# Cluster analysis as a tool for enhancing dialog between epidemiologists and clinicians ?

Vincent Auvigne, Lucas Léger

Ekipaj, 22 rue d'Assas, F-49000 Angers, France

vincent.auvigne@ekipaj.com

EKIPAJ Epidemiology and Data management

Both clinicians and epidemiologists have a role to play in the management of animal diseases, but sometimes encounter problems in understanding each other [1]
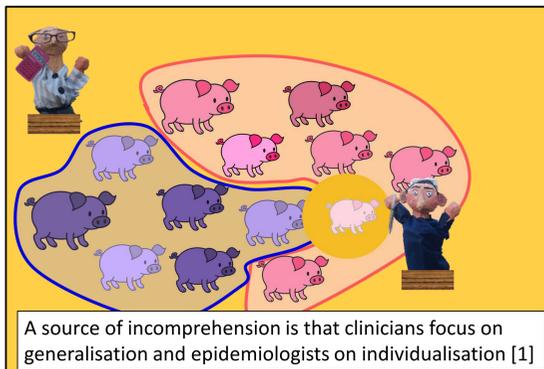
A source of incomprehension is that clinicians focus on generalisation and epidemiologists on individualisation [1]

In a cognitive perspective this incomprehension can be linked to our unwillingness to deduce the particular from the general and our willingness to infer the general from the particular [2].

In that perspective cluster analysis is a promising tool to enhance dialog between both characters as
- It's a thorough statistical method aiming at grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other clusters.
- It allows to represent sets by norms and prototypes, which is a feature of our heuristic [2]

The interest of cluster analysis was tested using a dataset of 80 ELISA titres in sow colostrum against five antigens of E. coli (F4ab, F4ac, F5, F6, F7).

Those seroprofiles were performed to assess quality of vaccination of sows.

|	F4ab	F4ac	F5	F6	LT
F4ab	1.00	0.58	0.46	0.79	0.68
F4ac	0.58	1.00	0.78	0.54	0.41
F5	0.46	0.78	1.00	0.57	0.65
F6	0.79	0.54	0.57	1.00	0.84
LT	0.68	0.41	0.65	0.84	1.00

Correlation coefficient between E. coli Titres in sows (colostrum)

Multiple correlation between titres make direct interpretation difficult : could cluster analysis be helpful ?

Constructing the model
A four step analysis without posing any prior hypothesis

Constructing, Step 1
- Perform Principal Components Analysis
- Create new non-correlated variables
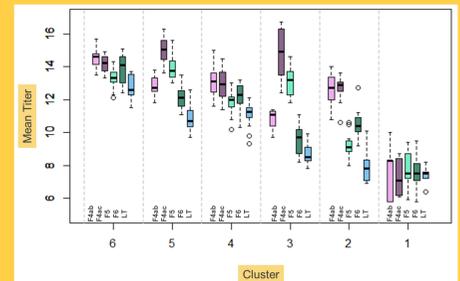- Select those of interest
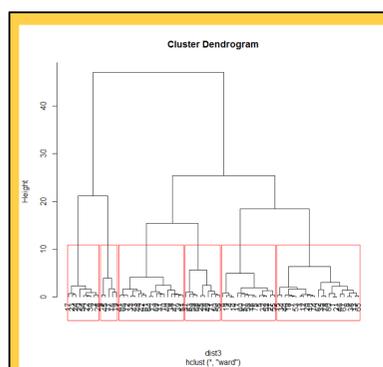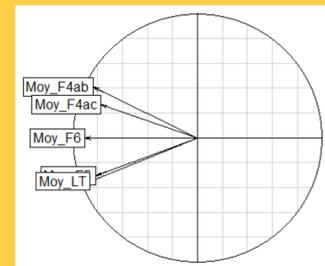
Constructing, Step 2
- Perform hierarchical clustering
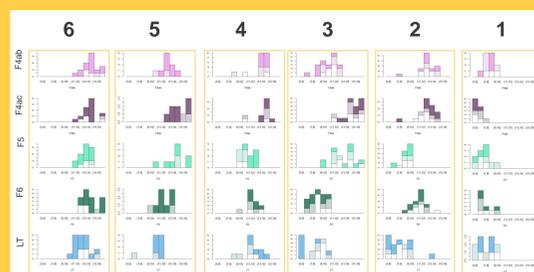- Choose number of clusters

