Mesure physiologique de joueurs de jeu vidéo (1)

Première partie : exploration et nettoyage des données

Godefroy Clair

Monday, July 13, 2015

Contents

```
Loading required package: knitr
Loading required package: plyr
Loading required package: ggplot2
Loading required package: reshape2
Loading required package: kohonen
Loading required package: class
Loading required package: MASS
```

1 L’expérience dont sont issues les données

Au cours de la période allant du 14 février 2014 au 9 mars 2014, dans le cadre de sa thèse sous la direction de Eric Gressier-Soudan, Viviane Gal a réalisé 67 expériences expériences au près de 24 individus. L’objectif de la thèse en question est de permettre l’établissement d’une cartographie des émotions à partir de biosignaux et d observations. Il est donc nécessaire de réaliser des mesures qui permettent de poser une classification associant les signaux et des états émotionnels. Ces mesures ont été réalisées au cours d’expériences mettant les participants aux prises avec différents média dont on peut supposer qu’elles suscitent chez eux un certain nombre d’émotions. Avec les mesures en question, l’objectif est de décrire ces émotions à l’aide d’outils statistiques, graphiques et autres algorithmes d’apprentissage.

1.1 Description de l’expérience

Durant chaque expérience, différentes tâches sont proposées aux participants : - Regarder un film : une succession d images et d extraits de films - Un test de concentration / méditation en utilisant les ondes cérébrales - Une expérience à base de sensations fortes, de stress et d’“ambiance de peur” - jouer à un jeu de stratégie pendant une durée de 20 minutes

Chaque expérimentation permet la capture de données de différents types (physiologiques, images, réponses). Les données physiologiques, mesurées à intervalle de temps régulier (31 millisecondes) consistent en un certain nombre de mesures réalisées par le biais de capteurs sur un individu pendant l’expérience. Les données physiologiques mesurées sont : 1. la température cutanée (T°) en degré celcius 2. la fréquence cardiaque (FC) en battement par minute 3. l’activité electrodermale (GSR) en microsiemens 4. la respiration (FR) en pourcentage de déformation du thorax.

1.2 Les appareils de mesure

Différents éléments ont été utilisés pour l'expérience : -3 capteurs sans fil, -1 convertisseur analogique numérique, -1 logiciel de capture.

Tout ces appareils/logiciels sont fournis par la société TEA. Les capteurs font partie de la gamme **T-Sens**. Nous allons maintenant les décrire plus en détail.

1.2.1 Le capteur sans fil "CFM" (réf. C2030)

Selon les spécifications fournies par le constructeur, il s'agit d'un "module à transmission sans fil pour la mesure de la fréquence cardiaque grâce à une ceinture thoracique"[1]. Voir image

Figure 1: Image du capteur CFM

Le tableau suivant est aussi fourni dans le document :

Caractéristiques du capteur :

Caractéristique	**Valeur**
Etalonnage	Pas d étalonnage
Unité	BPM
Plage mes.	10 220 BPM
Fréquence	16Hz
Résolution	1 BPM
Compatibilité	Polar Non codé ou équivalent

1.2.2 Le capteur sans fil "GSR + T°" (réf. C2034)

Toujours selon le constructeur, il s'agit d'un "module à transmission sans fil combinant à la fois la mesure de la conductivité électrodermale et de la température cutanée" - voir image . Il mesure l'activité electrodermale et la température à la surface de la peau grâce à "deux électrodes [qui] se fixent à l extrémité de deux doigts"[2]. En réalité, afin de ne pas gêner le participant dans ses mouvements, les expérimentateurs ont décidé - avec l'accord du fabriquant - de faire la mesure *aux doigts de pied*. L'objectif concernant la mesure de la conductivité électrodermale est de mesurer l'intensité du courant électrique à la surface de la peau, ce qui est généralement vu comme une bonne approximation de la transpiration.

[1] Voir http://teaergo.com/site/sites/default/files/docs/TSens_CFM-cardio_FR_V1.8.pdf
[2] Voir http://teaergo.com/drupal/sites/default/files/docs/TSens_FSR_FR_V1.8.pdf

Figure 2: Image du capteur GSR

Caractéristiques du capteur :

Caractéristique	**GSR**	**T°**
Nombre de voies	1	1
Etalonnage	Pas d étalonnage	Pas d étalonnage
Unité	µS (Siemens)	°C
Fréquence	32Hz	
Résolution	16 Bits	0.05°C
Plage de mesure	0 -30µS	-40°C à 120°C
Linéarité	N/A	0.5°C
Précision	N/A	0.5°C
Caractéristiques électriques :		
Alimentation Accumulateur	190mAh	
Caractéristiques mécaniques :		
Dimensions	52mm x 25mm x 14mm	
Longueur câbles + électrodes	200mm	
Poids	20g	
Conditions d utilisation		
T°	0°C à 40°C	
Humidité	< 60%	

1.2.3 Le capteur sans fil “Respirations” (réf. C2033)

le capteur est un “module à transmission sans fil pour la mesure des mouvements thoraciques ou abdominaux. Le module (ceinture inclue) fournit des informations permettant d analyser le rythme respiratoire et l amplitude des respirations”[3]. Voir image

[3]Voir http://teaergo.com/drupal/sites/default/files/docs/TSens_Respi_FR_V1.8.pdf

Figure 3: Image du capteur de respiration

Caractéristique	Valeur
Fréquence	32Hz
Résolution	0.01%
Caractéristiques électriques :	
Alimentation Accumulateur	190mAh
Autonomie	8h
Temps de chargement	3h
Caractéristiques mécaniques :	
Dimensions	52mm x 25mm x 14mm
Poids	20g
Longueur ceinture	1m
Plage de mesure	0-75%
Allongement maximal	70mm (75%)
Conditions d utilisation :	
T°	0°C à 40°C
Humidité	< 60%

1.3 Les données mesurées

1.3.1 La température cutanée

C'est la température à la surface du corps, prise par le biais d'un capteur aux orteils (pieds nus). Le fabriquant a expliqué lors d'une conversation avec Viviane Gal que cette température est bien influencée par la température du corps mais aussi par la température environnante. Il sera nécessaire de procéder à des mesures complémentaires pour mesurer l'impact théorique de la température ambiante sur la température fournie par le capteur et d'évaluer leur influence sur les mesures obtenues.

1.3.2 La respiration

La réspiration est mesurée par une ceinture abdominale sur laquelle est disposé un capteur mesurant la nombre de battement par minute. Là encore, lors de l'expérience, certains effets parasites ont pu perturber la bonne mesure de la respiration. En effet, la ceinture se place autour de l'abdomen et se serre à l'aide de crans. Or, il n'y a pas eu ni vérification sur le cran choisi, ni mesure de la circonférence de l'abdomen. Cela remet en question la possiblité de faire des comparaisons entre individus. En effet, si la ceinture était très serrée pour une personne par rapport à une autre, le même mouvement de l'abdomen pourrait ne pas donner la même mesure. Il est donc important de faire des mesures complémentaires pour vérifier l'influence de ce phénomène.

1.3.3 La fréquence cardiaque

La fréquence cardiaque est mesurée par une ceinture abdominale qui "sonde" la fréquence cardiaque par le biais d'un capteur.

1.3.4 L'activité electrodermale

L'activité electrodermale mesure l'insensité du courant à la surface de la peau. Elle donne une approximation de la transpiration d'une personne. La prise de mesure au pied doit là aussi amener à réfléchir à la pertinence des mesures. La transpiration au pied est-elle pertinente dans l'exercice en question ?

2 Les données brutes

Nous avons reçu de Viviane Gal un fichier par expérience. Ce fichier est un tableau au format “csv” comportant 5 colonnes : - la date au format “J/M/AAAA H:Mn:S.m” (J : jour, M : mois, A : année, H : heure, Mn : minutes, S : secondes, m : millisecondes) - la respiration : décimal positif ou négatif, 3 chiffres après la virgule - l’activité électrodermale : décimal positif ou négatif, 3 chiffres après la virgule - la température : décimal positif, 3 chiffres après la virgule - la fréquence cardiaque : entier positif

La figure 1 représente les premières lignes des données pour l’expérience AB.

	A	B	C	D	E	F
1	date	respiration	activite_electrodermale	temperature	frequence_cardiaque	
2	06/03/2014 15:19:29.756	-10,644	-44,157	29,417	74	
3	06/03/2014 15:19:29.787	-10,379	-44,157	29,417	74	
4	06/03/2014 15:19:29.819	-10,168	-44,157	29,417	74	
5	06/03/2014 15:19:29.850	-9,981	-44,157	29,417	74	
6	06/03/2014 15:19:29.881	-9,833	-44,169	29,417	74	
7	06/03/2014 15:19:29.912	-9,692	-44,169	29,417	74	
8	06/03/2014 15:19:29.944	-9,575	-44,169	29,417	74	
9	06/03/2014 15:19:29.975	-9,482	-44,169	29,417	74	
10	06/03/2014 15:19:30.006	-9,427	-44,169	29,417	74	
11	06/03/2014 15:19:30.037	-9,435	-44,169	29,417	74	
12	06/03/2014 15:19:30.069	-9,396	-44,169	29,43	74	
13	06/03/2014 15:19:30.100	-9,365	-44,169	29,417	74	
14	06/03/2014 15:19:30.131	-9,365	-44,169	29,417	74	
15	06/03/2014 15:19:30.162	-9,373	-44,169	29,417	74	
16	06/03/2014 15:19:30.194	-9,435	-44,169	29,417	74	
17	06/03/2014 15:19:30.225	-9,412	-44,169	29,417	74	
18	06/03/2014 15:19:30.256	-9,443	-44,169	29,417	74	
19	06/03/2014 15:19:30.287	-9,435	-44,169	29,417	74	
20	06/03/2014 15:19:30.319	-9,443	-44,169	29,43	74	
21	06/03/2014 15:19:30.350	-9,466	-44,169	29,417	74	
22	06/03/2014 15:19:30.381	-9,497	-44,169	29,417	74	
23	06/03/2014 15:19:30.412	-9,497	-44,169	29,417	74	
24	06/03/2014 15:19:30.444	-9,521	-44,169	29,417	74	

Figure 4: exemple d’un tableau de données reçu

Les fichiers qui nous ont été donnés pour le moment sont au nombre de 10. Le nom de chaque fichier est une concaténation des initiales de la personne ayant fait le test avec un numéro si la personne a fait plusieurs expériences et avec le terme “SurEchantillonHF”. Par exemple, pour l’individu A.B., le fichier sera “AB_SurEchantillonHF.csv”.

2.1 Modification “à la main”

Pour pouvoir importer les données de manière simple et “automatisée”, il faut uniformiser le contenu des fichiers. Par exemple, tous les fichiers ne contiennet pas obligatoirement les noms des variables, nous avons donc ajouter des nouveaux fichiers avec le nom des variables inséré dans la première ligne en pré-traitement. Nous avons ajouté des dates factices lorsqu’elle manquaient en “copiant-collant” les dates à partir du fichier “AB”(l’important étant que les intervalles de temps soit de 9 ms).

Le nom de ces fichiers est le nom du fichier original avec le suffixe “_header” avant le format.

3 Importation, vérification et nettoyage des données sous 'R'

Nous avons décidé de traiter les données sous le logiciel 'R' très pratique pour gérer des données au format csv, qui a surtout comme avantage : - d'être, ainsi que ces bibliothèques, sous licence libre (généralement GPL-2 ou 3[4], - d'avoir de bonnes bibliothèques pour les algorithmes d'apprentissage, de réseaux de neurones, etc., - d'avoir de bonnes bibliothèques de modélisations graphiques (notamment *ggplot2*, très flexible)

3.1 Bibliothèques requises pour le travail sous 'R'

3.2 Importation des données

3.2.1 L'objet choisi pour l'analyse sous R : le data frame

Les données qui sont traitées sous le logiciel "R" ont été mises sous la forme d'un *data frame*. Les *data frames* sont de simples tables de valeurs où chaque colonne correspond à un suite de valeur d'une variable d'un type donné (entier, booléén, chaîne de caractère,...) et chaque ligne a un n-upplet (une instance) de chacune des variables. Chaque colonne a de plus un nom et chaque ligne peut aussi en avoir un.

Dans cette partie, nous avons créé autant de "data frame" qu'il y a d'expériences réalisées (donc 10). Dans chaque data frame, une ligne correspond à un instant d'une expérience et chaque colonne indique les différentes mesures.

Nous utilisons une fonction load.file pour charger les données en mémoire dans des data frames (dont le nom sera data. concaté aux initiales de l'individu, par exemple data.AB pour l'expérience avec l'utilisateur A.B.)

3.3 Tranformation des données (nettoyage, mises en forme, décomposition, marquage)

chargement des données au format numérique (sauf 1ère colonne au format "date", POSIX). Les noms des variables sont : -date, -respiration, -activite.electrodermale, -temperature, -frequence.cardiaque

3.4 Vérification des données

Dimension des data frames :

Data frame	dimension
"AB"	48803 x 7
"DA"	27622 x 7
"CW"	48804 x 7
"FS1"	49518 x 7
"HL"	48762 x 7
"LM"	48719 x 7
"PCo"	48608 x 7
"PCo2"	27047 x 7
"PCo3"	11787 x 7
"ST"	48775 x 7
"CLP"	48719 x 7
"DE"	48803 x 7

[4]Voir http://www.r-project.org/Licenses pour plus de détails.

Représentation des 5 premières données pour l'expérience AB pour exemple :

```
[1] "     date       respiration  activite.electrodermale  temperature  frequence.cardiaque"
[2] "--------------------- ------------ ------------------------ ------------ --------------------"
[3] "2014-03-06 15:19:29      -10.644               -44.157      29.417                74"
[4] "2014-03-06 15:19:29      -10.379               -44.157      29.417                74"
[5] "2014-03-06 15:19:29      -10.168               -44.157      29.417                74"
[6] "2014-03-06 15:19:29       -9.981               -44.157      29.417                74"
[7] "2014-03-06 15:19:29       -9.833               -44.169      29.417                74"
[8] "2014-03-06 15:19:29       -9.692               -44.169      29.417                74"
```

Y a-t-il des données manquantes ?

Data frame	dimension
"AB"	aucune
"DA"	aucune
"CW"	aucune
"FS1"	aucune
"HL"	aucune
"LM"	aucune
"PCo"	existe(s)
"PCo2"	aucune
"PCo3"	aucune
"ST"	aucune
"CLP"	aucune
"DE"	aucune

Il existe effectivement des données manquantes pour PCo :

```
     date respiration activite.electrodermale temperature
NA   <NA>          NA                      NA          NA
NA.1 <NA>          NA                      NA          NA
     frequence.cardiaque quart.temps nom.experience
NA                    NA        <NA>           <NA>
NA.1                  NA        <NA>           <NA>
```

A quelles lignes et quelle(s) colonne(s) ?

```
      row col
[1,] 22340   2
[2,] 22345   2
```

Vu qu'il ne s'agit que de deux données sur plus de 27000, nous proposons de copier la valeur précédente à la place de la donnée manquante.

```
#remplace la valeur manquante par la valeur précédente de la même colonne
data.PCo[l.na[1],ncol.1] <- data.PCo[l.na[1]-1,ncol.1]
data.PCo[l.na[2],ncol.2] <- data.PCo[l.na[2]-1,ncol.2]
```

4 Exploration graphique et statistique des données

Nous allons utiliser la bibliothèque ggplot2[5] développée par Hadley Wickham et Winston Chang et qui implémente la théorie de la "grammaire des graphiques" de Leland Wilkinson[6] sous 'R'. Cette théorie permet de construire des graphiques de manière modulaire et en décomposant les différentes étapes de création d'un graphique. Ce qui permet 1) de facilement modifier des graphiques 2) de mentalement avoir une meilleure idée du processus de création d'un graphique et des différents choix, combinaisons disponibles.

4.1 A propos des outils graphiques de visualisation utilisés

L'objectif principal est de se donner une idée de la distribution et de la dispersion des données. En particulier, nous voulons comparer les valeurs mesurées à celles annoncées par le constructeur.

Nous allons représenter les données par expérience en utilisant plusieurs outils permettant d'améliorer la qualité des graphiques :

- nuage de points type "jitter" (ou simplement Jitter) : sur la base d'un 'scatter plot' classique, chaque donnée est déplacée horizontalement et verticalement selon une valeur déterminée - pour chaque donnée - par un tirage aléatoire uniforme dans un intervale). L'avantage de ce graphique est de donner facilement une idée de la densité des données. Cela est souvent particulièrement intéressant lorsque les données sont à la limite entre le discret et le continue ou lorsqu'elles sont condensées autour de certaines valeurs.

Exemple : Créons un ensemble de 1000 points dont l'abcisse est compris entre 1 et 10 et choisi par tirage uniforme et l'ordonnée est une fonction linéaire de l'abscisse auquel s'ajoute un "bruit" variant entre 0 et 5 suivant une loi du chi-deux. Cela a permet de donner un effet "longue traîne" au nuage de points.

```
set.seed(7)
nb.data=1000
x <- sample(1:10, nb.data, TRUE)
y <- 3*x + rchisq(nb.data, 0, 2)
```

Avec un "scatter plot" habituel, nous obtenons le graphique suivant :

```
qplot(x,y)
```

Par construction, l'abscisse d'un point ne peut prendre qu'un nombre limité de valeur, beaucoup de données sont donc "superposées". Il est difficile de se donner une idée de la variation de densité dans le nuage de point. En ajoutant un "bruit" uniforme à l'abscisse dans un intervale appropriée (ici 0.2), "jitter"" nous permet d'avoir une meilleure idée du nuage de point.

```
qplot(x,y,position = position_jitter(w = 0.2))
```

-"Alpha" : ce paramètre est aussi utiliser pour réduire la perte d'information due à la superposition de données. l'idée est de donner une certaine transparence aux formes représentant chaque donnée. Lorsque ces formes sont superposées sur le graphique, la couleur des pixels correspondant à la partie superposée sera plus foncée.

