

Mesures physiologiques de joueurs de jeu vidéo (2)

Deuxième partie : analyse statistique et fouille des données

Godefroy Clair

Monday, July 13, 2015

Contents

1	Introduction	2
2	Selection des données	2
3	Fouille des données	2
3.1	Quelques statistiques descriptives	2
3.2	Quelques représentations graphiques	3
3.2.1	Pour la respiration :	4
3.2.2	Pour l'activité électrodermale :	5
3.2.3	Pour la température :	6
3.2.4	Pour la fréquence cardiaque :	7
3.3	Première fouille des données	8
3.3.1	Cartes auto-adaptatives (SOM)	8
3.4	Dendrogrammes des neurones.	33
3.4.1	Dendrogrammes par régions	35
3.4.2	Dendrogramme des autres expériences	38
3.4.3	Nouvelle projection du temps sur les groupes de référents	39
4	Conclusion :	41

1 Introduction

Comme nous avions conclu dans la partie précédente en soulevant quelques doutes concernant *certaines* mesures issues des expériences, nous allons procéder à une selection des expériences qui nous semblent en accord avec les plages de valeurs attendues. Nous procéderons ensuite à une fouille statistique sur cette selection.

2 Selection des données

Nous avons à notre disposition 12 expériences nommées ‘AB’, ‘CLP’, ‘CW’, ‘DA’, ‘FS1’, ‘HL’, ‘LM’, ‘PCo’, ‘PCo2’, ‘PCo3’, ‘DE’ et ‘ST’. Nous allons mettre de côté celles dont les mesures nous paraissent trop incertaines pour pouvoir être utilisées. Dans un second temps, il sera possible d’envisager de faire une selection par variable : au lieu de supprimer une expérience complète, on ne supprime que les variables à écarter.

Nous proposons d’écarter les expériences suivantes :

- ‘FS1’ à cause de la variable respiration
- ‘LM’ à cause de la variable respiration
- ‘HL’ à cause de la variable activité électrodermale (transpiration)
- ‘ST’ à cause de la variable activité électrodermale (transpiration)

Il nous reste donc les expériences ‘AB’, ‘CLP’, ‘CW’, ‘DA’, ‘PCo’, ‘PCo2’, ‘PCo3’ et ‘DE’, soit 8 expériences.

```
df.selec <- df.all[!(df.all$nom.experience %in% c("FS1", "LM", "HL", "ST")),]  
list.expe.selec <- c("AB", "CLP", "CW", "DA", "PCo", "PCo2", "PCo3", "DE")
```

3 Fouille des données

Nous allons commencer par fournir quelques statistiques et graphiques pour se donner une vue d’ensemble.

3.1 Quelques statistiques descriptives

Quelques statistiques pour synthétiser l’ensemble données :

activite.electrodermale	temperature	frequence.cardiaque	quart.temps
Min. :-47.77	Min. :21.73	Min. : 48.00	1er :65349
1st Qu.:-44.78	1st Qu.:22.47	1st Qu.: 58.00	2eme:65348
Median :-42.80	Median :27.06	Median : 67.00	3eme:65344
Mean :-42.47	Mean :27.72	Mean : 67.33	4eme:65349
3rd Qu.:-41.32	3rd Qu.:32.80	3rd Qu.: 75.00	NA
Max. :-30.81	Max. :35.00	Max. :109.00	NA

Nous ajoutons la déviation standard par variable:

```
## activite.electrodermale          temperature      frequence.cardiaque  
##             3.204418            4.338599        10.565698  
## quart.temps                      1.118046
```

3.2 Quelques représentations graphiques

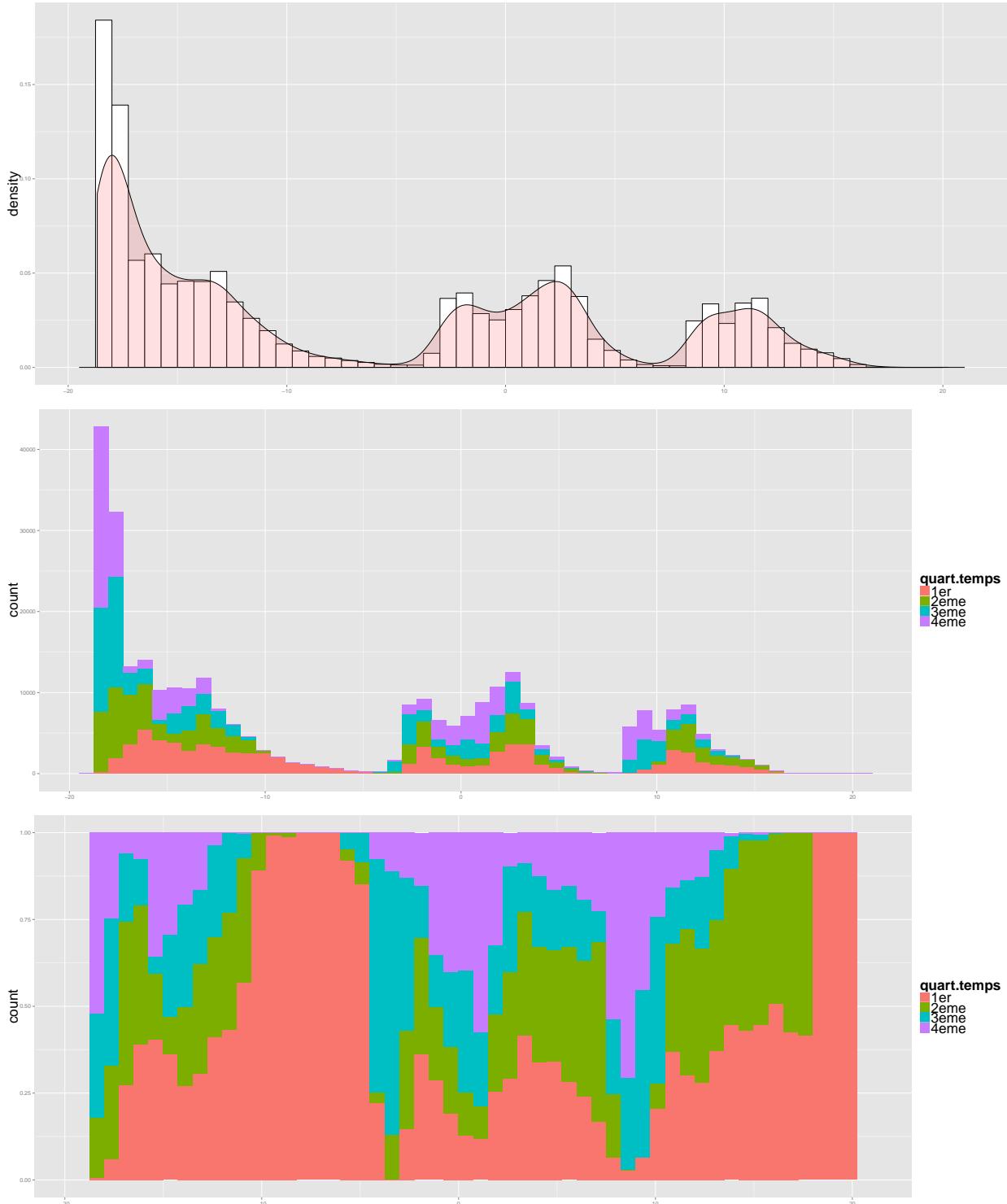
Nous allons proposer pour chaque variable trois histogrammes différents. Ils permettent de se donner une idée de la répartition globale des données :

-Le premier superpose à un histogramme classique, une courbe de densité. Il permet de “lisser” les variations et facilite le rapprochement avec une distribution de variable aléatoire.

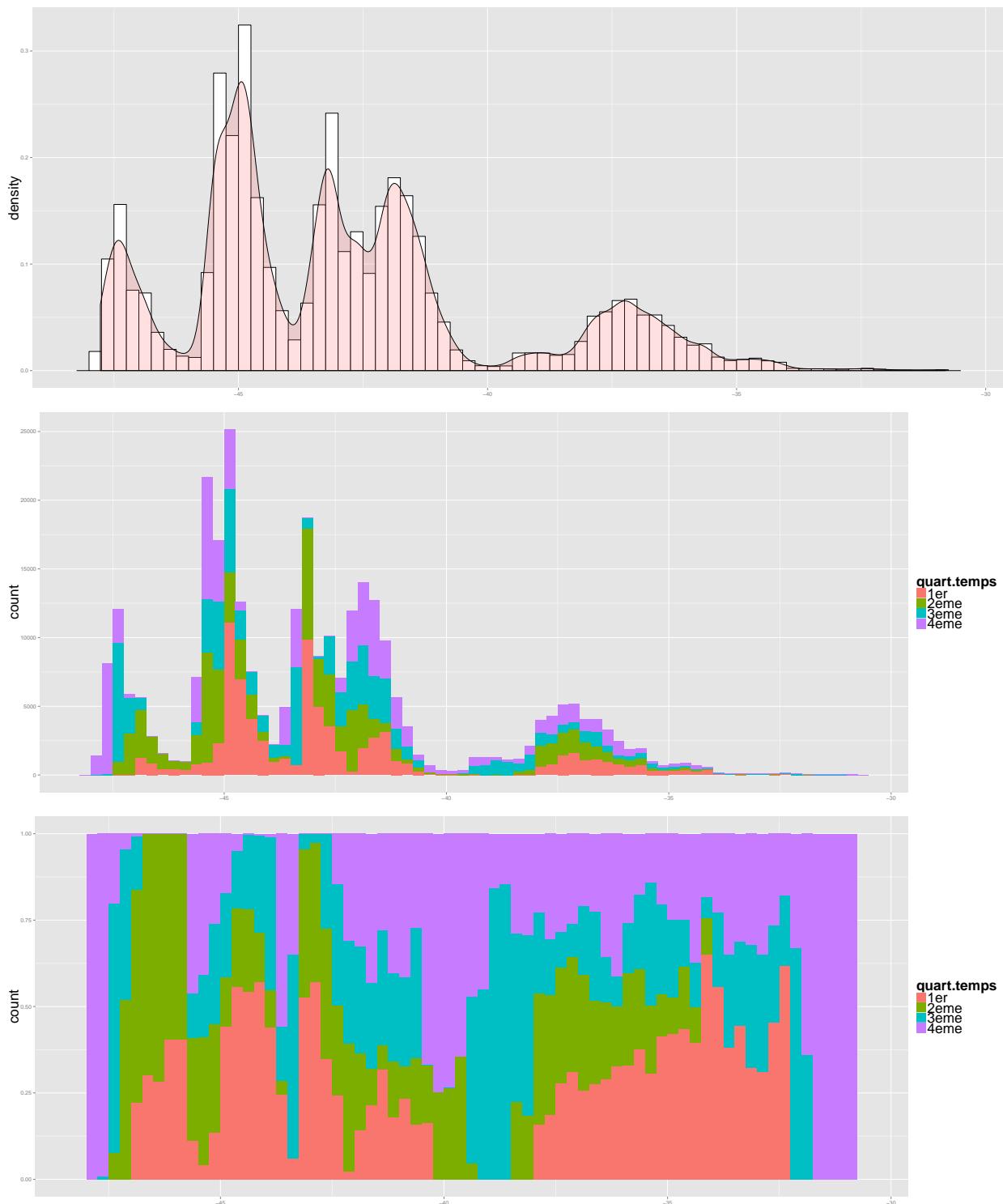
-Le second histogramme nous montre comment se décompose chacunr des barres de l'histogramme entre les 4 quart temps.

-Dans le troisième, les variations entre les barres sont gommées pour ne laisser voir que la répartition de des données entre les 4 goupes pour chaque intervalle.

