Consultations statistiques avec le logiciel 

Que signifie la taille des ellipses dans ade4 ?

Résumé

La question est souvent posée. Elle est revenue dans un message de Ken Rutledge qui avait demandé la figuration des groupes de points dans la représentation triangulaire. On donne la définition des ellipses d'inertie et quelques illustrations.

Plan

1. Une question graphique

Elle est posée par Ken Rutledge, avec beaucoup d'adresse, sur adelist :

```
I'm a new list serve member, using R-ADE4.
It is very helpful. Appreciate the work your group has finished.
This may have been discussed before -- and I need to know how to retrieve
that if so.
Can I create triangular plots with several groups included?  Red squares for
group one, blue circles for group two -- in the same triangular plot?
Thank you.
```

La solution est dans la fonction triangle `triangle.class`

Alias

```
triangle.class
```

Titre

```
Représentation triangulaire et groupes de points
```

Usage

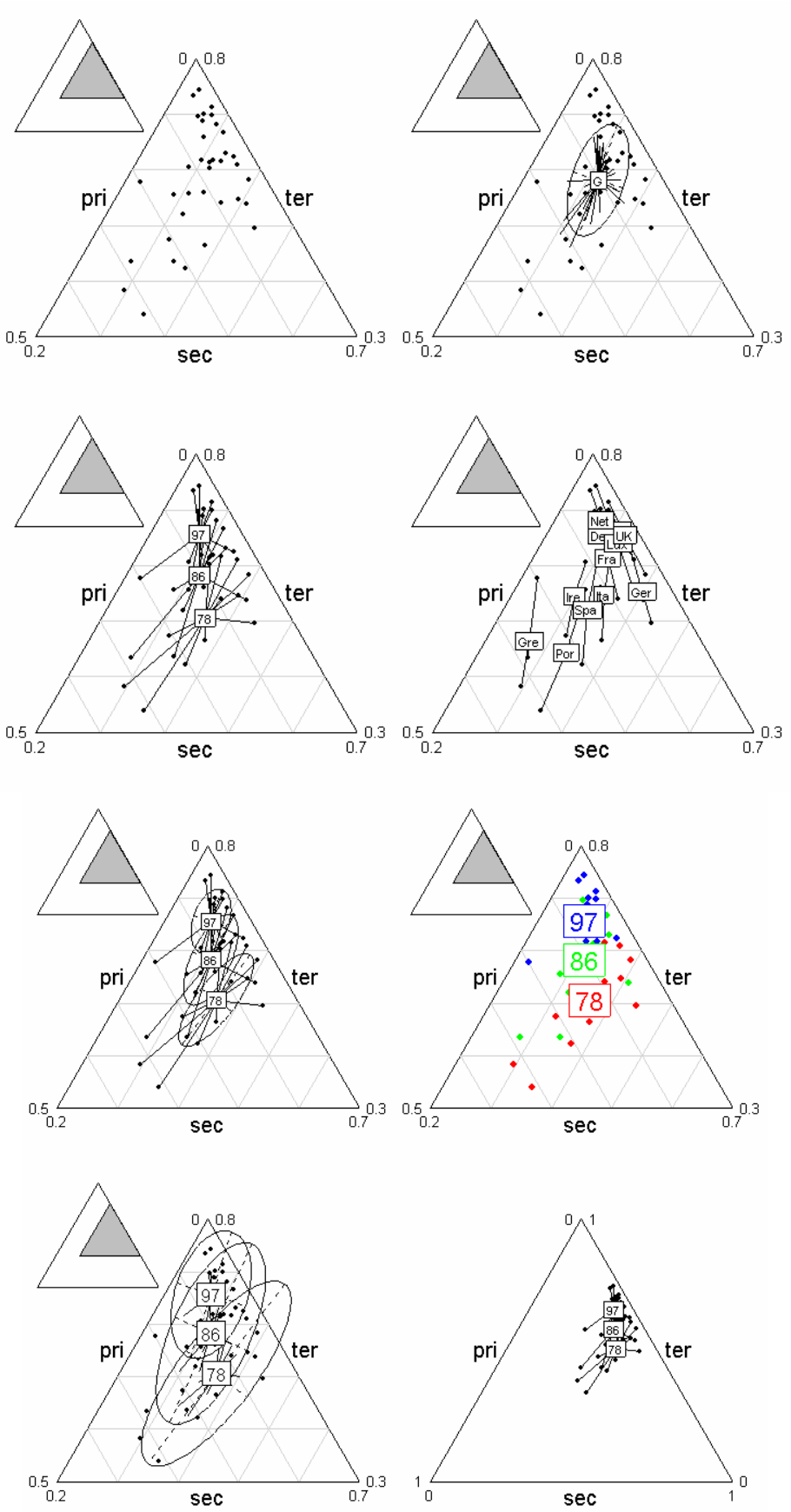
```
triangle.class(ta, fac, col = rep(1, length(levels(fac))), wt = rep(1, length(fac)),cstar
= 1, cellipse = 0, axesell = TRUE, label = levels(fac), clabel = 1, cpoint = 1, pch=20,
draw.line = TRUE, labeltriangle = TRUE, sub = "", csub = 1, possub = "bottomright",
show.position = TRUE, scale = TRUE, min3 = NULL, max3 = NULL)
```

Arguments

```
ta         data.frame à trois colonnes de nombres positifs ou nuls
fac        facteur de longueur le nombre de lignes de ta
col        vecteur de couleur  pour la représentation des groupes
wt         poids des lignes pour calculer les centres de gravité par classes
cstar      taille des étoiles, entre 0 (pas d'étoiles) et 1 (étoile complète) pour les
traits relaiant un point au centre de la classe
cellipse   taille des ellipses représentant les groupes
axesell    logique, si vrai les axes des ellipses sont tracés
label      vecteiur d'étiquettes des centres ds classes
clabel     taille des caractères des labels des classes (pas d'étiquettes si 0)
cpoint     taille des points (pas de représentation des points di 0)
pch        caractère de représenation des points
draw.line  logique, si vrai, les lignes du triangle sont tracées
labeltriangle   logique, si vrai les étiquettes des variables de ta sont placées sur les
côtés du triangle
sub        chaîne de charactères pour le titre de la figure
csub       taille des caractères pour le titre de la figure
possub     position de la chaîne de charactères pour le titre de la figure
("bottomright", "bottomlefet", "topright", "topleft")
show.position   logique, si vrai le sous-triangle contenant les données est replacé dans
le triangle complet
scale      logique, si faux, on utilise le triangle complet
min3       vecteur, si non nuls doit comporter trois nombres dans[0,1]
max3       vecteur, si non nuls doit comporter trois nombres dans[0,1]. min3 + max3 doit
faire c(1,1,1)
```

Exemples

```
data(euro123)
par(mfrow=c(2,2))
x=rbind.data.frame(euro123$in78,euro123$in86,euro123$in97)
triangle.plot(x)
triangle.class(x,as.factor(rep("G",36)),csta=0.5,cell=1)
triangle.class(x,euro123$plan$an)
triangle.class(x,euro123$plan$pays)
triangle.class(x,euro123$plan$an,cell=1,axesell=T)
triangle.class(x,euro123$plan$an,cell=0,csta=0,col=c("red","green","blue"),axesell=T,clab
=2, cpoi=2)
triangle.class(x,euro123$plan$an,cell=2,csta=0.5, axesell=T,clab=1.5)
triangle.class(x,euro123$plan$an,cell=0,csta=1, scale=F,draw.line=F,show.posi=F)
```


pri
ter
sec
0
0.8
0.5
0.2
0.3
0.7
G
97
86
78
Net
De
UK
Lux
Fra
Ire
Ita
Ger
Spa
Gre
Por
1

2. La taille des ellipses

D'une question à l'autre :

> *This works wonderfully. I thank you for the effort that was required. I am not familiar with these ellipses. Are they probability ellipses?*

Cela renvoie à une autre, encore :

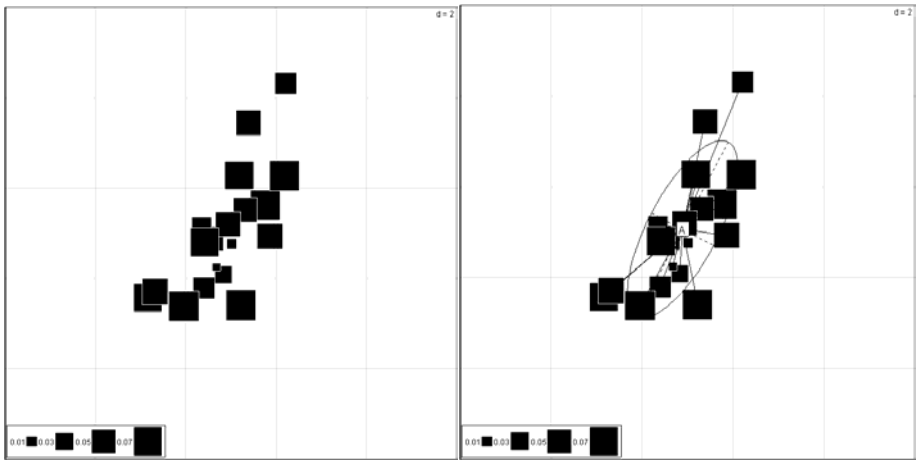
> *I'm a new user of R and this is the first time I adopt discriminant analysis. I'm using the ade4 package to distinguish treatments with three repetitions, from which color was measured based in 3 variable (L, a and b). I also analysed data in Systat 9.0 and results were different from that of R, mainly in ellipse size that was larger in Systat. My doubt is if this ellipse is a confidence interval, how to control its significance? I tried to change cellipse in s.class function, so that with value 6, the result was similar to that from Systat. But I think that the value must be calculated to provide the confidence ellipse. Then, I don't know what is correct. Please, could someone help me about this questions?*

C'est la même : quelle signification a la représentation d'une ellipse sur un nuage de points.

La question est juste. Il est nécessaire de préciser la taille des ellipses tracées dans ade4, que ce soit dans la version classique (ADE-4) ou dans le package pour R (ade4). Les utilisateurs du package peuvent faire l'expérience :

```
x = rnorm(20) + 5
y = x + rnorm(20)
xy = cbind.data.frame(x,y)
 poi = runif(20)
poi = poi/sum(poi)
s.value(xy, poi, xlim = c(0,10), ylim = c(0,10))
```

On voit un nuage de points pondéré.

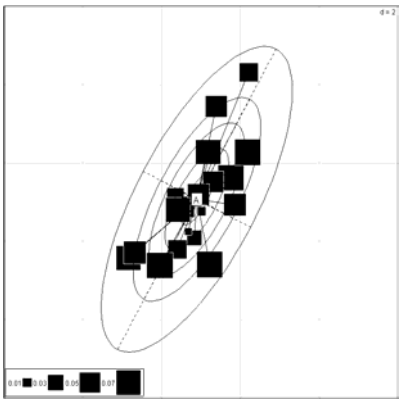


```
s.class(xy, as.factor(rep("A",20)), add.plot = T)
```

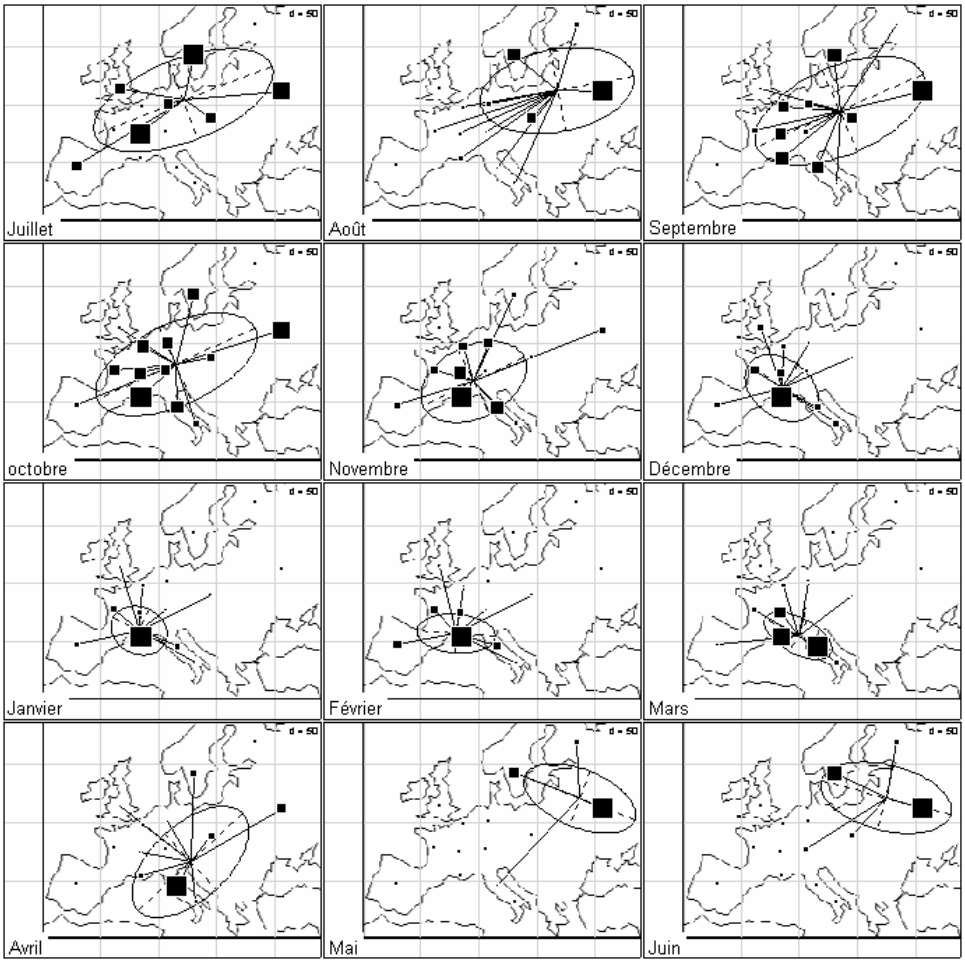
On voit le même nuage sur lequel s'ajoute une ellipse, une étoile et deux axes en pointillé. Dans ade4, l'ellipse est le résumé graphique de ce nuage pondéré et rien d'autre. Elle n'a rien à voir avec une ellipse de confiance, avec l'analyse discriminante, ou avec une ellipse de confiance en analyse discriminante. C'est un résumé d'un nuage de points.

Le centre de l'ellipse est le centre de gravité. L'étoile relie chaque point au centre de gravité. Les axes sont les axes principaux du nuage (Pearson 1901). La longueur des axes (par défaut k =1.5) égale k fois la racine des valeurs propres de la matrice de covariance, c'est-à-dire k fois l'écart-type des coordonnées des projections sur les axes.

```
s.class(xy, as.factor(rep("A",20)), add.plot = T, cell = 0.5)
s.class(xy, as.factor(rep("A",20)), add.plot = T, cell = 1)
s.class(xy, as.factor(rep("A",20)), add.plot = T, cell = 2)
s.class(xy, as.factor(rep("A",20)), add.plot = T, cell = 3)
```



On voit le même nuage sur lequel s'ajoute des ellipses de plus en plus grosses. On verra comment se dispersent les sarcelles d'hiver en Europe dans (Auda et al. 1983) http://pbil.univ-lyon1.fr/R/articles/arti033.pdf. Pour refaire la figure :




```
data(sarcelles)
if (require(pixmap, quietly=TRUE)) {
```

```
  bkgnd.pnm <- read.pnm(system.file("pictures/sarcelles.pnm", package =
"ade4"))
  par(mfrow = c(4,3))
  for(i in 1:12) {
  s.distri(sarcelles$xy, sarcelles$tab[,i], pixmap = bkgnd.pnm,          sub =
      sarcelles$col.names[i], clab = 0, csub = 2)
    s.value(sarcelles$xy, sarcelles$tab[,i], add.plot = TRUE, cleg = 0)
  }
}
```

Les ellipses de dispersion synthétisent *l'atterrissage des canards sur les cartes factorielles* dans Pirot (1981) et Pirot et al. (1984) reproduit à http://pbil.univ-lyon1.fr/R/articles/arti041.pdf

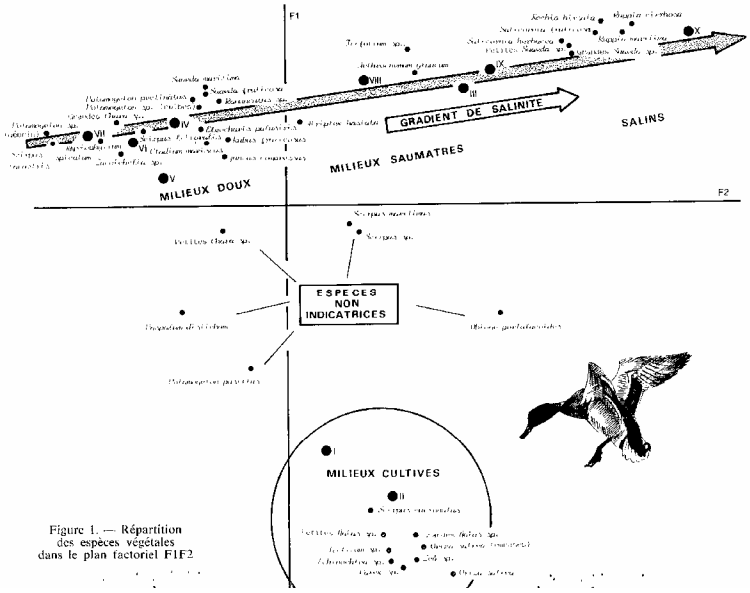
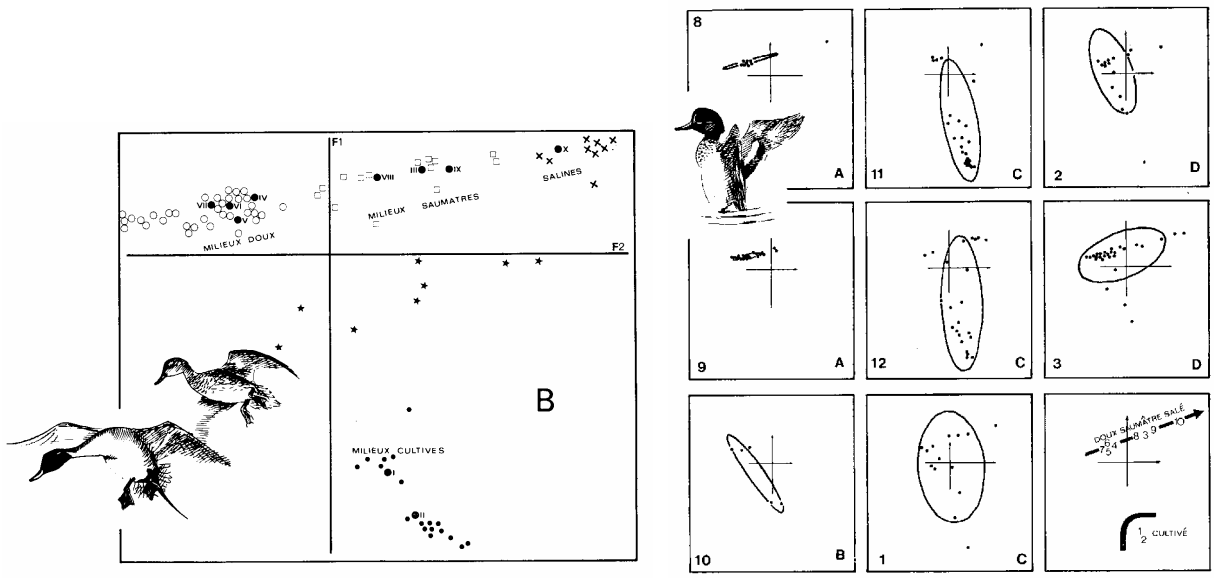


Figure 1. — Répartition des espèces végétales dans le plan factoriel F1F2



En haut, plan d'une analyse des correspondances croisant 40 espèces végétales et 10 types de milieu (abondance des graines dans des carottages de sédiments). On éliminera les espèces non indicatrices. A gauche, positions sur ce plan par moyenne conditionnelle de 82 pilets. A droite, positions sur ce plan par moyenne conditionnelle 180 sarcelles d'hiver répartis en 8 lots mensuels et donnant 4 modes de dispersion. Logiciel graphique de Y. Auda et dessins de S. Nicolle.

Il n' y a pas de règle pour définir la taille de l'ellipse. On peut simplement dire que si le nuage est un échantillon aléatoire simple d'une loi normale bivariée, la probabilité d'être dans l'ellipse de taille k est :

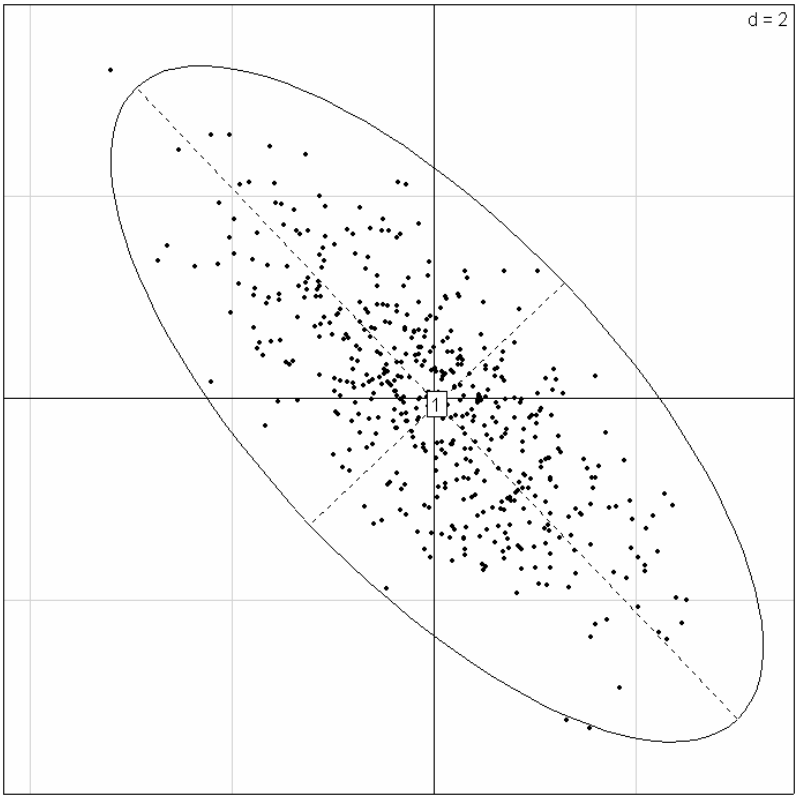
$$p = 1 - \exp\left(-k^2/2\right)$$

Dans ce cas, environ 67% des points sont dans l'ellipse $k = 1.5$ et 95% dans l'ellipse $k = 2.5$

```
1-exp(-0.5*(1.5)^2)
[1] 0.6753
1-exp(-0.5*(2.5)^2)
[1] 0.956
```

Les ellipses d'ade4 sont donc des résumés graphiques et non des régions de confiance.

```
library(MASS)
w=mvrnorm(500,c(0,0),matrix(c(1,-0.7,-0.7,1),2,2))
s.label(w,clab=0,cpoi=1.5)
1-exp(-0.5*(3.25)^2)
[1] 0.995
s.class(w,as.factor(rep(1,500)),cell=3.25,cstar=0)
```



On attendait 2.5 points en dehors (en moyenne !) et on en a eu 3. Le procédé est finalement très simple.

Auda, Y., D. Chessel, and A. Tamisier. 1983. La dispersion spatiale des Oiseaux au cours du cycle annuel : deux méthodes de description graphique. Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D **III**:387-392.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine **2**:559-572.

Pirot, J. Y. 1981. Partage alimentaire et spatial des zones humides camarguaises par 5 espèces de canards de surface en hivernage et en transit. Thèse de 3° cycle, Université Paris VI.

Pirot, J. Y., D. Chessel, and A. Tamisier. 1984. Exploitation alimentaire des zones humides de Camargue, delta du Rhône, France par cinq espèces de canards de surface hivernant: modélisation spatio-temporelle. La Terre et la Vie (Revue d'Ecologie) **39**:167-190.