En poursuivant notre exemple précédant, ajoutons un niveau de transparence à notre premier graphique :

[5] Voir notamment http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
[6] Voir http://www.springer.com/us/book/9780387245447

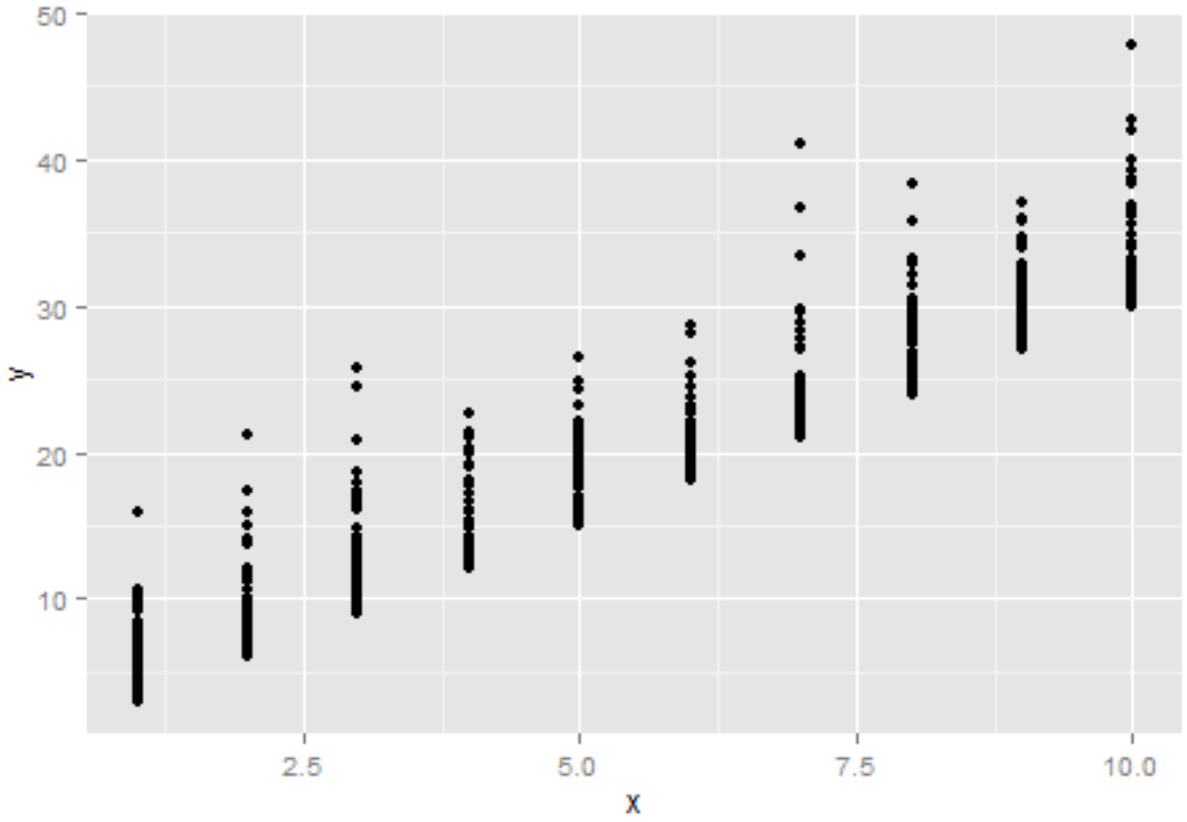


Figure 5:

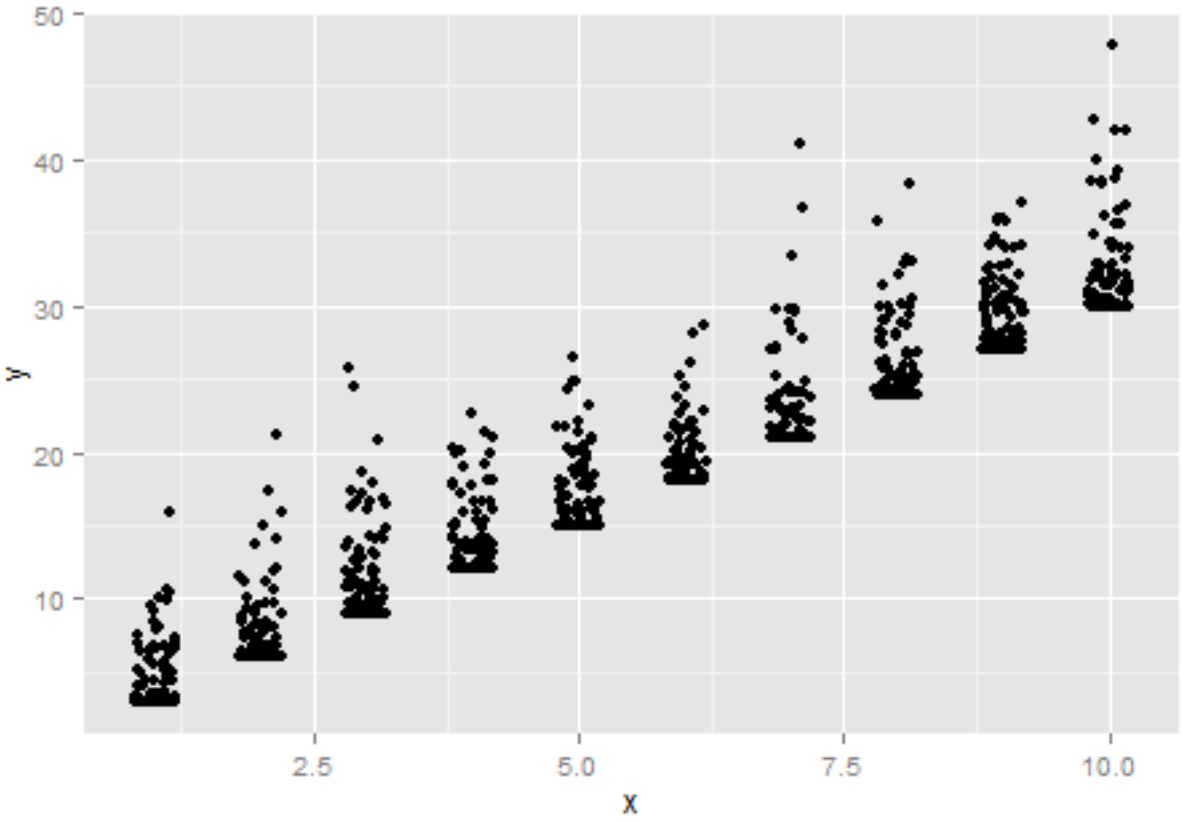


Figure 6:

```
qplot(x,y,alpha=I(0.25))
```

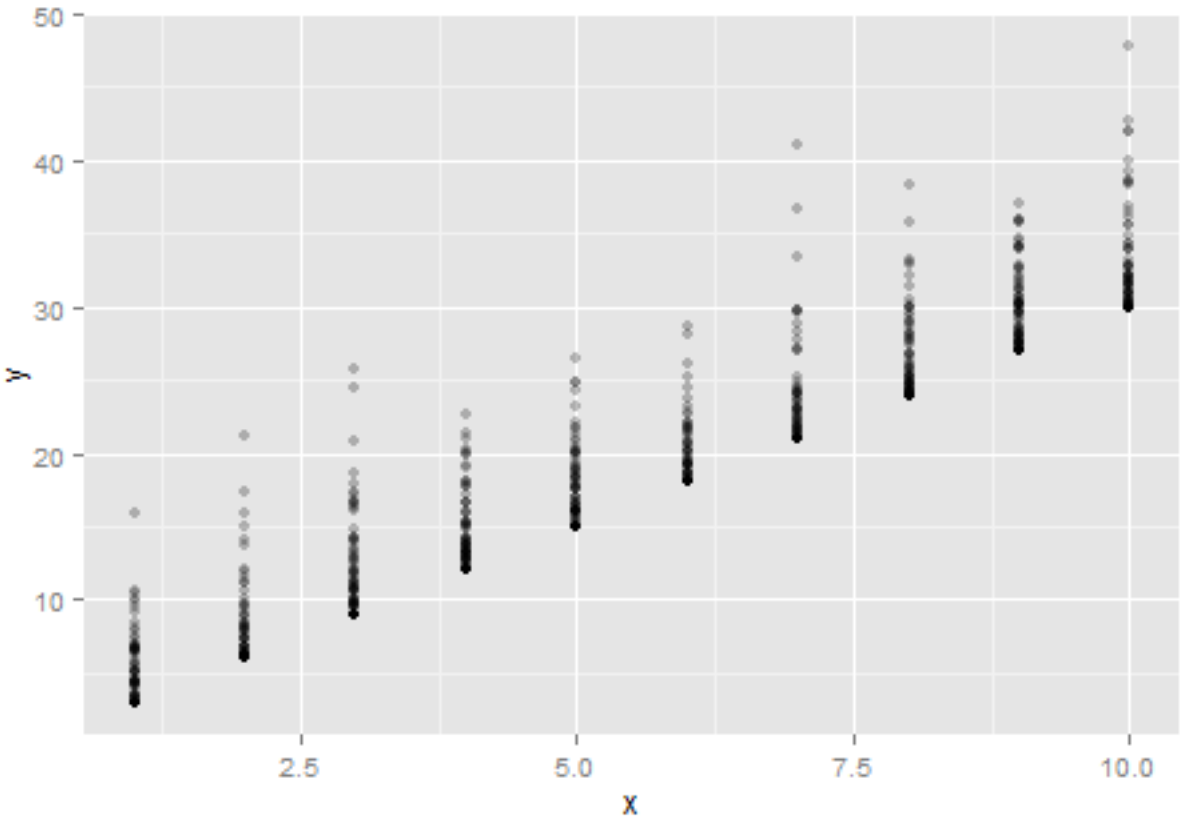


Figure 7:

Notons qu'en combinant "Alpha" à d'autres options des graphiques, il peut être utilisé dans d'autres contextes. Par exemple, lorsqu'on représente plusieurs nuages de points sur un même graphique, on peut utiliser alpha et les couleurs pour distinguer des nuages de points en partie superposés.

```
y1 <- 3*x + rchisq(nb.data, 0, 2)
y2 <- 3*x + rchisq(nb.data, 0, 2)+1
df = as.data.frame(list(rep(x,2),c(y1,y2),c(rep("groupe1",1000),rep("groupe2",1000))))
colnames(df) <- c("x","y", "groupe")
p1 <- ggplot(df, aes(x=x, y=y, color=groupe)) + geom_point(alpha=I(1)) + ggtitle("Nuage de points pour
p2 <- ggplot(df, aes(x=x, y=y, color=groupe)) + geom_point(alpha=I(.4)) + ggtitle("Avec \"alpha\"") + s
p3 <- ggplot(df, aes(x=x, y=y, color=groupe)) + geom_point(position = position_jitter(w = 0.1, h = 0.1)
multiplot(p1,p2,p3,cols = 1)
```

- "Boxplot" ou "boîte à moustache", c'est une manière beaucoup plus fréquente de représenter un ensemble de données numériques. Dans un graphique dont l'échelle horizontale est mise en correspondance avec l'ensemble des valeurs des données, on dessine un rectangle allant horizonalement du premier quartile au troisième quartile et coupé par la médiane en son milieu. Les premier et neuvième déciles sont représentés par deux traits reliés au rectangle par deux segments. Enfin, chacune des données non comprises entre ces déciles est représentées par des points (notamment les maxima et "outliers" ou données aberrantes) sur le graphique. La grande simplicité et concisions que les boxplots apportent

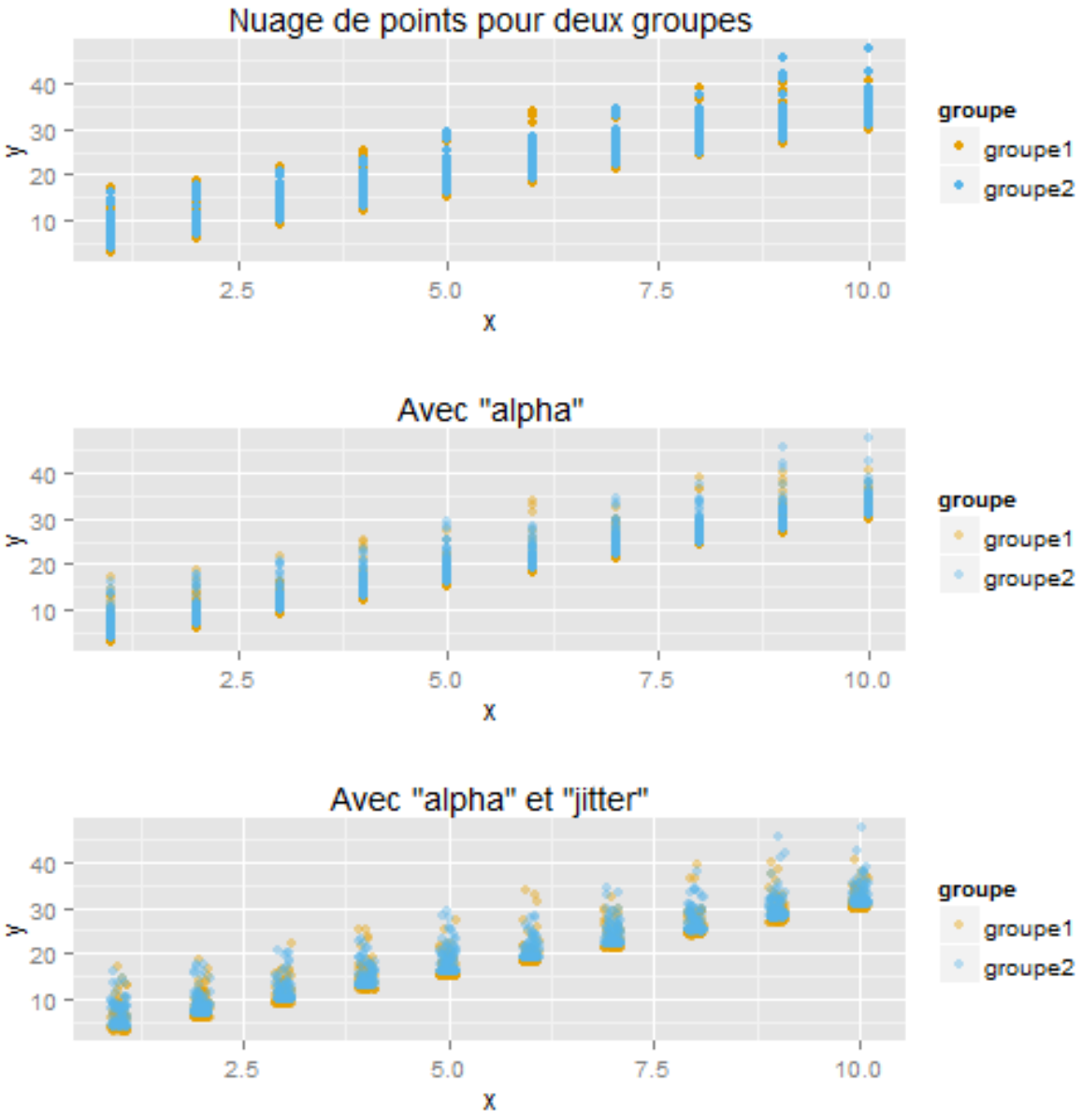

Nuage de points pour deux groupes
y
x
groupe
groupe1
groupe2
Avec "alpha"
Avec "alpha" et "jitter"

Figure 8:

sont particulièrement intéressants pour procéder à des comparaisons entre différentes distributions de données. Pour se donner une première idée de la forme des donnes et procéder à des comparaisons entre expériences, une combinaison de graphique de type “Jitter” et “Boplot” nous semble particulièrement adaptée.

4.2 Représentations graphiques des variables par expériences

Pour chaque variable successivement, nous allons donc proposer un nuage de points type “jitter” et un “boxplot” afin d’observer si la variation et le spectre des données correspond bien à ce qui est a priori attendu.

4.2.1 respiration : graphiques “Boxplot” & Nuage de points type “jitter”

Les deux graphiques suivant représentent les données mesurées pour la respiration (en ordonnée) par expérience (en abscisse, chacun des expérience).

Comme nous l’avons vu, le constructeur annonce des valeurs entre 0% et75%, or nous voyions que les données vont en tout de -20 à 20. Ces données ne semblent donc pas compatibles. Lors d’une discussion à ce propos avec des représentants du fabriquant des capteurs, ceux-ci nous ont expliqué que les données négatives devaient être inversées. Mais il reste un autre problème : certaines expériences ont des valeurs à la fois positives et négatives (ce que montre la figure 5). Des analyses et tests supplémentaires sont donc nécessaires avant de pouvoir exploiter ces données. On peut aussi noter que l’expérience ‘LM’, les données sont pratiquement constantes pendant toute l’expérience. Ce que confirme le fait que sur `rlength(data.LM$respiration)`, 46765 sont constantes. Comme les données sont au millième de %, cela ne peut s’expliquer que par un problème lors de la prise de mesure.

4.2.2 activité electrodermale : graphiques “Boxplot” & “Jitter”

Les deux graphiques suivant représentent les données mesurées pour l’activité electrodermale (en ordonnée) par expérience (en abscisse, chacun des expérience).

Comme nous l’avons déjà dit, la plage de valeur annoncée par le constructeur est 0-30 µS. Or nous voyions ici des valeurs allant de -30 à -50. Selon un représentant du fabriquant, cela pourrait être du à un décallage fortuit du à l’algorithme du logiciel utilisé pour capter les mesures. Selon lui, il faudrait décaller les données de +69. Cela semble assez cohérent avec nos observations. Par ailleurs, pour un certain nombre d’expériences, nous avons à nouveau une variation insuffisante qui laisse penser que certaines mesures sont imparfaites. En particulier ‘HL’ (une valeur unique : -48.936) et ‘ST’ (nombre de valeurs : 1) mais ‘DA’ est aussi douteux.

Par, un phénomène intéressant est l’existence d’une “traîne” vers le haut pour la plupart des expériences : il serait intéressant de regarder à quel momment c’est traîne est créée.

4.2.3 température: graphiques “Boxplot” & “Jitter”

Les deux graphiques suivant représentent les données mesurées pour la température (en ordonnée) par expérience (en abscisse, chacun des expérience).

