# Fouille de données avec R

*B. ROUDIER*

## Table des matières





Courivaud Raphaël

# Exercice 1

Nous avons dans ce cas, des données concernant différentes traitements du cancer du poumon. Dans ces données nous connaissons le type de cellules atteintes par le cancer mais aussi s'il s'agit d'un traitement de référence ou d'un traitement de test. De plus nous avons un nombre de jours de survie après que le patient ait commencé le traitement et un score de Karnofsky qui représente la gravité de la maladie.

### Effectuez une ACP, interprétez les axes, et les variables dans le(s) plan(s) d'analyse.

On veut réaliser une Analyse en composantes principales. Elle permet de créer des axes permettant la visualisation de plusieurs variables rapportées sur deux axes. Cependant ces variables doivent être absolument quantitatives. Les axes peuvent être créés dans tous les cas mais ils possèdent un intérêt s'ils représentent et gardent une certaine qualité sur les données initiales. Pour cela il faut que les données soient corrélées entre elles, ou qu'elles possèdent un lien.

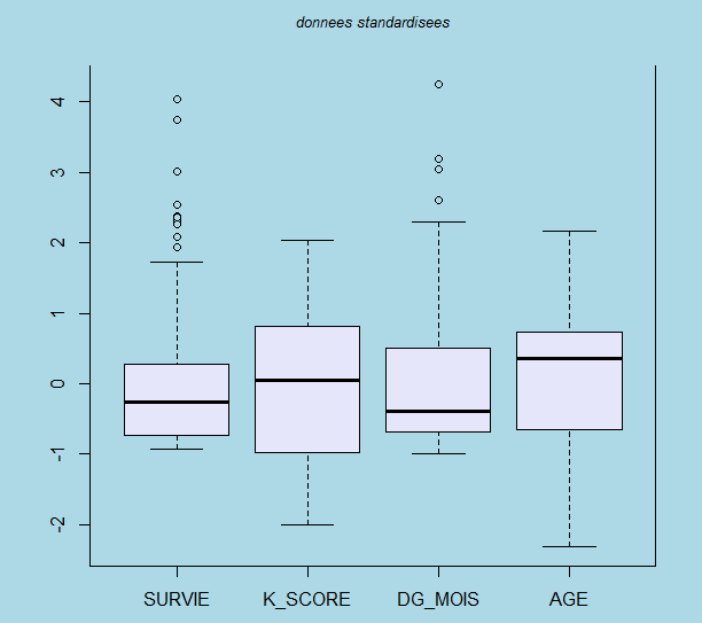



Figure 1 Boxplot exercice1

On peut dans un premier temps observer la matrice de corrélation de toutes les valeurs dont nous disposons.

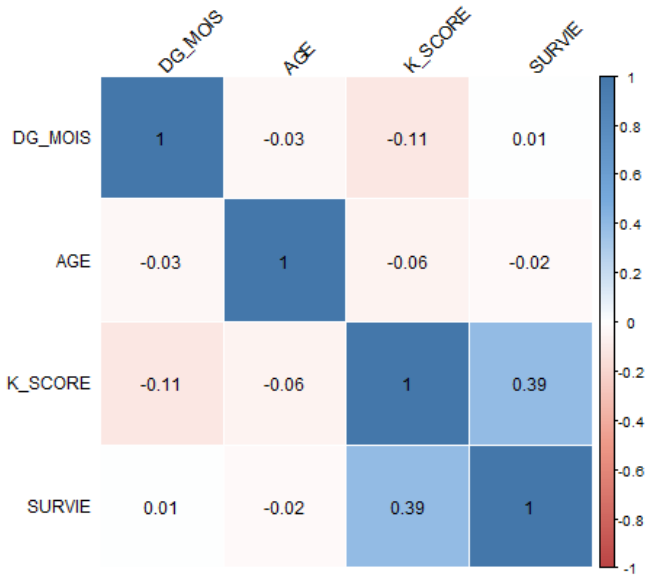



Figure 2 Corrélation exercice 1

On voit tout de suite ici que les variables ne sont pas du tout corrélées. On peut déjà se demander s'il est intéressant de se lancer dans une ACP à ce niveau-là.

On peut voir la qualité des différents axes en représentant l'axe de Kaiser. Il montre l'inertie cumulé pour le nombre d'axe qu'on souhaite afficher.

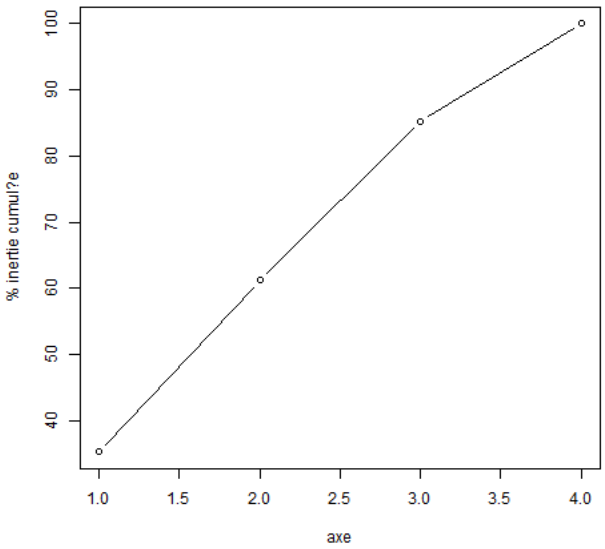



Figure 3 Inertie exercice 1

On voit qu'ici avec deux axes on n'obtient que 60% de qualité de représentation, qui sur quatre axes ne fait pas beaucoup.

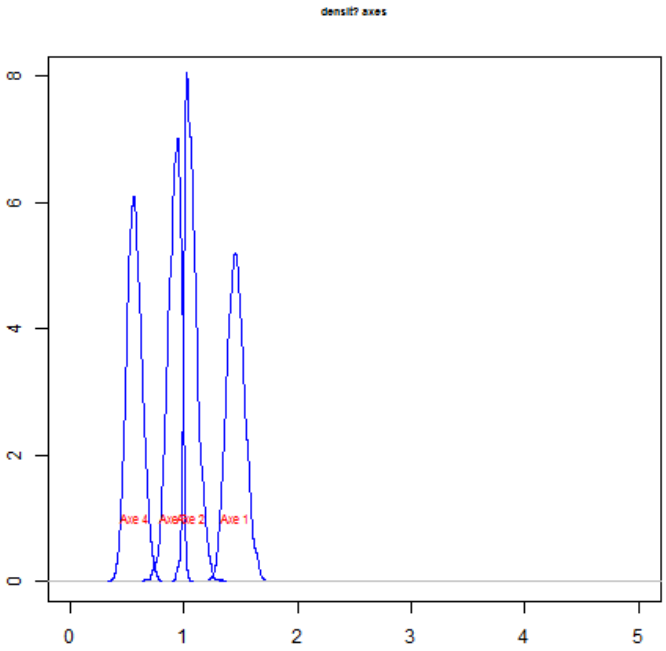



Figure 4 Représentation des axes exercice 1

On voit bien que les axes sont très regroupés, donc qu'ils ne permettent pas de représenter des informations très pertinentes sur nos données brutes.

Pour réaliser une ACP qui permettra la plus grande fiabilité au niveau des données observées on décide de réaliser un Bootstrap. Le Bootstrap permet de réaliser un échantillon aléatoire de données qui permet d'avoir un grand nombre d'échantillons avec un nombre plutôt limité de données. On en réalise un très grand nombre (Ici 2000). On peut alors afficher la distribution du tirage aléatoire.

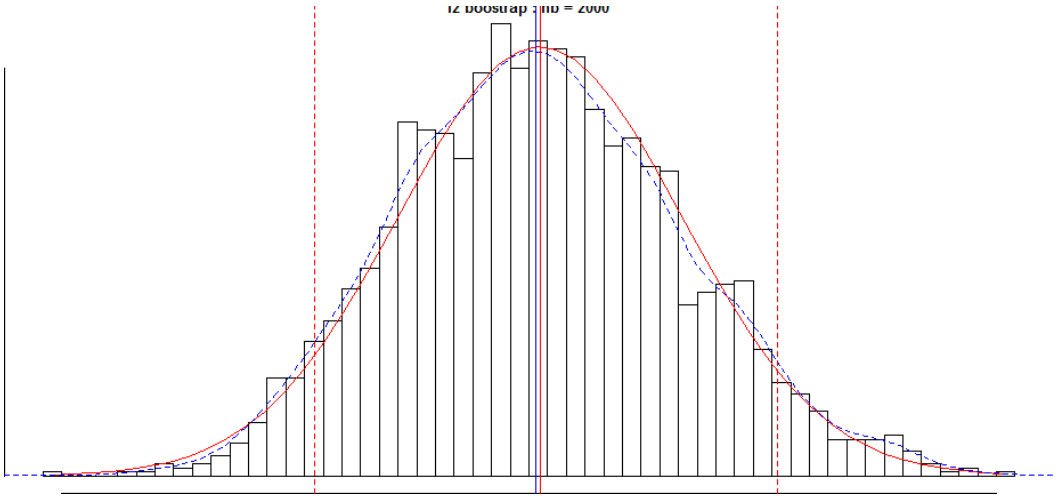


Figure 5 Distribution bootstrap exercice 1

On réalise alors une ACP sur tous les échantillons réalisé et on en fait la moyenne.

On peut maintenant afficher les différents axes que nous avons calculés. Nous ne gardons que les deux premiers pour avoir un affichage lisible.

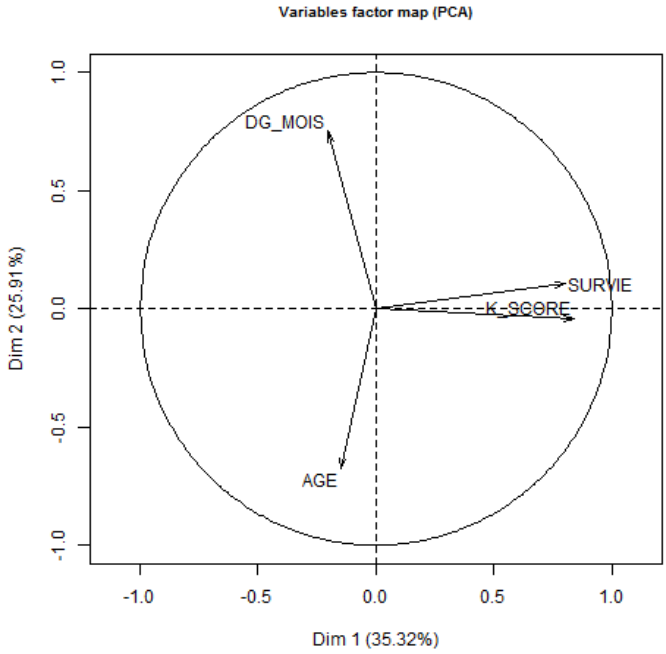



Figure 6 Représentation des variables exercice 1

On avait vu que le taux de survie ainsi que le K-score avant une corrélation d'environ 0.40 ce qui prouve le rapprochement qui est fait par l'ACP. Les deux autres variables n'avaient aucune corrélation donc elles sont presque perpendiculaires.

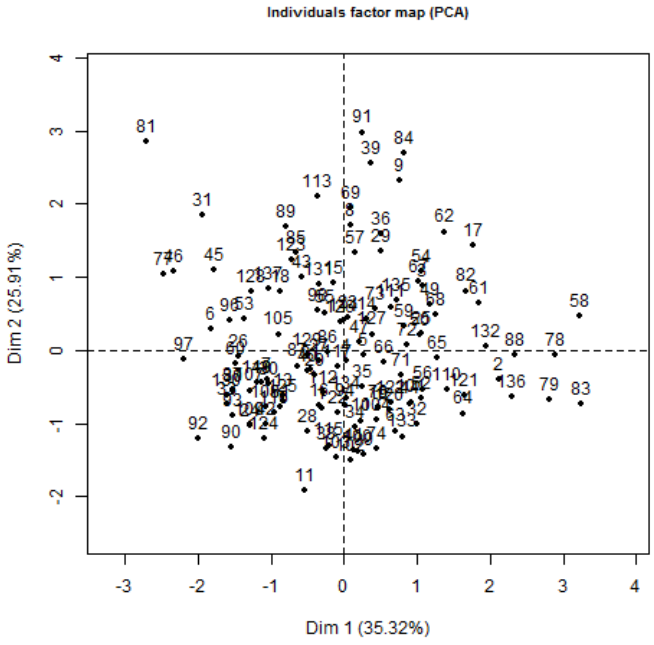



Figure 7 Positionnement des individus sur les deux axes exercice 1

### Positionner les variables catégorielles. Qu'observe-t-on ?

On souhaite positionner les variables catégorielles en fonction des nouveaux axes calculés. Dans un premier temps on affiche les différents types de cellules atteintes A, B, C, D. On voit qu'ils sont vraiment très regroupés et que les individus ne sont pas facilement affectables à chaque catégorie.

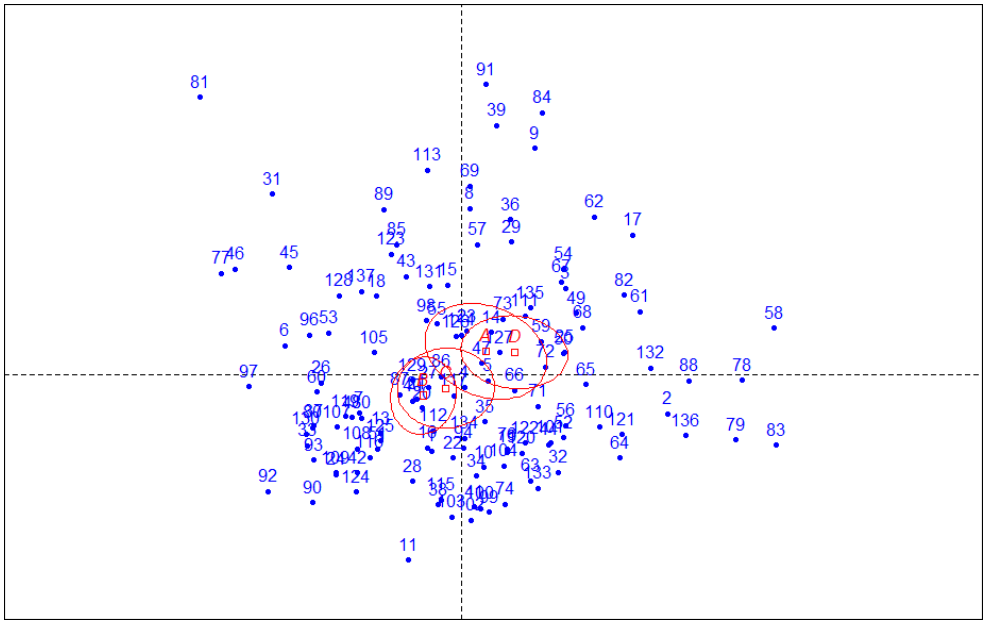



Figure 8 Positionnement des types de cellules affectées

De même pour la distinction entre un traitement de test et un traitement de référence. Ils sont encore plus durs à distinguer l'un de l'autre.

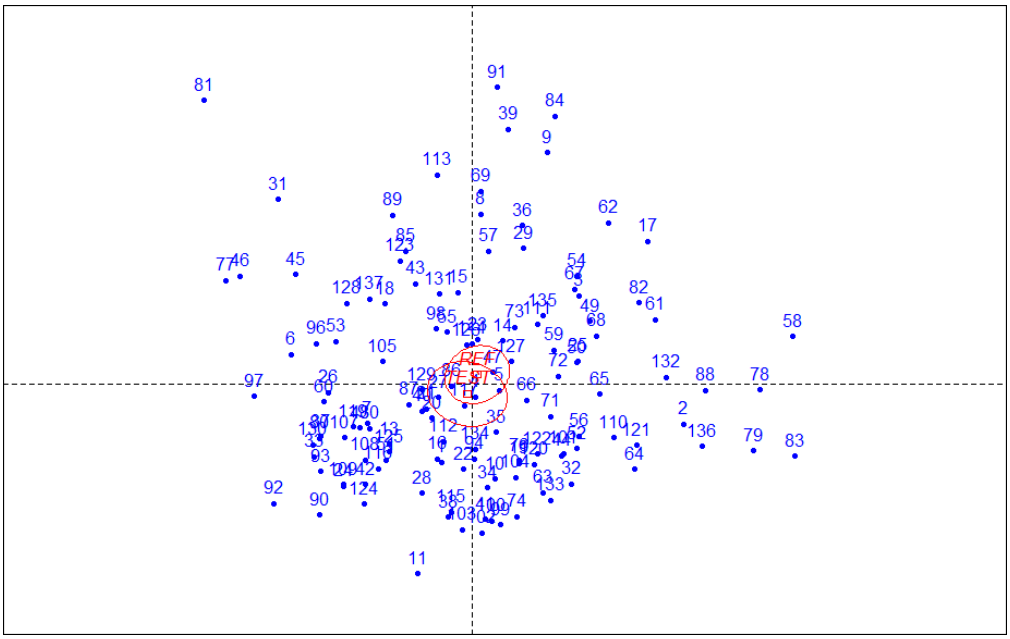



Figure 9 Positionnement des deux traitements

Pour cela nous souhaitons calculer les « distances euclidiennes » entre les individus et les centres des traitements. Pour cela nous devons dans un premier temps calculer la moyenne afin de retrouver

les coordonnées des centres des deux traitements. Dans un deuxième temps nous calculons alors la distance de chaque point avec les centroïdes et affectons une classe à chaque individu avec la plus proche.

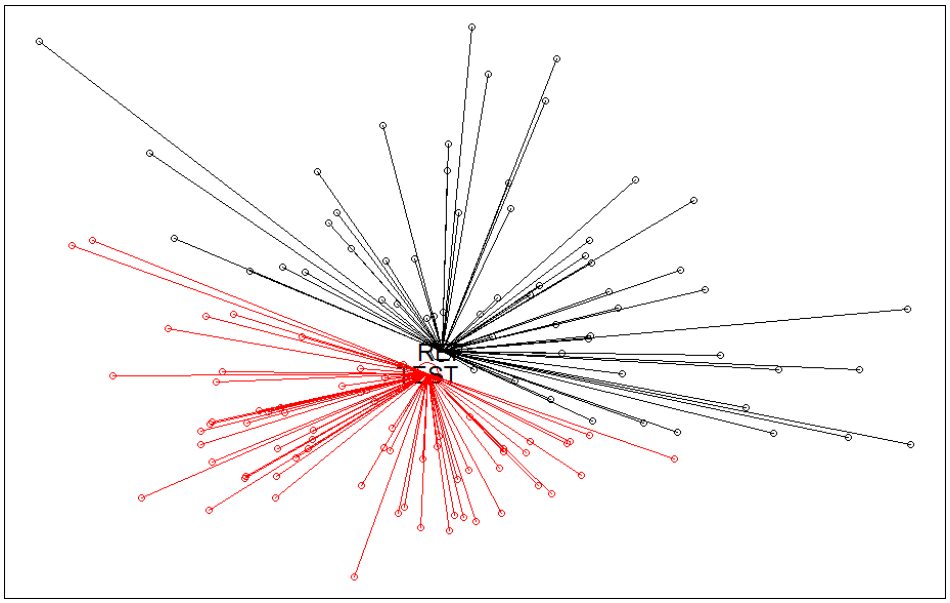


Figure 10 Calcule des distances exercice 1

Nous pouvons maintenant comparer les points réaffectés avec les données initiales et repérer les points bien affecté ou non.