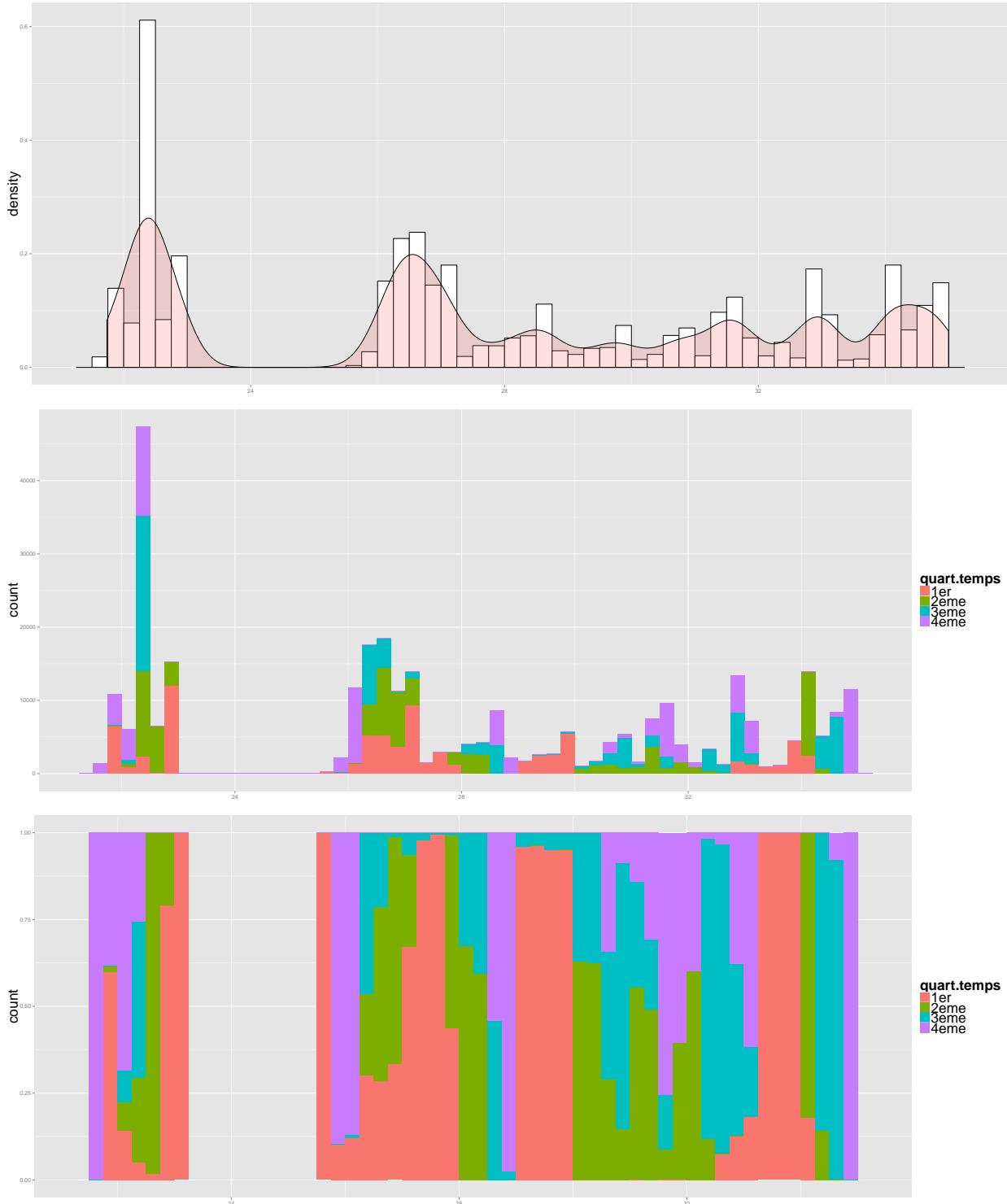
3.2.1 Pour la respiration :



3.2.2 Pour l'activité electrodermale :

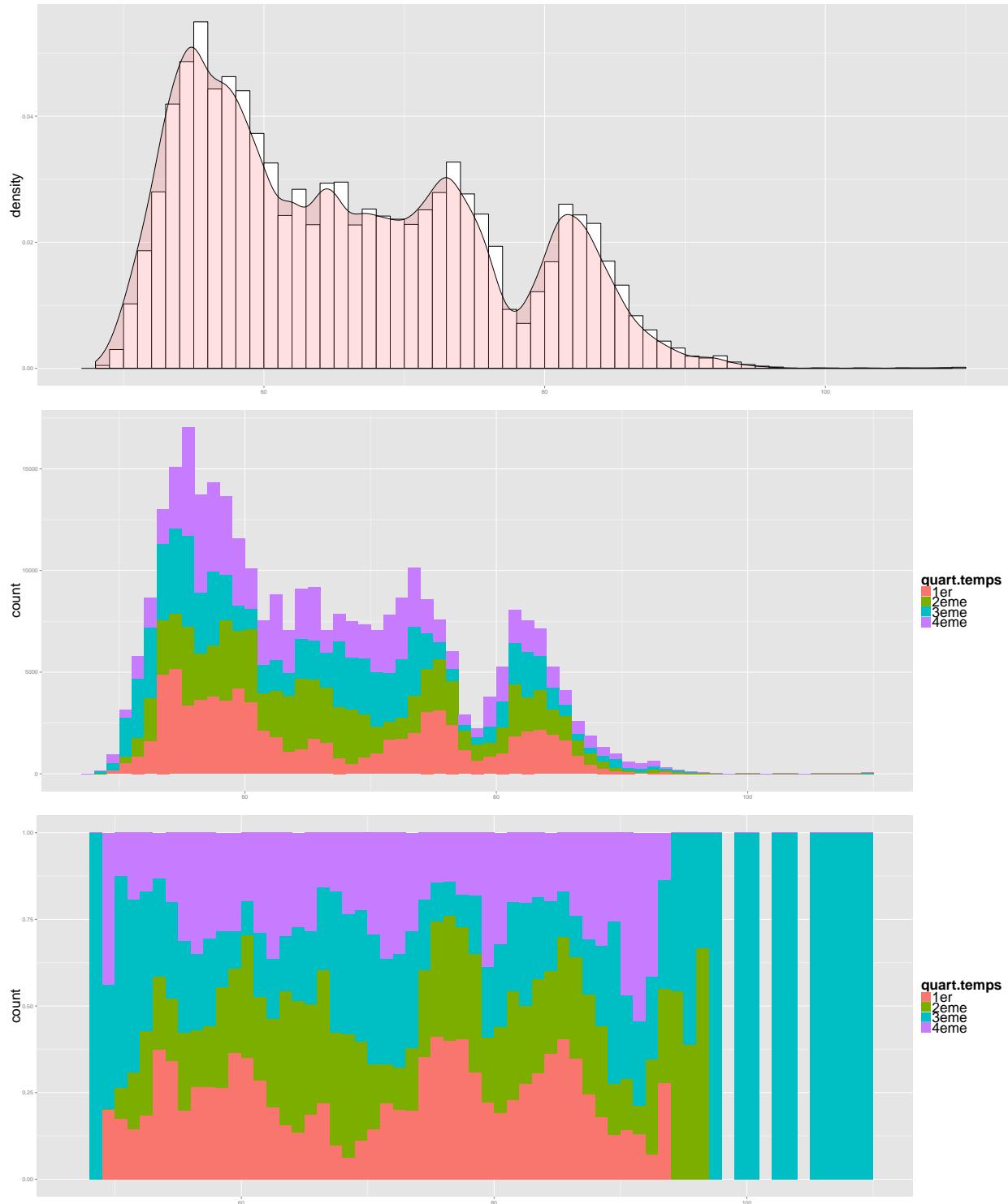


3.2.3 Pour la température :



3.2.4 Pour la fréquence cardiaque :

Différents histogrammes pour la fréquence cardiaque :



3.3 Première fouille des données

3.3.1 Cartes auto-adaptatives (SOM)

Nous allons procéder à une simulation grâce à la bibliothèque “kohonen” implémentée sous ‘R’. Après avoir centré et réduit l’ensemble des données, nous allons créer une carte de Kohonen de taille 40x40 (ie une grille de 40 neurones - ou référents - par 40) avec une topologie hexagonale.

La bibliothèque Kohonen sous R¹ permet de créer une telle carte grâce à la fonction “som”². La distance utilisée est la distance euclidienne. Le nombre d’itération peut être fixé par le modélisateur. Par défaut, la “courbe d’apprentissage”, indiquant la pondération donnée au poids que chaque nouvelle unité d’un neurone sur la “similarité” de ce neurone, diminue linéairement de 0.05 à 0.01 à chaque itération et le “radius” de voisinage : la taille initiale du voisinage de chaque neurone et la fonction caractérisant son évolution.

Comme nous voulons donner autant de poids à toutes les variables, nous centrons et réduisons les 4 variables.

```
#obtenir un dégradé de couleurs de bleu à rouge
coolBlueHotRed <- function(n, alpha = 1) {rainbow(n, end=4/6, alpha=alpha)[n:1]}

#les variables sont centrées et réduites
df.sc <- scale(df.selec[,2:5])

#verification en prenant les 1ères lignes
kable(head(df.sc))
```

respiration	activite.electrodermale	temperature	frequence.cardiaque
-0.3523725	-0.3967906	0.3522169	0.7847289
-0.3279903	-0.3967906	0.3522169	0.7847289
-0.3085767	-0.3967906	0.3522169	0.7847289
-0.2913712	-0.3967906	0.3522169	0.7847289
-0.2777540	-0.4005354	0.3522169	0.7847289
-0.2647809	-0.4005354	0.3522169	0.7847289

```
set.seed(77)
#calcul de l'a cart' algorithme d'attribution des données aux neurones
#rlen permet de préciser le nombre d'itérations
som1 <- som(data = df.sc, grid = somgrid(30, 30, "hexagonal"), rlen=75)
```

```
## Warning in aperm.default(X, c(s.call, s.ans)): Reached total allocation of
## 8089Mb: see help(memory.size)
```

```
## Warning in aperm.default(X, c(s.call, s.ans)): Reached total allocation of
## 8089Mb: see help(memory.size)
```

```
## Warning in aperm.default(X, c(s.call, s.ans)): Reached total allocation of
## 8089Mb: see help(memory.size)
```

```
## Warning in aperm.default(X, c(s.call, s.ans)): Reached total allocation of
## 8089Mb: see help(memory.size)
```

¹voir <https://cran.r-project.org/web/packages/kohonen/kohonen.pdf>

²voir <http://www.jstatsoft.org/v21/i05/paper>

Nous pouvons aussi voir une synthèse du résultat de l'algorithme

```
summary(som1)

## som map of size 30x30 with a hexagonal topology.
## Training data included; dimension is 310193 by 4
## Mean distance to the closest unit in the map: 0.004325715

#enregistrement des données
som1.save <- paste(data.path , "som1.Rda", sep = "/")
save(som1,file=som1.save)
```

3.3.1.1 Vérification Un certain nombre de graphiques sont disponibles pour aider à juger du bon déroulement de l'algorithme et ensuite permettre l'interprétation et la visualisation des résultats. La plupart ont une base commune : chaque neurone est représenté sur la carte par un disque ayant certaines caractéristiques esthétiques (couleurs, transparencies...) et géométriques (courbes, camemberts...) qui permettent de visualiser certaines propriétés ou valeurs associées à ce neurone. A noter que, par convention, on considère que la lecture se fait de gauche à droite et de bas en haut. Ainsi, le *premier* référent est celui en bas à gauche et le *dernier* se situe en haut à droite.

Le premier graphique est le “Training Progress”. Il donne l'évolution de la distance moyenne des vecteurs de données au neurone auquel elles ont été attribuées.

```
plot(som1, type="changes", main="")
```

L'apprentissage semble s'être correctement déroulé : après avoir continuement diminuée, la distance moyenne des données aux neurone a atteint un plateau un peu avant la 40ème itération. Il ne semble donc pas nécessaire de faire plus d'itérations.

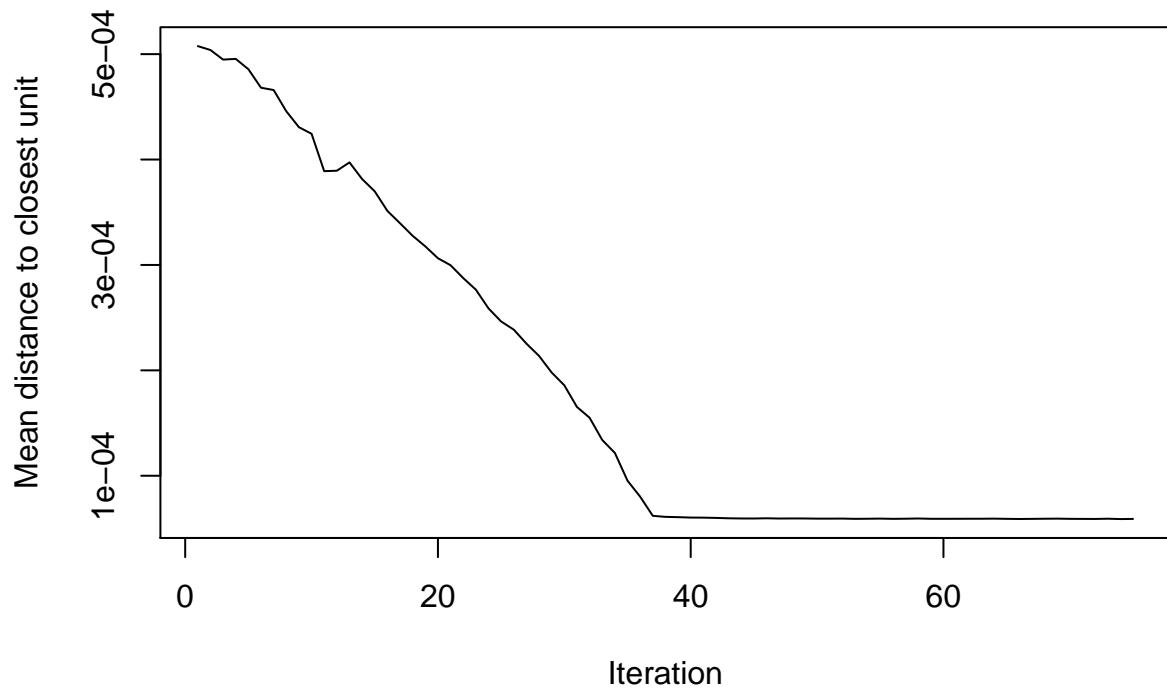


Figure 1: carte du progrès d'apprentissage

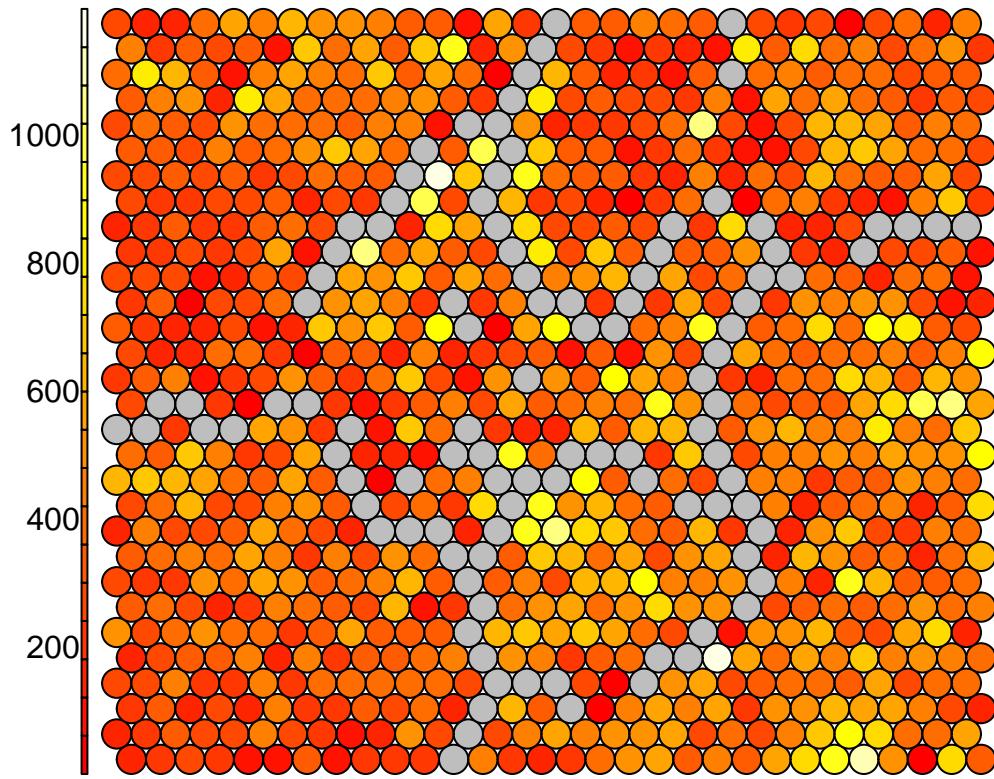


Figure 2: carte de comptages des données captées

Une autre carte utile pour voir comment s'est déroulé l'apprentissage est la carte “Node Counts” qui permet de visualiser comment se sont distribués les données entre les neurones : y a-t-il des neurones qui ont capté une grande partie des données ou la répartition est plus ou moins égalitaire ? y a-t-il des neurones qui ne captent pas de données ?

```
plot(som1, type="count", main= "")
```