Concernant la température, on voit que la dispersion des données est généralement faible. Les spécifications du capteurs annoncent un spectre de mesure de -40°C à 120°C, avec une précision de 0.05°C. Les données sont toutes bien à l’intérieur de cette plage. On remarque en particulier sur le “box plot” que les données sont (à quelques “outliers” près) très rapprochées. Cela correspond ici à ce qui est attendu puisque la mesure au corps ne varie pas beaucoup. Ainsi, une étude récente a donné lieu à une mesure de la température d’une personne selon trois méthodes (buccale, anale et cutannée) et la variation moyenne de la température du

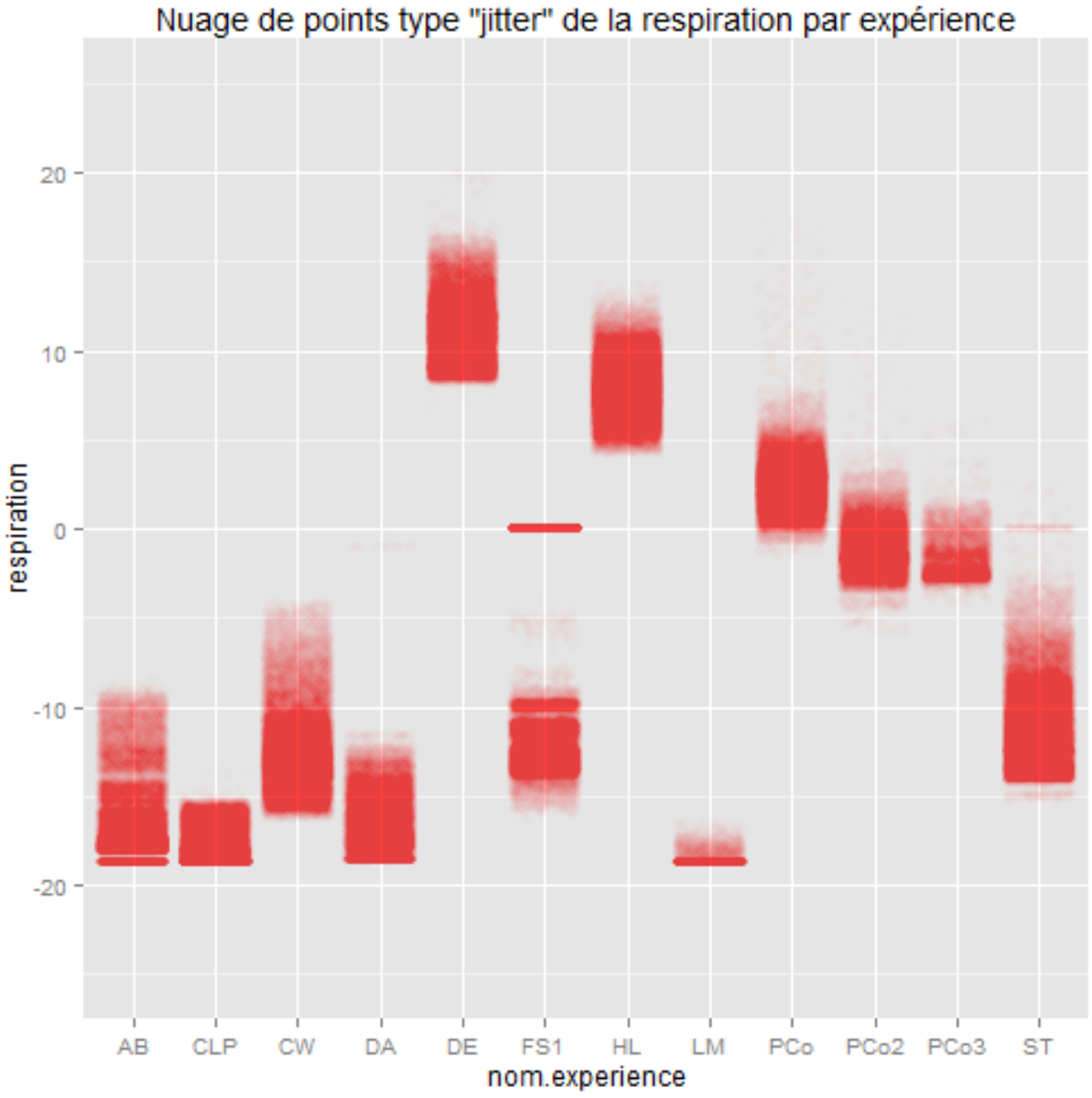

Nuage de points type "jitter" de la respiration par expérience
respiration
20
10
0
-10
-20
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 9:

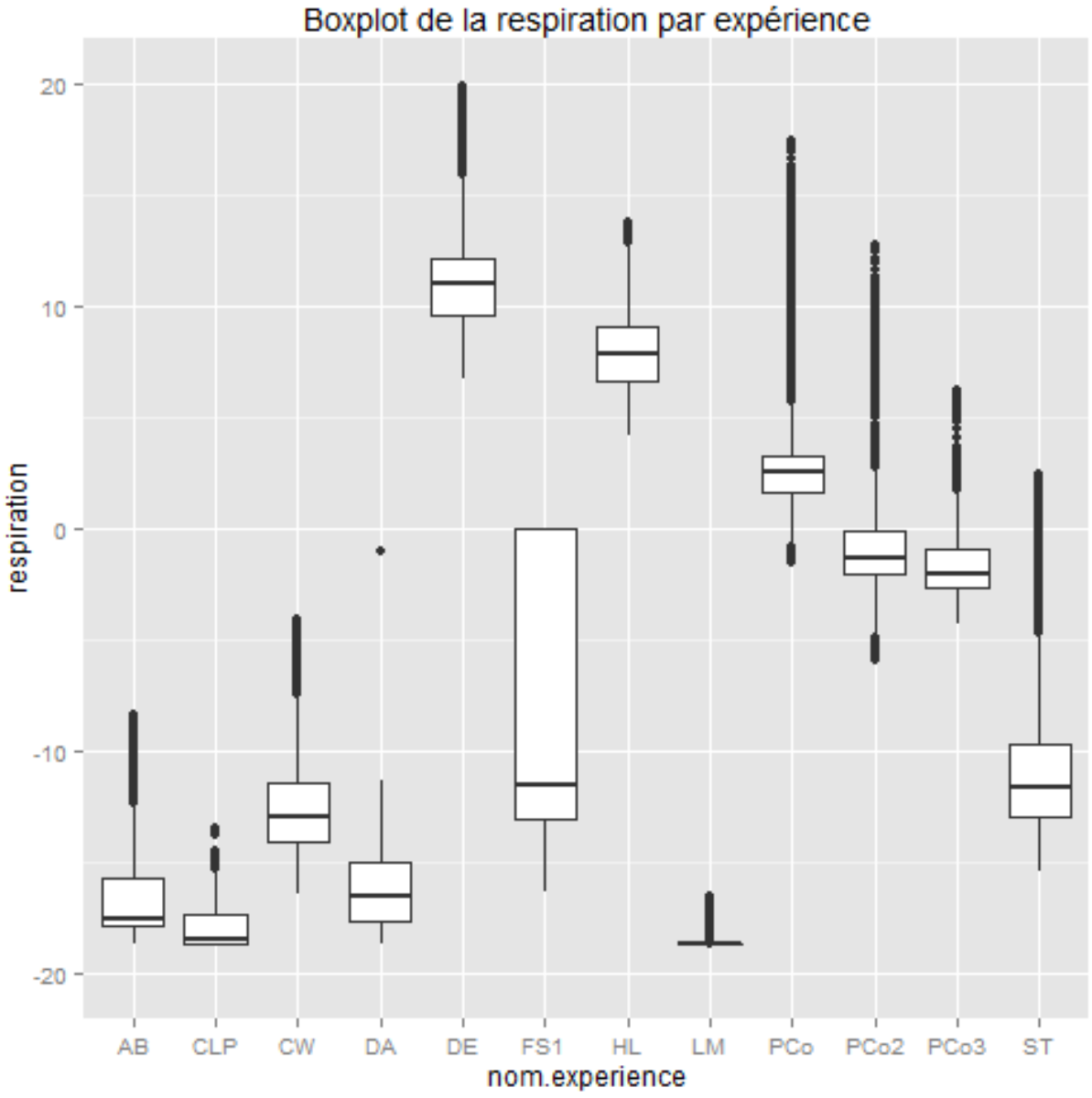

Boxplot de la respiration par expérience
respiration
20
10
0
-10
-20
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 10:

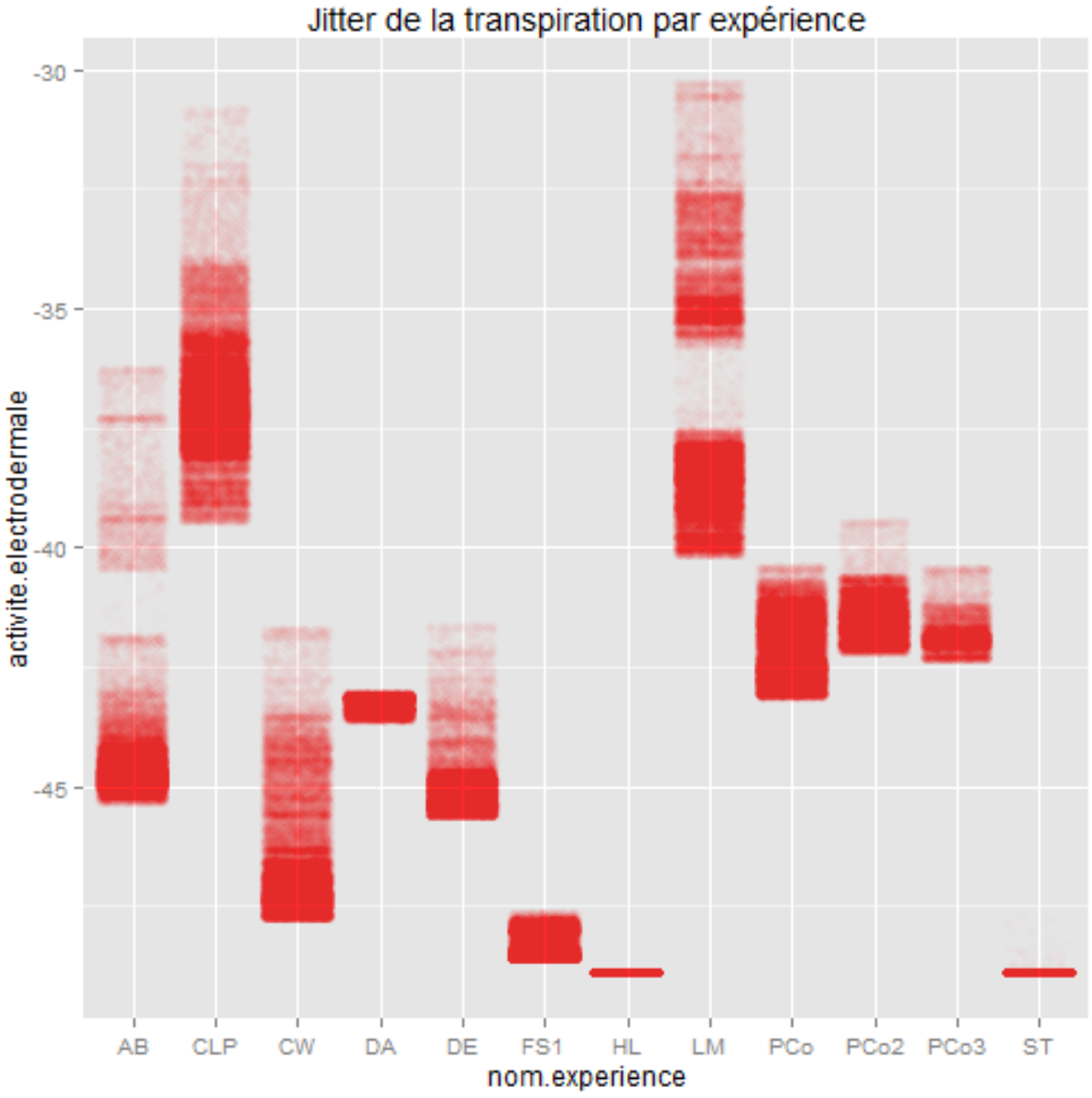

Jitter de la transpiration par expérience
activite.electrodermale
-30
-35
-40
-45
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 11:

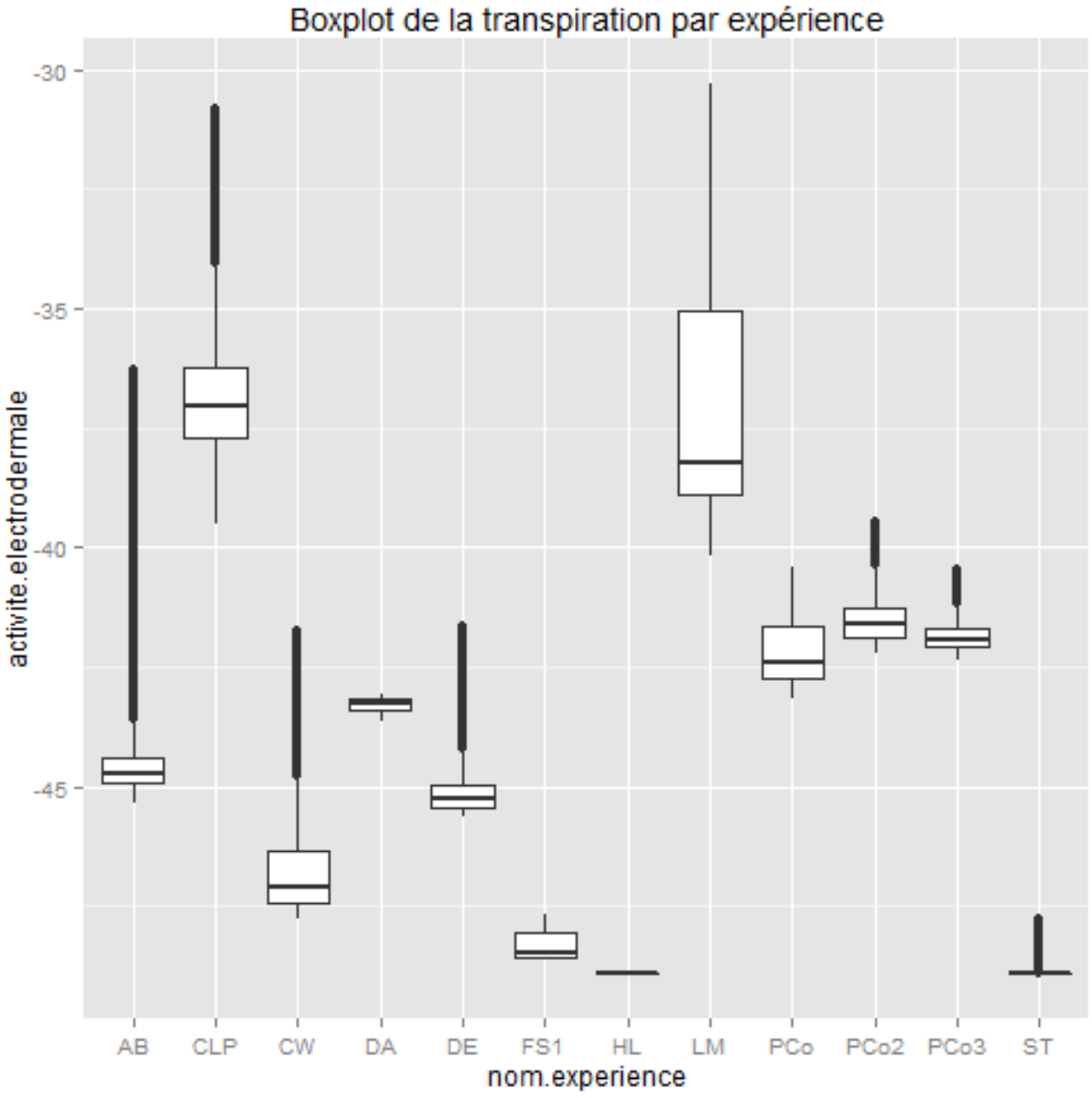

Boxplot de la transpiration par expérience
activite.electrodermale
-30
-35
-40
-45
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 12:

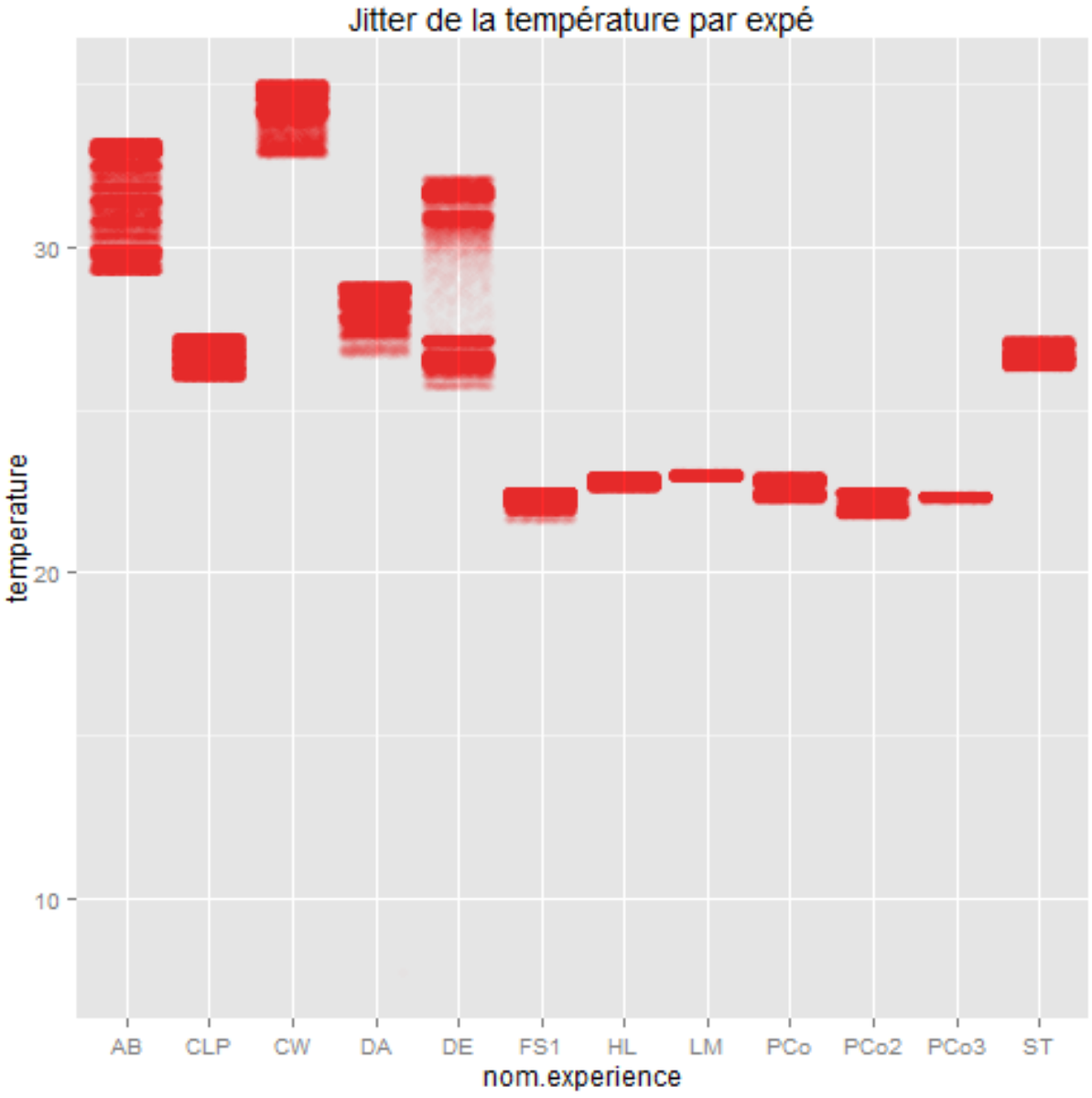

Jitter de la température par expé
temperature
30
20
10
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 13:

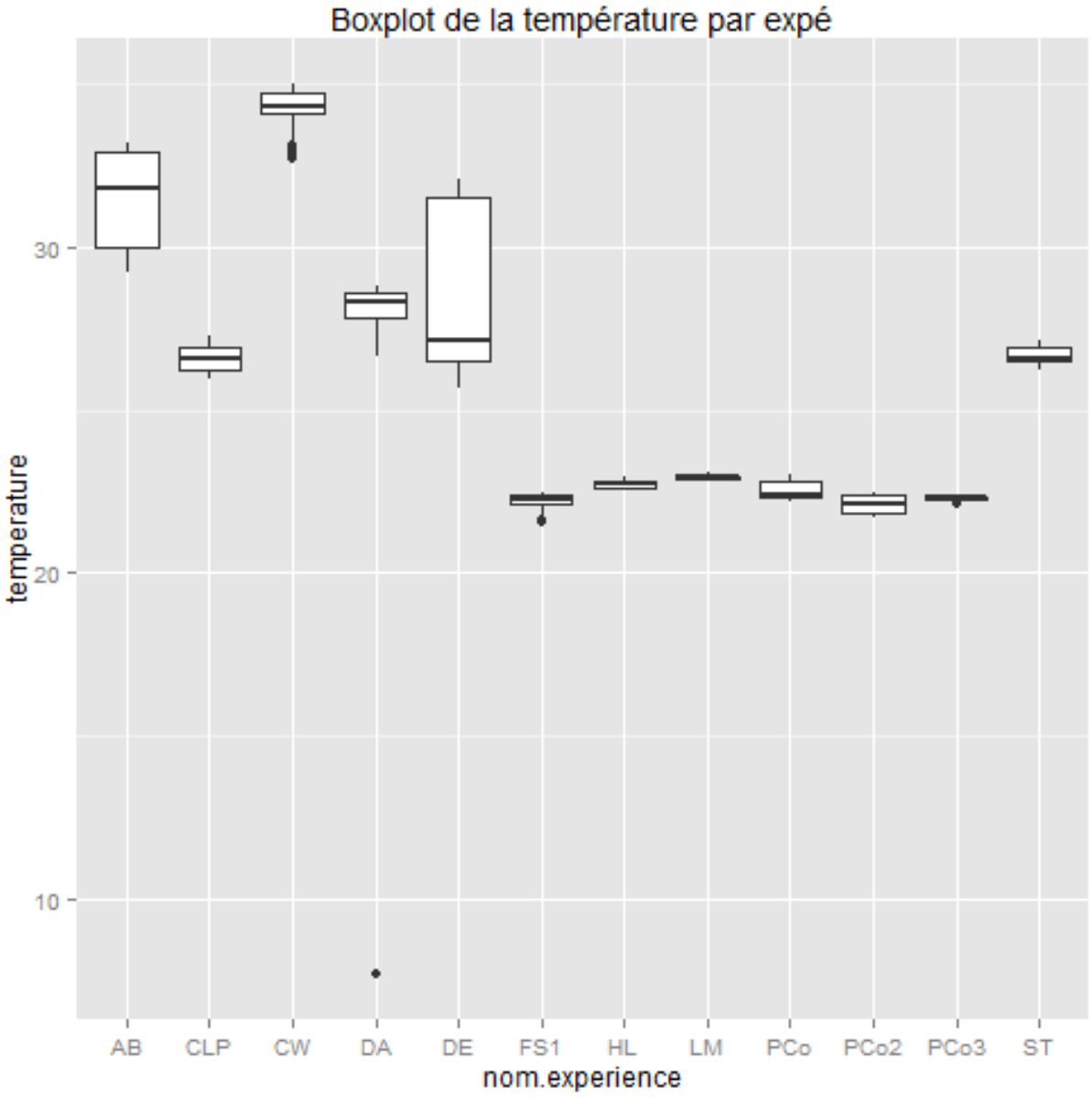

Boxplot de la température par expé
temperature
30
20
10
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 14:

corps humain est d'environ 1 degré[7]. Toujours dans cette étude, pour la température cutanée, la variation maximum est de 2 degrés environ. Seule l'expérience 'AB' propose une amplitude importante : 4r qui demandera une enquête plus appronfondie.

L'expérience 'DA' semble avoir une donnée de température extrêmement éloignée des autres. 2014-03-06 15:19:47, -12.953, -43.118, 7.703, 64, 1er, DA

Nous pouvons supposer qu'il s'agit d'une erreur puisque c'est la seule température mesurée en dessous de 20 degré toute expérience confondue et que l'ensemble des valeurs de 'AD' pour la variable température est compris dans l'intervalle : 26.665, 28.79

Nous allons remplacer cette donnée aberrante par la température mesurée à l'instant précédant :

```
#obtenir l'index de la ligne de la donnée aberrant
num.row<- which(df.all$temperature< 20)
#remplacer cette valeur par la valeur de la ligne au-dessus
df.all[num.row,"temperature"] <- df.all[num.row-1,"temperature"]

#idem pour data.DA
num.row<- which(data.DA$temperature <20)
data.DA[num.row,"temperature"] <-data.DA[num.row-1,"temperature"]
```

Verification, nombre de données inférieures à 20 :

```
## [1] "aucune"
```

4.2.4 fréquence cardiaque : graphiques "Boxplot" & "Jitter"

Pour finir, les deux graphiques suivant représentent les données mesurées pour la fréquence cardiaque (en ordonnée) par expérience (en abscisse, chacun des expérience).

Concernant la mesure cardiaque, la plage de mesure annoncée est de 10 à 220 BPM avec une résolution de 1 BPM. Les données sont compatibles avec ces spécifications. Nous remarquons une plus grande variation de la fréquence cardiaque, ce qui là encore est en accord avec les résultats trouvés par ailleurs. D'une personne à l'autre, d'un moment à l'autre, les variations des battements par minute peuvent être importantes (un sportif peut ainsi réduire son rythme cardiaque par 2).

[7] Voir graphique page 282, http://www.altmedrev.com/publications/11/4/278.pdf

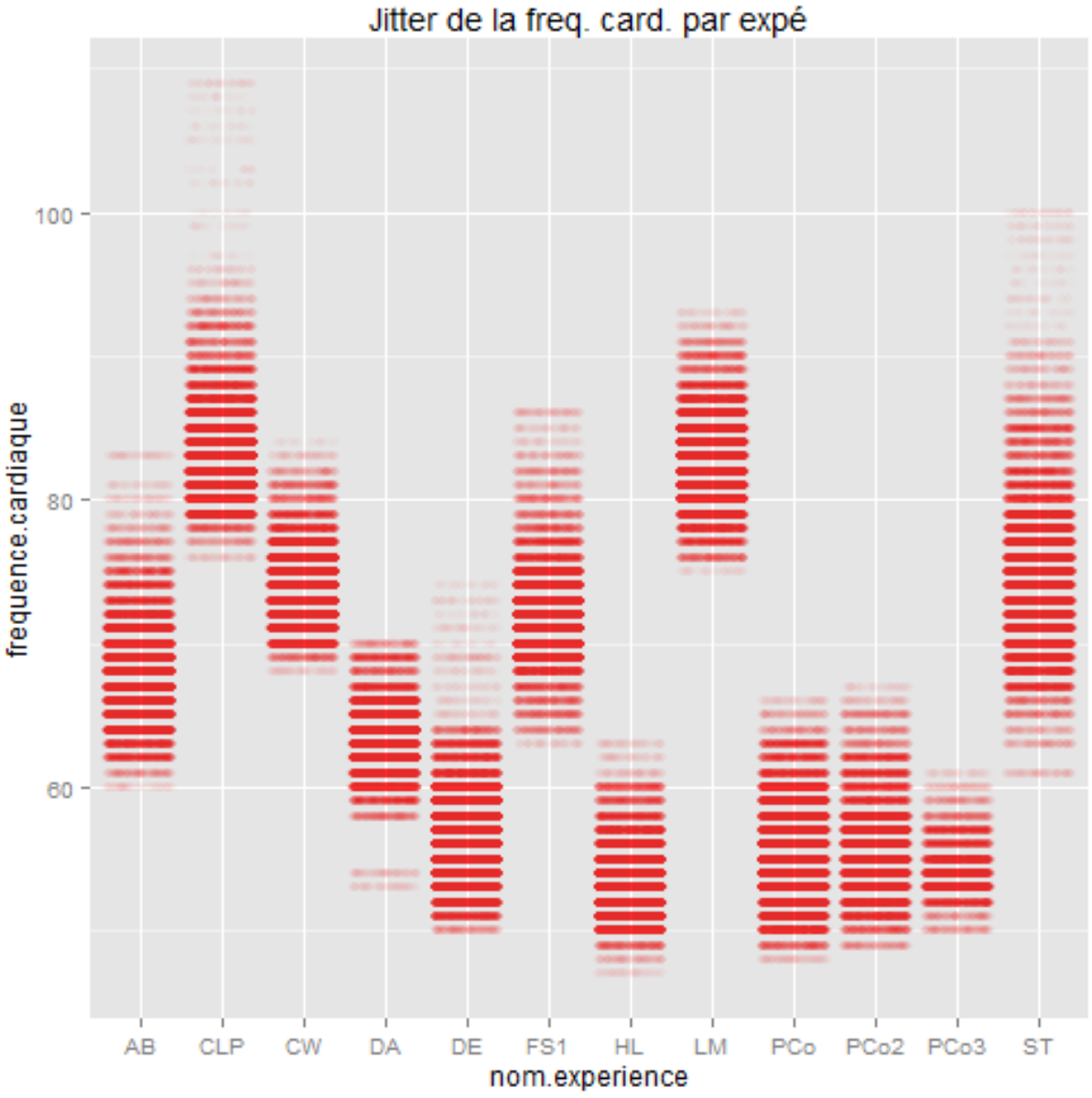

Jitter de la freq. card. par expé
frequence.cardiaque
100
80
60
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

Figure 15:

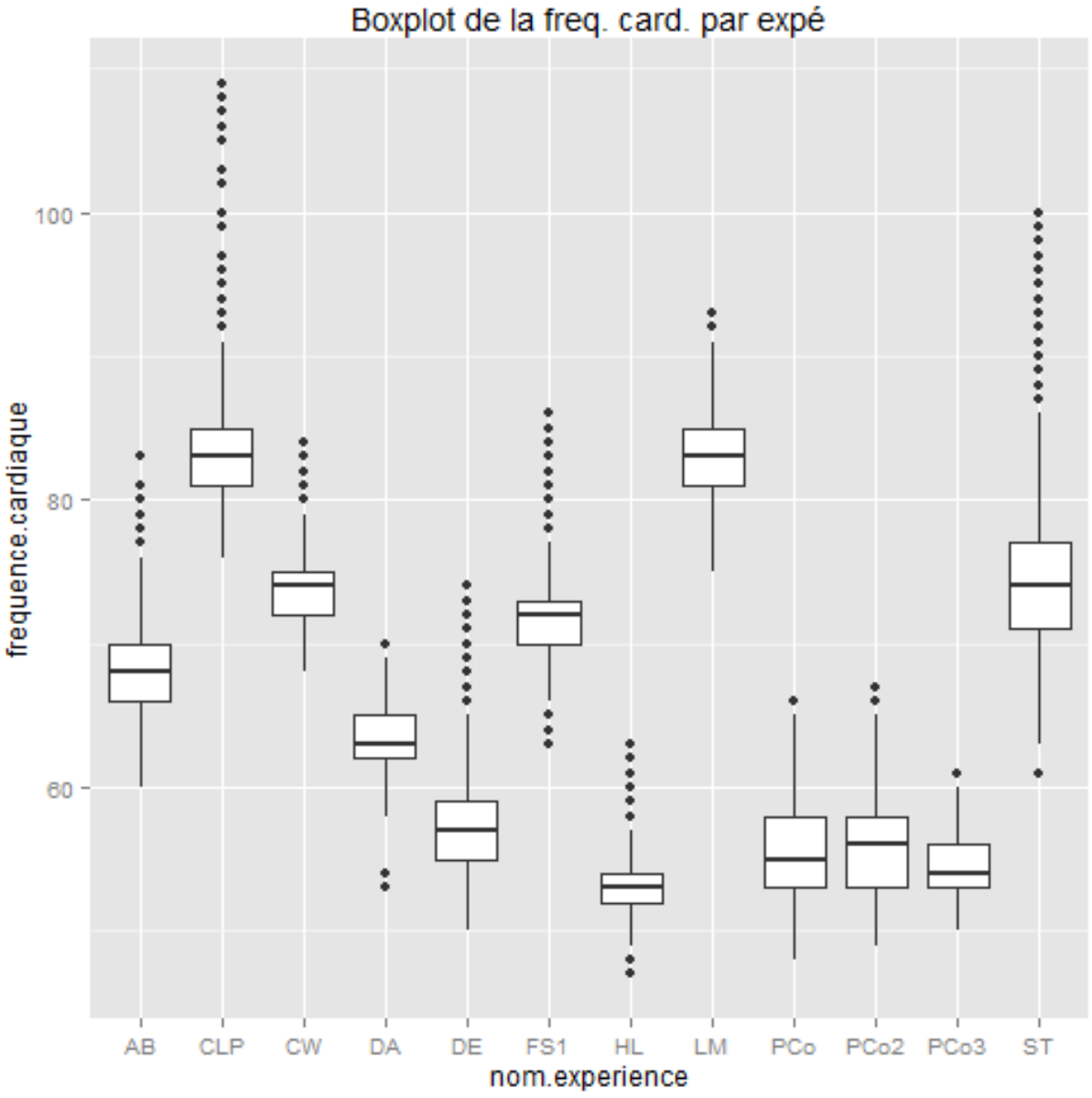
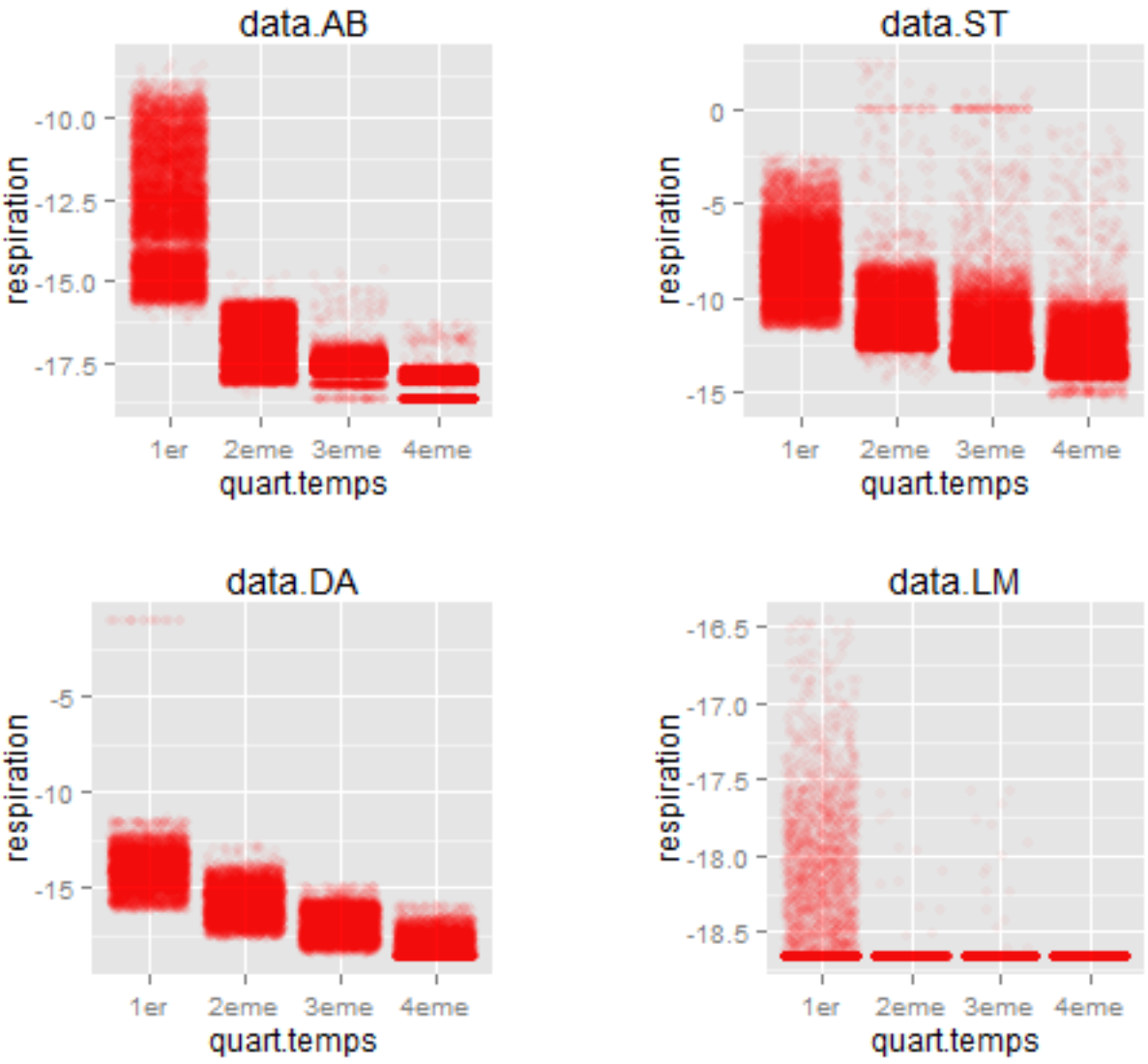

Boxplot de la freq. card. par expé
frequence.cardiaque
100
80
60
AB
CLP
CW
DA
DE
FS1
HL
LM
PCo
PCo2
PCo3
ST
nom.experience

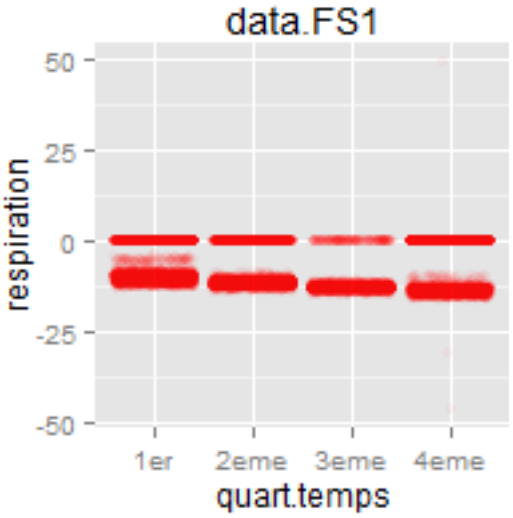
Figure 16:

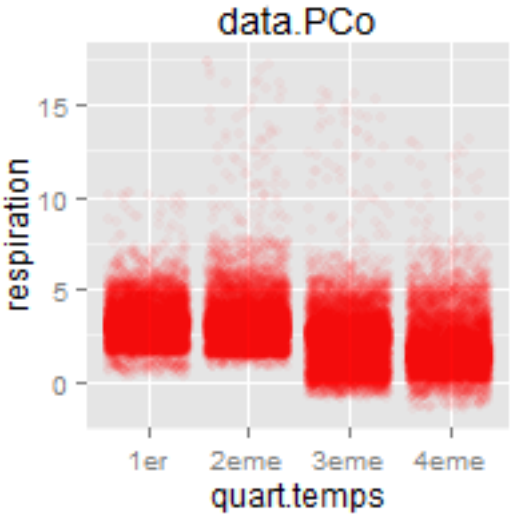
4.3 Représentations graphiques des données par expérience et par quart-temps.

