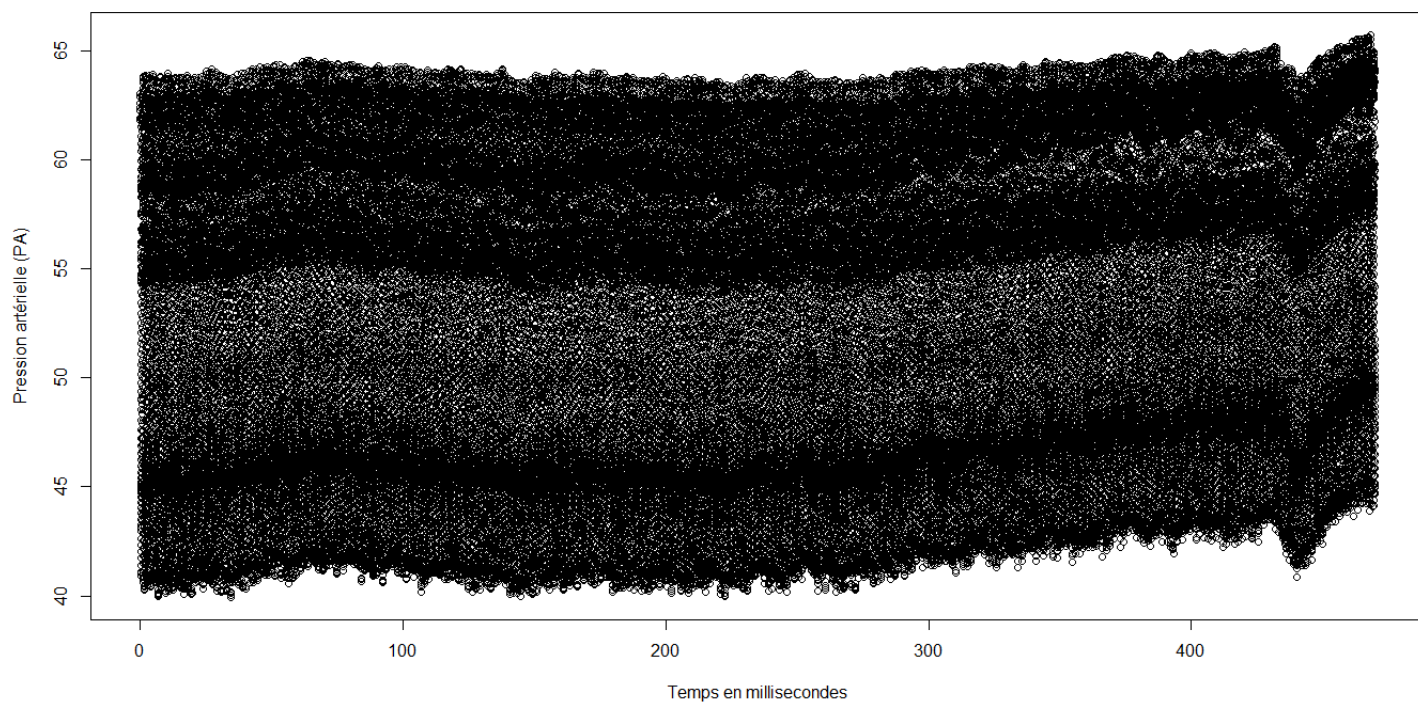


Figure 3 : Représentation graphique du couple pression artérielle/sans injection pour l'individu 2 après recodage

I1_SI	36660 obs. of 6 variables
I2_SI	343960 obs. of 7 variables

↓ Recodage

I1_SI	36660 obs. of 6 variables
I2_SI	93961 obs. of 6 variables

Figure 4 : Vérification du nombre d'observations après recodage

- **Interprétation des analyses temporelles**

Afin de pouvoir poursuivre vers mon objectif de classification, j'ai effectué une double représentation graphique de mes données pour observer la distribution de ces dernières. Pour cela, j'ai observé l'aspect global de chaque courbe de données et je l'ai comparé à son histogramme afin d'observer la répartition des données ainsi que la présence ou non de valeurs extrêmes qui pourraient biaiser la tendance centrale de chaque série et fausser les conclusions de nos calculs. On peut observer dans les figures ci-dessous, représentatives de l'individu 12 avec adrénaline, qu'il n'y a pas la présence de valeurs dites extrêmes. En effet, si l'on prend la **Figure 5**, basée sur la pression artérielle, on remarque que le début de la courbe (de 0 à 20 secondes), représentatif de l'état basal de l'individu, oscille entre 20 et 60 sur un "court moment" avant d'avoir un "bref" pic d'augmentation (de 20 à 45 secondes) jusqu'à 100 puis de nouveau une oscillation entre 80 et 120 jusqu'à la fin de la série. Si l'on observe maintenant l'histogramme de cette même figure, on remarque que la majorité de données se situent entre 80 et 120 de pression artérielle ce qui est logique au vu de nos observations précédentes, l'histogramme est donc bien représentatif de la distribution des données et il ne semble pas y avoir de valeurs aberrantes.

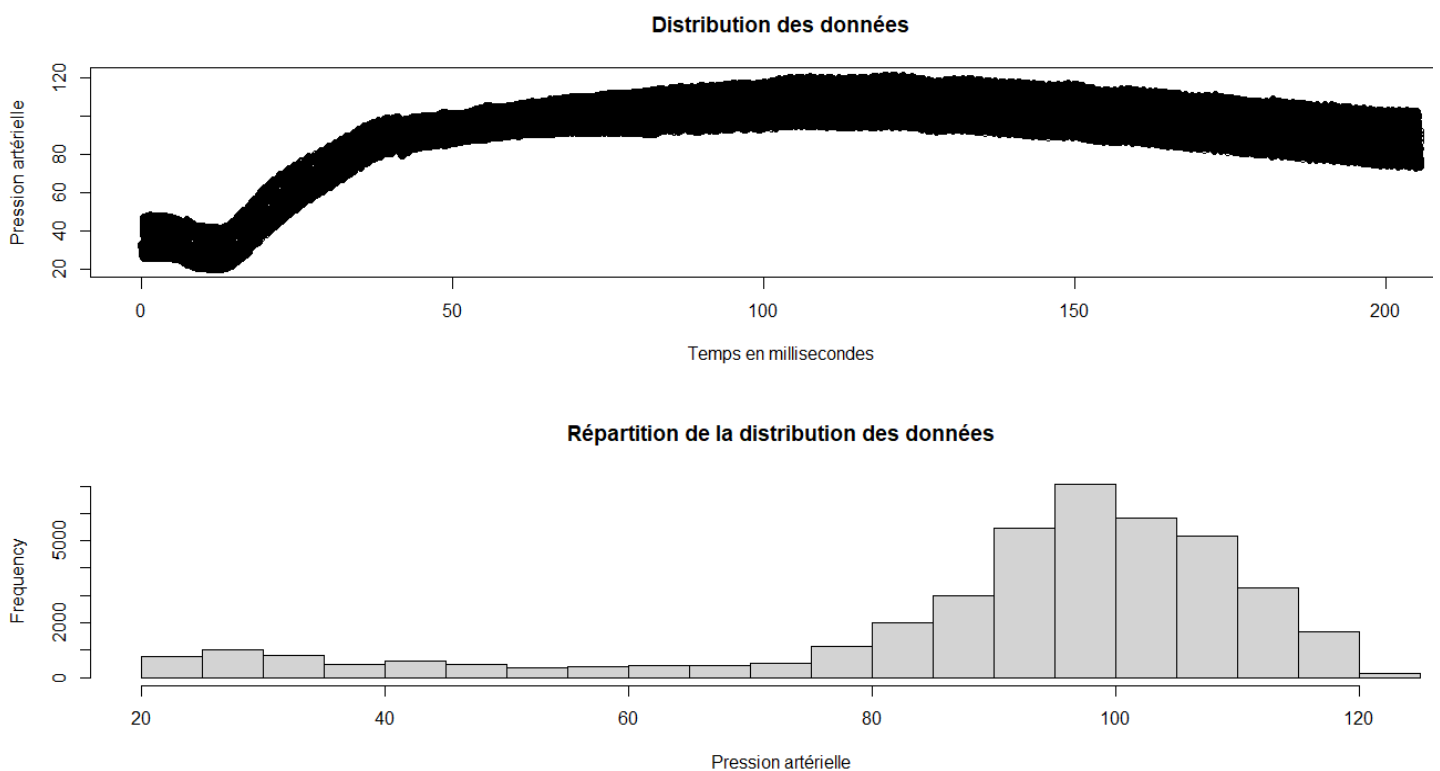


Figure 5 : Double représentation graphique du couple pression artérielle/adrénaline de l'individu 12

Si l'on observe maintenant la **Figure 6**, représentative de la fréquence cardiaque de l'individu 12 avec adrénaline, on remarque que la début de la courbe (de 0 à 20 secondes), représentative de l'état basal, oscille entre 105 et 140 avant d'avoir une "brève" augmentation (de 20 à 30 secondes) jusqu'à 190 puis une oscillation (de 30 à 80 secondes) entre 160 et 180 pour enfin terminer par une lente diminution de 160 à 100 et ce jusqu'à la fin de la série. Si l'on regarde maintenant l'histogramme de cette même figure, on observe que la grande majorité des données se situe entre 100 et 190 de fréquence cardiaque, ce qui est concordant avec la représentation de la courbe, car elle se situe entre 100 et 190.



On remarque cependant la présence d'une valeur anormale entre 230 et 240, visible sur la courbe ainsi que sur l'histogramme. Or une "valeur dite extrême" est une valeur considérée comme impossible au vu du contexte dans lequel elle est présente. Ici, au vu de la différence de fréquence cardiaque, cette valeur n'est pas considérée comme extrême, car elle peut être possible, elle est juste aberrante, c'est-à-dire que "c'est une valeur distante des autres observations effectuées sur le même phénomène" [6].

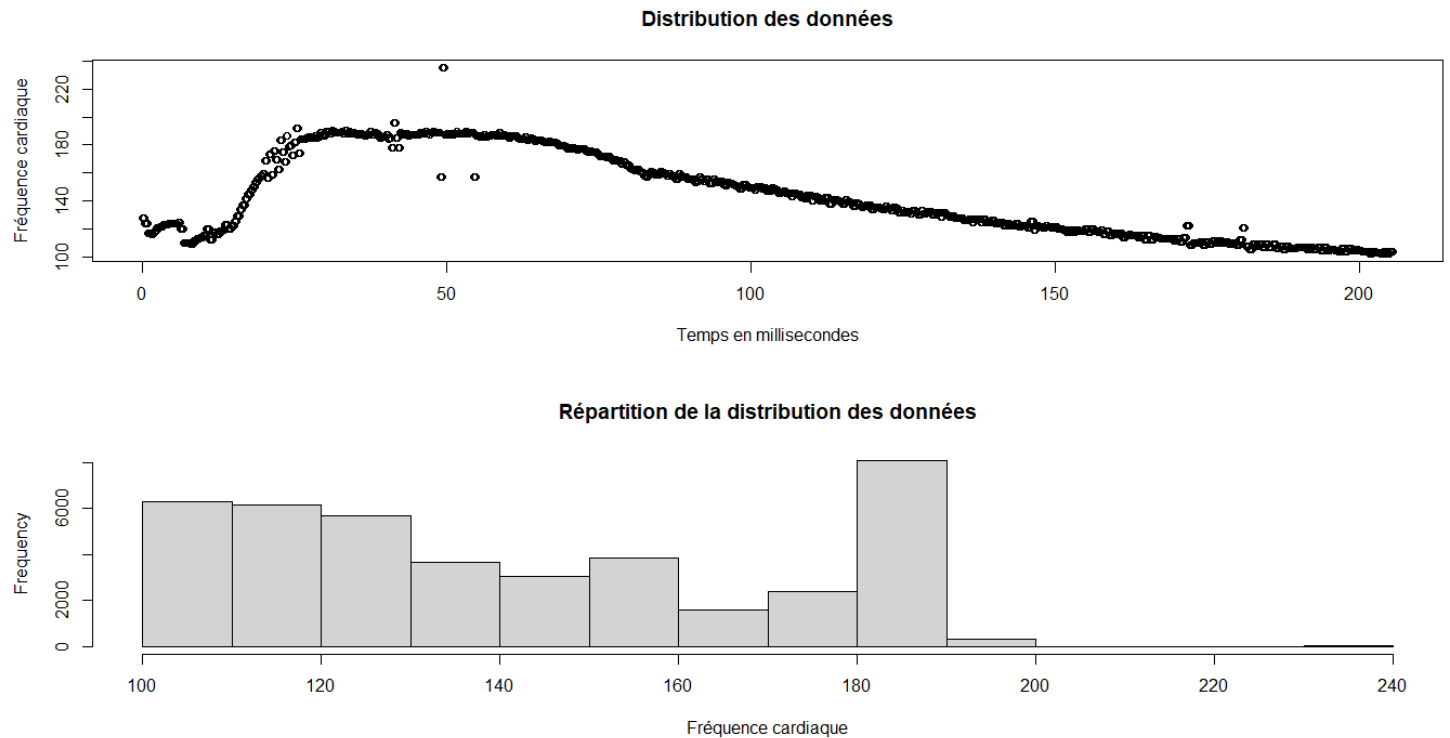


Figure 6 : Double représentation graphique du couple fréquence cardiaque/adrénaline de l'individu 12

Les résultats obtenus pour la diurèse, ici de l'individu 12 avec adrénaline, ont une signification différente des autres variables précédemment interprétées. En effet, comme je l'ai déjà dit dans "Matériel et Méthodes", ce n'est pas l'amplitude des données qui nous intéresse, mais bien la fréquence d'apparition des pics de diurèse. Dans la figure ci-dessous, on observe qu'avant l'injection (de 0 à 20 secondes) et l'effet observable (à 90 secondes) de l'adrénaline, il y a une faible présence de pics avec 2 pics pour 50 secondes. On observe cependant très bien le moment où l'adrénaline injectée fait effet, car la fréquence d'apparition des pics augmente fortement passant de 2 pics/50 secondes à 8 pics/50 secondes pour finir à 16 pics/50 secondes. L'adrénaline augmente donc la diurèse. Si l'on regarde maintenant l'histogramme de cette même figure (cf. **Figure 7**), on remarque bien qu'entre chaque pic il y a un retour à 0 montrant un bon paramétrage de la sonde de mesure et son efficacité de détection. Les résultats peuvent donc être facilement utilisés.

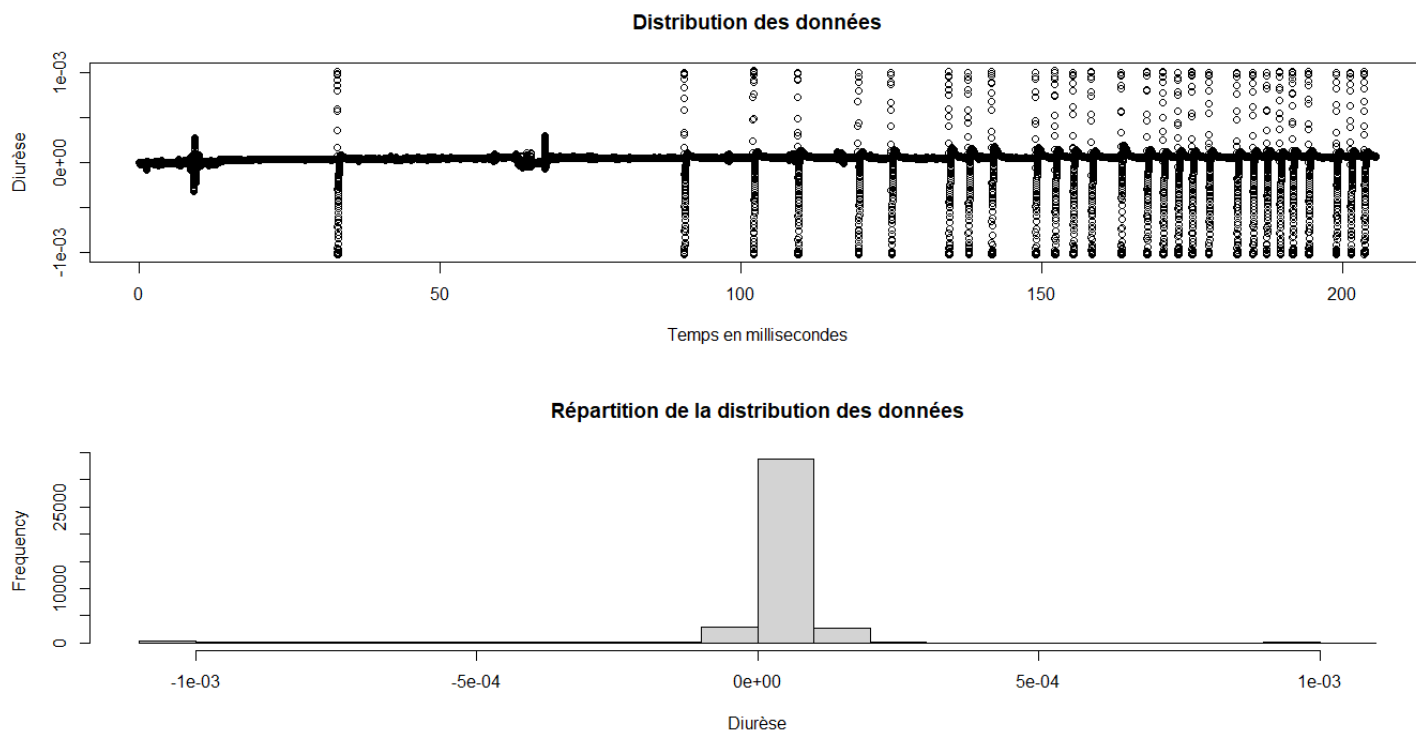


Figure 7 : Double représentation graphique du couple diurèse/adréaline de l'individu 12

Après avoir étudié les différentes représentations des différentes variables, j'ai effectué une analyse temporelle de chaque couple variable/molécules pour chaque individu à l'aide du test Ljung-box avec un alpha de 5%. On observe dans la figure ci-dessous (cf. **Figure 8**), que toutes les p-values sont inférieures à 0.05, ce qui signifie que les résultats sont significatifs et que l'on peut rejeter l'hypothèse nulle. Les résidus ne sont donc pas distribués indépendamment, ils présentent une corrélation en série.

```
$`test PA`  
Box-Ljung test  
data: x$PA  
X-squared = 41032, df = 1, p-value < 2.2e-16
```

```
$`test FC`  
Box-Ljung test  
data: x$FC  
X-squared = 41102, df = 1, p-value < 2.2e-16
```

```
$`test Diurèse`  
Box-Ljung test  
data: x$Diurèse  
X-squared = 40136, df = 1, p-value < 2.2e-16
```

Figure 8 : Résultats du test de Ljung-Box des 3 couples variables/adréaline de l'individu 12

- **Tendance et saisonnalité**

Suite à ces analyses, j'ai pu identifier la tendance et la saisonnalité de mes séries de données. Pour la tendance, j'ai observé par représentation graphique la distribution des moyennes de 6 cycles en fonction du temps. Si la distribution s'effectue dans un intervalle plutôt précis, et que l'on ne distingue pas d'augmentation ou de diminution marquée, alors il n'y a pas de tendance, ce qui est le cas dans la **Figure 9**, représentative de la pression artérielle de l'individu 21 sans injection, où les points représentatifs des moyennes varient entre 52 et 57 de pression artérielle et sont relativement dispersés en "ligne droite".

A contrario, si l'on remarque une augmentation ou une diminution dans un intervalle plus large, alors on peut identifier une tendance, ce qui est le cas dans la **Figure 10**, représentative de la pression artérielle de l'individu 21 avec ocytocine. On observe bien dans cette dernière figure, une augmentation de 48 à 70 (du temps 0 à 80) suivie d'une diminution plus étendue de 70 à 53 (du temps 80 à 350). De plus, les données se situent entre 48 et 70 de pression artérielle, offrant un "large" intervalle.

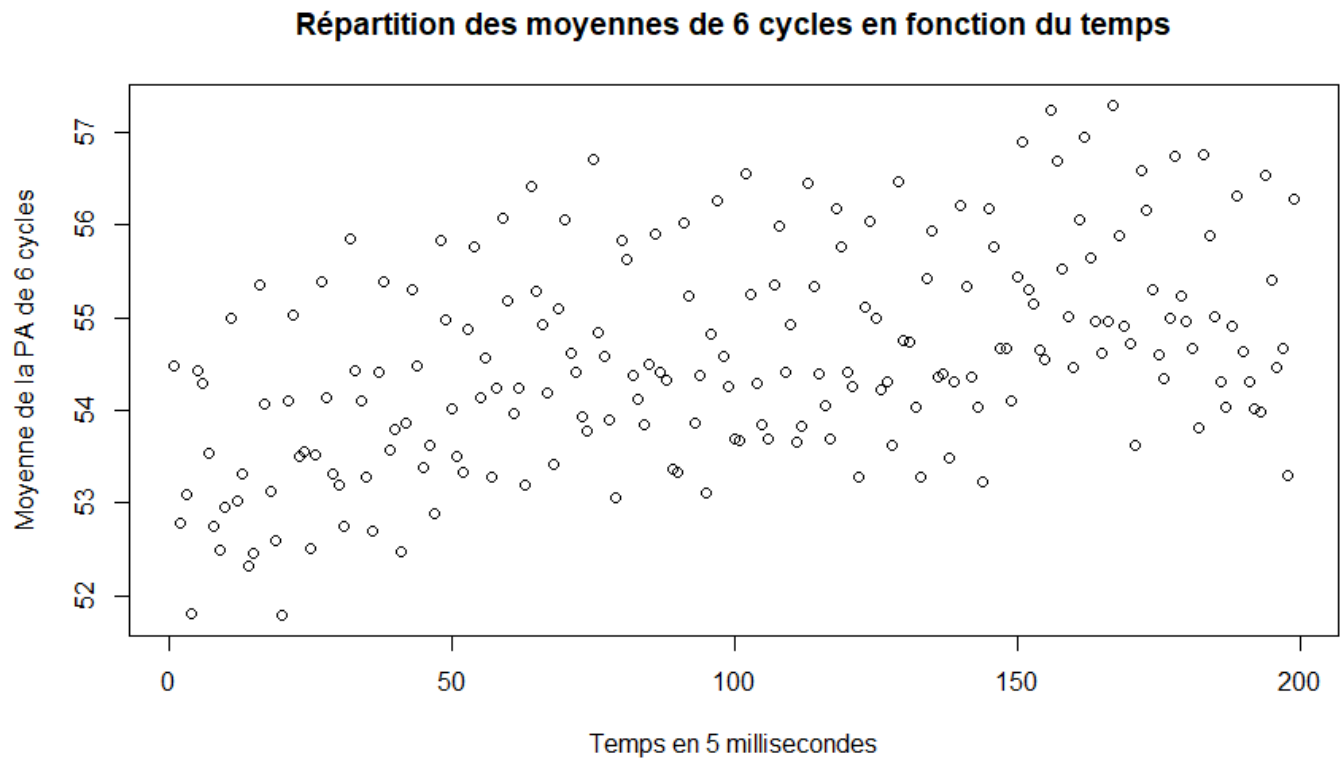


Figure 9 : Répartition des données moyennées du couple pression artérielle/sans injection de l'individu 21 (absence de tendance)

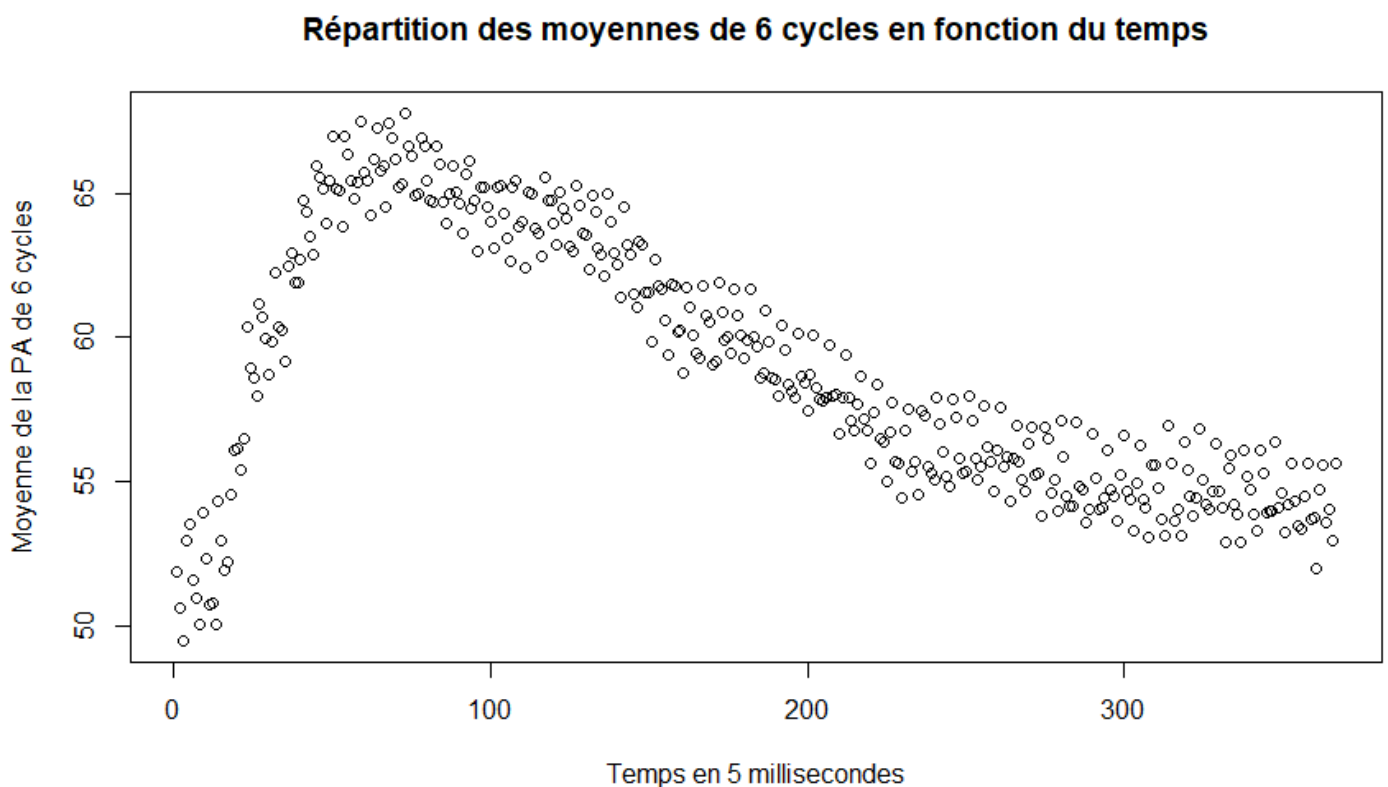


Figure 10 : Répartition des données moyennées du couple pression artérielle/ocytocine de l'individu 21 (présence de tendance)

Pour la saisonnalité, j'ai d'abord réalisé une superposition du 1er et dernier cycle afin d'observer la répétition et la superposition du motif puis j'ai réalisé la superposition de plusieurs cycles pour vérifier que le motif restait le même tout au long de ma série de données. On peut observer dans la **Figure 11**, représentative d'un cycle de pression artérielle pour l'individu 1 sans injection, que le 1er cycle en noir et le dernier cycle en bleu se superposent quasiment parfaitement, le motif semble donc rester le même tout au long de ma série. De plus, on observe sur la **Figure 12**, représentative de la pression artérielle de l'individu 1 sans injection, que les 12 premiers cycles semblent bien se superposer tout comme les 12 seconds (cf. **Figure 13**) semblant confirmer la présence de saisonnalité.

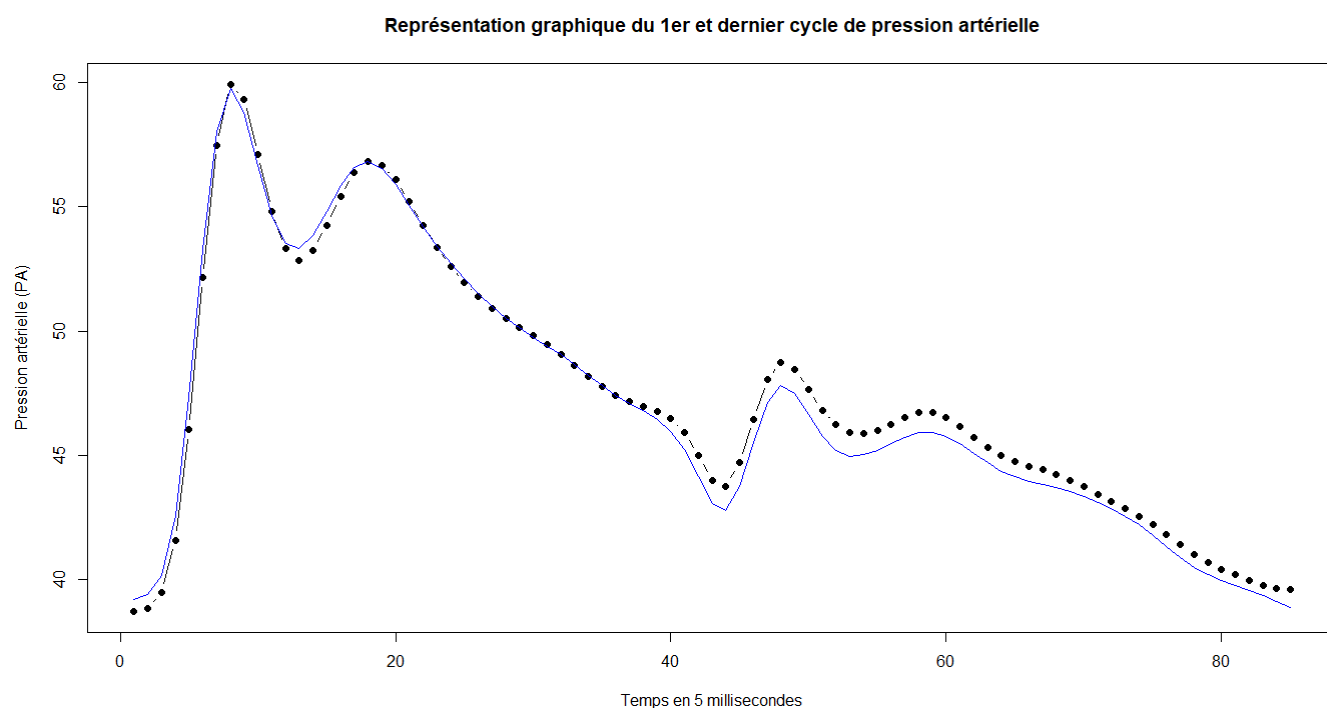


Figure 11 : Superposition du 1er et dernier cycle du couple pression artérielle/sans injection de l'individu 1

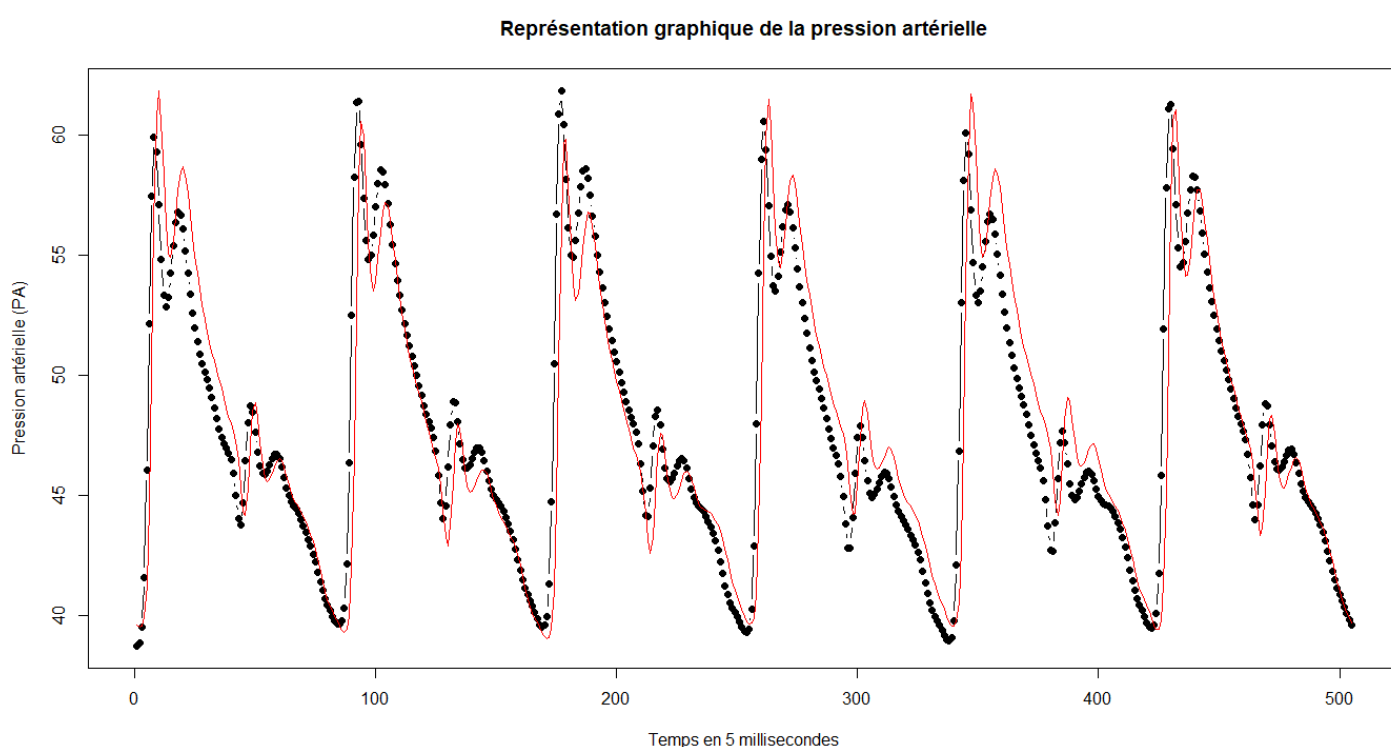


Figure 12 : Superposition des 12 premiers cycles du couples pression artérielle/sans injection de l'individu 1

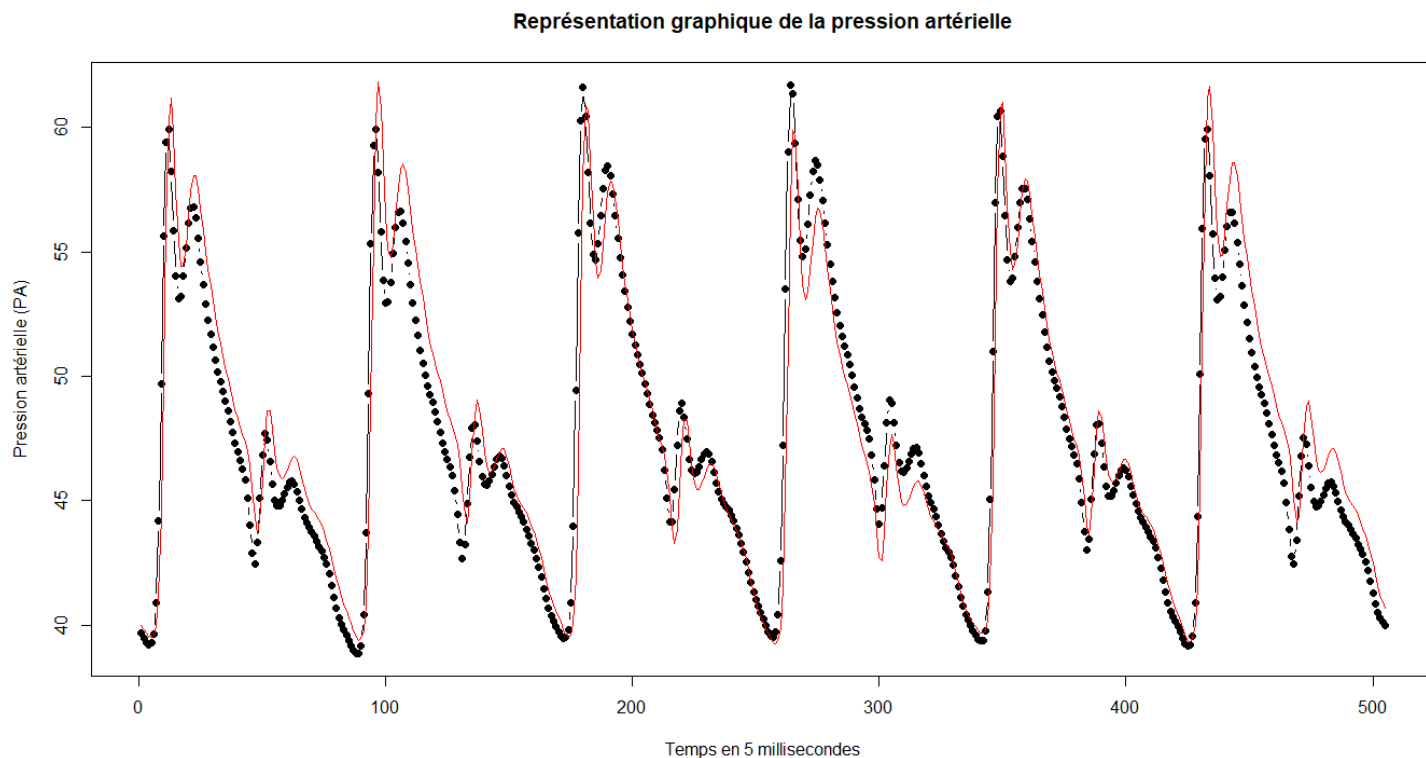


Figure 13 : Superposition de 12 seconds cycles du couple pression artérielle/sans injection de l'individu 1

Enfin, afin d'avoir une véritable confirmation de la présence de saisonnalité, j'ai effectué la fonction acf afin de visualiser l'autocorrélation entre les données d'une même série sur 6 cycles avec un affichage de 3 cycles. On peut alors observer sur la **Figure 14**, représentative de la pression artérielle de l'individu 1 sans injection, qu'il y a effectivement une saisonnalité dans la séquence, car il y a présence d'une corrélation de 100% entre la valeur initiale (à 0) et la valeur à 0, une corrélation de plus de 80% entre la valeur initiale et la valeur à 80 ainsi qu'une corrélation de plus de 60% entre la valeur initiale et la valeur à 170. De plus, on remarque bien la représentation des 3 cycles sur la figure. Ce pourcentage de corrélation et cette représentativité des cycles montrent effectivement la présence de saisonnalité, ce qui a également été confirmé lors de la réalisation de la fonction acf sur d'autres intervalles de 6 cycles. J'ai choisi d'utiliser l'observation de 3 cycles, car l'on peut observer qu'à partir du 4ème cycle la corrélation diminue et arrive aux alentours des 50 % ce qui est moins représentatif de l'autocorrélation des données.

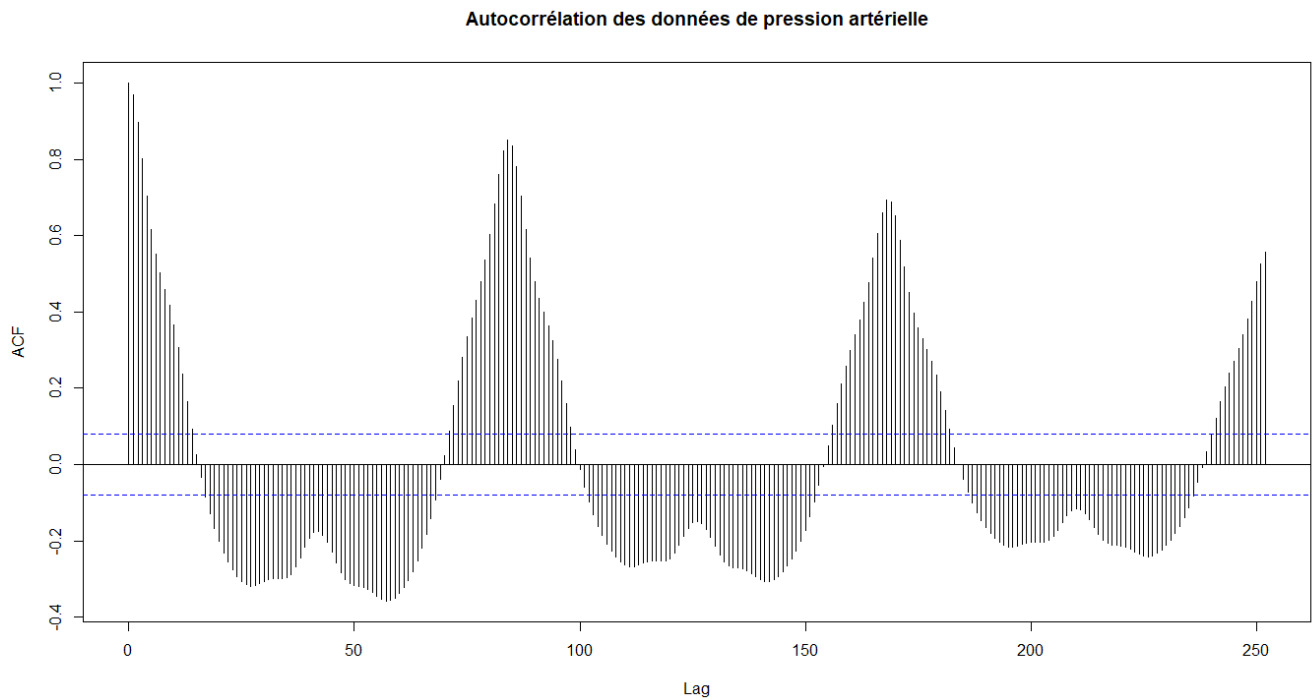


Figure 14 : Résultats de la fonction `acf()` pour le couple pression artérielle/sans injection de l'individu 1

- **Observation de l'enveloppe**

Suite à la vérification de la tendance et de la saisonnalité de chacune de mes séries de données, j'ai ensuite isolé l'enveloppe représentative des maximums et minimums des séries et l'ai représenté graphiquement par superposition des plots afin d'observer l'intervalle de variation des données. On observe dans la figure ci-dessous, représentative de la pression artérielle de l'individu 7 avec acétylcholine, la présence de 2 courbes possédant une tendance similaire : la courbe rouge représentative des maximums et la courbe bleue représentative des minimums. On remarque également que ces 2 courbes possèdent un écart ayant toujours la même grandeur, ce qui signifie que tout au long de ma série des données l'écart entre les valeurs maximales et minimales reste le même, les données sont bien réparties.

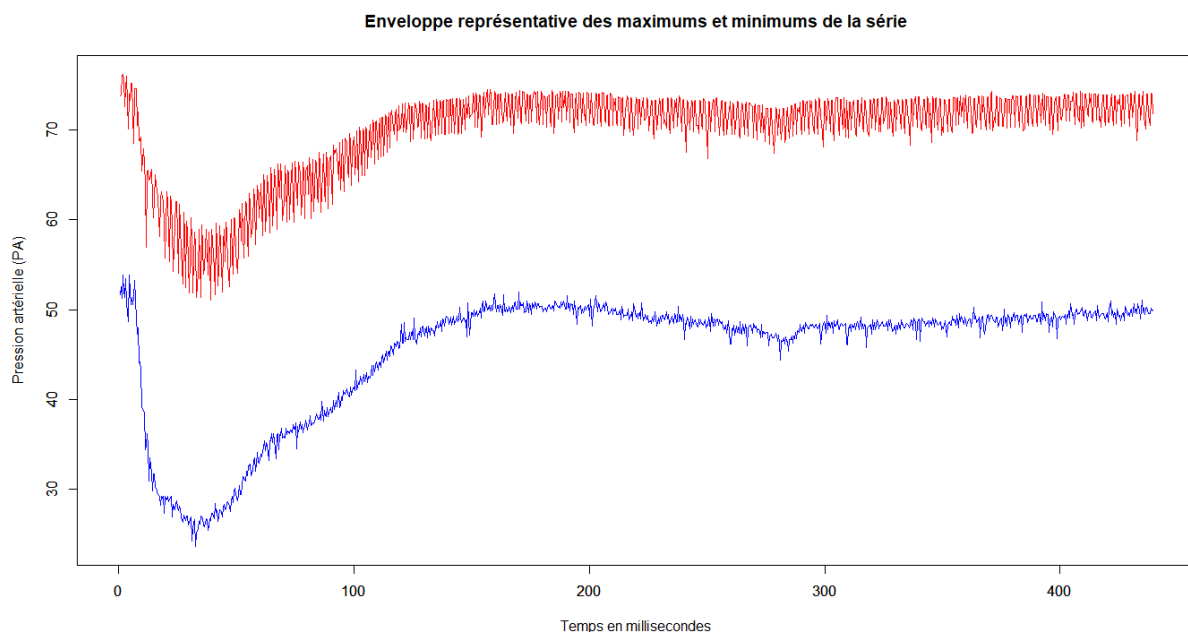


Figure 15 : Représentation graphique de l'enveloppe du couple pression artérielle/acétylcholine de l'individu 7

- **Diurèse et fréquence d'apparition**

La dernière étape “d’analyse” des variables, a été la définition de la fréquence d’apparition des pics de la variable diurèse. Pour la réaliser, j’ai tout d’abord voulu observer le nombre de pics présents sur les 6 premiers cycles, ainsi que la hauteur maximale des pics, à l’aide d’une représentation graphique. On peut observer dans la figure ci-dessous, que pour l’individu 13 avec desmopressine, il y a présence d’un seul pic sur les 6 premiers cycles avec une hauteur maximale de 0,001.

Suite à cela, j’ai appliqué mon code (cf. **Figures 16 et 17**) pour déterminer le nombre de pics présents tous les 6 cycles, ainsi que leur fréquence d’apparition également tous les 6 cycles, directement dans un tableau comme dans la **Figure 17**. En comparant les 2 figures, on observe bien que dans le tableau il y a un seul pic sur les 6 premiers cycles. Cette vérification peut également être confirmée lorsque l’on observe les autres cycles (6 par 6).

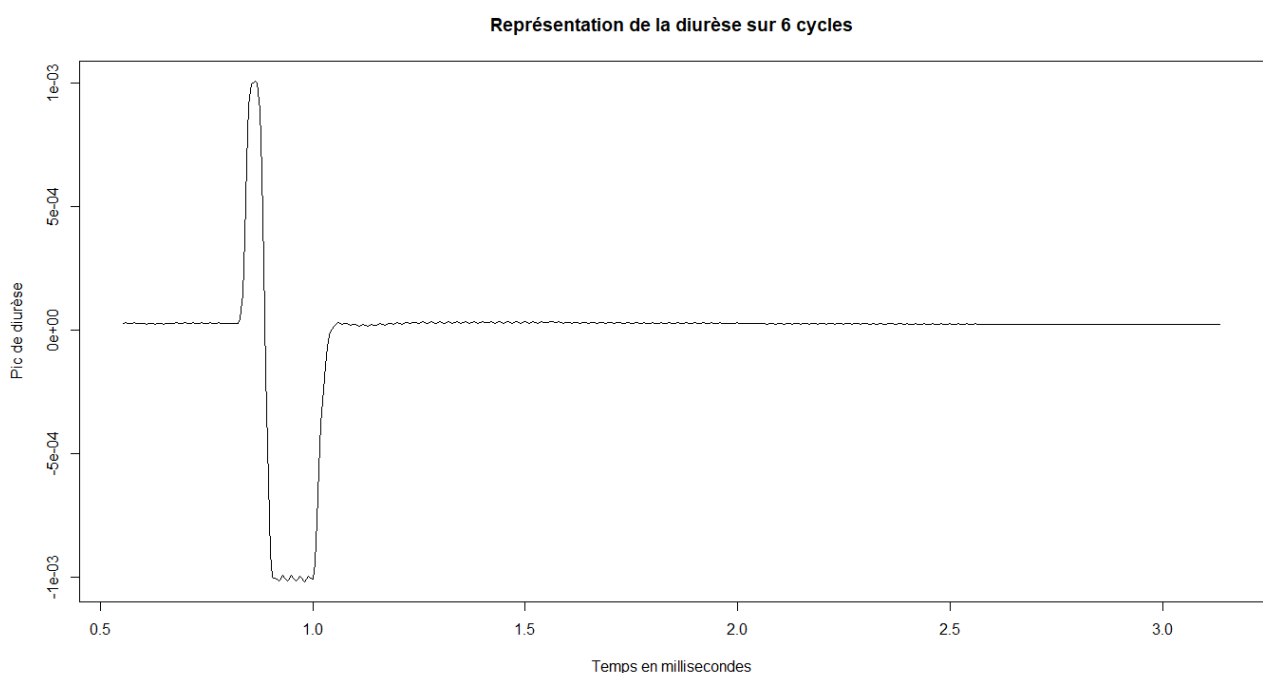


Figure 16 : Représentation du nombre de pics de diurèse sur 6 cycles de l’individu 13 avec desmopressine

▲	PERIODE ▾	HAUTEUR ▾	NB ▾	FREQ ▾
1	0.43	0.00100591	1	0.3875969
2	0.86	0.00000000	NA	NA
3	1.29	0.00000000	NA	NA
4	1.72	0.00000000	NA	NA
5	2.15	0.00000000	NA	NA
6	2.58	0.00000000	NA	NA
7	3.01	0.00100106	2	0.7751938
8	3.44	0.00000000	NA	NA
9	3.87	0.00100106	NA	NA
10	4.30	0.00000000	NA	NA
11	4.73	0.00000000	NA	NA
12	5.16	0.00000000	NA	NA
13	5.59	0.00000000	2	0.7751938

Figure 17 : Dataframe du nombre de pics de diurèse et de leur fréquence d’apparition sur 6 cycles de l’individu 13 avec desmopressine

• Classification DTW

La classification a été ma dernière étape ainsi que mon objectif final. Pour pouvoir la réaliser, j'ai tout d'abord identifié l'individu avec le moins d'observations pour réaliser mon tableau nécessaire à cette dernière. Ici, c'est l'individu 2 qui possède le moins de données avec 7400 observations. Suite à cela, j'ai identifié qu'elles étaient les molécules injectées à chaque individu et j'ai regroupé les résultats dans le tableau ci-dessous.

Individus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sans injection	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Acétylcholine	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Adrénaline	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Angiotensine II	X	X		X	X			X	X	X	X			X	X	X	X	X	X	X	X
Ocytocine	X	X	X	X	X	X	X	X	X	X	X	X			X	X	X	X	X	X	X
Desmopressine			X	X	X	X		X	X	X	X		X		X	X	X		X	X	X

Figure 18 : Identification des molécules injectées selon l'individu

Après avoir effectué ces relevés de données, et créé les tableaux nécessaires, j'ai pu effectuer la classification par dtw à l'aide de mon code (cf. **Figure Annexe 7**). On observe que, dans la figure ci-dessous représentative de la classification de la diurèse des individus avec desmopressine, nous avons bien la présence de 5 clusters d'individus.

Le cluster 1, contenant les individus 10, 9, 11, 3, 5, 8, 6, 4 et 13, possède une diurèse qui sur les 7400 premières observations n'a pas de pics. Le cluster 1 contient également l'individu 16 possédant une diurèse avec des pics plutôt espacés et peu nombreux avec un intervalle peu régulier.

Le cluster 2, contenant les individus 21 et 19, possède une diurèse avec de nombreux pics répartis en 2 groupes plus ou moins similaires. Ce qui est intéressant dans ce cluster, c'est que l'on remarque que les individus 21 et 19 possèdent une séquence identique pouvant provenir d'une même collecte donnée ou d'un duplicata de séquence. La classification permet donc un test de vérification et de similarité des données.

Le cluster 3, contenant l'individu 20, possède une diurèse avec de nombreux pics et un intervalle non-régulier.

Le cluster 4, contenant l'individu 17, possède une diurèse avec plusieurs pics et un intervalle plus ou moins régulier.

Enfin le cluster 5, contenant l'individu 15, possède une diurèse avec de nombreux pics de forme différente ainsi qu'un intervalle plus ou moins régulier.

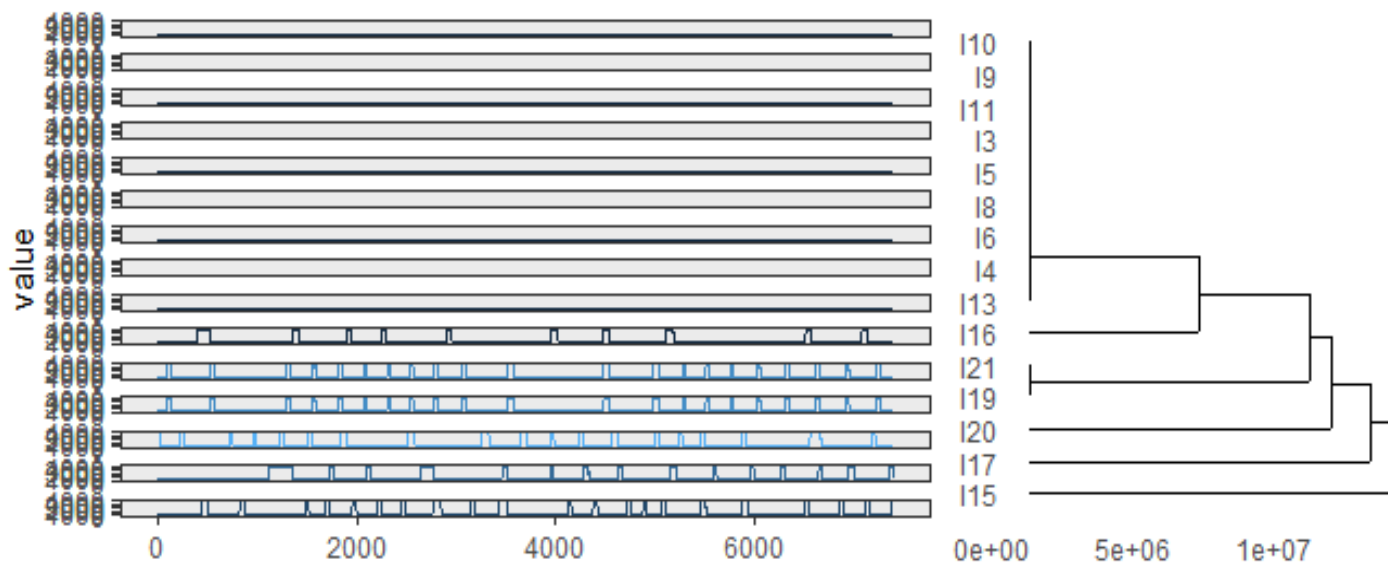


Figure 19 : Classification par DTW du couple diurèse/desmopressine des différents individus

Conclusion / Discussion

• Conclusion du stage

A la fin de ce stage, j'ai réussi à réaliser 18 classifications de données (6 molécules x 3 variables) sur un échantillon de 21 individus. Ces classifications, en plus d'avoir un effet de clustering, permettent un test de vérification et de similarité des données.

• Difficultés rencontrées

Durant la réalisation de ce stage, j'ai rencontré plusieurs grosses difficultés. La première difficulté rencontrée a été la détermination de la saisonnalité. En effet, étant donné la taille des séries, il était difficile de visualiser la présence ou non de motifs se répétant à intervalle régulier. J'avais tout d'abord utilisé la fonction *decompose* qui me permettait directement de décomposer ma série en 4 courbes : données observées, tendance, saisonnalité et données aléatoires. Cependant, cette fonction n'était pas adaptée à mon projet et les résultats obtenus n'étaient pas exploitables. A l'aide de mon responsable de master, Mr. HARDOIN Jean-Benoît, j'ai donc pris contact avec Mme VIBET Marie-Anne, enseignante-chercheuse possédant des connaissances approfondies sur l'analyse des séries temporelles. Après plusieurs échanges et une visioconférence, nous avons déterminé que, pour vérifier la présence de saisonnalité dans mon projet, il serait plus efficace de passer par une comparaison des motifs représentatifs de plusieurs cycles et de réaliser une ACF (Auto-Correlation Function).

Ma deuxième grosse difficulté a été la détermination de la fréquence diurèse. En effet, je suis passée par plusieurs méthodes de codage et de calcul (cf. **Figure Annexe 8**) mais aucune n'était vraiment aboutie et ne convenait aux résultats souhaités.

- **Points à améliorer**

Le travail effectué dans le cadre de mon stage n'est qu'un prémice du travail nécessaire à la réalisation de ce projet. En effet, plusieurs choses peuvent être améliorées : l'automatisation de certaines "fonctions", le lissage des séries afin d'obtenir des données subissant moins les autres paramètres biologiques comme par exemple la mesure de la pression artérielle pouvant être impactée par la respiration de l'animal incluant une autre saisonnalité à la série de données ou encore tenter une autre approche pour l'analyse de la variable diurèse.

- **Conclusion personnelle**

Ce stage a été pour moi très instructif, d'une part par le fait de manipuler des données réelles et biologiques issues de TP réalisées plus ou moins récemment ; et de l'autre, par le fait de manipuler de "nouveaux matériaux" non vus en cours, en TD ou en TP. Cette dernière partie m'a permis de réaliser les recherches et de répondre à la plupart de mes questions par moi-même ainsi que d'élargir ma base de connaissance dans le logiciel RStudio et ce qu'il est possible de faire avec. J'ai découvert et testé de nombreuses fonctions, fonctionnant plus ou moins bien, et j'ai pu interpréter les résultats obtenus de ces dernières.

Durant ce stage, j'ai également eu des contacts avec Mme VIBET Marie-Anne par le biais de mon référent de master Mr HARDOIN Jean-Benoît. Elle m'a notamment permis de comprendre l'utilité de la fonction acf et d'identifier les meilleurs moyens d'observer la tendance et la saisonnalité de mes séries. Elle a également eu la gentillesse et la patience, de m'accorder de son temps notamment lors d'une visio-conférence pour que cela soit plus simple à comprendre lors des explications et échanges.

Bibliographie / Sitographie

- [1] <https://www.youtube.com/watch?v=5mi0oyz8Dks>
- [2] <https://www.statology.org/ljung-box-test/>
- [3] <http://www.jybaudot.fr/Previsions/trend.html>
- [4] <https://www.lokad.com/fr/d%C3%A9finition-saisonnalit%C3%A9>
- [5] <https://damien-datasci-blog.netlify.app/post/time-series-clustering-with-dynamic-time-warp/>
- [6] https://fr.wikipedia.org/wiki/Donn%C3%A9e_aberrante

Annexes

Figure A1 : Code de recodage du temps

Lignes 252 à 324 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A2 : Représentation d'un cycle diastole-diastole

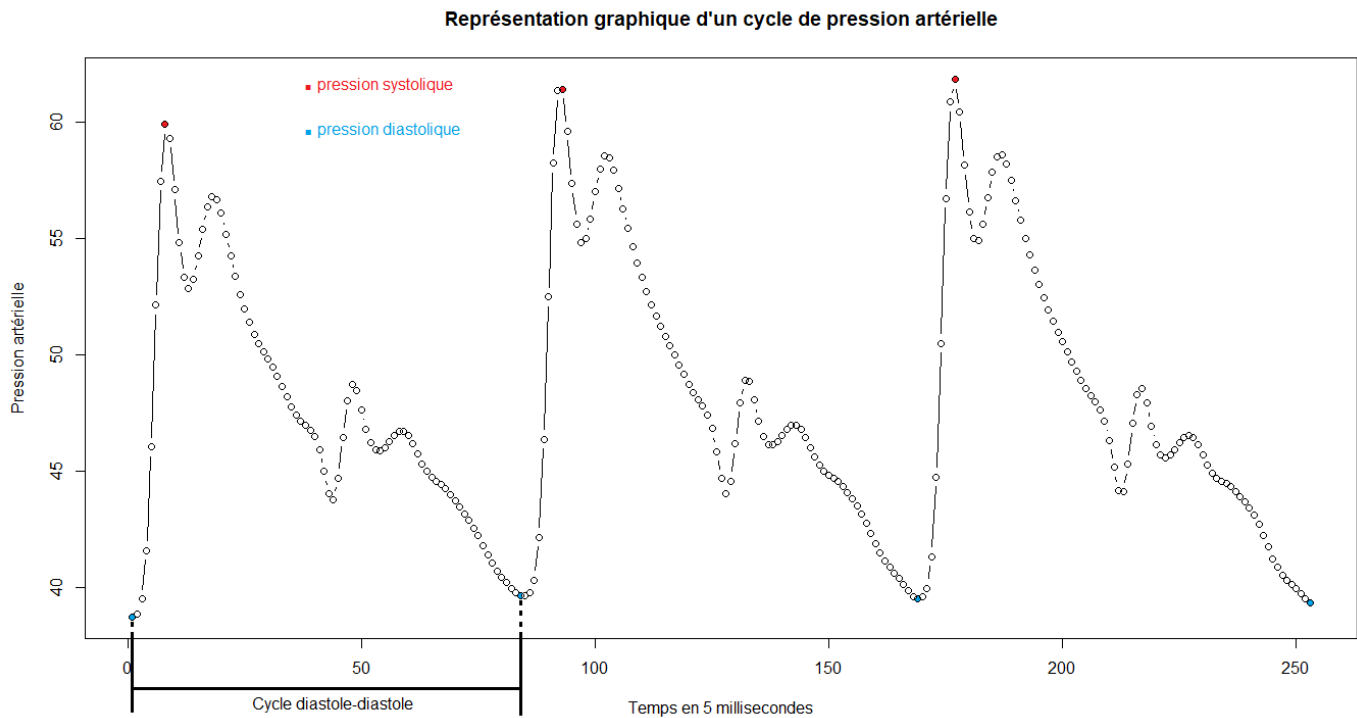


Figure A3 : Code de l'isolement du 1er cycle diastole-diastole

Lignes 382 à 675 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A4 : Code de corrélation par acf() et superposition des graphiques

Lignes 756 à 808 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A5 : Code d'obtention de l'enveloppe

Lignes 703 à 755 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A6 : Code du calcul de la fréquence diurèse


Lignes 809 à 878 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A7 : Code utilisé pour la classification par dtw

Lignes 938 à 1166 du document “BARBEY_Cassandra_M1BB_Script optimisé TER.R”

Figure A8 : Méthodes de calcul de la fréquence diurèse

1ère méthode

Mon objectif était d'obtenir la fréquence des pics présents lors de la diurèse. Pour cela, j'ai d'abord cherché à obtenir la distance entre 2 s. J'ai donc réalisé une boucle afin d'obtenir la position du 1er pic et une deuxième boucle pour obtenir la position du 2ème pic puis j'ai calculé la distance (le nombre de valeurs) présente entre les 2. J'ai ensuite voulu récupérer toutes les valeurs maximales représentatives des pics de diurèse. Pour cela, j'ai tout d'abord créé un tableau où je pourrais enregistrer mes points maximums, puis j'ai créé une boucle me permettant d'extraire ces points et de les enregistrer dans mon tableau.

A partir de là, j'ai calculé ma fréquence en divisant la taille de mon tableau contenant les points avec la taille de mon tableau contenant les données de ma série initiale.

Cette méthode n'est cependant pas très précise, car il se trouve que la distance entre 2 pics varie et ce tout au long d'une même série de données. J'ai donc essayé de chercher une autre méthode.

2ème méthode

Mon objectif était d'obtenir la fréquence des pics présents lors de la diurèse. Afin d'avoir plus de précision dans mes estimations, j'ai décidé d'obtenir la fréquence des pics tous les 6 cycles diastole-diastole.

J'ai tout d'abord fait la représentation graphique des 6 premiers cycles afin d'obtenir le nombre de pics présents sur cette période. J'ai ensuite cherché à automatiser le calcul du nombre de pics tous les 6 cycles à l'aide d'une boucle. Pour obtenir le nombre de pics, j'ai pensé à faire un `table()` pour obtenir le nombre de valeurs (représentatives des pics) étant supérieures ou égales à la médiane de mon dataframe contenant toutes les valeurs maximales de la diurèse (`DIU_AS_SI`). Cependant, j'ai vite remarqué que cela ne convenait pas, car je n'obtiens qu'une seule valeur alors que j'avais 2 pics sur mon plot. J'ai donc décidé de faire varier ma limite et ai essayé d'utiliser les quantiles. J'ai réussi à l'aide du quantile 45% à obtenir 2 pics, le seul souci est qu'il n'y a pas d'automatisation et que ce quantile ne fonctionne pas pour toutes les périodes, il faut donc trouver le bon à chaque fois, ce qui est une grosse perte de temps.

Code 2ème méthode :

```
plot(AS_SI$Diurèse[c(num:(num+(84*6)))]  
table(AS_SI$Diurèse[c(num:(difference*6))]>=quantile(DIU_AS_SI$MAX,probs=c(0.45)))
```