Nous voyions (figure 2) que la plupart des neurones se sont vu attribuer un nombre de données compris entre 100 et 1000. Autre point intéressant, un certain nombre de neurones n'ont aucune donnée associée (en gris sur le graphique) et l'ensemble de ces neurones “vides” semblent former des frontières qui séparent les données en 5 groupes.

Une autre carte intéressante est la carte dite de “qualité” (figure 3) qui montre la distance de chaque neurone aux données qui lui ont été attribuées.

```
plot(som1, type="quality", palette.name = coolBlueHotRed, main = "")
```

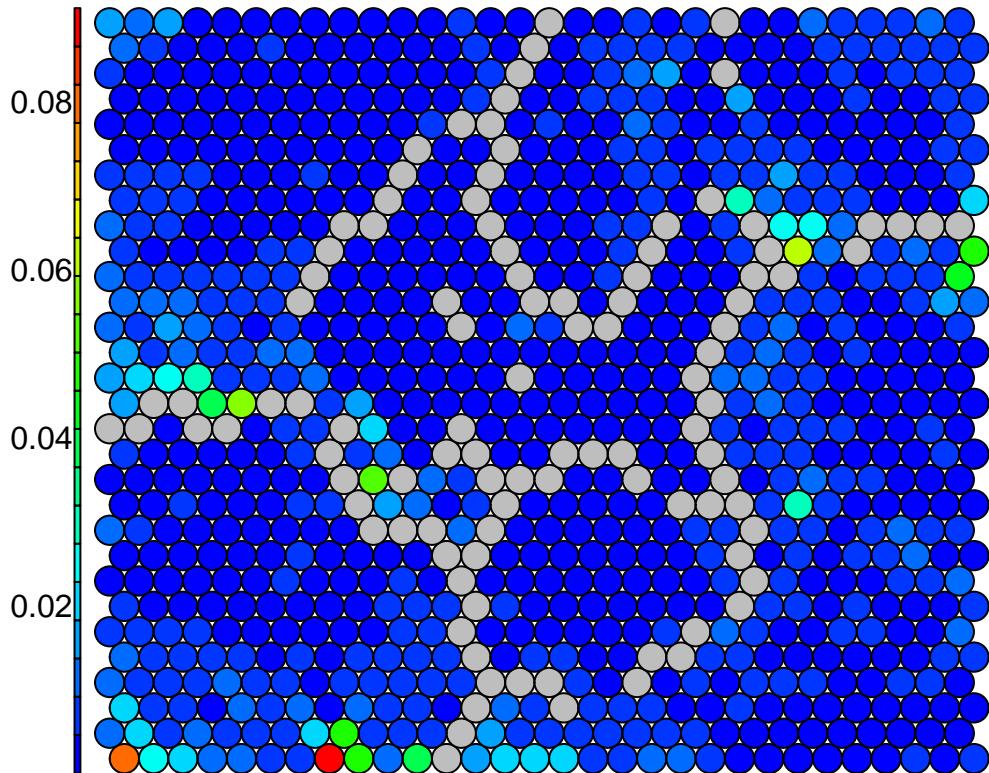


Figure 3: carte de “qualité” : distance moyenne des données aux neurones

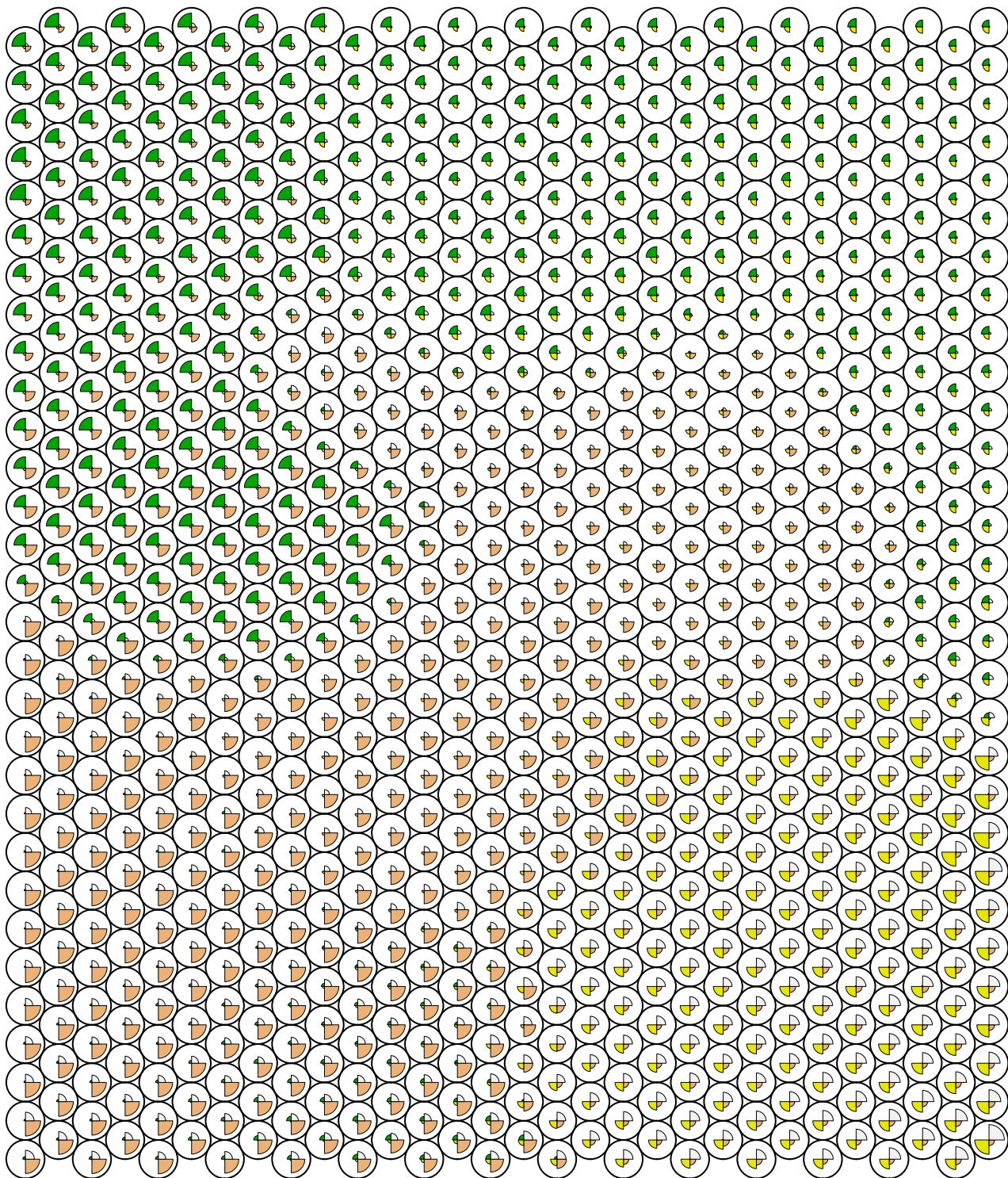
Nous voyions que la distance est *faible*, assez *uniforme* et montrant peu de cas “aberrants” ; ce qui laisse là aussi présager d’une bonne répartition des données. On remarque aussi que les neurones qui sur la carte précédente n’avaient pas de données associées n’ont pas de distance (couleur grise).

Il est intéressant de comparer aux premières cartes de Kohonen réalisées à partir d’expériences particulières³ avec des données non-traitées. On voit que, si dans le cas présent la distance est globalement bien inférieure (<0.01) et que seuls deux neurones voient cette distance moyenne être supérieure à 0.05 (pour les expériences AB et LM, nous trouvions 10% des neurones ayant une distance moyenne à ses données de plus de 0.5 (et presque 5% avec une distance supérieure à 1).

³voir les fichiers *experience_AB.html* et *experience_LM.html* dans le dossier joint en annexe.

3.3.1.2 Interprétation La carte suivante (figure 4) est un premier élément d’interprétation de la carte : cette carte dite des “codebook vectors” indique quelles sont les caractéristiques de la donnée moyenne associées à chaque neurone. On peut ainsi parler de la carte d’identité de chaque neurone. Pour représenter cela dans le graphe, tout neurone de la carte (représenté par le biais d’un disque) contient des “camemberts” plus ou moins larges qui se partagent ainsi la surface de ces disques. Ils représentent la valeur moyenne des données captées par le neurone. Ainsi, si la surface du camembert représentant la température est la plus grande (comme c’est le cas au nord-est de la carte), alors cela signifie que, aux instants capturées par ce neurone, la température cutanée était relativement élevée.

Ainsi, nous voyions quand dans une zone nord-ouest de la carte, la température est élevée.



Du fait du nombre de neurones, la carte est difficilement lisible. On peut utiliser certaines “astuces” pour remédier à ce problème.

D’abord, on peut travailler avec une carte plus petite. (voir figure 5)

respiration	activite.electrodermale	temperature	frequence.cardiaque
-0.3523725	-0.3967906	0.3522169	0.7847289
-0.3279903	-0.3967906	0.3522169	0.7847289
-0.3085767	-0.3967906	0.3522169	0.7847289
-0.2913712	-0.3967906	0.3522169	0.7847289
-0.2777540	-0.4005354	0.3522169	0.7847289
-0.2647809	-0.4005354	0.3522169	0.7847289

La carte suivante (figure 6) a été créée avec une topologie de 100 (ie 10 * 10) neurones.

```
plot(som2, main = "carte SOM")
```

Vu le petit nombre de variables, nous pouvons aussi proposer des cartes “heatmaps” qui permettent de regarder la valeur moyenne associée à chaque neurone valeur par valeur :

Heatmap de la respiration : voir figure 6

Heatmap de l’activité électrodermale : voir figure 7

Heatmap de la température : voir figure 8

Heatmap de la fréquence cardiaque : voir figure 9

Synthèse des heatmaps :

La température est un des facteurs les plus clivants, la respiration varie aussi beaucoup entre un gros quart “sud-est” et le reste de la carte. Il est intéressant de superposer température et respiration (voir figure 10) : les deux cartes semblent varier à l’opposé l’une de l’autre.

Heatmap de la respiration :

visualisation des véritables valeurs associées aux neurones

Cela consiste à présenter la variable avant la normalisation. Cela demande sous R un travail préalable

Sur la figure 11, on peut voir que les 4 cartes se superposent et que l’on peut définir un certain nombre de zones où les variables varient peu.