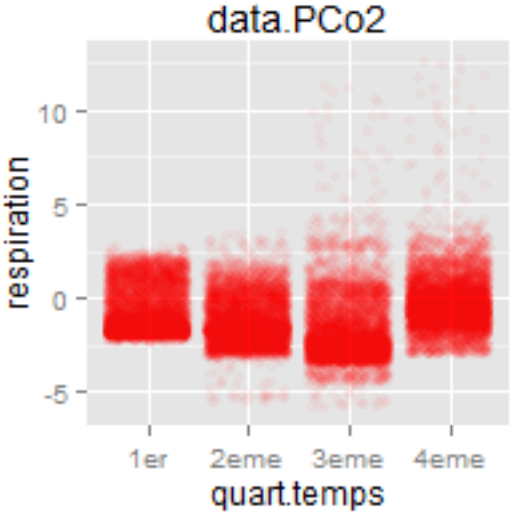
Nous présentons brièvemment un certain nombre de représentations graphiques complémentaires par expérience et par valeur. Nous avons là encore utilisé les graphiques de type “Jitter” avec de la transparence mais cette fois-ci en découpant les données par quart temps grâce à une donnée catégorielle afin d’avoir une idée de l’évolution dans le temps. L’idée est de voir si cette évolution montre des phases où les données sont constantes et - afin de préparer la suite - de voir si cette évolution montre des régularités ou des tendances intéressantes.

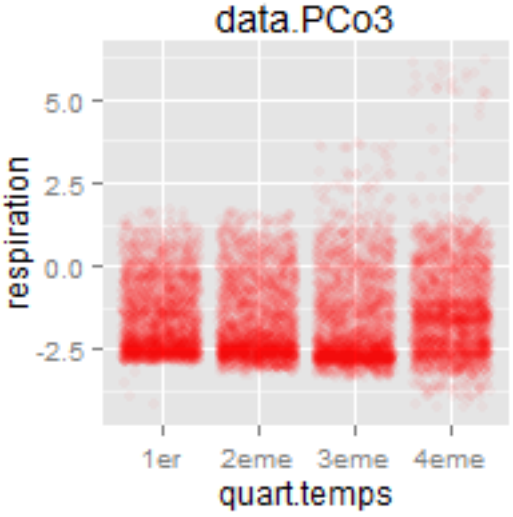
4.3.1 respiration : graphiques “jitter”

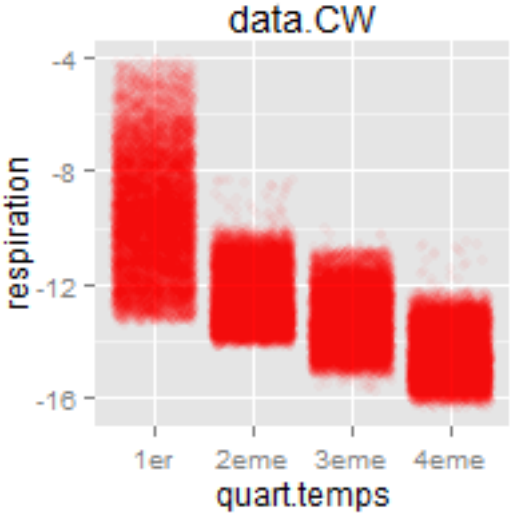


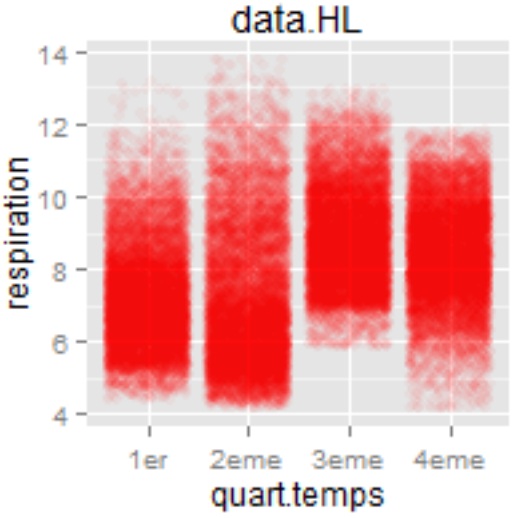

data.FS1
respiration
50
25
0
-25
-50
1er
2eme
3eme
4eme
quart.temps

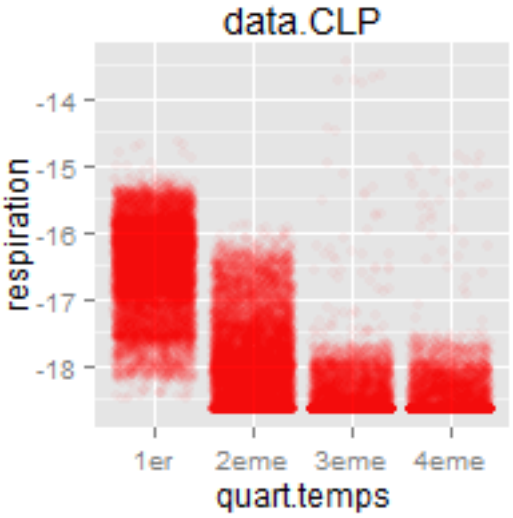

data.PCo
respiration
15
10
5
0
1er
2eme
3eme
4eme
quart.temps

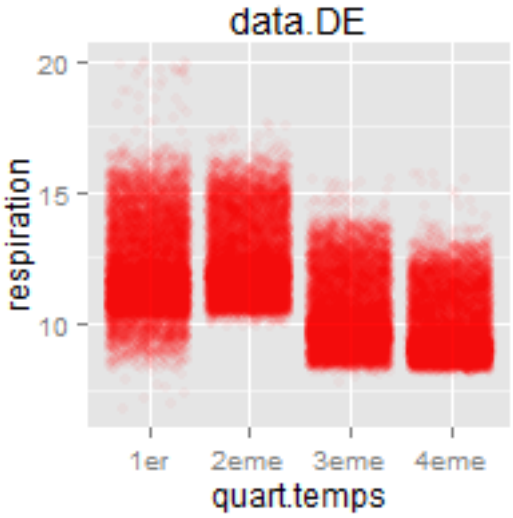

data.PCo2
respiration
10
5
0
-5
1er
2eme
3eme
4eme
quart.temps


data.PCo3
respiration
5.0
2.5
0.0
-2.5
1er
2eme
3eme
4eme
quart.temps


data.CW
respiration
-4
-8
-12
-16
1er
2eme
3eme
4eme
quart.temps


data.HL
respiration
14
12
10
8
6
4
1er
2eme
3eme
4eme
quart.temps


data.CLP
respiration
-14
-15
-16
-17
-18
1er
2eme
3eme
4eme
quart.temps


data.DE
respiration
20
15
10
1er
2eme
3eme
4eme
quart.temps

4.3.1.1 Analyse. Pour la respiration, nous voyions une confirmation d'un certain nombre d'anomalies que nous avons detecté. Ainsi, concernant LM, nous pouvons observer que les valeurs sont constantes pendant la majorité de l'expérience. FS1 montre des variations assez étonnantes. Puisqu'il y a un grand nombre de valeur nulle : 14588 , soit 3133.8360296 % du totale. 49.408

Comme pour 'DA' un peu plus haut, nous remarquons des données aberrantes : *Avec une unique valeur positive : 2014-03-06 15:41:34, 49.408, -47.848, 22.073, 73, 4eme, FS1* Avec deux valeurs largement hors du spectre :

```
##Lignes des données de respi de FS1 < 18
which( df.all$nom.experience == 'FS1' & df.all$respiration< -18)
```

```
## [1] 216302 216843
```

Là aussi, nous allons les modifier par la valeur qui les précède :

```
#obtenir l'index de la ligne de la donnée aberrant
num.row<- which(df.all$nom.experience == 'FS1' & df.all$respiration > 0 )
#remplacer cette valeur par la valeur de la ligne au-dessus
df.all[num.row,"respiration"] <- df.all[num.row-1,"respiration"]

#obtenir l'index de la ligne de la donnée aberrant
num.row<- which(df.all$nom.experience == 'FS1' & df.all$respiration < -18)
#remplacer cette valeur par la valeur de la ligne au-dessus
df.all[num.row,"respiration"] <- df.all[num.row-1,"respiration"]
```

Vérification, nombre de données > 0 ou < -18 après modification :

```
## [1] "aucune"
```

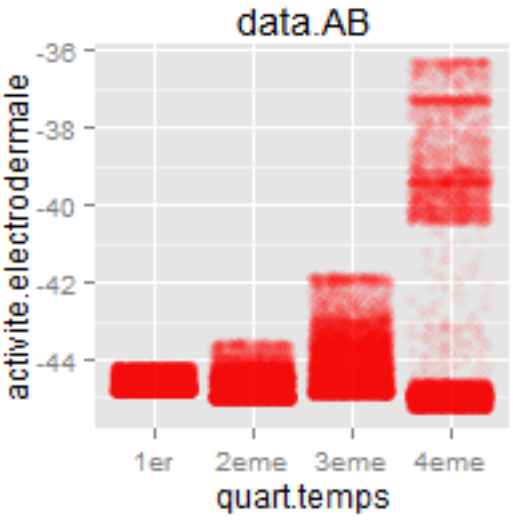
4.3.1.2 Conclusion concernant la respiration :

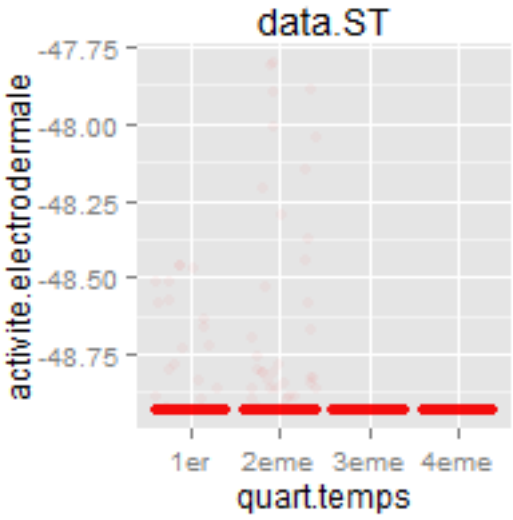
En conclusion, nous pouvons dire que les expériences LM et FS1 sont particulièrement peu crédibles en ce qui concerne les valeurs fournies. L'autre problème concerne plus la différence dans les spectres des valeurs d'une expérience à l'autre.

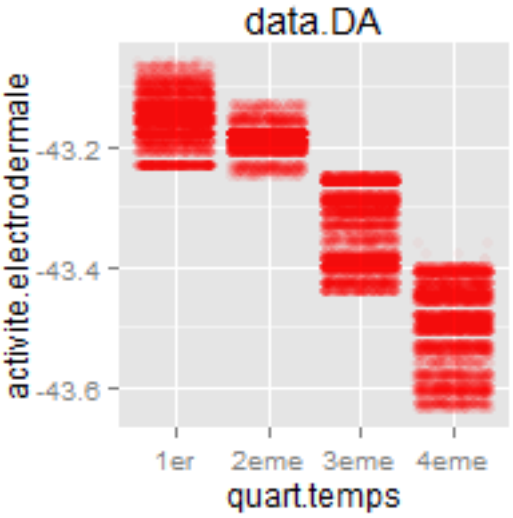
Intervales par expériences pour la respiration :

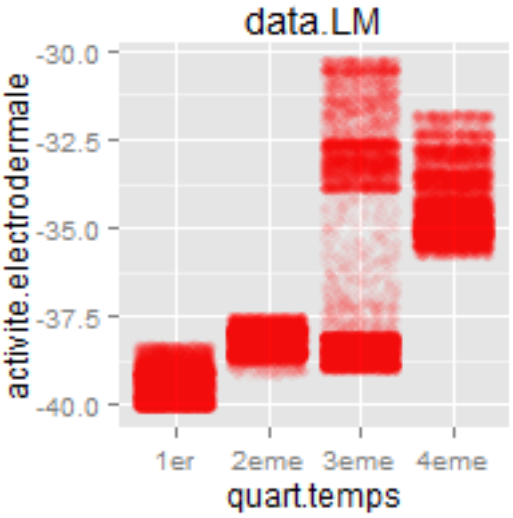
	min	max
AB	-18.670	-8.304
ST	-15.418	2.468
DA	-18.660	-1.000
LM	-18.670	-16.455
FS1	-16.338	0.000
PCo	-1.565	17.405
PCo2	-5.863	12.725
PCo3	-4.295	6.227
CW	-16.416	-4.045
HL	4.145	13.778
CLP	-18.670	-13.436

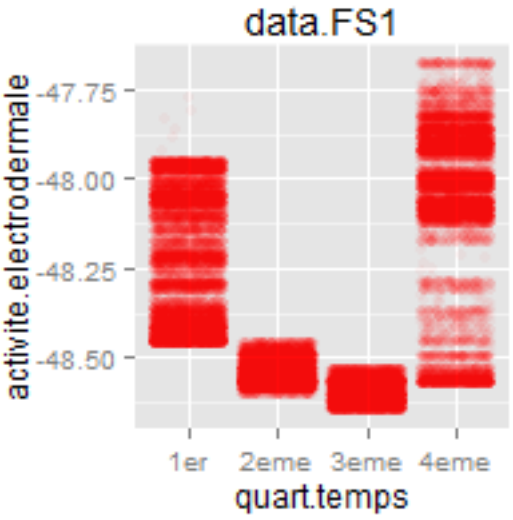
4.3.2 transpiration (activité electroderm.) : graphiques "jitter"

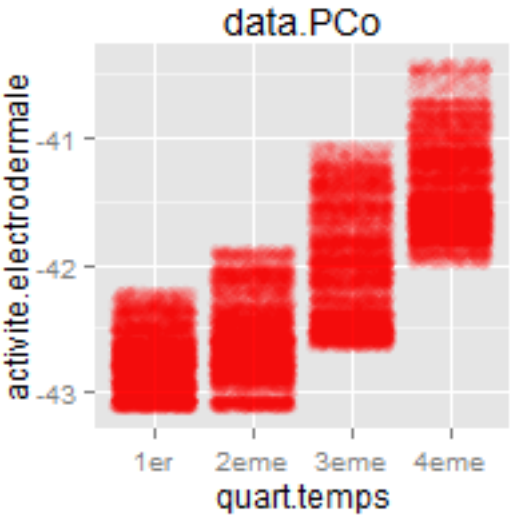


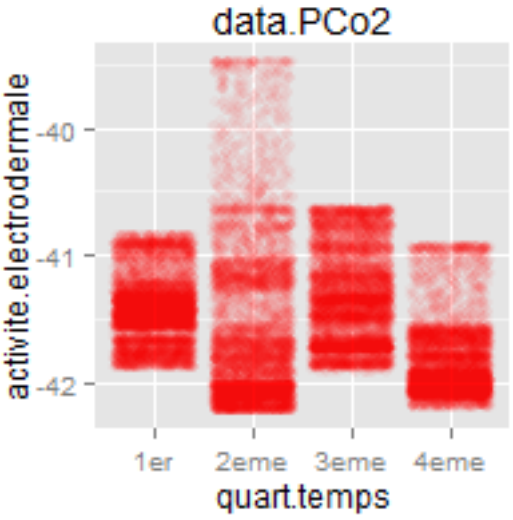


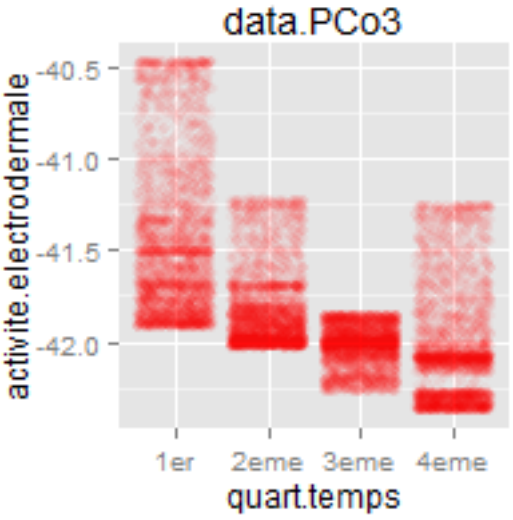

data.DA
activite.electrodermale
-43.2
-43.4
-43.6
1er
2eme
3eme
4eme
quart.temps

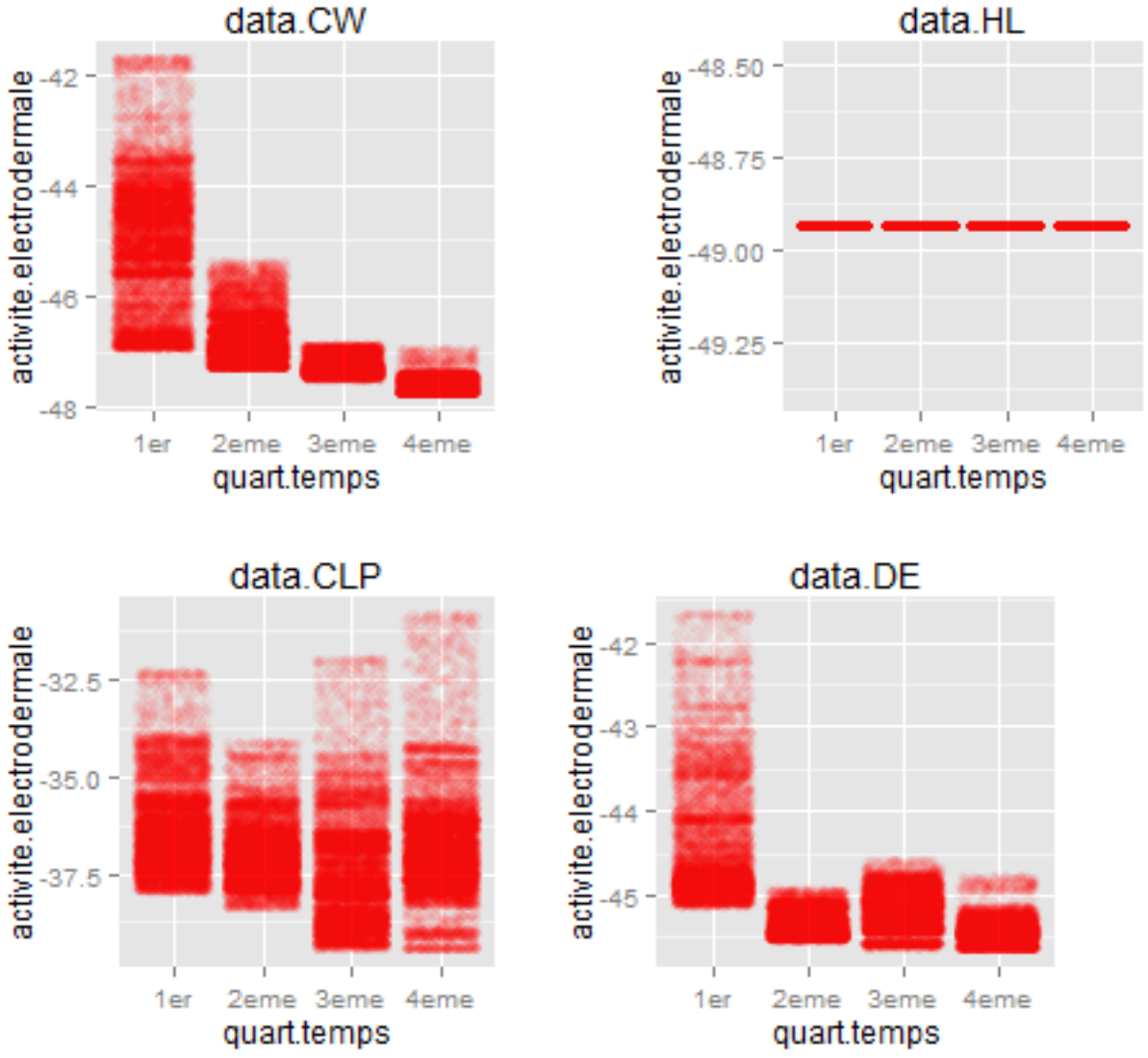

data.LM
activite.electrodermale
-30.0
-32.5
-35.0
-37.5
-40.0
1er
2eme
3eme
4eme
quart.temps