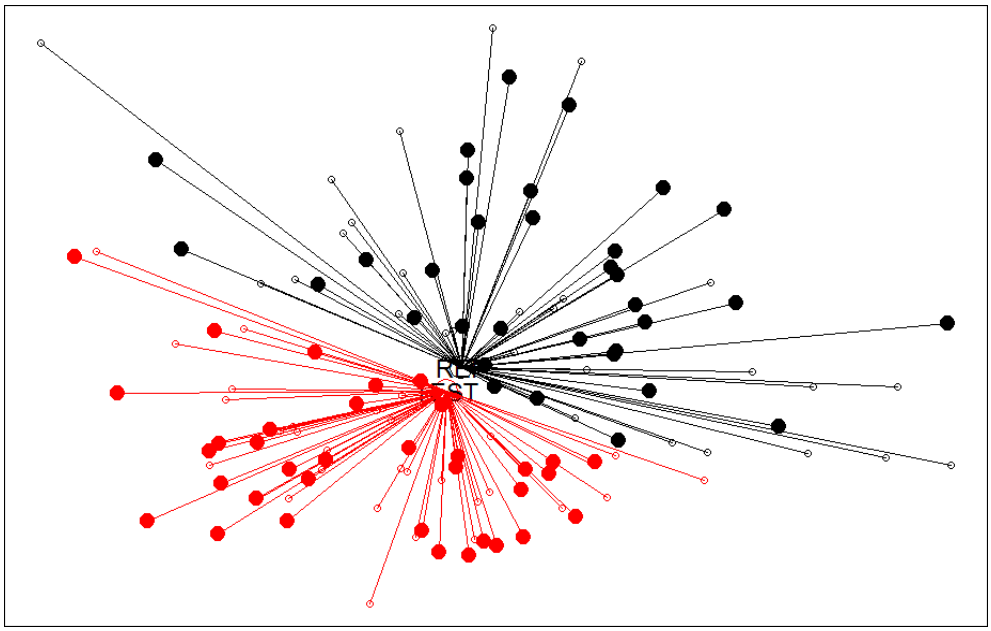


Figure 11 Individus correctements affectés exercice 1

On voit en gras les points bien placés et en petit les points qui ont été mal placé par les calculs de l'ACP.

### Concluez sur l'opportunité d'utilisation de l'ACP dans cette étude. Quelles sont les variables qui sont les plus discriminantes ?

On voit très bien que beaucoup de point ont été mal placés. On peut très facilement les compter en comparant les deux colonnes. On trouve que 63 ont été bien placé et 68 mal placés. Ce qui fait environ 50% d'erreur ce qui est énorme.

L'intérêt de l'ACP sur ce type de données est alors très limité. Cependant les deux variables qui sont le plus discriminantes sont le nombre de jours de survie et le K-score. En effet, ces deux variables sont très représentatives de l'efficacité d'un traitement.

On peut d'ailleurs afficher quelques graphiques permettant d'expliquer ces différentes variables.

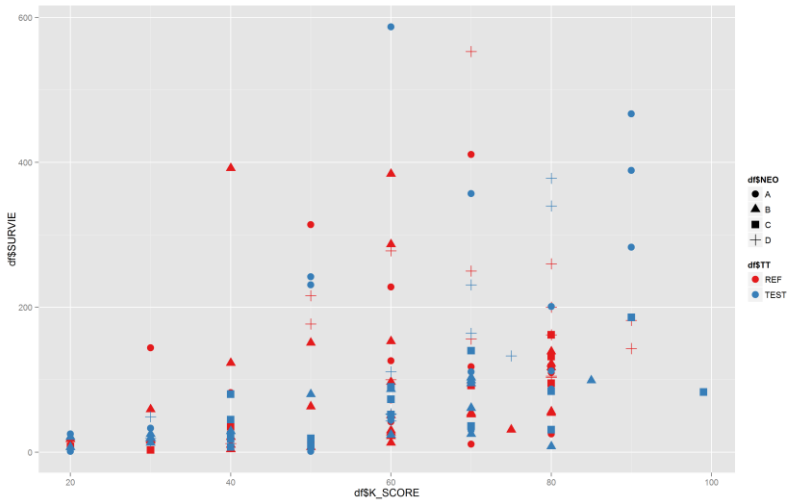



Figure 12 Nombre de jours de survie en fonction du K-Score des patients

On affiche ici Le K score en fonction du nombre de jour de survie pour chaque patient. On applique différentes formes en fonction des différents traitements et deux couleurs différentes en fonction de si il s'agit d'un traitement de référence ou un traitement de test.

On peut aussi compter le nombre de jour de survie pour les différentes cellules atteintes A, B, C, D.

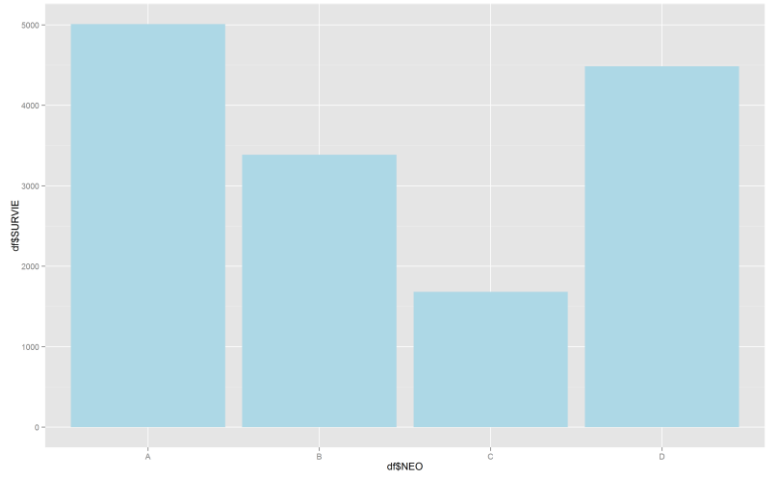



Figure 13 Jours de survie en fonctions des cellules affectés

Cela montre bien que les différentes cellules n'ont pas la même gravité sur les patients. Ici on observe la somme de tous les jours de survie des patients ayant ces différentes cellules atteintes. On voit que le type de cellule B est celui ou les patients ont le plus survécu alors que les cellules de type C les patients n'ont pas vécu longtemps après le début du traitement.

On peut aussi effectuer la même opération sur la différence entre les traitements de références et les traitements de test.

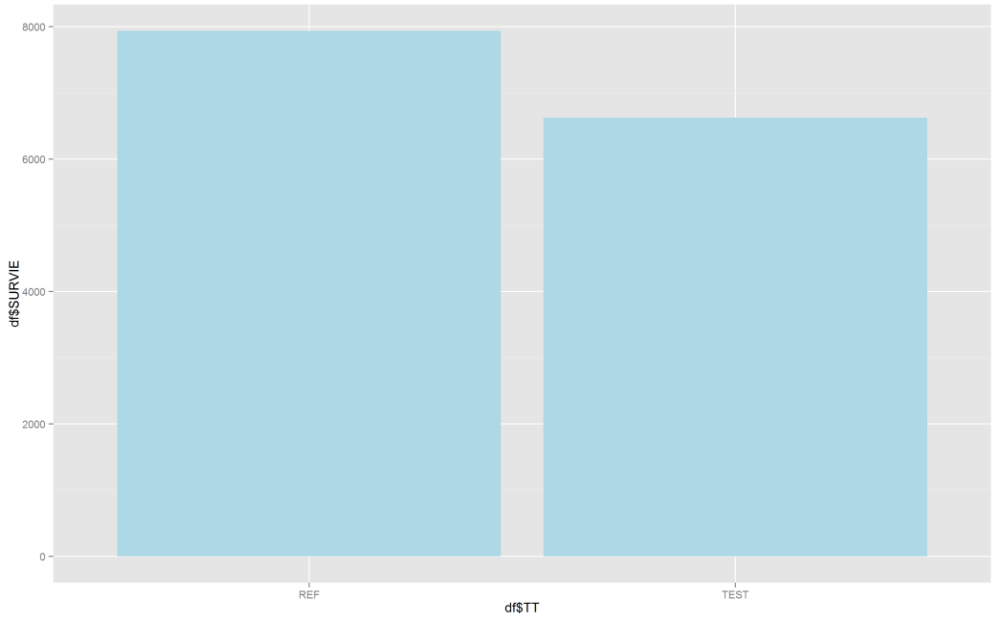



Figure 14 Nombre de jours de survie en fonction du traitement

On voit dans l'ensemble que les traitements de références sont plus efficaces que les traitements de tests.

# Exercice 2

Nous avons à notre disposition des données sur des marqueurs biochimiques de femmes présentant un Kyste ovarien. On veut grâce à ces marqueurs biochimiques déterminer les types de kystes. On souhaite encore réaliser une ACP. Pour pouvoir représenter les différentes variables sur deux axes.

On peut réaliser dans un premier temps un petit aperçu de nos données standardisées.

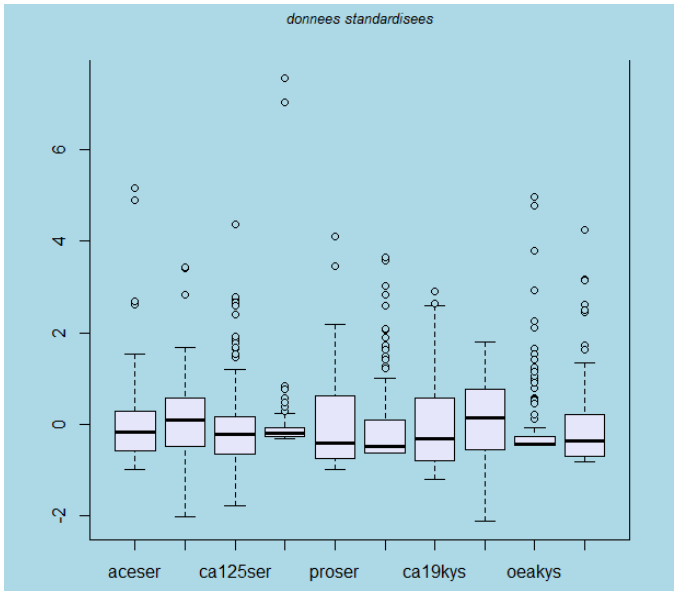



Figure 15 Boxplots exercice 2

On peut maintenant regarder les corrélations entre ces différents marqueurs biochimiques.

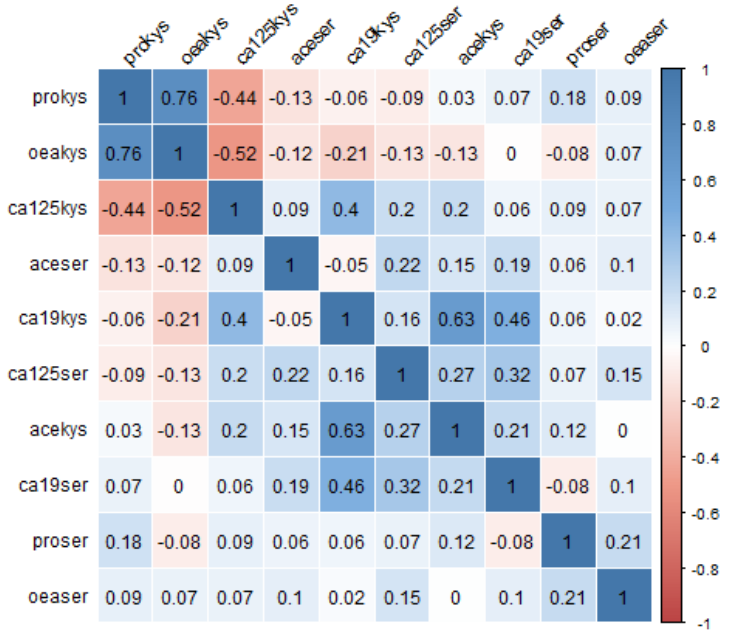



Figure 16 Corrélation exercice 2

On observe une forte corrélation entre le prokys et oeakys mais aussi entre acekys et le ca19kys. Nous allons maintenant réaliser l'ACP pour voir comment peut-on placer ces différents points sur les axes.

On observe dans un premier temps que l'inertie n'est pas très bonne du tout. En effet avec deux axes on observe une qualité de représentation d'environ 45% ce qui n'est pas du tout précis.

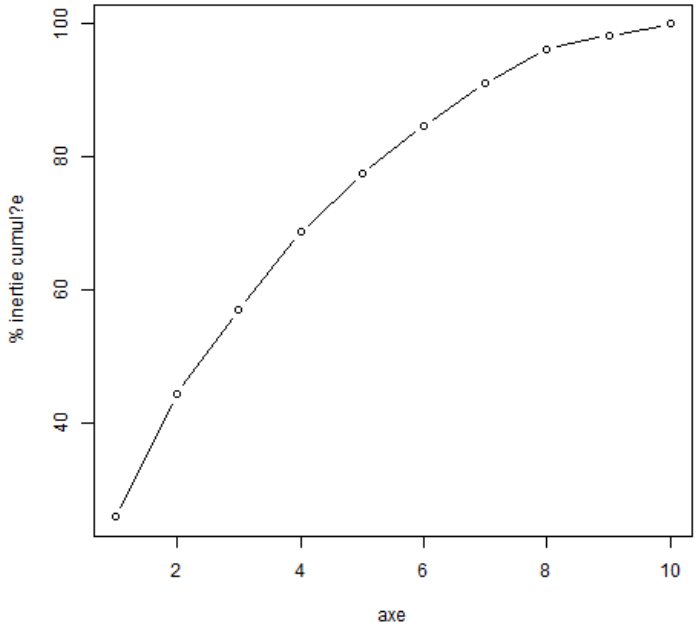



Figure 17 Inertie exercice 2

Les différentes données ne se prêtent pas trop à une ACP dans ce cas. On voit d'ailleurs les différents axes ici sont très rapprochés. Cela ne permet pas d'avoir une représentation très fiable des données.

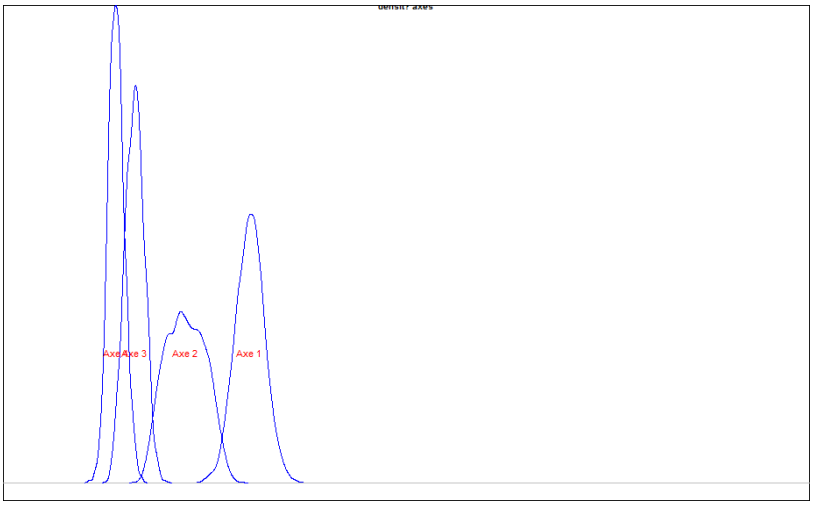



Figure 18 Positionnement des axes exercice 2

On réalise de même que dans le premier exercice un Bootstrap pour obtenir des échantillons avec un tirage aléatoire.

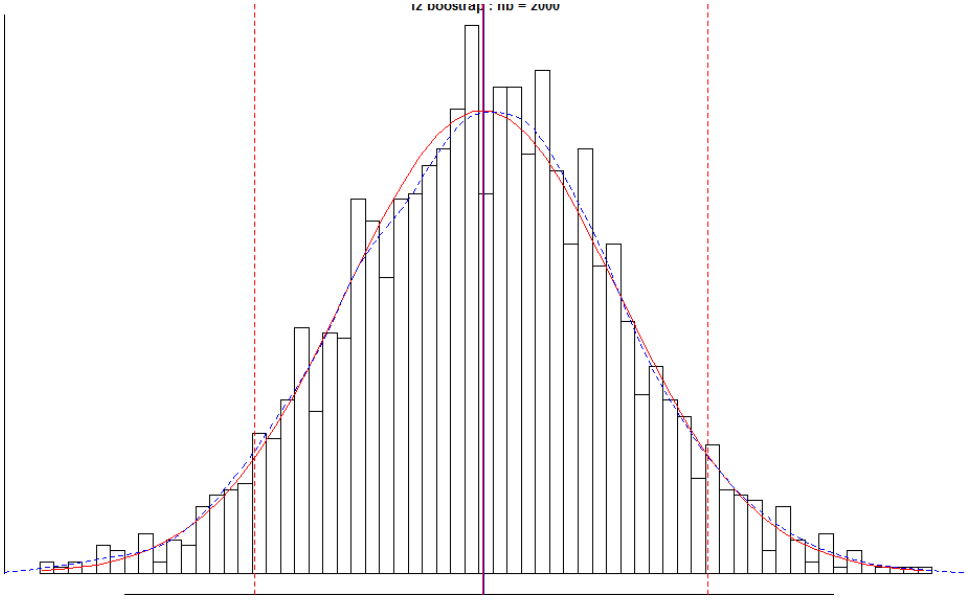


Figure 19 Bootstrap exercice 2

On réalise alors 2000 ACP. Et on prend la moyenne de chaque pour obtenir un résultat le plus fiable possible. On peut afficher la disposition des axes les uns par rapport aux autres.

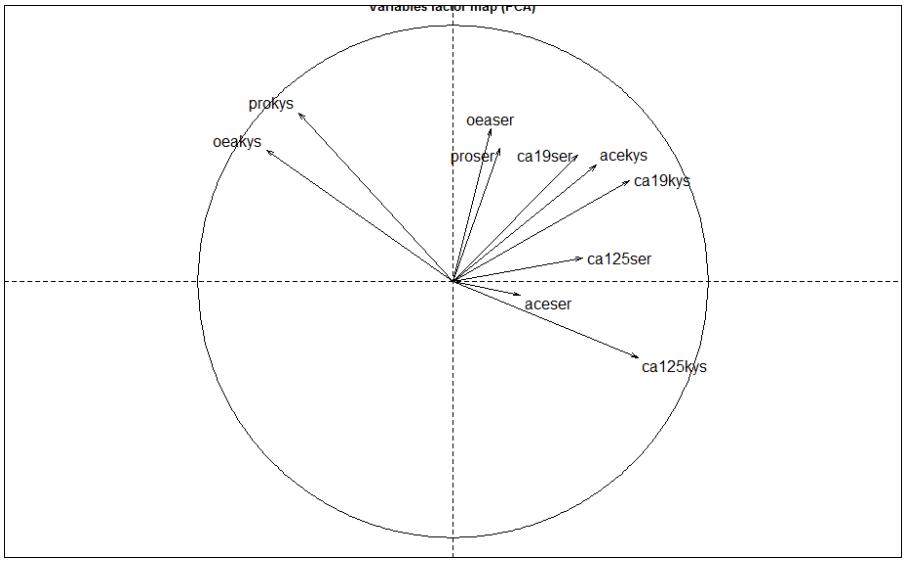



Figure 20 Positionnement des variables exercice 2

On voit, comme escompté, que l'oeakys et prokys sortent du lot. Mais tous les autres marqueurs biochimiques sont tassés et ne se différentient pas les uns des autres. Cela va être très difficile de les représenter avec une qualité suffisante.

On peut aussi vérifier la qualité en essayant de réaffecter avec les nouveaux axes et les nouvelles données calculées avec l'ACP les différents types de kyste en fonction des deux axes fournis par l'ACP.

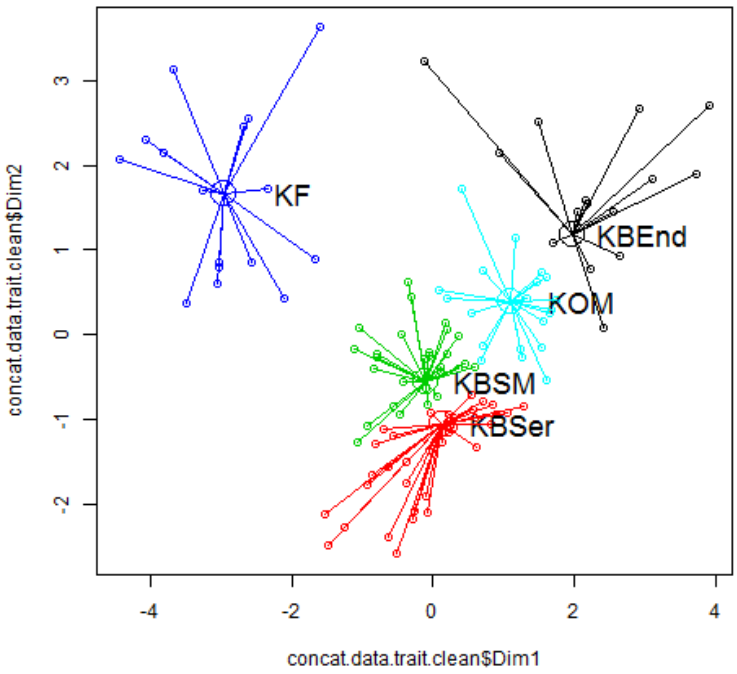



Figure 21 Calcul des distances exercice 2

On peut essayer de retrouver les données bien positionnées et celles qui ont été mal positionnées à cause de l'incertitude amené par l'ACP.

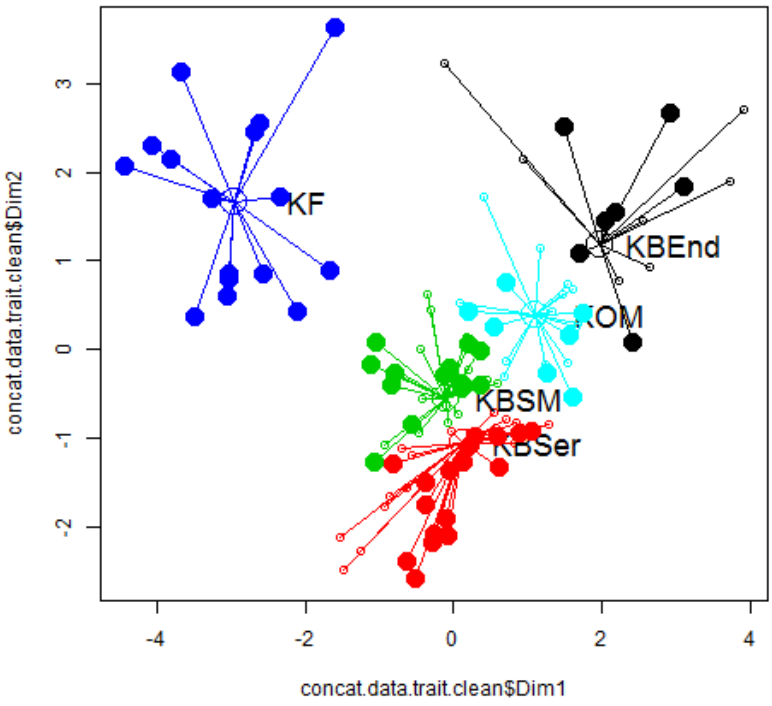



Figure 22 Individus correctement positionnés exercice 2

On voit en gras les points qui ont été bien positionnés. On voit que pour les kystes malins ont été très bien positionnés. Cependant pour les kystes bénins. On voit de nombreuses erreurs. On voit que 54 kystes ont été mal positionnés et 61 ont bien été positionnés dans les bonnes classes.

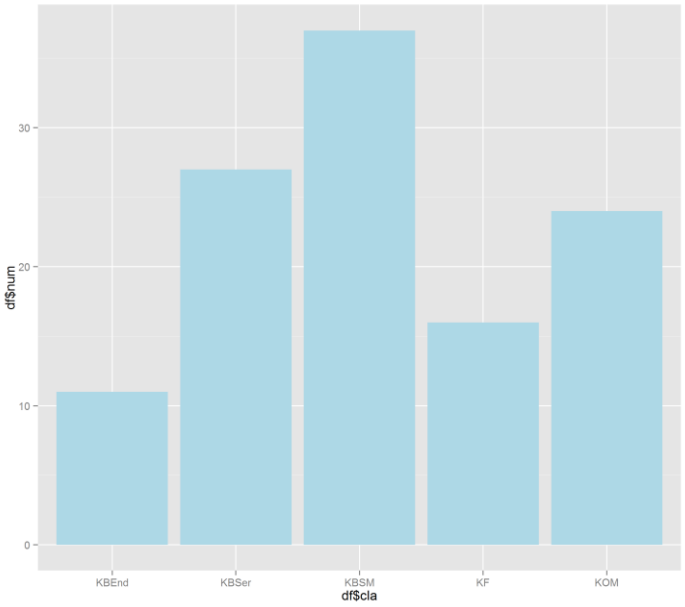



Figure 23 Nombre de formes de kystes représentés

On voit ici le nombre d'échantillons selon chaque classe de kyste. On voit que les kystes malins sont en minorité avec les kystes bénins endométrioides. La classe la plus représentée est la classe des kystes bénins séro-mucineux.

## Conclusion :

Malgré la perte très importante d'information due à l'ACP. On peut supposer que le but de cette analyse est de détecter plus facilement les kystes malins qui sont les plus dangereux. Or cette ACP permet très bien de déterminer si un kyste est malin ou bénin. On a plus de mal cependant pour différentier toutes les catégories de kystes bénins. Mais cette différentiation est peut-être un peu moins importante. Ici l'ACP à quand même un grand intérêt et possède une qualité de représentation assez élevée pour notre analyse.

# Exercice 3

Nous avons à notre disposition des données démographiques concernant les Etats Unis en 2000 et 2001. Dans ce cas-là, L'ACP n'est pas une bonne solution puisque ce sont des données de dénombrement. Nous pouvons réaliser une Analyse Factorielle de Correspondances, mais nous devons avoir des données de dénombrement soient uniquement positives. Or l'immigration locale est négative dans certains états. Pour régler ce problème nous aurions pu créer deux colonnes avec les départs et les arrivées pour chaque état mais il nous manque des données pour déduire les informations manquantes.

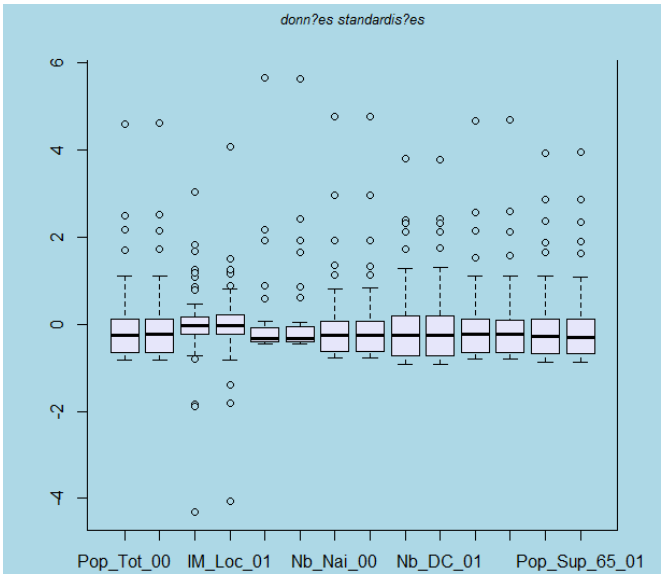



Figure 24 Boxplot exercice 3

Nous pouvons alors transformer ces données en données catégorielles. Nous utilisons les histogrammes de chaque variable pour définir des catégories à peu près équitables. Avec ces variables, nous allons réaliser une analyse des correspondances multiples.

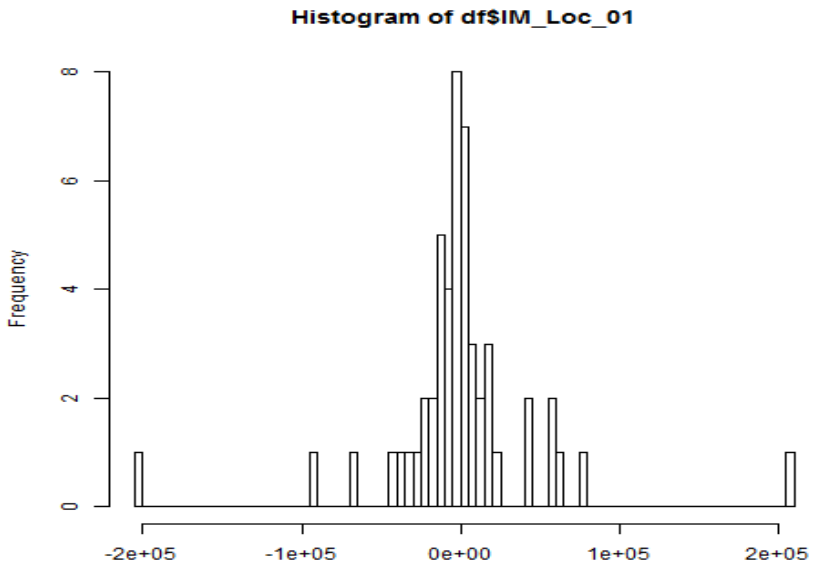



Figure 25 Histogram de l'immigration locale

## Etat démographique au sein des différents états de l'Union en 2001

On observe la corrélation entre les différentes variables mises à notre disposition. On voit que ces variables sont toutes très liées entre elles. Sauf l'immigration locale. Comme ce sont des données de dénombrement il est normal qu'un état possédant une très grande population possède aussi un nombre proportionnel de naissance et de décès par exemple.

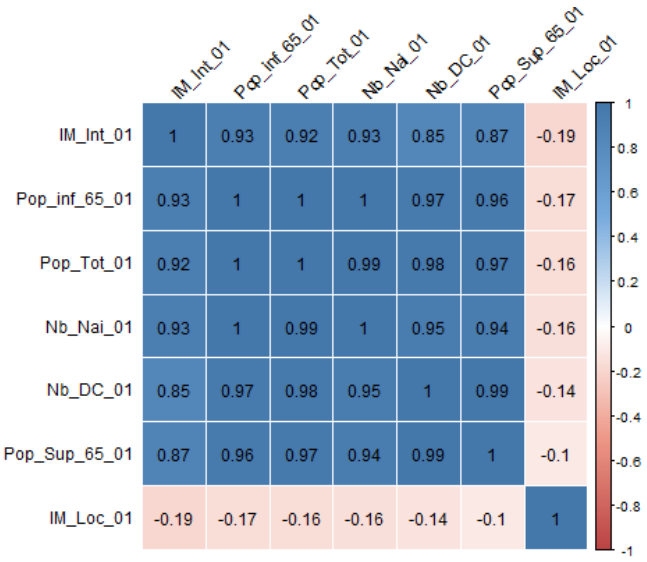
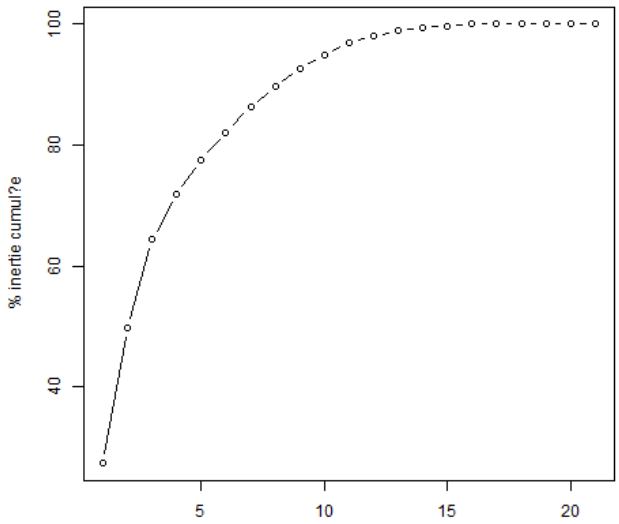
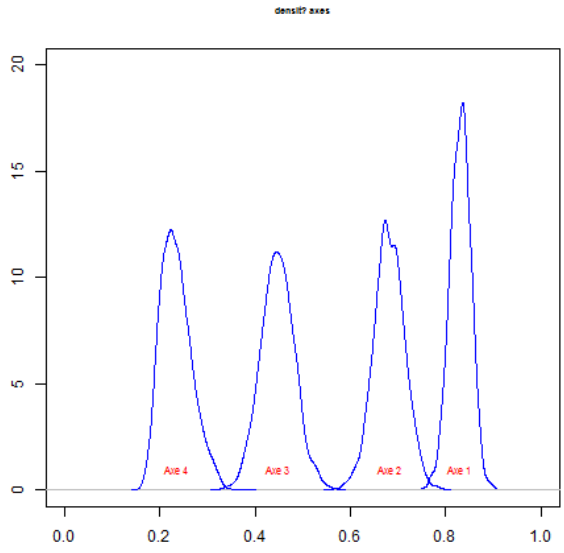



Figure 26 Corrélation pour les variables de 2001 exercice 3

On peut afficher la qualité de représentation de l'AFCM avec l'axe des inerties cumulées ainsi que les différents axes calculés.




On voit ici par contre que la qualité de représentation des données sur deux axes n'est encore pas très fiable. En effet, on récupère qu'environ 50% de l'information avec deux axes.

On réalise encore un Bootstrap pour les même raisons que pour les deux premiers exercices.

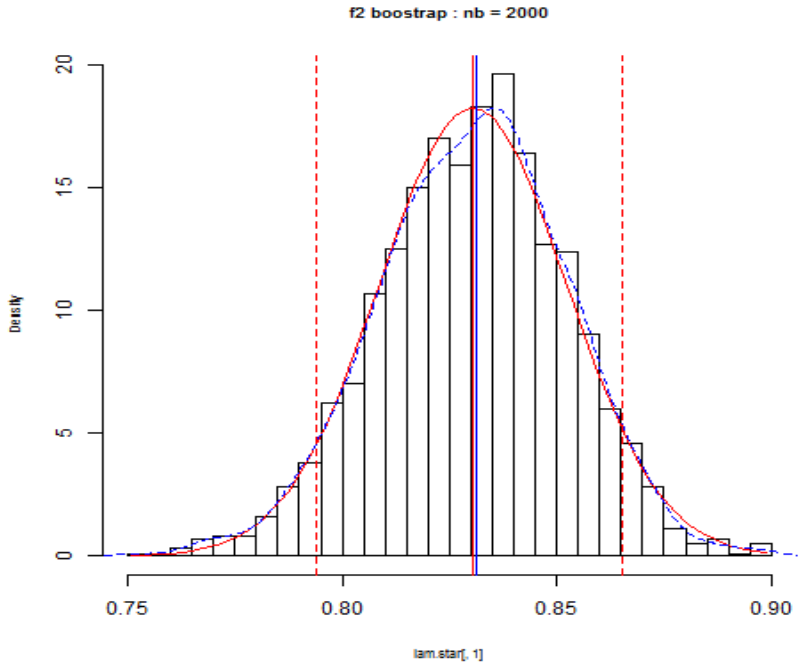



Figure 27 Bootstrap exercice 3

On peut maintenant afficher les résultats de l'AFCM. On représente les ellipses des différentes catégories que l'on avait créées plus tôt. Tout d'abord celles représentant l'immigration Internationale et Locale ainsi que la population.

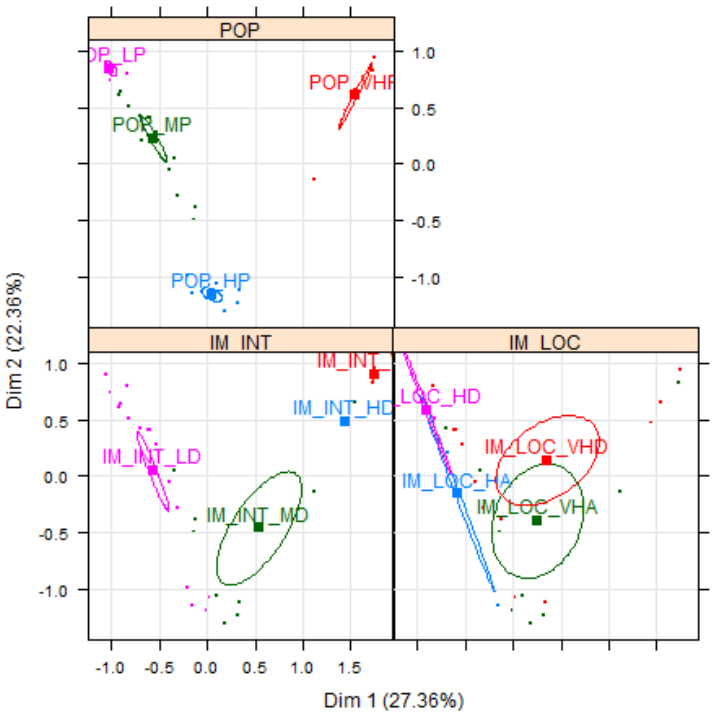



Figure 28 Ellipse Population Immigration locale et internationale

Pour la population j'ai choisi les labels : LP (Low Population), MP (Medium Population), HP (High Population) et Very High Population. De même pour les catégories des variables représentant la population inférieure et supérieure à 65 ans. Pour l'immigration j'ai dû régler le problème des valeurs négatives. J'ai donc utilisé les labels VHD (Very High Departure), jusqu'à VHA (Very High Arrival).

On voit ici que les états sont très différents les uns des autres sur ces différentes variables.

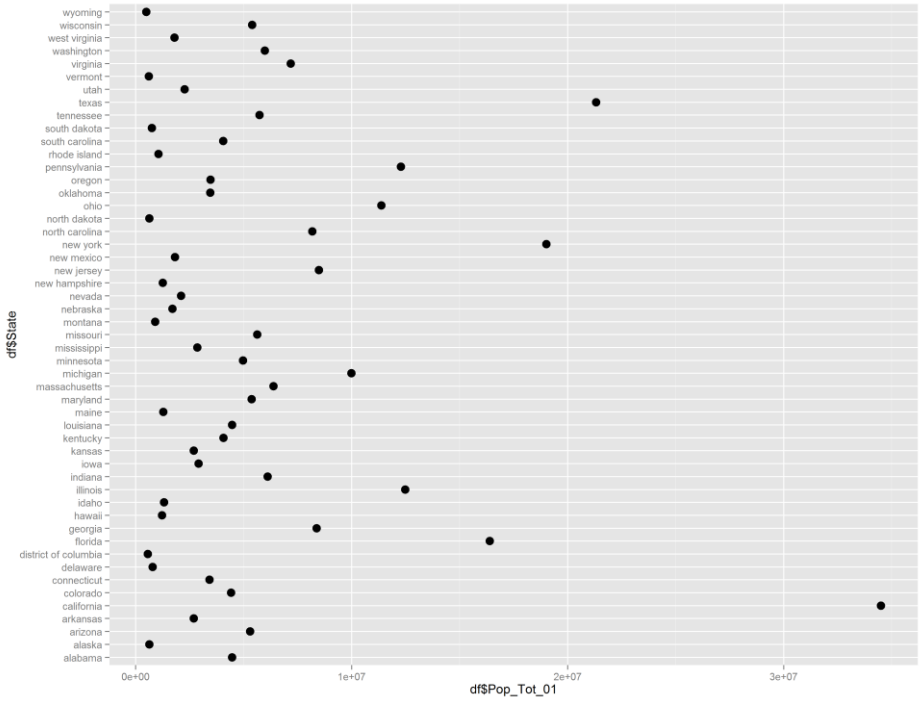



Figure 29 Population 2001 pour chaque état

On voit ici que la Californie est l'état le plus peuplé des états unis et que les états sont de population très variable. On peut aussi afficher les différentes catégorie créées pour les naissances et les décès.

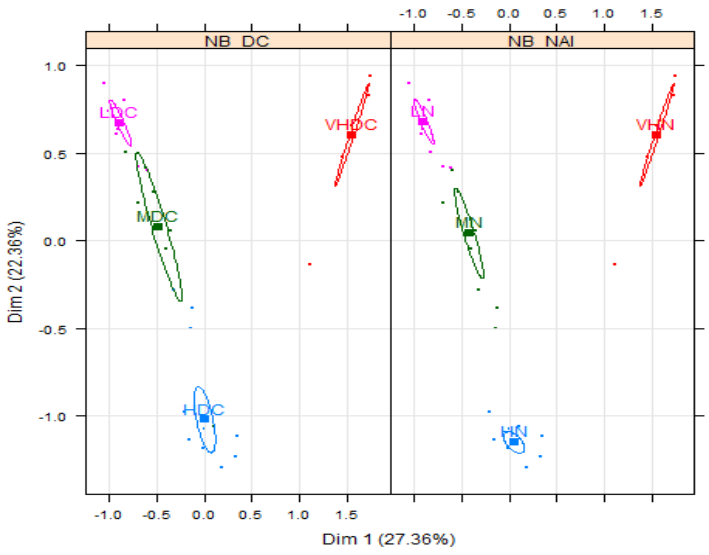



Figure 30 Ellipse Nombre de naissance et Décès

Et enfin, les populations inférieures et supérieures à 65 ans.

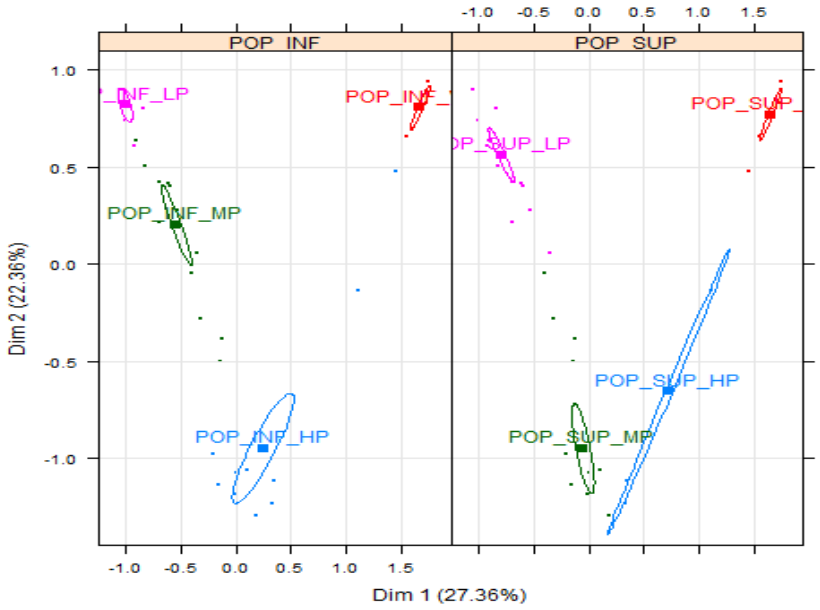



Figure 31 Ellipse Population inférieur et superieur à 65 ans

Cependant on voit ici que des allures et des catégories, mais non n'avons pas de détail des états. Nous ne pouvons donc pas faire d'analyse poussée.

On peut regarder très rapidement les états en fonction de l'immigration locale et internationale.

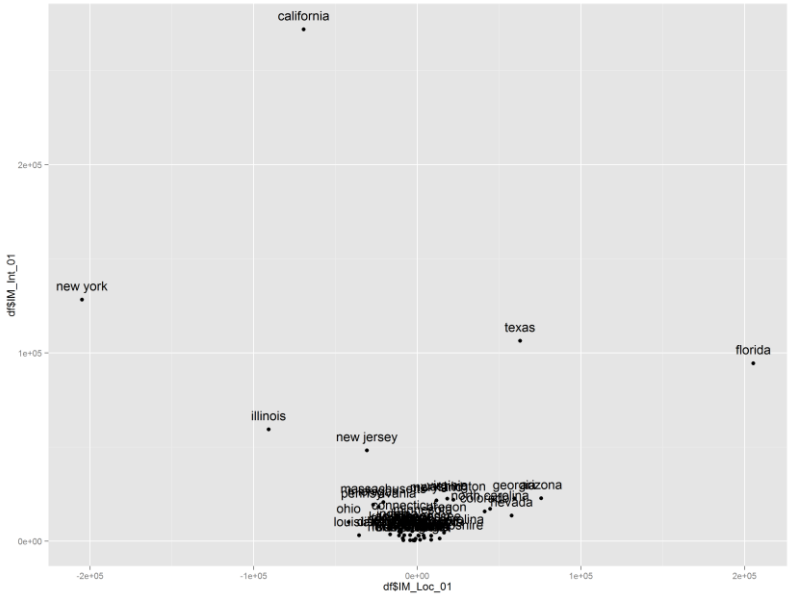



Figure 32 Immigration internationale en fonction de la locale

On voit ici que la plupart des états possèdent une très faible immigration que ce soit internationale ou locale. Cependant quelques Etats sortent du lot. La Californie tout d'abord. Qui possède une très forte immigration internationale. New York, quant à lui possède une immigration internationale assez élevé mais une émigration assez forte et non une immigration contrairement à la Floride. On

sait que la Floride devient un état de plus en plus vieux car de nombreux retraités des différents états mais aussi du monde entier viennent se retirer dans cette région où il fait plutôt bon vivre.

On peut regarder aussi très rapidement, les états les plus « vieux », là où la population de plus de 65 ans est très importante. Cependant cette population est très liée à la population totale des états. Mais on peut en retirer quelques informations quand même.

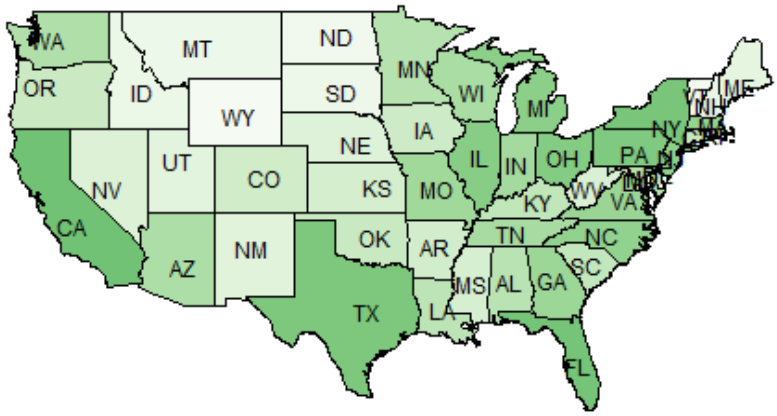



Figure 33 Population supérieur à 65 ans

On voit ici que les états du sud ainsi que les états de l'ouest possèdent une population assez âgée. Comme nous avons dit plutôt pour passer une retraite au soleil.

On peut afficher finalement le résultat de l'Analyse Factorielle des Correspondances Multiples.

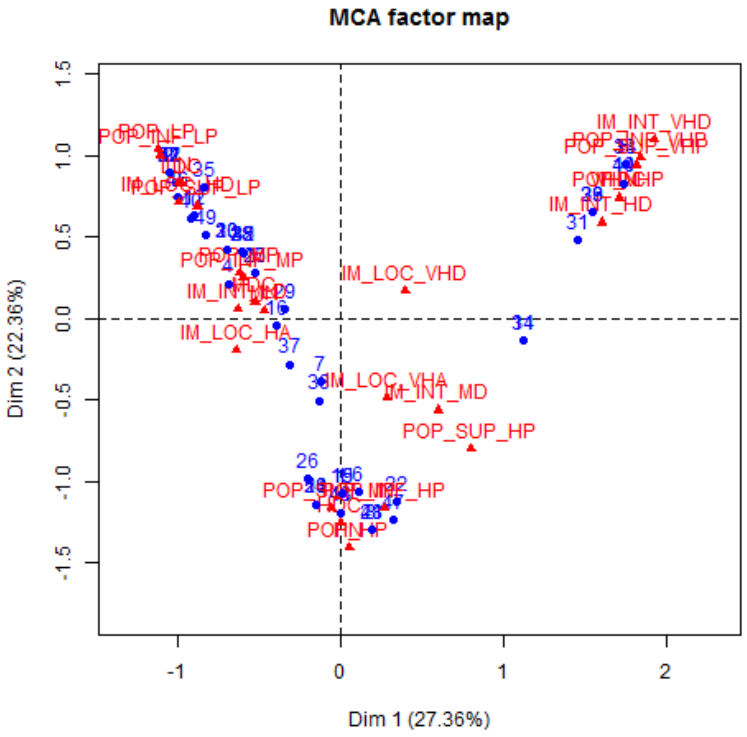



Figure 34 Résultat de l'AFCM

## Quelles sont les variations des données démographiques des états entre 2000 et 2001

On s'intéresse maintenant aux différences pour chaque état entre 2000 et 2001. On pourra voir que malgré la très courte durée, on observe quand même de nombreuses différences démographiques. On peut tout d'abord s'intéresser à la population au sein de chaque état.

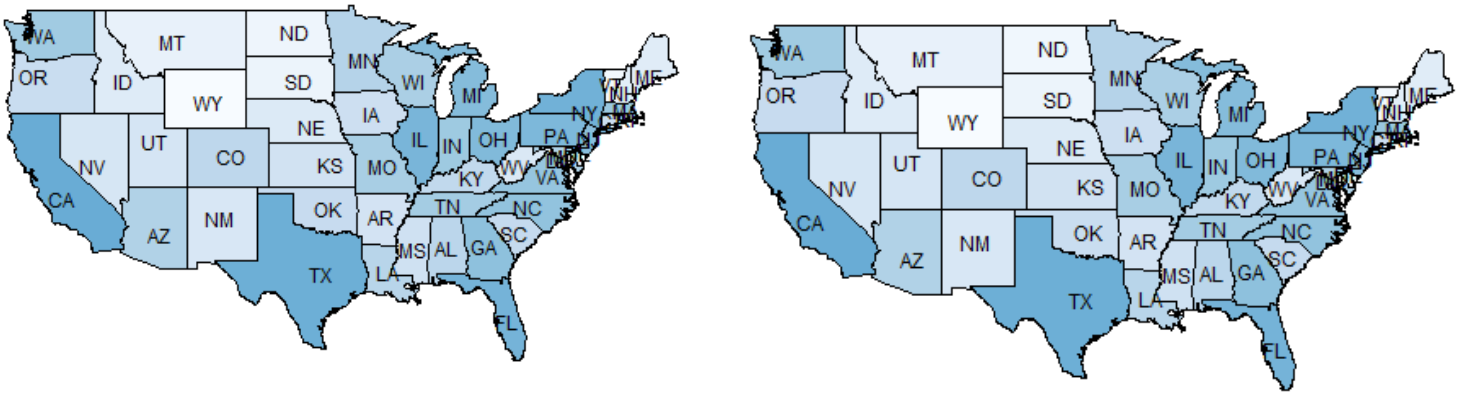



Figure 35 Population en 2000 et 2001

Au niveau de la population il n'y a pas de grands bouleversements dans le classement des états les plus peuplés. Car ici sur ces cartes nous voyons en fonction des couleurs le classement des états les plus peuplé pour chaque année. Les deux couleurs ne sont pas à la même échelle. On voit seulement la différence des états les uns par rapport aux autres. Cependant on peut tout de même constater une différence entre les années. On fait la différence de la population entre les deux années et trace la distribution de cette nouvelle variable.

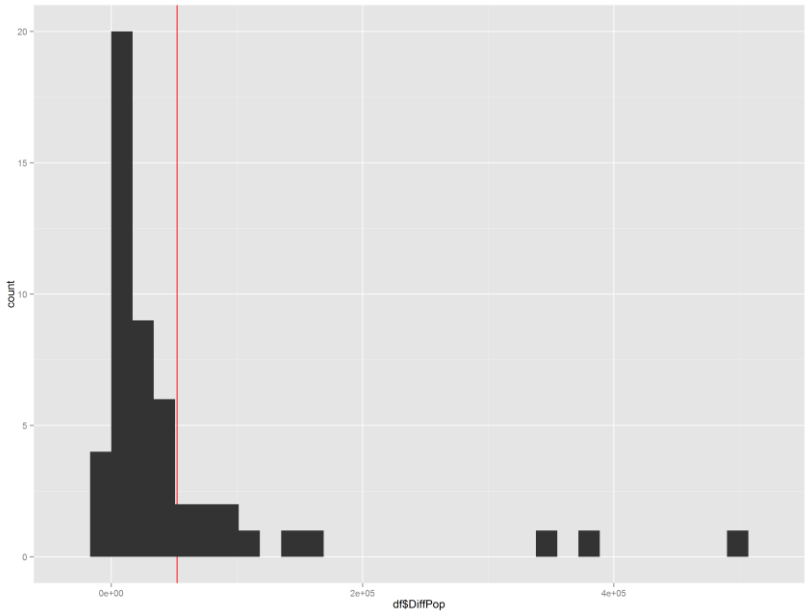
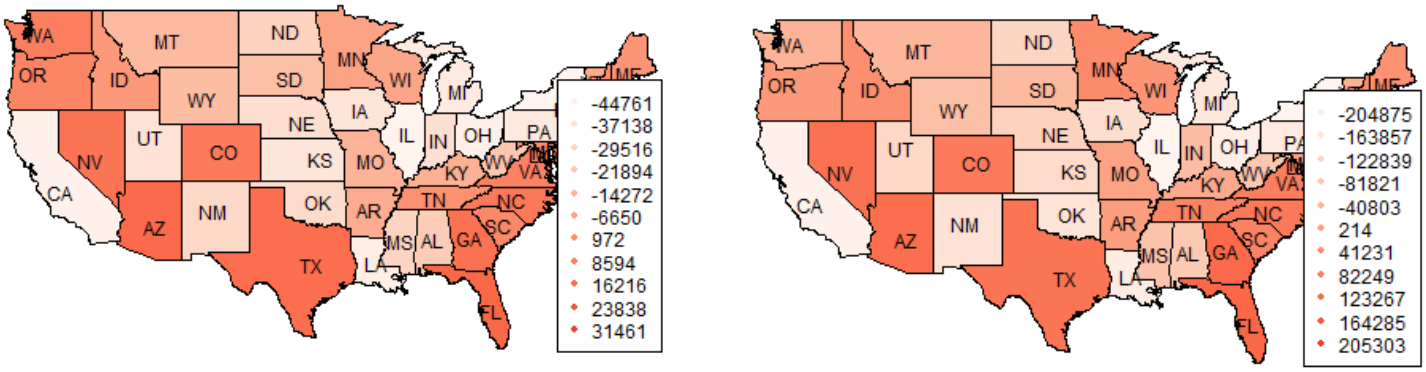



Figure 36 Distribution de la différence de population entre 2001 et 2000

On voit quand même une très légère augmentation générale de la population puisque la moyenne est positive et qu'il n'y a que quatre états ayant une démographie négative entre ces deux années.

On s'intéresse maintenant à l'immigration locale des états. On voit ici que malgré la forte population de la Californie il y très peu voir une légère émigration des américains en Californie. Contrairement au Texas et à tous les états du sud est des états unis comme la Floride.




On voit aussi que de 2000 à 2001 l'immigration ainsi que l'émigration ont fortement augmentés. Notamment dans l'état de l'Oregon et celui de Washington. On voit que l'immigration à baisser en 2001. On voit aussi que dans les états du sud des états unis comme le Nevada, le Colorado, le nouveau Mexique ou l'Arizona par exemple l'immigration est restée plus ou moins constante.

Nous pouvons aussi nous intéresser à l'immigration internationale.

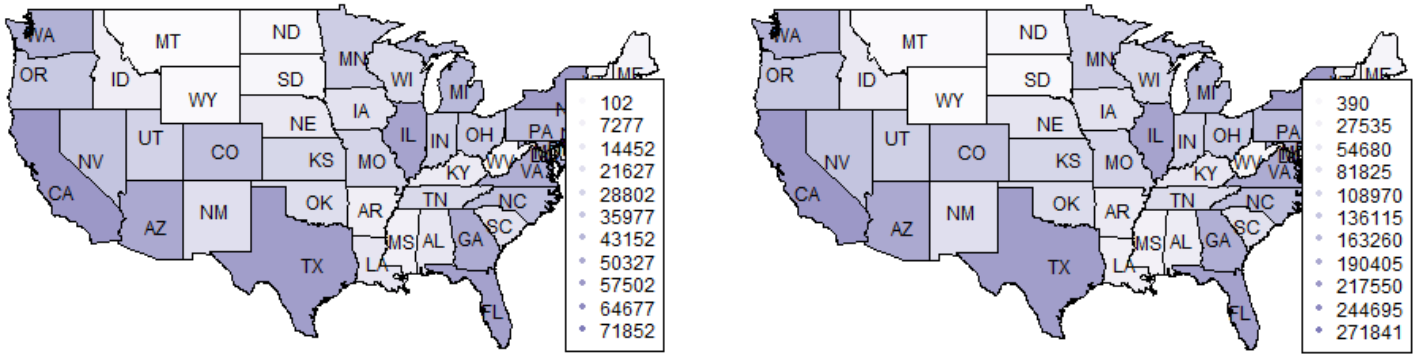



Figure 38 Immigration Locale des états

On voit que les états les plus touchés par cette immigration sont les états des extrémités du territoire américain. On voit que les états du Montana, du Wyoming, du Dakota du Nord et du Sud, ainsi que le Nebraska sont très peu touchés par cette immigration. Contrairement à la Californie, la Floride, le Texas ou même l'état de Washington. La différence entre les deux années n'est pas flagrante. On voit que les états n'ont pas eu de bouleversement de ce côté-là.