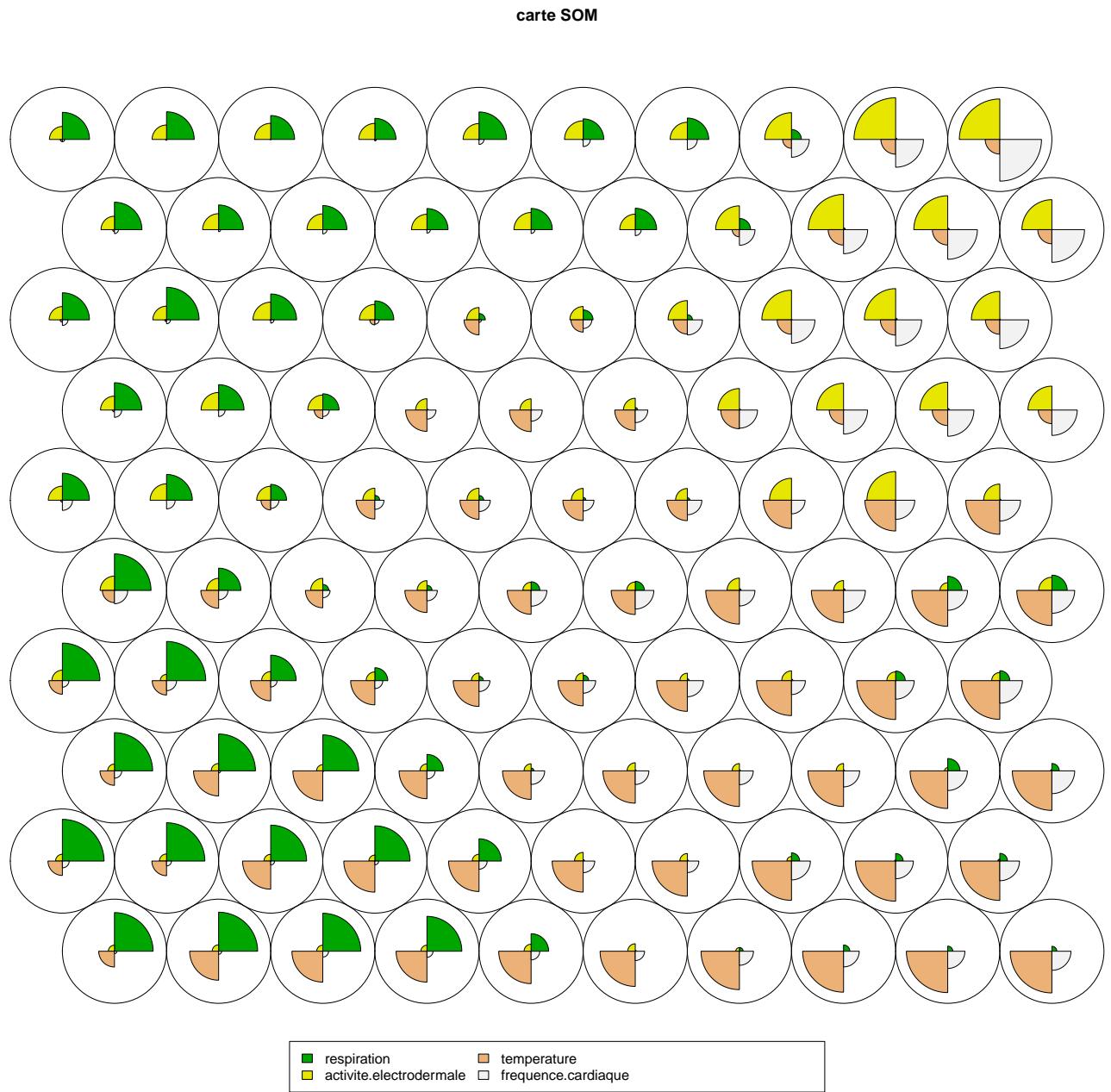


Figure 5: carte des neurones (10x10)