data.FS1
activite.electrodermale
-47.75
-48.00
-48.25
-48.50
1er
2eme
3eme
4eme
quart.temps


data.PCo
activite.electrodermale
-41
-42
-43
1er
2eme
3eme
4eme
quart.temps


data.PCo2
activite.electrodermale
-40
-41
-42
1er
2eme
3eme
4eme
quart.temps


data.PCo3
activite.electrodermale
-40.5
-41.0
-41.5
-42.0
1er
2eme
3eme
4eme
quart.temps



4.3.2.1 Analyse. \

Concernant la transpiration, il semble confirmer les valeurs mesurées pour 'ST' et 'HL' sont pratiquement toutes nulles.

4.3.2.2 Conclusion concernant la transpiration : \

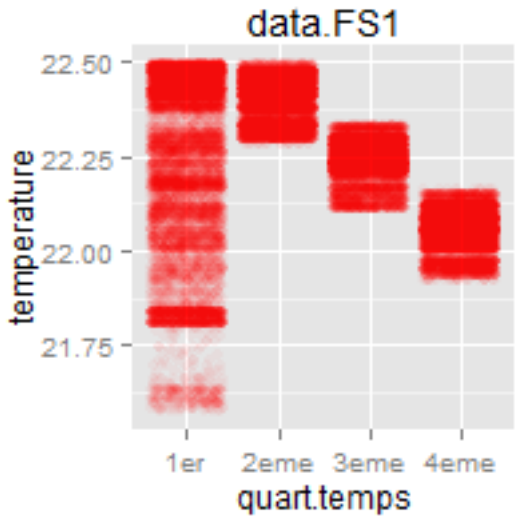
Ici aussi, il y a deux expériences qui semblent n'avoir pas "réussi" (ici ST et HL). Concernant le problème les intervales dans lesquels varient valeurs des différentes expériences, les choses semblent plus cohérentes (voir tableau ci-dessous).

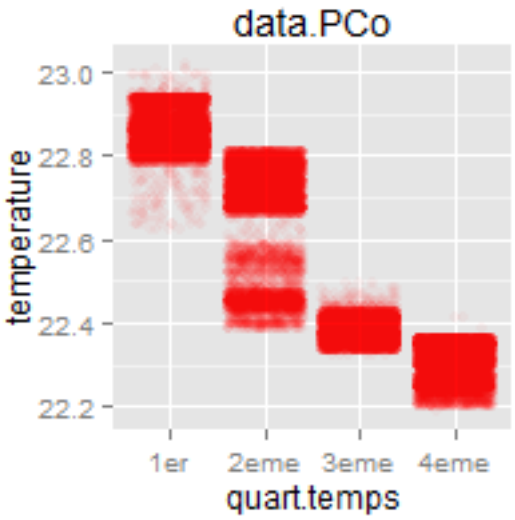
Intervales par expérience pour la transpiration :

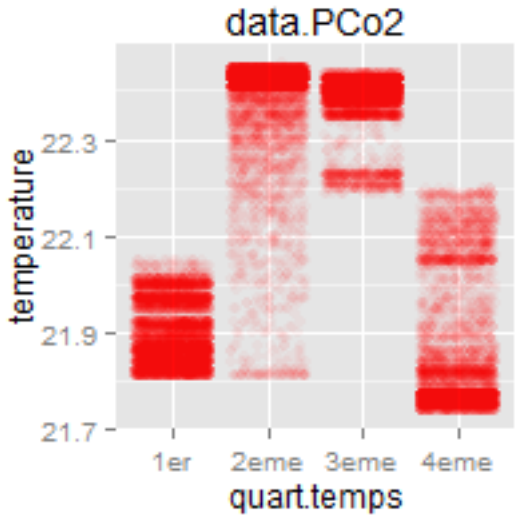
	min	max
AB	-45.345	-36.299
ST	-48.936	-47.798
DA	-43.640	-43.056

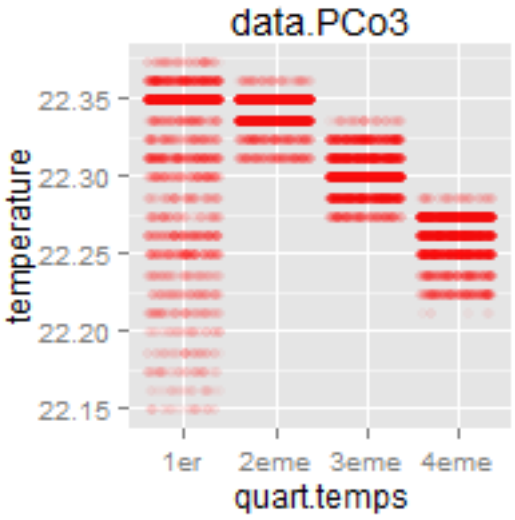
	min	max
LM	-40.165	-30.280
FS1	-48.649	-47.673
PCo	-43.156	-40.403
PCo2	-42.217	-39.477
PCo3	-42.380	-40.465
CW	-47.773	-41.729
HL	-48.936	-48.936
CLP	-39.489	-30.806

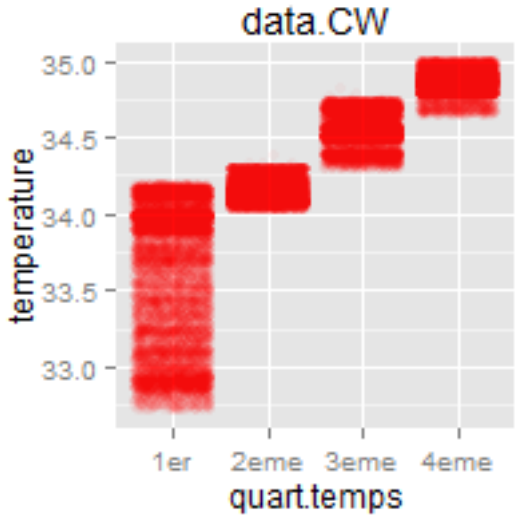
4.3.3 Température : graphiques “jitter”

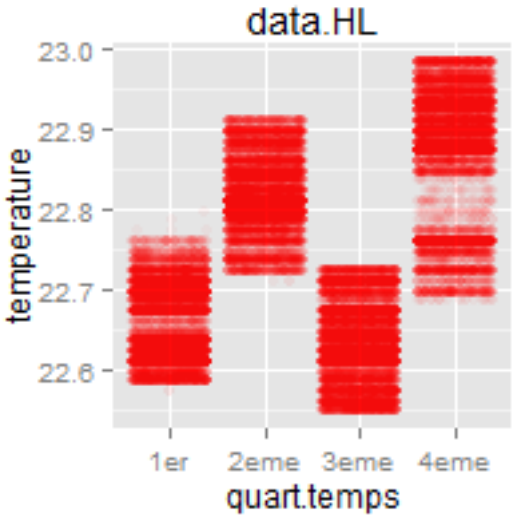

data.FS1
temperature
22.50
22.25
22.00
21.75
1er
2eme
3eme
4eme
quart.temps

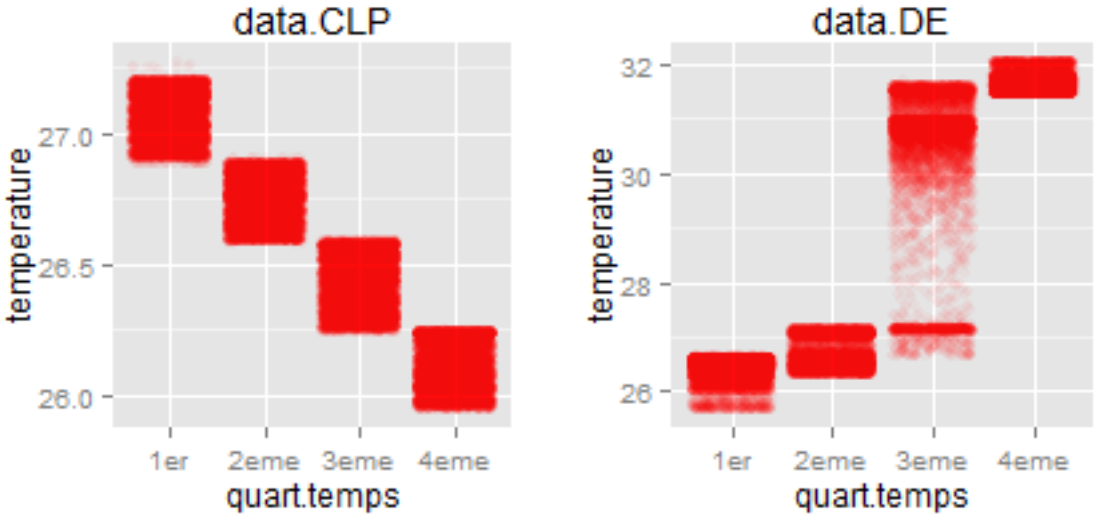

data.PCo
temperature
23.0
22.8
22.6
22.4
22.2
1er
2eme
3eme
4eme
quart.temps


data.PCo2
temperature
22.3
22.1
21.9
21.7
1er
2eme
3eme
4eme
quart.temps


data.PCo3
temperature
22.35
22.30
22.25
22.20
22.15
1er
2eme
3eme
4eme
quart.temps


data.CW
temperature
35.0
34.5
34.0
33.5
33.0
1er
2eme
3eme
4eme
quart.temps


data.HL
temperature
23.0
22.9
22.8
22.7
22.6
1er
2eme
3eme
4eme
quart.temps



4.3.3.1 Analyse \

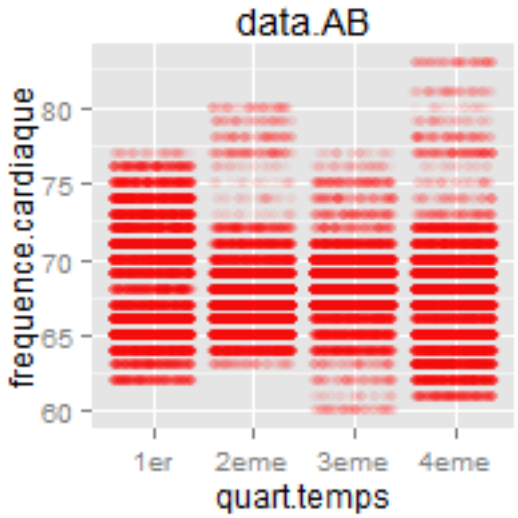
En dehors des problèmes corrigés, il ne semble pas y avoir de données aberrantes ou opposées aux spécifications du fabriquant du matériel de mesure. Les seules questions concernent la variation pour certaines expériences ainsi que nous l'avons déjà souligé ('AB' qui varie de 4 degrés et dans une moindre mesure 'CW' qui varie de 2.29).

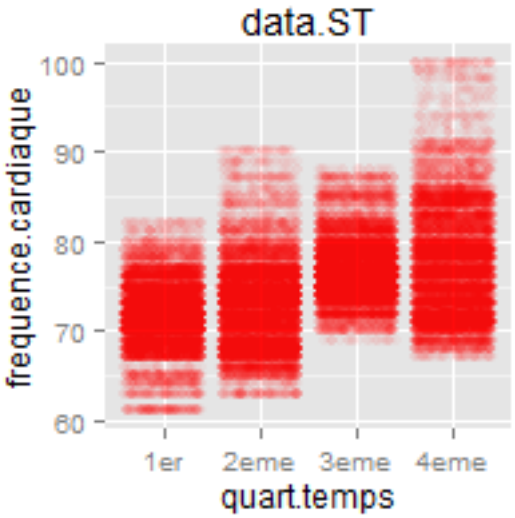
4.3.3.2 Conclusion concernant la température : \

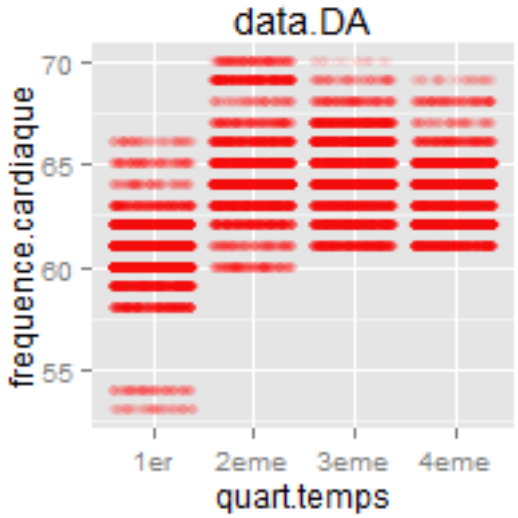
En conclusion, les données peuvent être utlisées pour des fouilles de données. Pour information, les intervales par expériences pour la température :

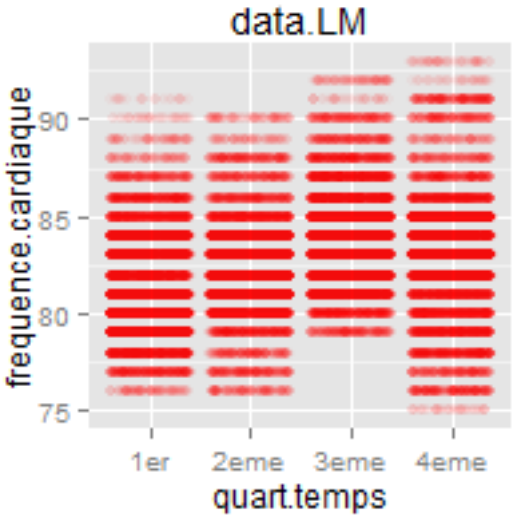
	min	max
AB	29.180	33.184
ST	26.214	27.128
DA	26.665	28.790
LM	22.898	23.061
FS1	21.572	22.498
PCo	22.185	23.023
PCo2	21.735	22.460
PCo3	22.148	22.373
CW	32.708	34.998
HL	22.548	22.986
CLP	25.939	27.278

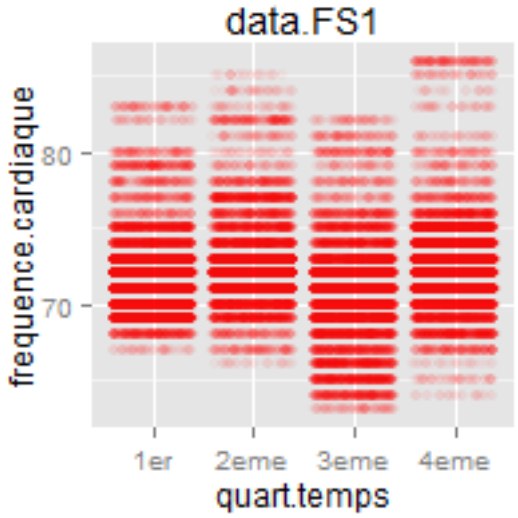
4.3.4 fréquence cardiaque : graphique “jitter”