Cluster Dendrogram
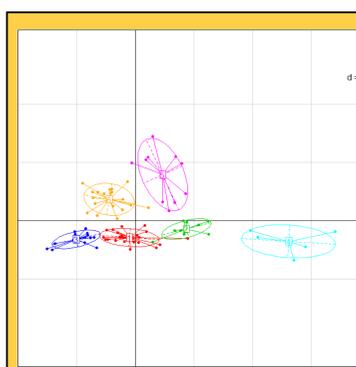
Constructing, Step 3
- Perform K-means clustering
- Validate clustering

Constructing, Step 4
Describe clusters and discuss biological meaning

The nearest profile to the centre of each cluster can been chosen as a prototype to describe the cluster

Using the model
Once the model constructed, it can be used to classify new seroprofiles i.e. determining which cluster fit the best to this new profile.
- The distance from the new seroprofile to the centre of each group is calculated
- A logistic regression model is used to assess the probability of belonging to each group

We hope such an approach help reaching the necessary balance between clinical and epidemiological vision of our world.

E. coli seroprofile classification (sow colostrum)

Enter the Elisa mean titer for all five antigens of the seroprofile to classified

The sero profile to classified is plotted on the Principal Component Analysis graphs

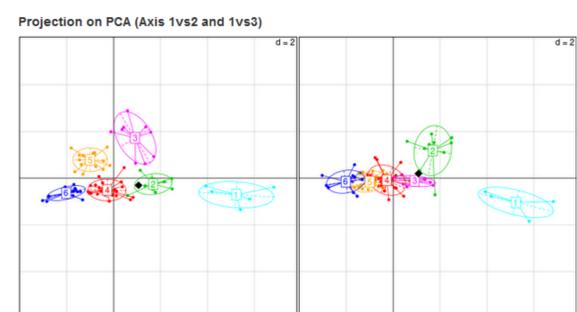The probability of belonging to each cluster is given by the bar chart
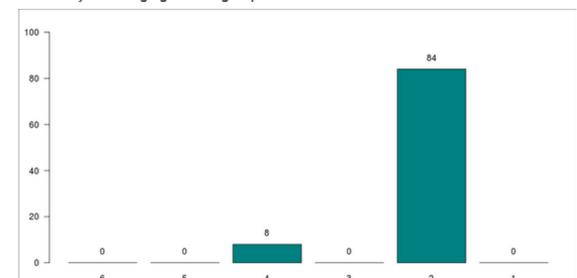
F4ab: 13
F4ac: 11
F5: 12
F6: 11
LT: 8.5

Designed by Ekipaj for MSD Animal Health

EKIPAJ

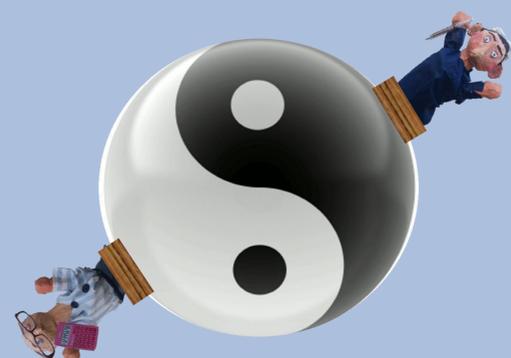Projection on PCA (Axis 1vs2 and 1vs3)

Probability of belonging to each groups

To perform the classification of new seroprofiles a web application in R using Shiny package from RStudio [3] is implemented. No installation is needed

Bibliography
1. - AUVIGNE V., BELLOC C.: [The clinician and the epidemiologist: two paradigms, one dialogue]. Epidémiol. et santé anim. 2012, 141-147.
2. - KAHNEMAN D.: Thinking, Fast and Slow. 512 pages, Penguin, 2012.
3. - RStudio - Shiny [Internet]. Available from: http://www.rstudio.com/shiny/