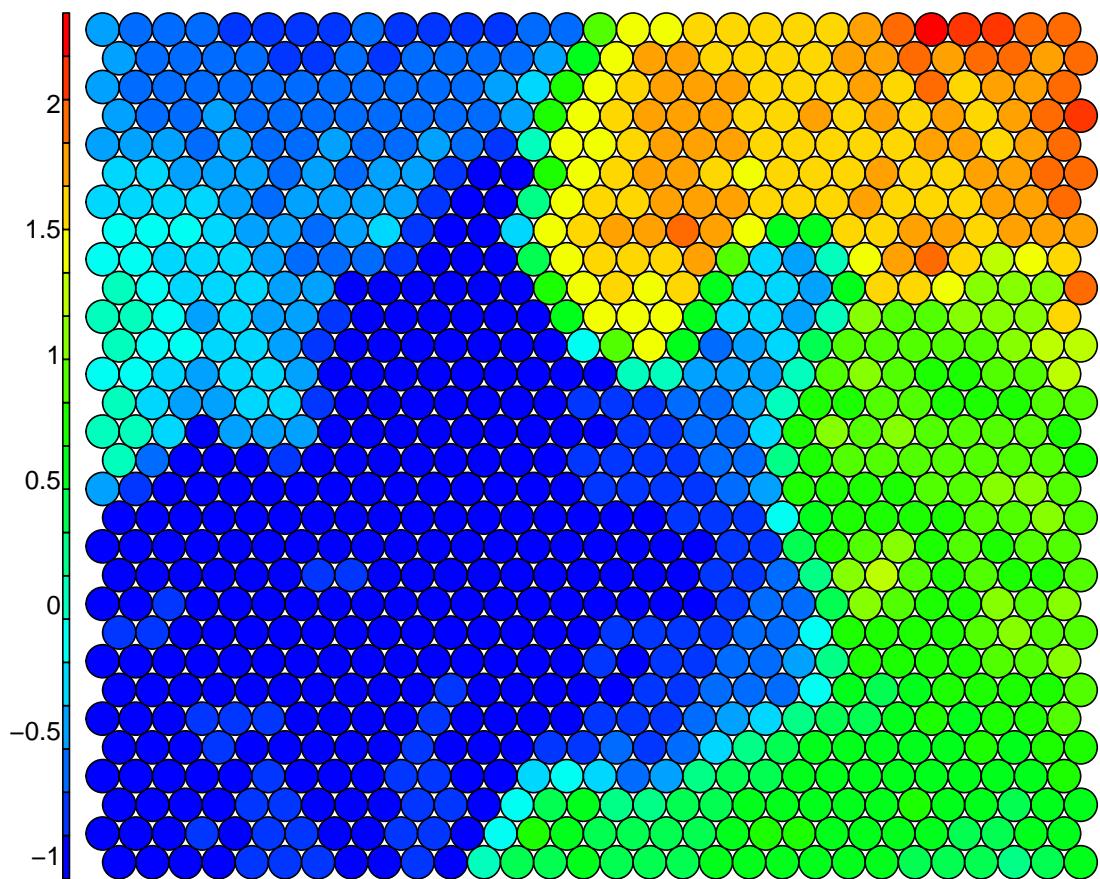


Figure 6: carte heatmap de la respiration

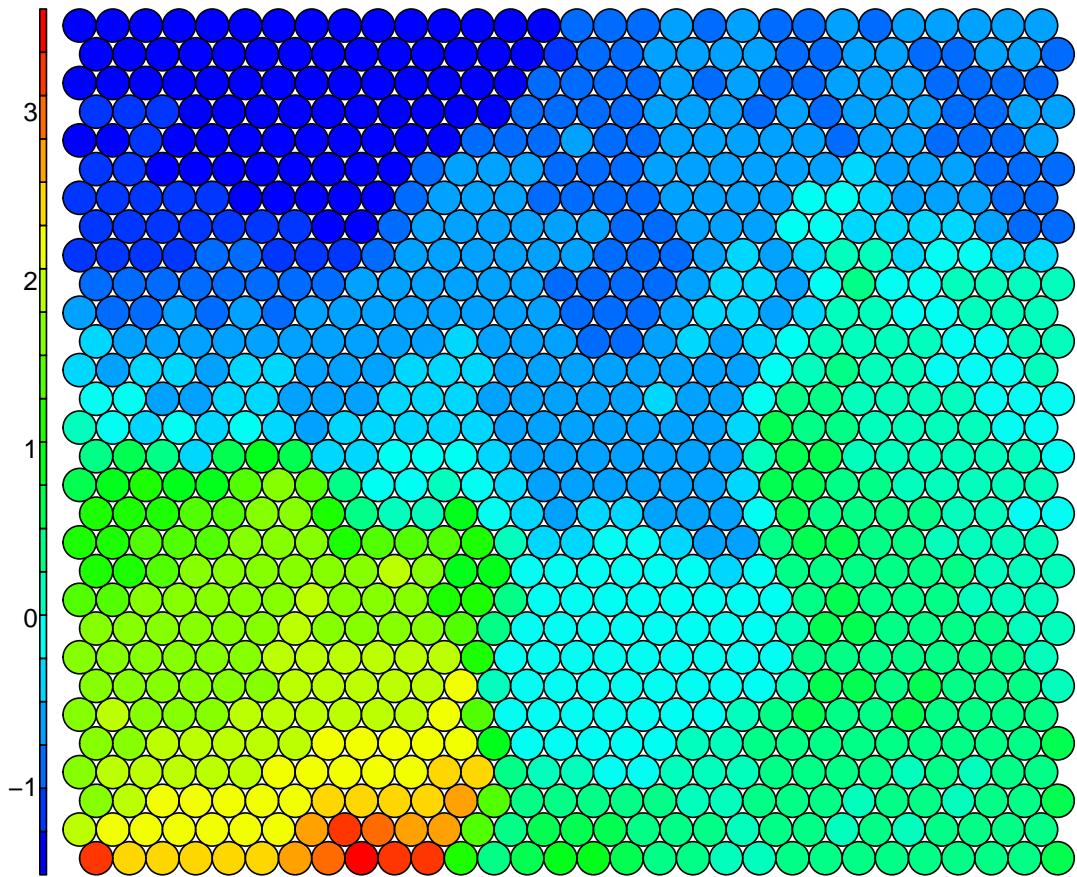


Figure 7: carte heatmap de l'activité électrodermale

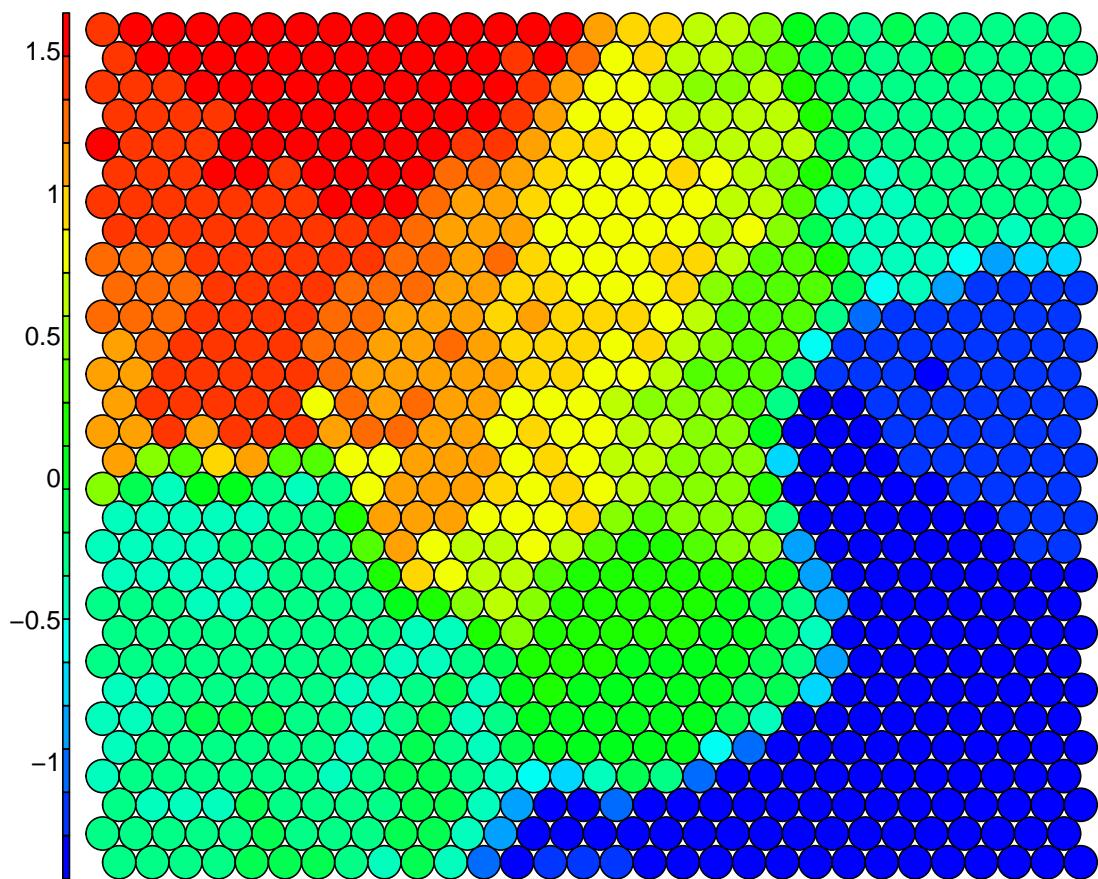


Figure 8: carte heatmap de la température

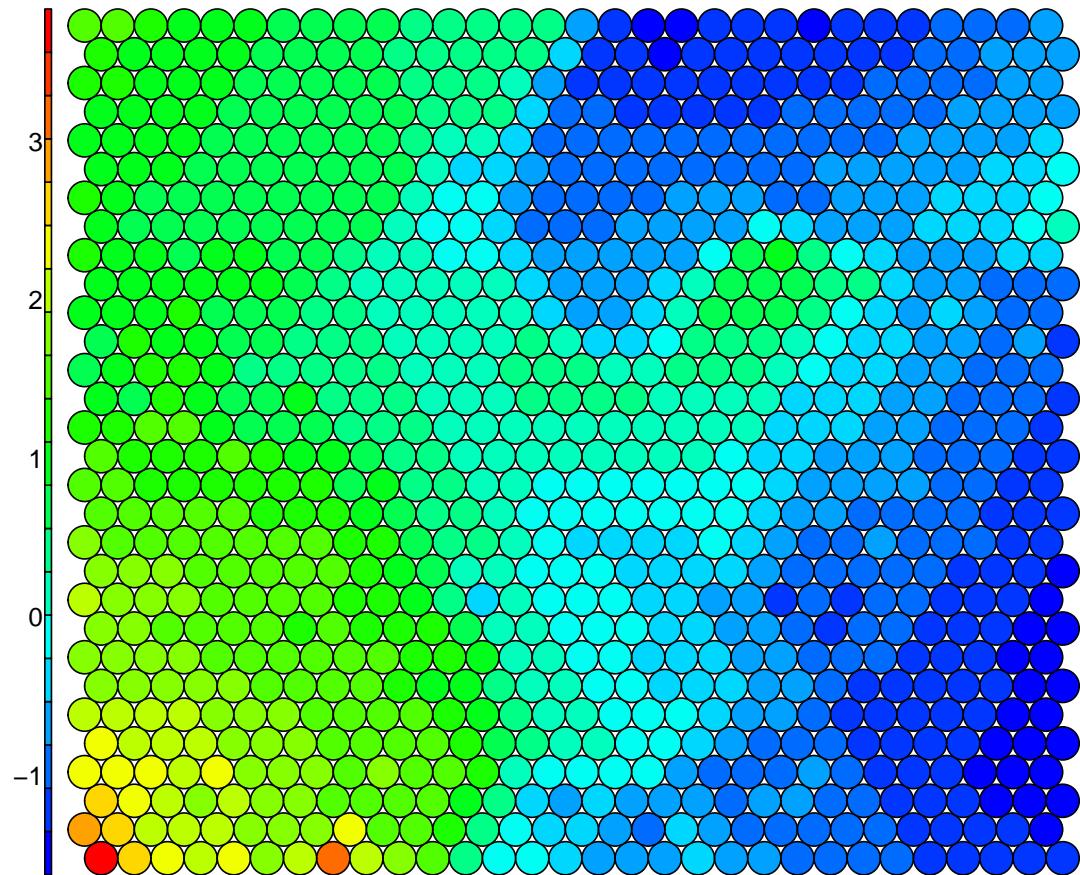


Figure 9: carte heatmap de la fréquence cardiaque

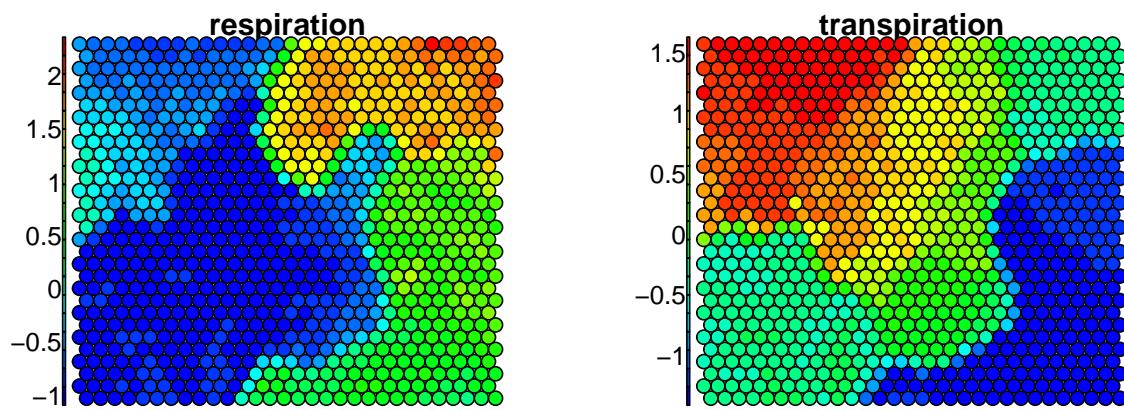


Figure 10: comparaison des “heatmaps” de transpiration et respiration

Voyons avec la figure 12, comment les différentes expériences se partagent la carte en observant quelle est l’expérience majoritaire pour chaque neurone, celle dont les données sont les plus nombreux à lui avoir été associées.

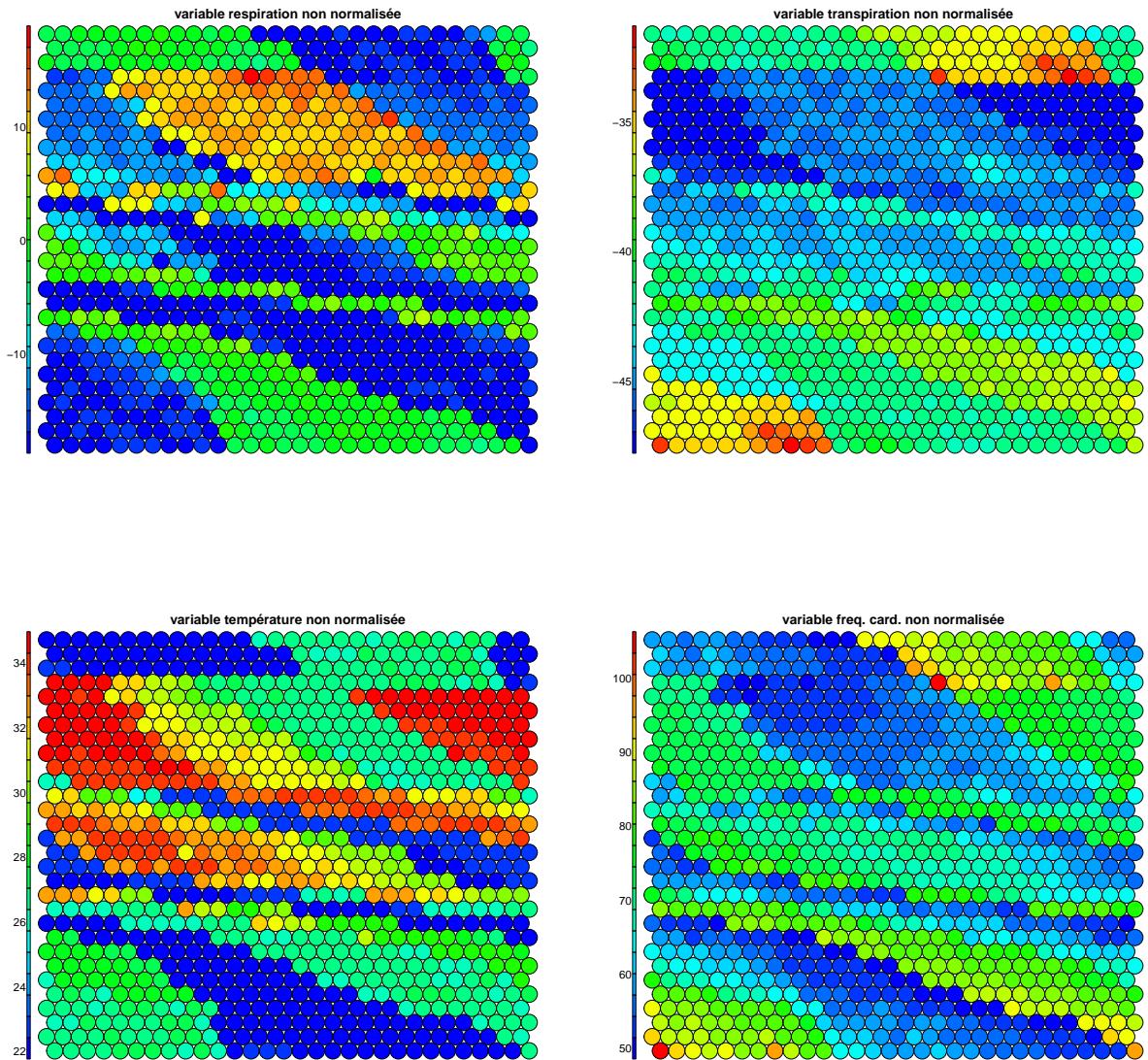


Figure 11: heatmaps des variables non-normalisées

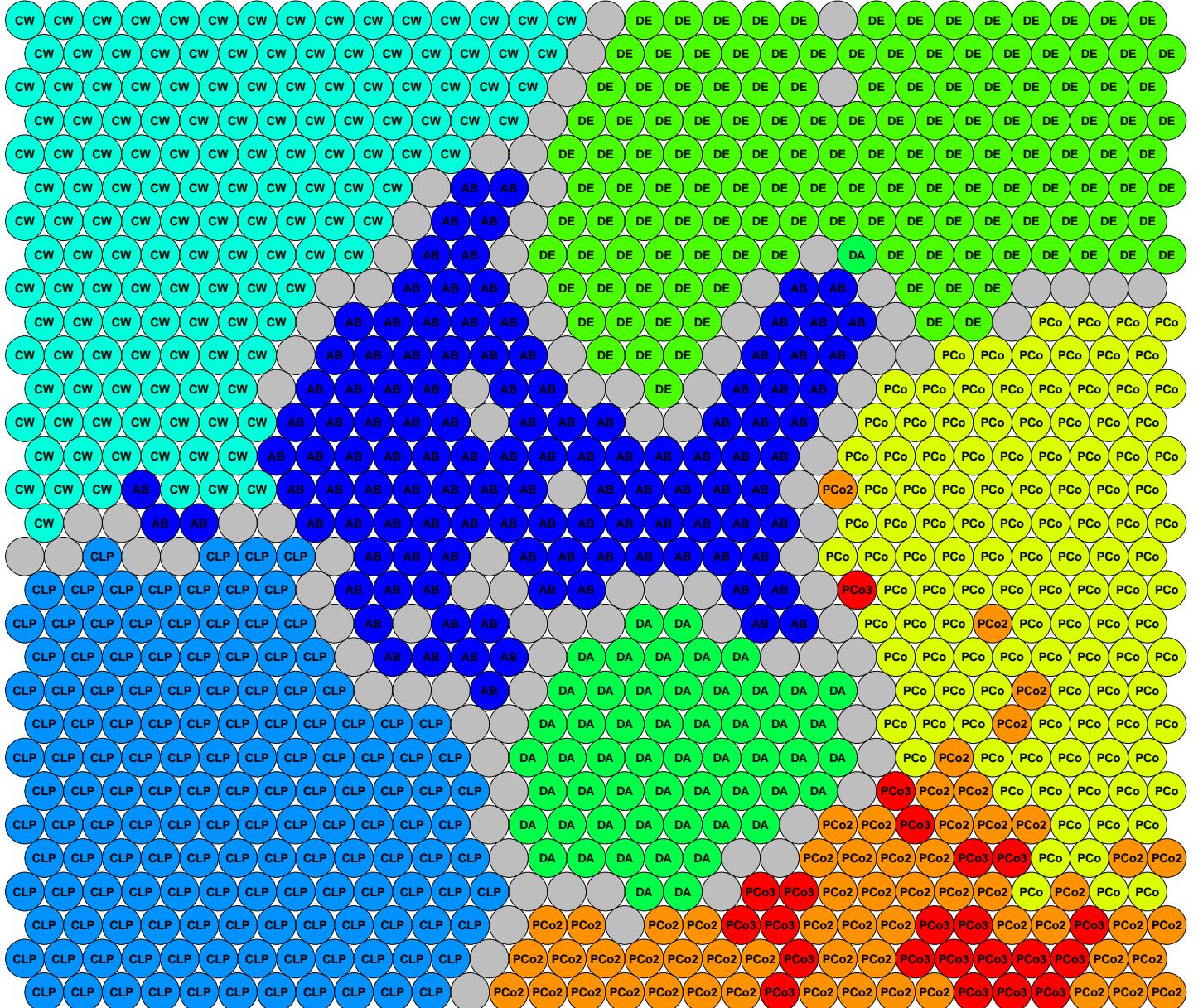


Figure 12: expérience majoritaire par neurone

On remarque que chaque expérience se confine dans une partie de la carte, ce qui laisserait entendre que l'apprentissage a pris en compte les particularités de chaque expérience dans son mapping des données aux neurones.

Nous pouvons confirmer cette impression grâce au tableau suivant qui, pour un échantillon de neurones tirés uniformément parmi les 900, donne la répartition des expériences dont sont issues les données que chacun a capté. Nous en tirons 45, soit 5%. La première colonne nous donne le numéro du neurone tiré au hasard parmi les 900, les suivantes, le nombre de données issus de chaque expérience.

	AB	CLP	CW	DA	PCo	PCo2	PCo3	DE
633	0	0	358	0	0	0	0	0
298	0	0	0	0	99	46	44	0
714	0	0	0	0	0	0	0	280
482	0	0	163	0	0	0	0	0
817	0	0	553	0	0	0	0	0
527	0	0	0	0	0	0	0	0
92	0	300	0	0	0	0	0	0
579	546	0	0	0	0	0	0	0
556	0	0	0	0	0	0	0	0
503	0	0	0	0	226	124	21	0
253	0	0	0	0	0	0	0	0
218	0	304	0	0	0	0	0	0
3	0	294	0	0	0	0	0	0
754	0	0	298	0	0	0	0	0
709	0	0	0	0	0	0	0	203
541	0	0	233	0	0	0	0	0
90	0	0	0	0	0	136	0	0
259	0	0	0	274	0	0	0	0
839	0	0	0	0	0	0	0	320
148	0	0	0	0	347	134	29	0
426	0	0	0	0	0	0	0	0
810	0	0	0	0	0	0	0	256
859	0	0	0	0	0	0	0	194
231	0	0	0	429	0	0	0	0
273	0	251	0	0	0	0	0	0
604	0	0	228	0	0	0	0	0
46	0	0	0	0	0	553	0	0
318	0	0	0	393	0	0	0	0
507	0	0	0	0	440	0	0	0
460	376	0	0	0	0	0	0	0
358	0	0	0	0	464	3	14	0
453	0	0	448	0	0	0	0	0
627	0	0	0	0	249	0	0	0
83	0	0	0	0	159	242	89	0
440	488	0	0	0	0	0	0	0
330	0	0	0	0	777	0	0	0
79	0	0	0	0	0	163	396	0
233	0	0	0	0	0	0	0	0
494	300	0	0	0	0	0	0	0
873	0	0	141	0	0	0	0	0
865	0	0	0	0	0	0	0	384
558	0	0	0	0	0	0	0	0
522	892	0	0	0	0	0	0	0
875	0	0	546	0	0	0	0	0

	AB	CLP	CW	DA	PCo	PCo2	PCo3	DE
669	0	0	208	0	0	0	0	0

Nous observions bien qu'une grande partie des neurones ont capté des données liées à une seule expérience. En fait, par un calcul simple, on peut trouver que 797, soit 89% des neurones n'ont capté que les données d'une expérience.

Par ailleurs, concernant le graphique montrant l'expérience majoritaire par neurone (figure 12), il est intéressant de noter que les zones des expériences PCo, PCo3 et Pco3 (qui ont été réalisées sur une même personne) sont assez fortement entremêlés ou en tout cas pas aussi démarqués qu'avec d'autres individus.

Dans ce cadre, il est aussi intéressant de reprendre le tableau précédent mais cette fois en regardant les neurones associés à des données non-issues d'une unique expérience.

	AB	CLP	CW	DA	PCo	PCo2	PCo3	DE
17	0	0	0	0	0	371	25	0
18	0	0	0	0	0	450	34	0
19	0	0	0	0	0	426	140	0
20	0	0	0	0	0	145	166	0
25	0	0	0	0	0	309	594	0
26	0	0	0	0	0	90	1000	0
27	0	0	0	0	0	168	366	0
29	0	0	0	0	0	365	357	0
30	0	0	0	0	0	317	2	0
47	0	0	0	0	0	418	36	0
48	0	0	0	0	0	452	40	0
50	0	0	0	0	0	557	2	0
51	0	0	0	0	2	52	210	0
54	0	0	0	0	0	18	428	0
55	0	0	0	0	0	288	378	0
56	0	0	0	0	0	410	437	0
57	0	0	0	0	0	40	732	0
58	0	0	0	0	0	36	355	0
59	0	0	0	0	0	252	128	0
75	0	0	0	0	62	248	0	0
77	0	0	0	0	0	28	2	0
78	0	0	0	0	0	408	17	0
79	0	0	0	0	0	163	396	0
80	0	0	0	0	0	48	389	0
83	0	0	0	0	159	242	89	0
84	0	0	0	0	77	197	243	0
85	0	0	0	0	105	149	386	0
86	0	0	0	0	0	339	257	0
87	0	0	0	0	0	534	30	0
88	0	0	0	0	0	20	403	0
110	0	0	0	0	0	73	422	0
111	0	0	0	0	0	49	501	0
113	0	0	0	0	1	322	19	0
117	0	0	0	0	331	133	131	0
119	0	0	0	0	266	34	99	0
120	0	0	0	0	378	19	0	0
141	0	0	0	0	0	1123	75	0
142	0	0	0	0	0	527	13	0

	AB	CLP	CW	DA	PCo	PCo2	PCo3	DE
143	0	0	0	0	0	338	43	0
145	0	0	0	0	0	132	303	0
146	0	0	0	0	164	82	422	0
147	0	0	0	0	208	157	102	0
148	0	0	0	0	347	134	29	0
149	0	0	0	0	82	345	0	0
150	0	0	0	0	180	189	0	0
173	0	0	0	0	0	302	184	0
174	0	0	0	0	0	114	386	0
176	0	0	0	0	88	167	47	0
177	0	0	0	0	0	163	100	0
178	0	0	0	0	452	79	44	0
179	0	0	0	0	709	12	0	0
203	0	0	0	0	0	116	133	0
204	0	0	0	0	0	202	116	0
205	0	0	0	0	3	209	203	0
206	0	0	0	0	189	113	0	0
207	0	0	0	0	284	99	23	0
208	0	0	0	0	408	104	25	0
209	0	0	0	0	381	19	17	0
210	0	0	0	0	373	47	0	0
234	0	0	0	0	267	159	48	0
236	0	0	0	0	815	65	32	0
237	0	0	0	0	468	102	83	0
240	0	0	0	0	351	10	0	0
264	0	0	0	0	541	78	5	0
265	0	0	0	0	310	122	53	0
267	0	0	0	0	388	20	0	0
268	0	0	0	0	154	3	0	0
269	0	0	0	0	378	7	5	0
294	0	0	0	0	136	22	0	0
295	0	0	0	0	466	19	0	0
296	0	0	0	0	583	54	49	0
298	0	0	0	0	99	46	44	0
299	0	0	0	0	395	16	16	0
300	0	0	0	0	366	3	0	0
323	0	0	0	0	388	39	0	0
324	0	0	0	0	105	26	0	0
325	0	0	0	0	358	2	0	0
326	0	0	0	0	296	52	49	0
327	0	0	0	0	343	48	65	0
353	0	0	0	0	342	101	0	0
354	0	0	0	0	402	38	0	0
357	0	0	0	0	312	41	0	0
358	0	0	0	0	464	3	14	0
382	0	0	0	0	46	83	123	0
383	0	0	0	0	376	45	0	0
384	0	0	0	0	391	81	0	0
385	0	0	0	0	483	33	10	0
386	0	0	0	0	394	43	0	0
387	0	0	0	0	538	24	0	0
388	0	0	0	0	526	1	0	0

	AB	CLP	CW	DA	PCo	PCo2	PCo3	DE
412	0	0	0	0	174	113	69	0
413	0	0	0	0	495	34	0	0
414	0	0	0	0	583	28	0	0
415	0	0	0	0	290	121	15	0
416	0	0	0	0	434	12	0	0
442	0	0	0	0	392	20	0	0
443	0	0	0	0	276	6	0	0
444	0	0	0	0	310	18	0	0
445	0	0	0	0	322	0	7	0
472	0	0	0	0	65	173	0	0
502	0	0	0	0	538	17	0	0
503	0	0	0	0	226	124	21	0
600	0	0	0	0	77	19	0	0

Nous voyions là-aussi que les neurones ayant captées des données issues de différentes expériences sont pour leur grande majorité des données des expériences de l'individu “PCo”. Ainsi, les neurones mettent ensemble les données d' expériences issues d'un même individu et seulement celle-ci. Ainsi, il semble que les cartes de Kohonen soit capable avec ces données de distinguer des individus.

L'intérêt à terme de ce fait peut être grand (sur la question de la prédiction des particularités émotionnelles *individuelles* notamment) mais puisque nous cherchons à associer des neurones non par une personne particulière mais plus à un type de mesure (qui serait lié à une émotion), il nous faudra aborder le problème d'une manière différente.

Comme nous avons un certain nombre d'expériences non encore exploitées, nous pensons qu'en les intégrant à ces cartes, nous espérons que les différentes zones ne correspondent pas uniquement à un individu mais à un type d'individu. Le fait que les expériences issues d'un même individu soit “reconnues” pas la carte comme étant des données similaires nous confortent un peu dans cette possibilité.

3.3.1.3 Projection par quart-temps Si nous arrivons à découper la carte entre zones représentant des individus types, il faut ensuite caractériser pour chaque zone à quels périodes correspond chaque neurone : y a-t-il des périodes prolongées qui sont captées par certains neurones ?

Dans un premier temps, pour le savoir, nous pouvons faire une projection par quart-temps en ajoutant sur la carte le quart-temps majoritaire dont sont issues les expériences.

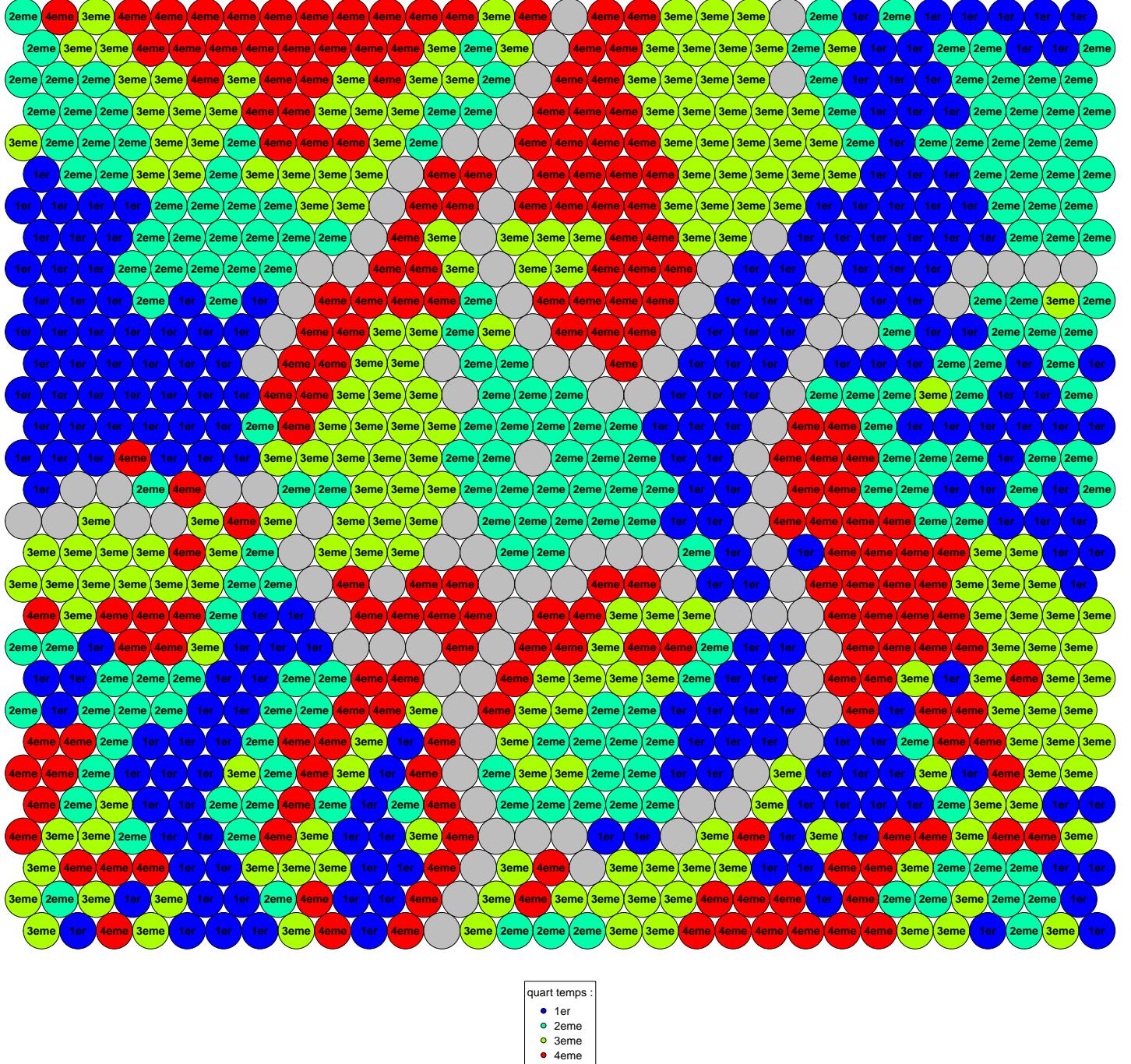


Figure 13: quart temps majoritaire par neurone

Nous pouvons recouper expériences et quart temps :

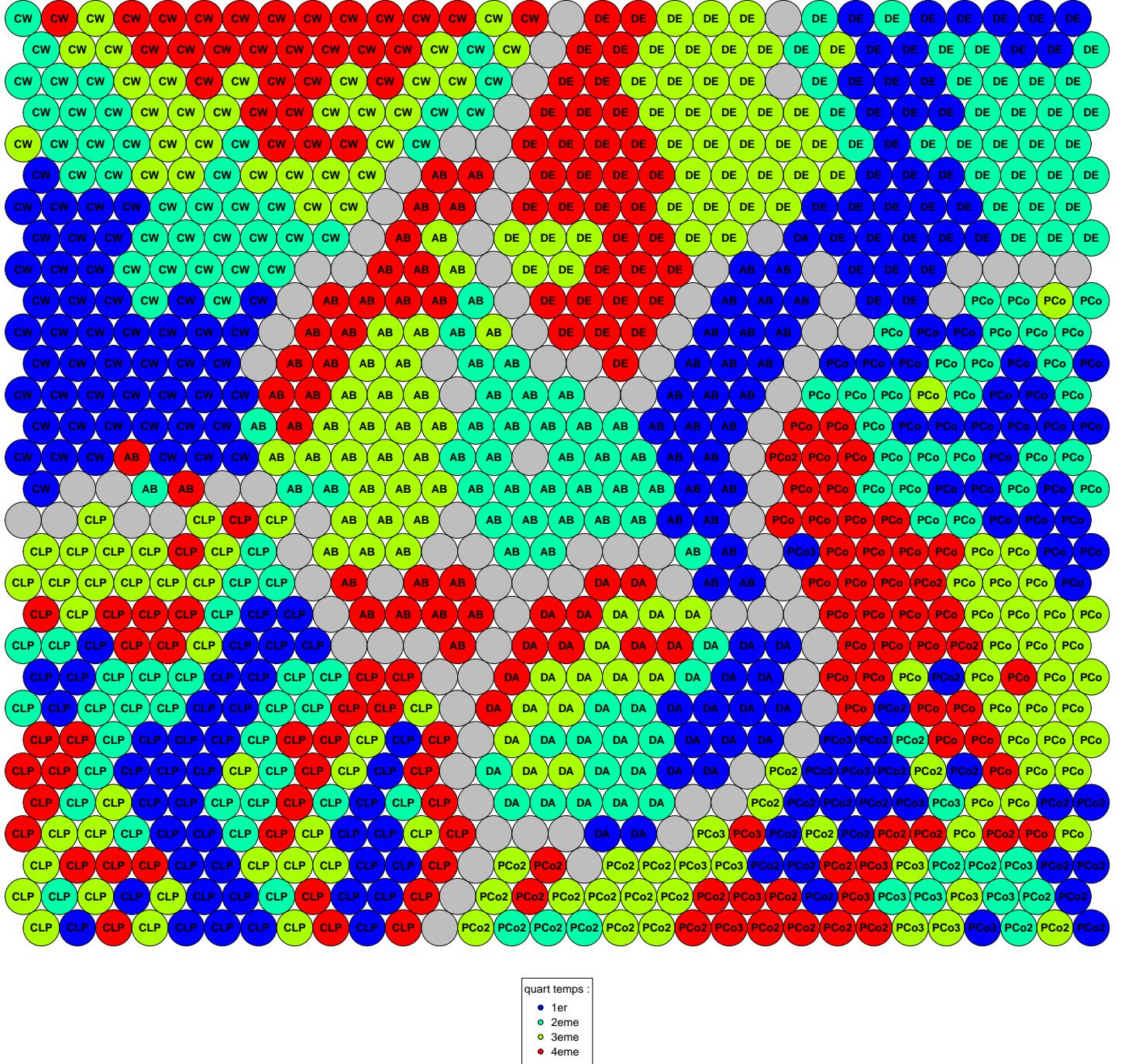


Figure 14: expérience et quart temps (couleur) majoritaire par neurone

Parmi les 900 neurones, combien de neurones captent des données issues de plusieurs quart temps ?

Il y a 458 neurones qui captent des données issues d'au moins deux quart-temps. Ce qui est donc la majorité...

3.3.1.4 Projection du temps sur les référents Sur chacun des graphiques qui forment la figure 15, nous pouvons voir comment évolue les données d'une expérience entre les différentes neurones.

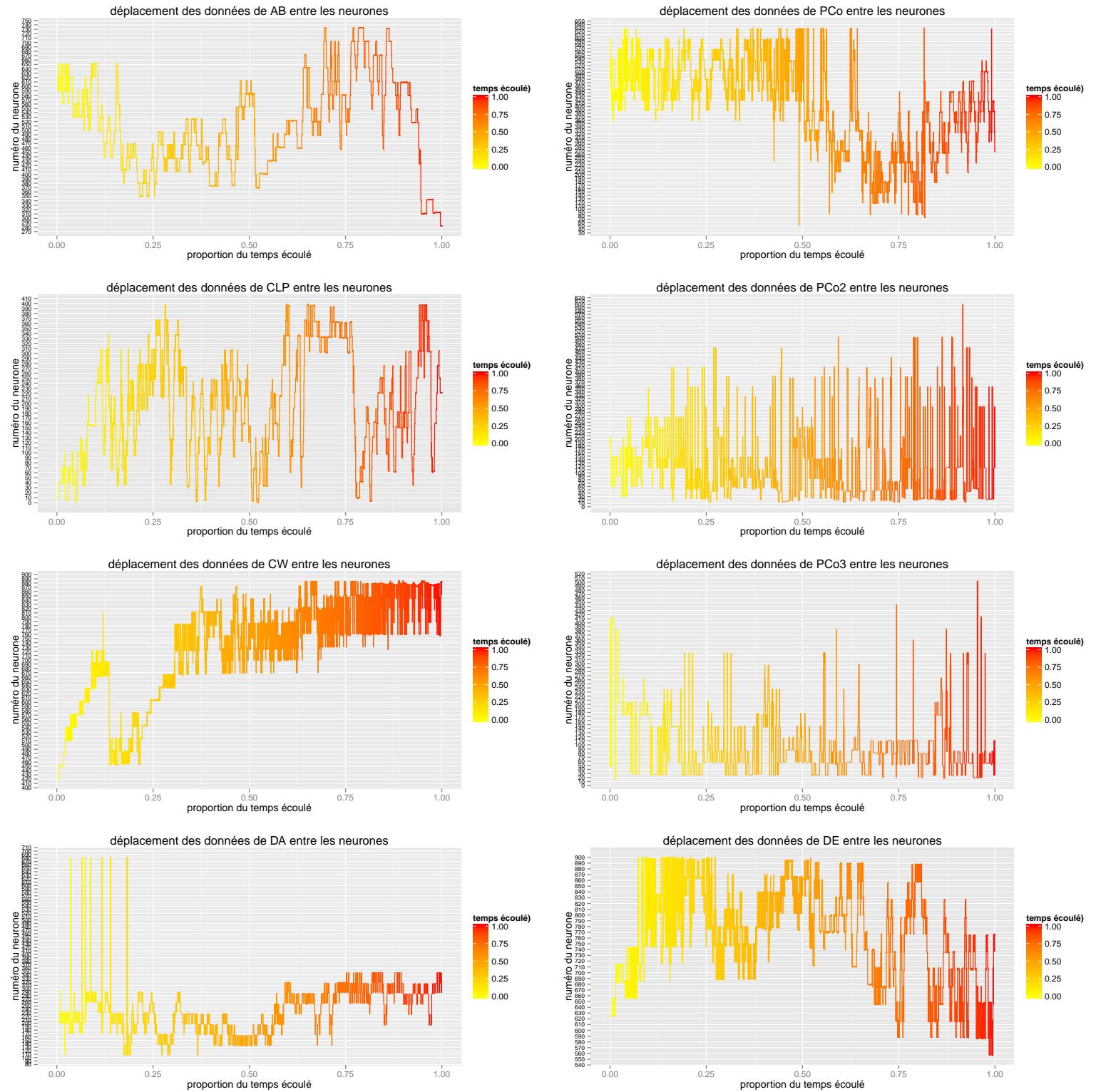


Figure 15: déplacement des données entre les neurones par expérience

Grâce à cette représentation, nous pouvons suivre l'évolution des référents dans le temps. Nous voyions que chacune des expériences produit des évolutions assez différentes. -Pour certains, en particulier "PCo", "DE" ou "CLP", il existe de longues périodes où les données sont confinées à une région de la carte donnée entrecoupé de brèves périodes où semblent se produire des "sauts qualitatifs" : les données se déplacent soudainement dans une autre zone de la carte. -Pour d'autres, comme les expériences "AB" ou "DA", l'évolution est plus lente même si on peut observer encore des périodes où les données sont bien localisées à un endroit de la carte. -Pour finir, certaines expériences ("PCo2" en particulier) montre une grande variabilité et ne semble se fixer à aucun moment dans une période donnée. Comme nous avons vu qu'il existait des régions assez distinctes dans la carte, chacune correspondant à une certaine combinaison de mesures, il va être intéressant de recouper les données en fonctions de ces régions. A l'oeil, il est cependant difficile de bien distinguer ces régions - surtout si on considère une unique expérience (on peut parler de "sous-régions"). Il est donc nécessaire de trouver des moyens de redécouper la carte afin de distinguer les différentes régions. Il nous semble que les dendrogrammes pourraient être intéressants pour cela.

3.4 Dendrogrammes des neurones.

Nous pouvons d'abord procéder à une classification de l'ensemble des neurones. Avec les dendrogrammes, il est possible de réaliser des rapprochements de certains vecteurs multimensionnelles. Un dendrogramme permet de clusteriser les données en fonction de la "proximité" des données. Différentes distances sont utilisées pour calculer la distance en dimension (en l'occurrence 4), nous utiliserons simplement la distance euclidienne généralisée.

Nous allons utiliser les fonctions de la bibliothèque ggplot pour afficher un dendrogramme basé sur un échantillon le plus large possible sous la contrainte qu'il puisse être intégré dans un document pdf ou html.

Dans un premier temps, comme nous travaillons avec l'ensemble des référents, nous allons travailler sur 300 neurones de notre "codebook" tirées au hasard (soit 1/3 de l'ensemble des neurones).

Remarque : cette partie n'est pas terminée, nous sommes entrain d'implémenter de nouvelles fonctions permettant de travailler avec des dendrogrammes sur des données plus larges mais nous ne pouvons pour l'instant l'intégrer à ce travail.

```
## Warning in sample.df$row.names <- rownames(sample.df): Conversion
## automatique de LHS en liste
```

Sur ce graphique, le dendrogramme (inversé), nous trouvons les neurones en ordonnée et la distance entre les neurones ou les groupes de neurones en abscisse. Par exemple, au niveau de la ligne rouge en pointillé, où la distance est de 2, tous les référents qui sont déjà regroupés par des segments verticaux ont une distance inférieure à 2. Nous voyions dans le cas présent que nous avons 3 groupes de référents pour lesquels la distance de chaque élément du groupe pris deux à deux est inférieure à 2.

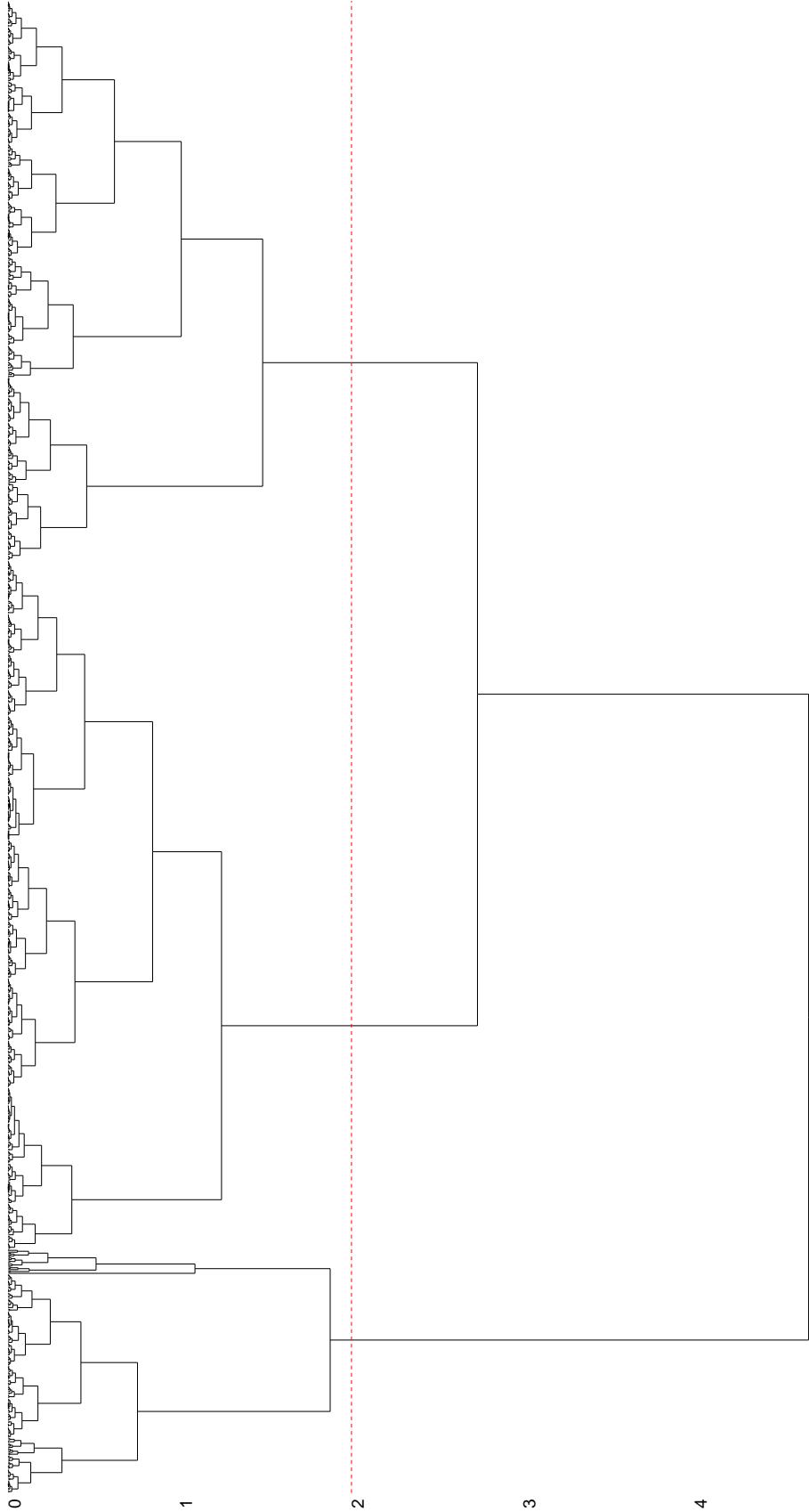


Figure 16: dendrogramme de tous les référents
34

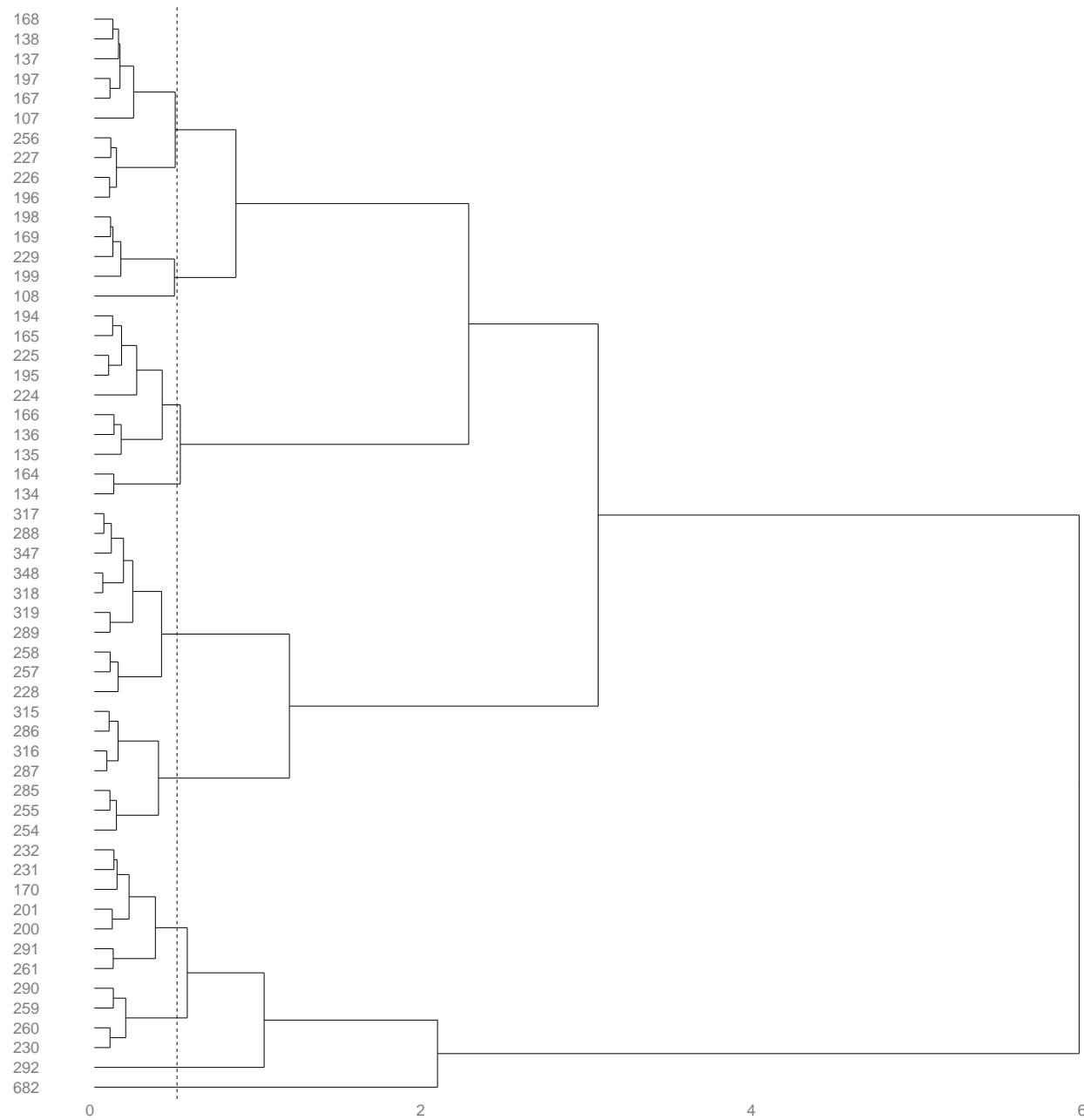
Cette carte est difficile à interpréter car il y a encore trop de référents. On voit tout de même des régions qui se forment. De plus, il me faut ajouter les légendes... (ce travail est en cours...)

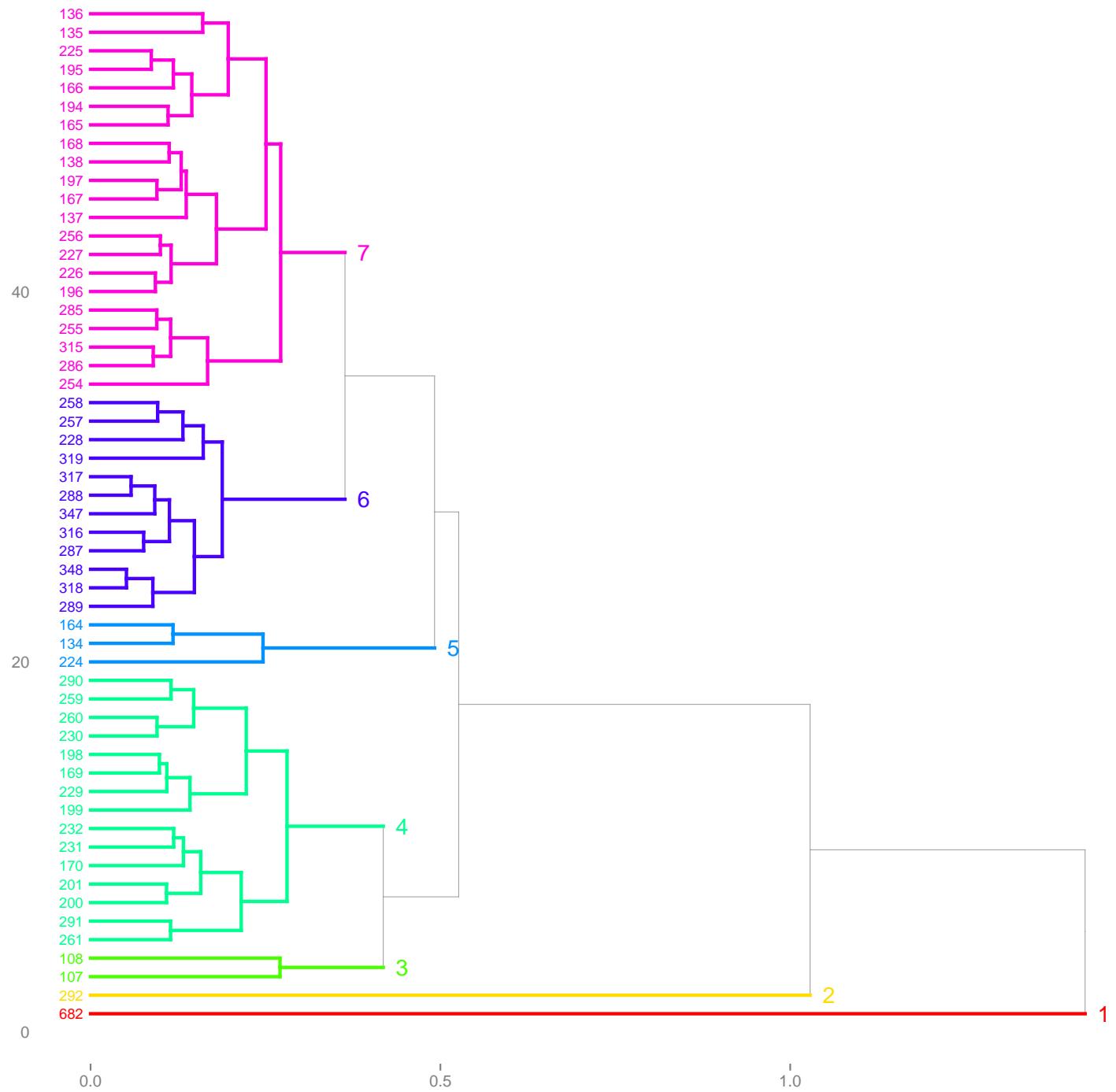
3.4.1 Dendrogrammes par régions

3.4.1.1 Un exemple : l'expérience “DA” NOus allons faire un première étude sur une région en particulier. Nous avons choisi “DA” car l'évolution des données dans la carte des neurones alterne longue période dans certaines régions et courtes périodes de changement brusques.

La région “DA” est constituée de 55 neurones.

Il est ainsi possible de constituer 7 groupes (ou “clusters”) de neurones à partir des 7 segments qui viennent couper la ligne pointillée au niveau de la distance 0.5. Les neurones 47 et 55 constituent des groupes à eux tous seuls.





3.4.2 Dendrogramme des autres expériences

A SUIVRE...

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

3.4.3 Nouvelle projection du temps sur les groupes de référents

3.4.3.1 DA Sur chacun des graphiques qui forment la figure 15, nous pouvons voir comment évoluent les données d'une expérience entre les différentes neurones.

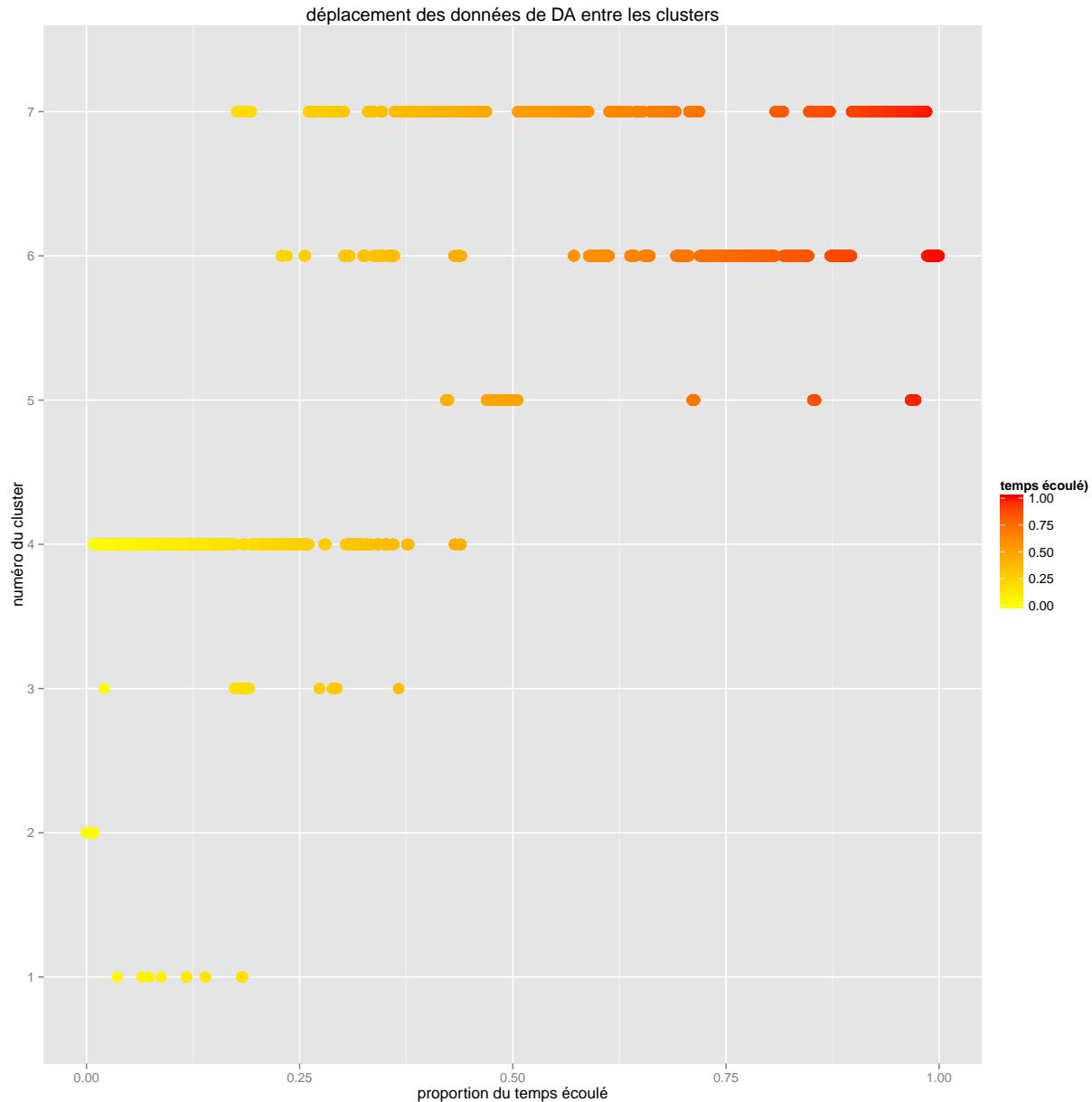


Figure 19: déplacement des données entre les groupes de neurones par expérience

Nous pouvons voir au travers ce graphique plusieurs nouveaux détails. Il serait notamment intéressant de se pencher sur les groupes 1, 2 et 3 qui sont composés d'un ou deux neurones. Par exemple, on peut voir que le cluster n°2 qui ne contient qu'un neurone et ne capte les données correspondant aux premiers instants de l'expérience. Il est donc possible que cela corresponde simplement à une période d'initialisation et que l'on décide de mettre de côté les premières secondes de l'expérience car elles ne sont pas pertinentes pour ce que l'on recherche.

Grâce à cette représentation, nous pouvons mieux suivre l'évolution des référents dans le temps entre des

clusters de neurones similaires. Nous avons la confirmation que les données restent sur dans des zones de référents pendant un certain temps puis se déplacent brusquement. Nous pouvons reproduire maintenant cette méthode sur l'ensemble des données.

4 Conclusion :

Nous avons montré que les cartes de Kohonen permettent de distinguer les individus sans pour autant savoir si elle donne la possibilité de distinguer différentes périodes (et donc différentes émotions exprimées pendant ces périodes) pour chacun de ces individus. Par ailleurs, puisqu'il existe de nombreuses autres expériences réalisées, il faut tester si certains individus n'ont pas des caractéristiques assez similaires pour être captés par les mêmes neurones. Le moyen de tester cela sera d'ajouter de nouvelles expériences afin de voir si les cartes sont capables d'en regrouper ensemble. Il faudra aussi peut-être procéder à des rectifications supplémentaires sur les données ou au moins certaines variables (prise en compte d'offset d'une expérience à l'autre...).

Par ailleurs, il reste un travail d'étiquetage des référents qui doit être réalisé avec l'expert afin de déterminer dans chaque expérience si le passage d'un groupe de neurones à un autre correspond à des changements émotionnelles pertinents. Les derniers graphiques, réalisés grâce aux dendrogrammes et à la clusterisation des neurones sur l'expérience "DA" peuvent former une base pour présenter à cet expert des axes pour qu'il puisse étiqueter les régions de la carte et finalement les différentes périodes d'une expérience. Il faut maintenant généraliser la méthode à l'ensemble des expériences et donner des moyens de visualiser le codebook des "super-référents" que sont les clusters. Ces clusters pourront aussi nous aider à poursuivre notre travail de nettoyage en nous permettant de mieux distinguer les données aberrantes ou dignes d'intérêt.