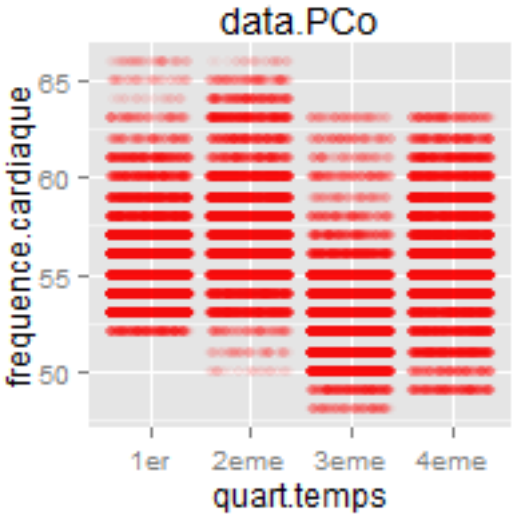


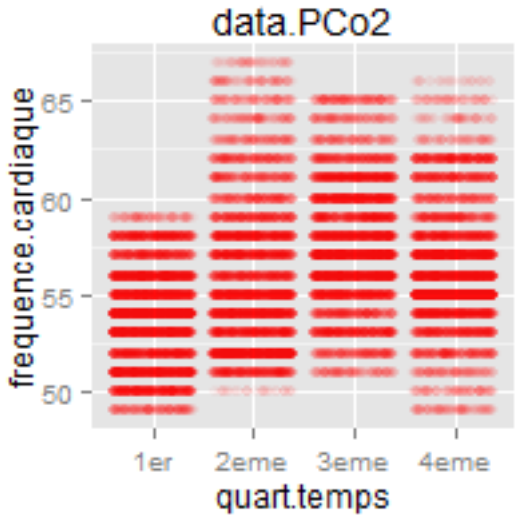


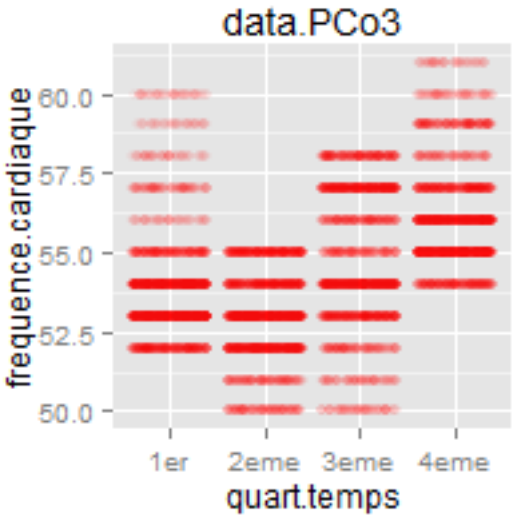


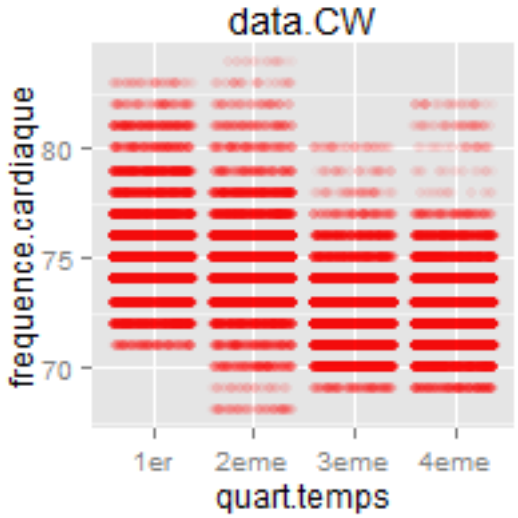


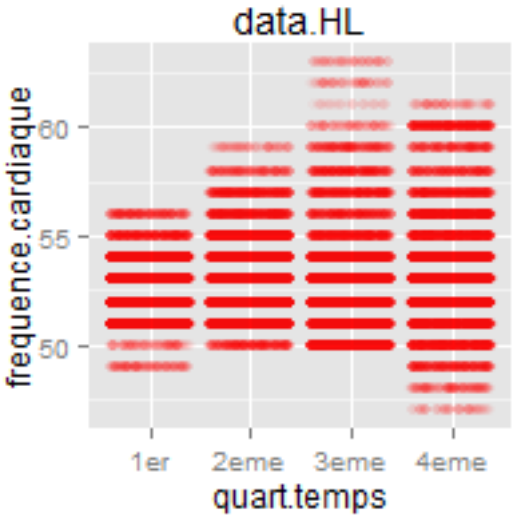

data.FS1
frequence.cardiaque
80
70
1er
2eme
3eme
4eme
quart.temps

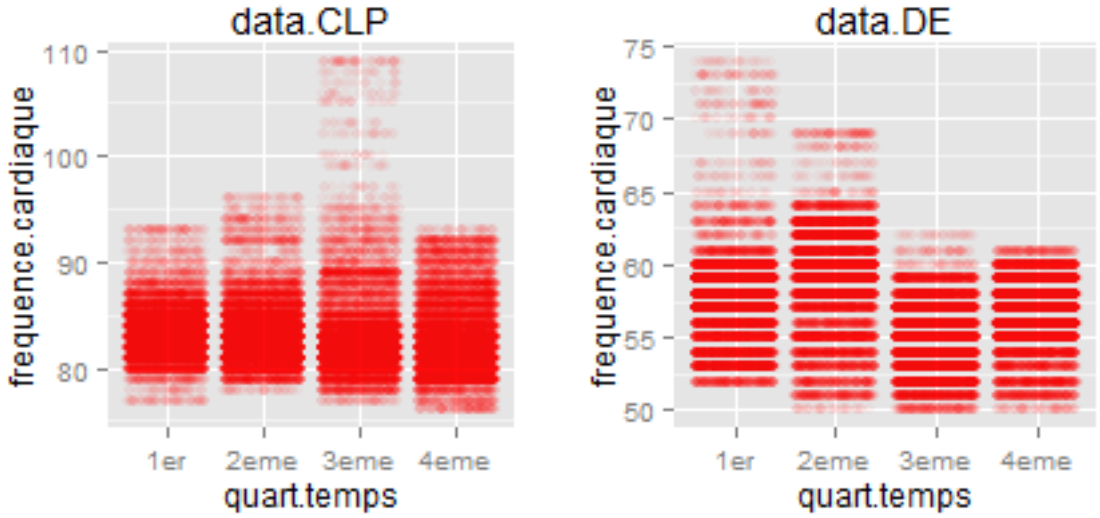

data.PCo
frequence.cardiaque
65
60
55
50
1er
2eme
3eme
4eme
quart.temps


data.PCo2
frequence.cardiaque
65
60
55
50
1er
2eme
3eme
4eme
quart.temps


data.PCo3
frequence.cardiaque
60.0
57.5
55.0
52.5
50.0
1er
2eme
3eme
4eme
quart.temps


data.CW
frequence.cardiaque
80
75
70
1er
2eme
3eme
4eme
quart.temps


data.HL
frequence.cardiaque
60
55
50
1er
2eme
3eme
4eme
quart.temps



4.3.4.1 Conclusion concernant la fréquence cardiaque : \

Les "rayures" sur les graphiques s'expliquent par le fait que les fréquences cardiaques soient exprimés en nombre entier (positif).

Intervales par expériences pour la température :

	min	max	diff max-min
AB	60	83	23
ST	61	100	39
DA	53	70	17
LM	75	93	18
FS1	63	86	23
PCo	48	66	18
PCo2	49	67	18
PCo3	50	61	11
CW	68	84	16
HL	47	63	16
CLP	76	109	33

4.3.5 Nuage de points par expérience et par pair de variable

Pour finir, nous ajoutons en annexe les graphiques des “matrices de nuages de points” ou “scatter plots” permettent de représenter pour chaque expérience des couples de variables dans un plan.

5 Conclusion :

Il semble qu’avant de pouvoir traiter les données, un certain nombre de tests doivent être réalisées. Par ailleurs, des questions doivent être posées au fabriquant des capteurs concernant la possiblite de devoir ajouter un “offset” aux mesures brutes. Il existe plusieurs incertitudes concernant la qualité des données mesurées ; malgré tout, nous pouvons procéder à de premières analyses plus poussées en reprenant la méthodologie du travail préliminaire (voir le document en annexe).

Nous allons maintenant procéder à un nettoyage des données en nous basant sur les différentes remarques faites dans la première partie et en utilisant le fait que les algorithmes que nous utilisons peuvent aussi permettre d’“isoler” les données “abhérantes” (comme les cartes topologiques de Kohonen). Nous espérons ainsi pouvoir tester des premiers resultats.

Par ailleurs, nous sommes en même temps en contact avec des représentants de la société fabriquant les capteurs. et nous avons procéder avec Viviane Gal à un certain nombre de tests avec le matériel pour mieux comprendre les données reçues, les possibles anomalies et trancher concernant un certain nombre de questions (influence de la température ambiante, vérifer s’il y a peu avoir des problèmes de déconnexion…) qui se sont posées à nous dans l’exploration des données. L’analyse de ces tests est en cours.

```
#enregistrement des données
rda.save <- paste(data.path ,"data-frame-all-expe.Rda",sep = "/")
save(df.all,file=rda.save)
```

5.1 Annexe :

5.1.1 Nuages de points par pair de variables (pour les données de type réels)

A l'intérieur d'une matrice, on trouve donc plusieurs nuages de points avec en abscisse une des quatres mesures (respiration, activité électrodermale, température, fréquence cardiaque) et en ordonnée une autres de ces quatres mesures, différente de celle choisie pour l'abscisse. Ainsi, pour l'expérience 'AB' (figure 13), le nuage de points situé en dessous du carré intitulé "respiration" dans la matrice scatter plot représente dans le plan l'ensemble des couples "respiration" et "activité electrodermale" pour chaque instant écoulé durant l'expérience.

Par ailleurs, en utilisant un dégradé de couleurs du jaune vers le rouge, nous avons représenté l'évolution dans le temps de ces couples : plus on approche de la fin de l'expérience, plus la couleur du point devient rouge.

5.1.1.1 Pour AB : Voir graphique ci-dessous

5.1.1.2 Pour ST : Voir graphique ci-dessous

5.1.1.3 Pour LM : Voir graphique ci-dessous

5.1.1.4 Pour FS1 : Voir graphique ci-dessous

5.1.1.4.1 Pour PCo : Voir graphique ci-dessous

5.1.1.4.2 Pour PCo2 : Voir graphique ci-dessous

5.1.1.4.3 Pour PCo3 : Voir graphique ci-dessous

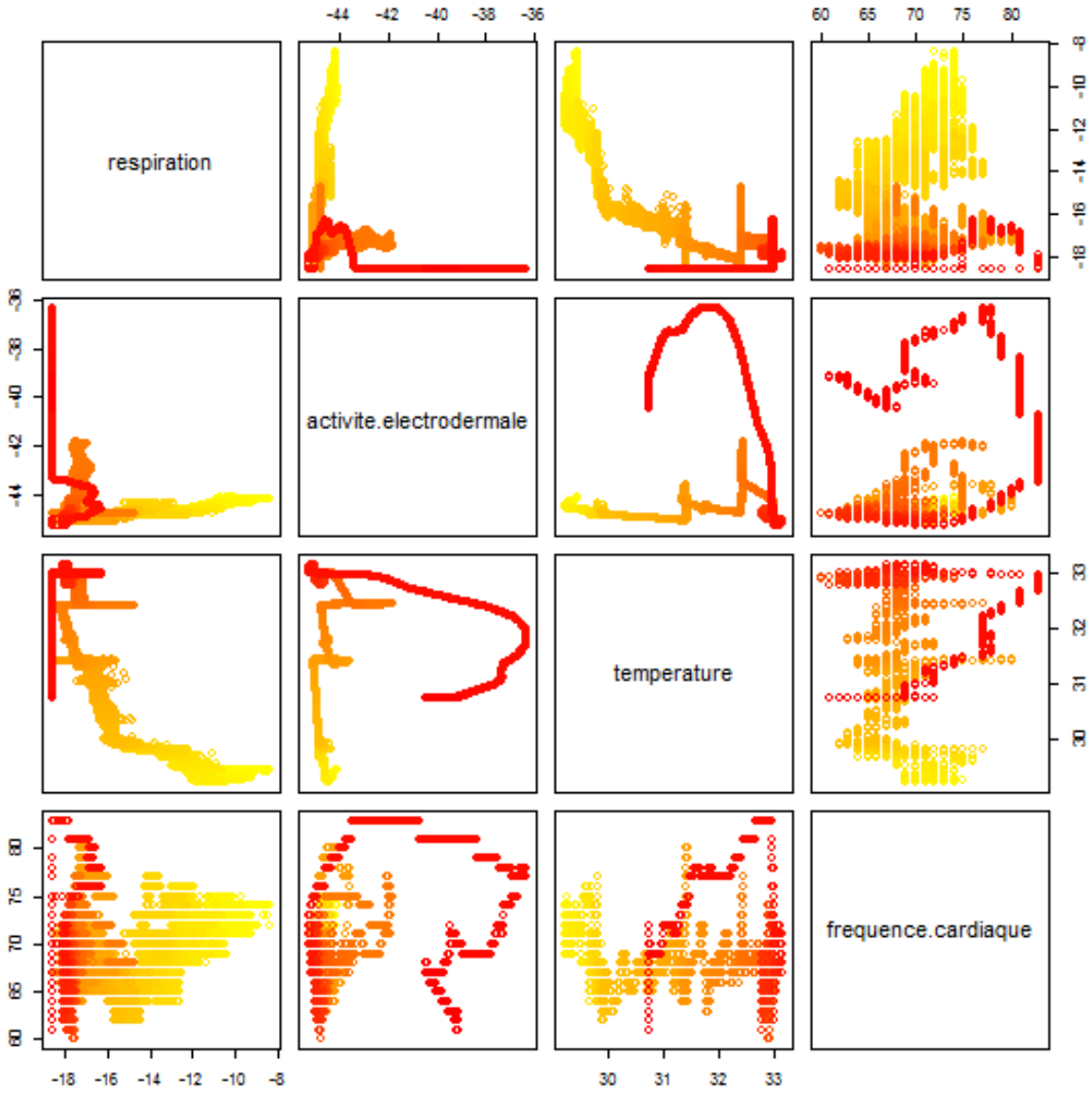

respiration
activite.electrodermale
temperature
frequence.cardiaque

Figure 17:

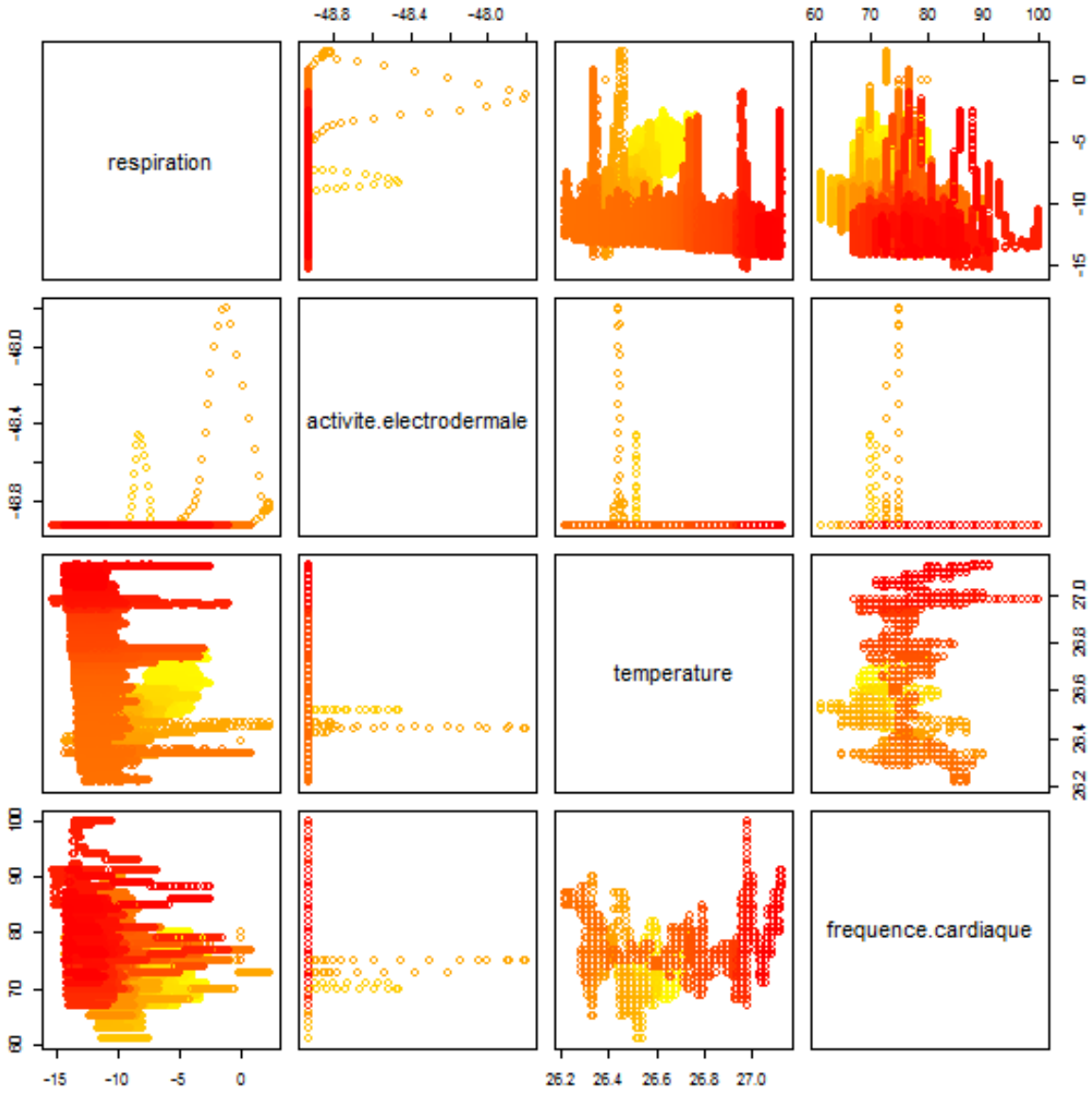

respiration
activite.electrodermale
temperature
frequence.cardiaque
-48.8
-48.4
-48.0
60
70
80
90
100
0
-5
-10
-15
26.2
26.4
26.6
26.8
27.0

Figure 18:

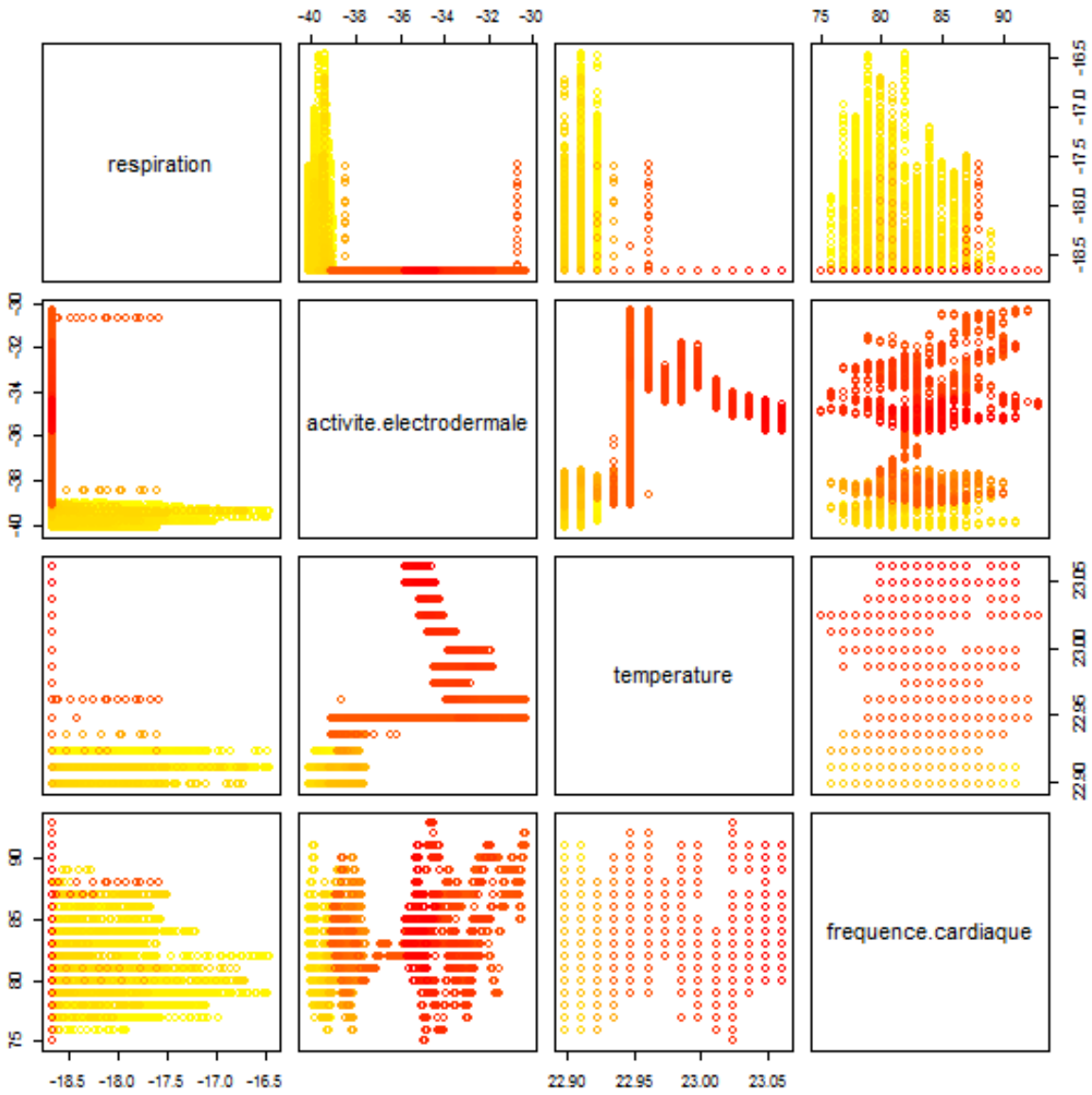

respiration
activite.electrodermale
temperature
frequence.cardiaque
-40 -38 -36 -34 -32 -30
75 80 85 90
-16.5 -17.0 -17.5 -18.0 -18.5
-30 -32 -34 -36 -38 -40
23.05 23.00 22.95 22.90
90 85 80 75
-18.5 -18.0 -17.5 -17.0 -16.5
22.90 22.95 23.00 23.05

Figure 19:

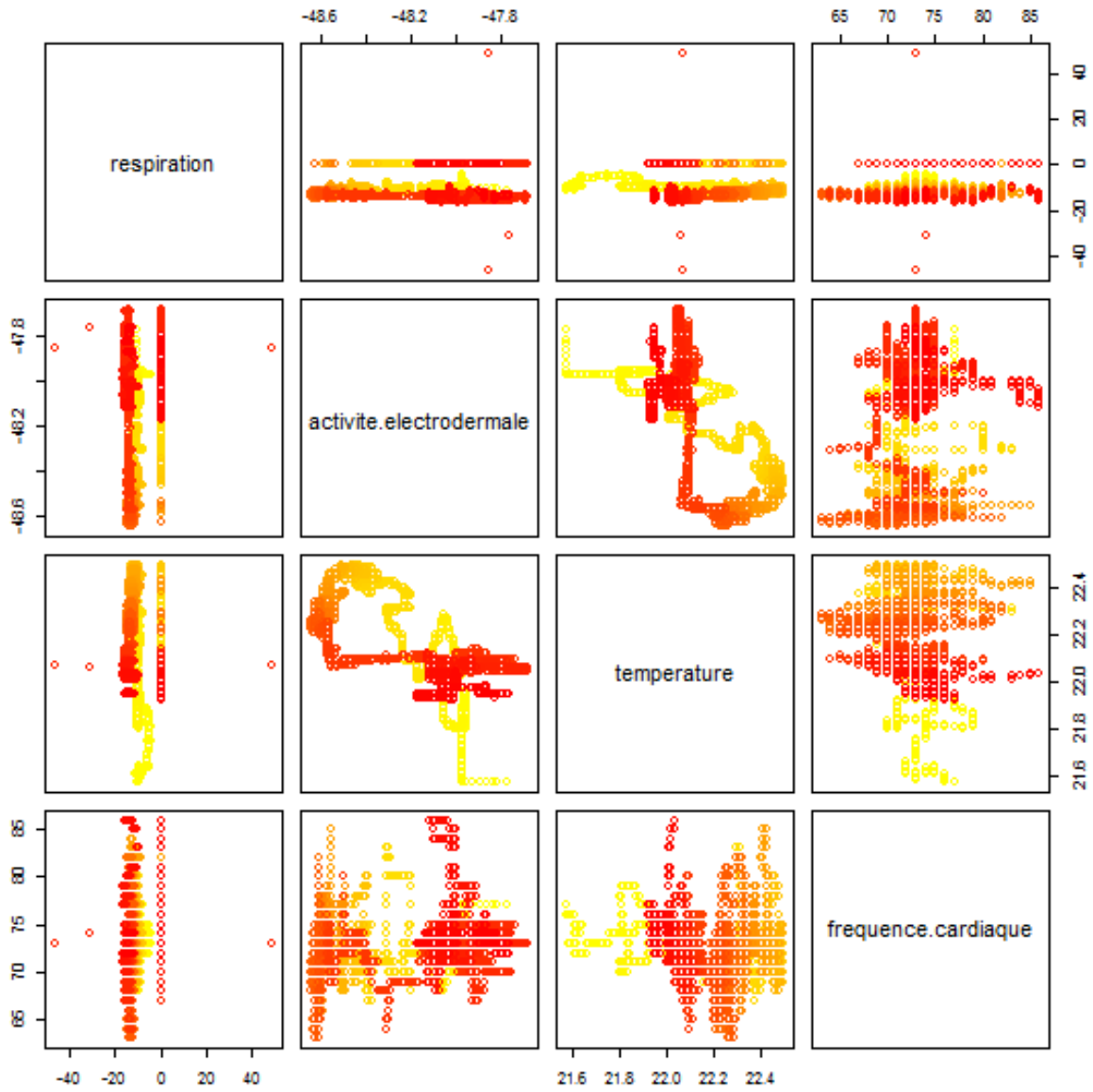

respiration
activite.electrodermale
temperature
frequence.cardiaque
-48.6
-48.2
-47.8
65
70
75
80
85
40
20
0
-20
-40
-47.8
-48.2
-48.6
22.4
22.2
22.0
21.8
21.6
85
80
75
70
65
-40
-20
0
20
40
21.6
21.8
22.0
22.2
22.4

Figure 20:

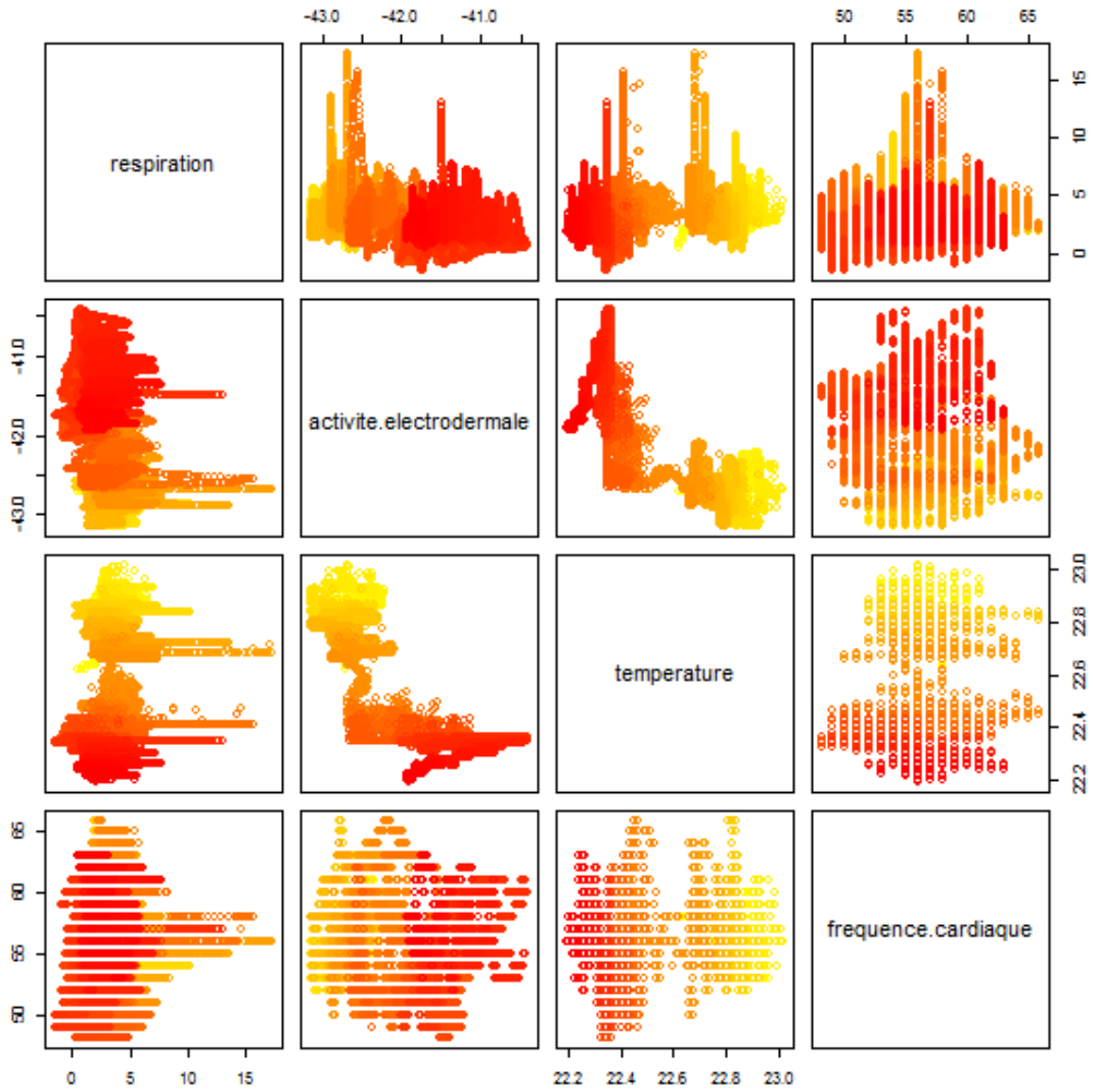

respiration
activite.electrodermale
temperature
frequence.cardiaque
-43.0
-42.0
-41.0
50
55
60
65
0
5
10
15
22.2
22.4
22.6
22.8
23.0

Figure 21:

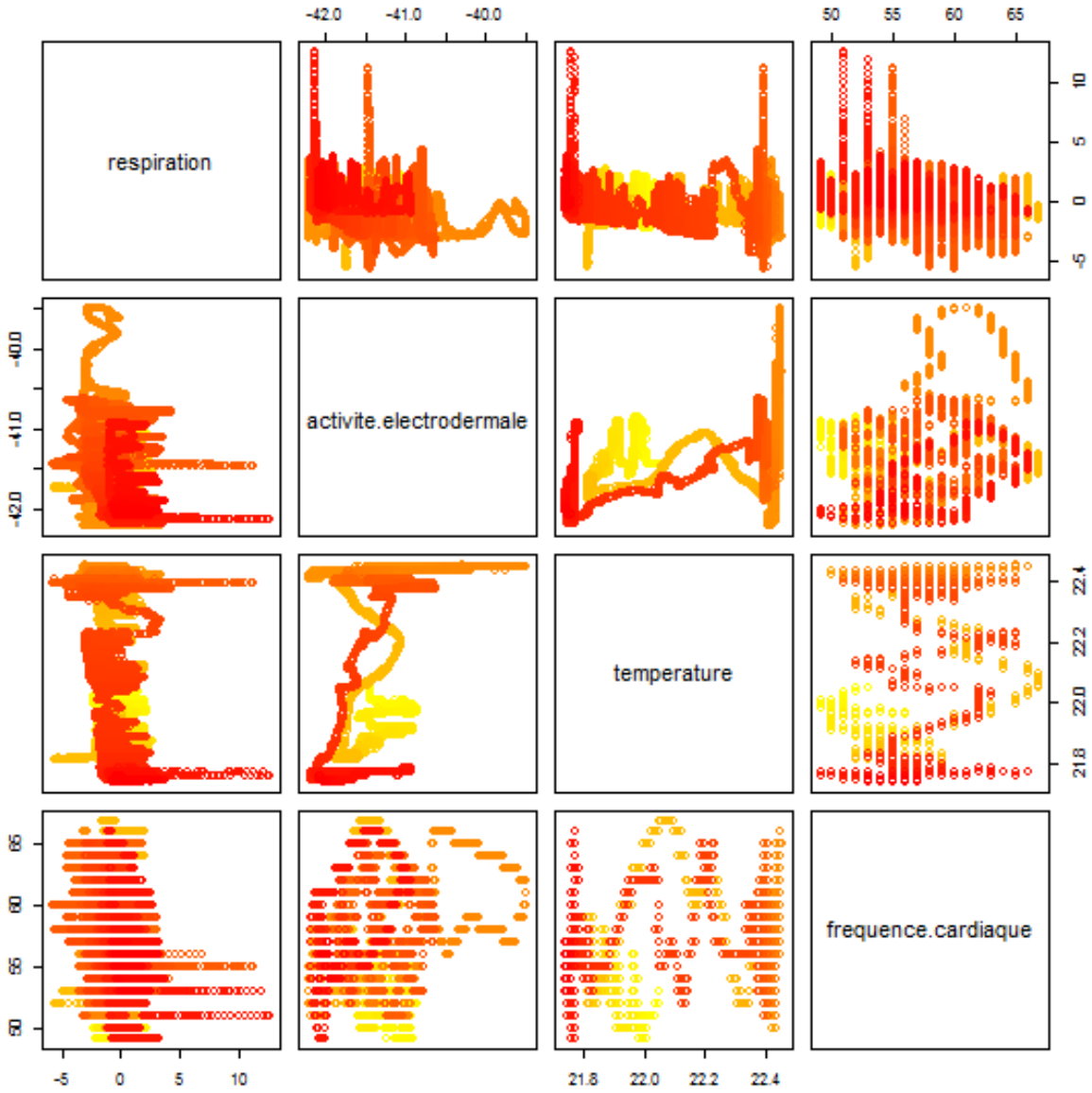

respiration
activite.electrodermale
temperature
frequence.cardiaque

Figure 22:

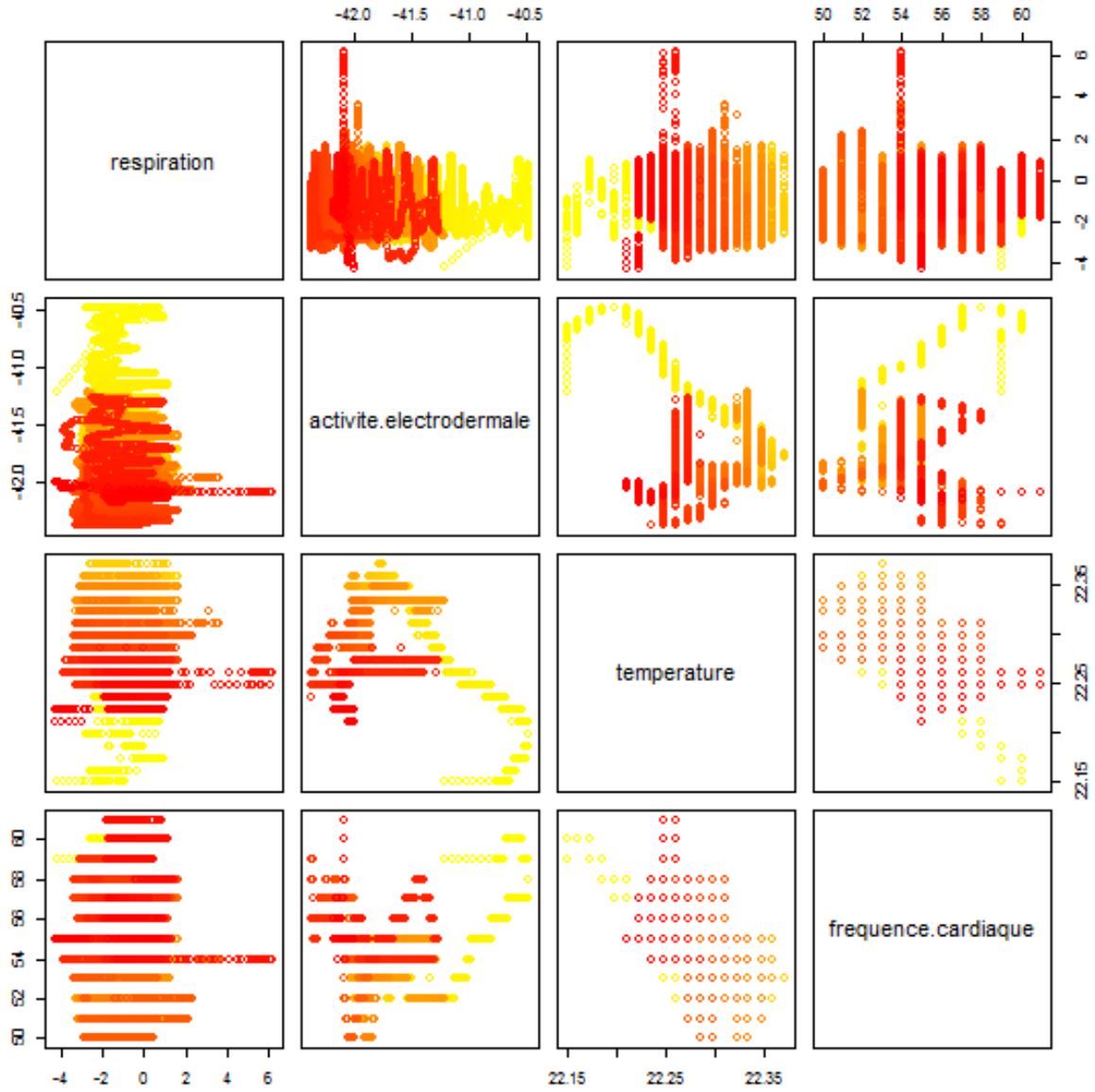

respiration
activite.electrodermale
temperature
frequence.cardiaque
-42.0
-41.5
-41.0
-40.5
50
52
54
56
58
60
-4
-2
0
2
4
6
22.15
22.25
22.35

Figure 23: