

[See all articles](#)

January 5, 2022 • 10 min read

# How useful are PCA statistics for Agronomy?

Selecting crosses, evaluating cultivars or inputs, involve **comparing materials on several traits**. And you wonder: how can I compare yield, flowering date, and drought tolerance at the same time? How to draw some conclusions from all these variables?

Then you ask yourself what to do to **explore the links between traits and visualize properly the similarities** between individuals or treatments. I am sure you have read or heard of Principal Component Analysis (PCA). And you wonder: can PCA help me choose which quantitative variables to analyze, and take the good decisions for my agronomy research? The answer to this question will allow you to elucidate or bring to light your dark spots on the PCA.

**Principal Component Analysis (PCA)** is one of the multivariate statistical methods, useful when confronted with  $n$  individuals observed on  $p$  quantitative variables. What is the purpose of this statistical tool? What are the limits of classical PCA for multiple comparisons, and what are the models that allow to go further?

*"It is obviously more tedious if not impossible to summarize a set of agronomic data visibly by describing agronomic characteristics under these conditions, multivariate methods*



*should be used"*

Romain Lucas Glele Kakai<sup>1</sup>.

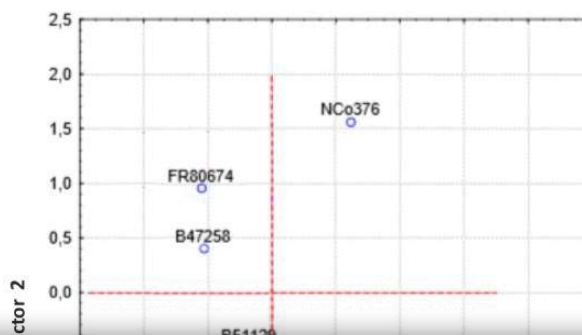
## What is the purpose of Principal Component Analysis (PCA)?

PCA is an extremely powerful statistical tool for synthesizing information, very useful when there is a **large amount of quantitative data** to be processed and interpreted. When the variables are quantitative, we can perform a Principal Component Analysis.

The forms of PCA visualization for better decision-making in agronomy are usually a **cloud of points** or **in the form of a circle of correlations between variables**.

Here is an example of the use of PCA in the context of **screening promising commercial varieties of sugar cane** in Northern Ivory Coast (*Picture 01*). We are interested in 5 varieties of sugar cane and a control. Five pieces of information were identified for each cultivar: Cane yield, sugar yield, saccharin content, smut and stem borer sensitivity.

The following graph represents our data, to see if there are significant differences between the varieties, or if we have broadly the same characteristics for all 5.



Software ▼

Your project

Customers

Try  
Bloomeo

Resources ▼

About us

sugar cane.

Factor 1: 54.08% of the total variance (saccharin richness  $r = 0.77$ , cane yields  $r = 0.95$  and sugar  $r = 0.99$ ).

Factor 2: 27.15% of the total variance (charcoal  $r = 0.66$  and bored internode %  $r = 0.78$ ).

For this, we have carried out a PCA with 2 principal components, to represent the 5 variables in only two dimensions. Horizontally, the factor 1 represents cane yield, saccharin and sugar richness; vertically, the second component displays the sensitivity of varieties to charcoal and stem borer. These two dimensions alone summarize **more than 85% of the information**.

You will notice in the figure 1 that four groups of varieties have been distinguished: At a glance you can see the control NCo376 at the top right, showing good agro-technical qualities but highly sensible to diseases. Then two groups give low yields: SP70-1143 with high sensibility, and a group of 2 varieties more resistant to diseases. Last but not least, the light blue group is made of the 2 most interesting varieties **FR80674** and **B47258**, with a high sucrose content, high cane and sugar yields, and very resistant to stem borer and smut!

The PCA algorithm performs on the individuals / variables matrix various operations (data centering-reduction, diagonalization of the correlation matrix, extraction of eigenvalues and eigenvectors, etc.) in order to combine the initial dimensions to obtain a reduced number of variables: **the principal components (PCs)**. This way, you can explain as much genetic variability (usually variance) as possible with as few PCs as possible.

*“Grouping a set of data promotes accuracy by broadening their scope”*

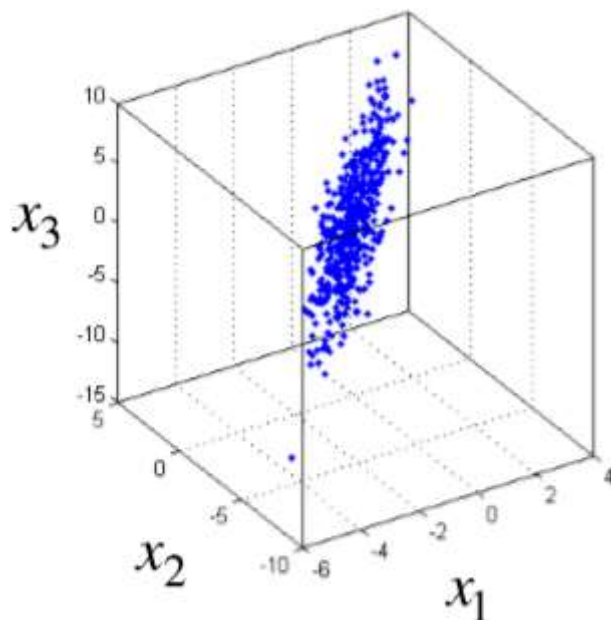
Erwann Lagabriele<sup>2</sup>.

## What are the limits of classical PCA for multiple comparisons ?

**PCA only works with quantitative variables** such as yield, height, water and amyloidosis content: You can't study colors, disease, drought, submergence. In short all that is numerical variables: counts, percents, or numbers.

In addition, if the data does not have an underlying structure, the dimension cannot be reduced, as shown in *picture 02*.





**Picture 02:** Three-dimensional data.

The principal components are the underlying structure in the data. They represent the directions in which the data has maximum variance and also the directions in which the data is the most spread out.

The main drawback of principal component analysis (PCA), especially for applications in high dimensions, is that the **principal components are linear combinations of all input variables**. Therefore, the results may be very sensitive to the presence of even a few atypical observations in the data. When, for example, you have data on lines with the greatest variance, PCA will largely bias interpretations of analyzes involving a common distribution. Without meeting a set of a priori requirements regarding data structure, the ordered axis plot approach is likely to produce misleading results<sup>3</sup>.

This is the case, for example, of the yield of a crop which is a function of several other components such as tillering, the quantity of grain per tiller and the weight of 1000 grain. You will notice through this example that the yield is a variable with a large dimension because it depends on the other components listed above.

Due to binding and correlated selection, the evolutionary response of any phenotypic trait can only be properly understood in the context of other traits<sup>4</sup>.

*“The limitations of Principal Component Analysis come from the fact that it is a projection method, and that the loss of information induced by the projection can lead to erroneous interpretations”*

Magloire Oteyami <sup>3</sup>.

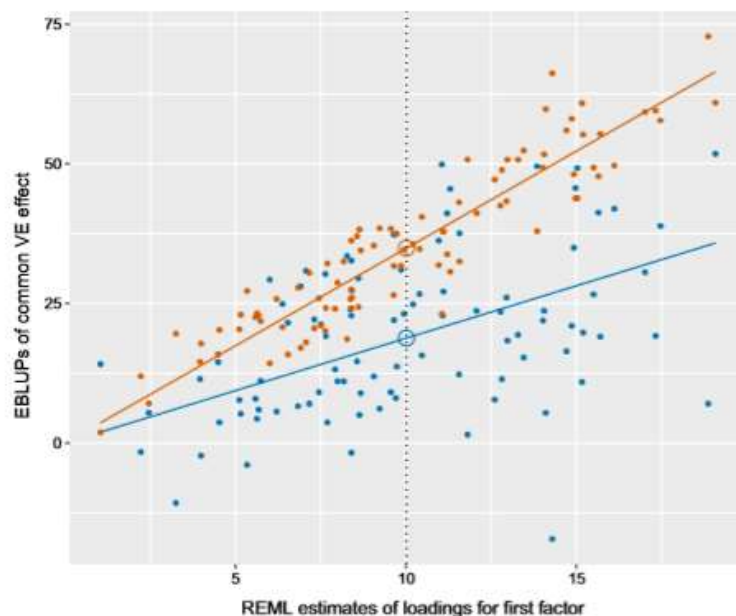


o further with Multiple Correspondence Analysis



Linear mixed model is used for evaluations of genotypes under various environmental conditions, to identify genotypes with superior performance in all area and under conditions or sets of conditions such as abiotic or biotic stresses. Know that here that the environment designates the unfavorable area for the production of rice, or corn, sorghum etc. This is due to the impact of climate variability or climate change which manifests itself in agriculture by drought, flooding, insect pests, diseases.

How to proceed with the selection of superior varieties while using data from multi-environment trials while being based on mixed models of factor analysis? *Pictures 4a and 4b* gives the answer to this question.

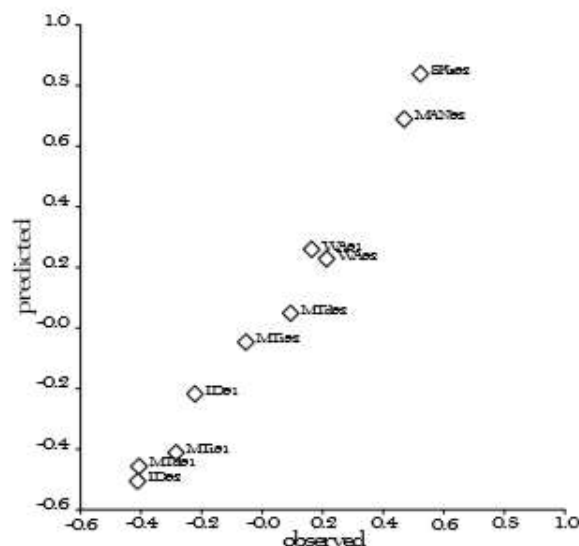


**Picture 04a:** Differential response of varieties to environments.

Indeed, Figure 4a is translated according to the following terms: Superimposed first latent regression plots for two varieties, V6 (colored blue) and V1 (colored orange). Slopes of the solid lines are given by the EBLUPs of the (rotated) variety scores for the first factor. The open circles are the overall performance measure for each variety, namely the value on the regression line at the mean value of the estimated loadings for the first factor (vertical dotted line).

How is the correlation between the different parameters for different area when using Linear Mixed Model? Let's see an example with this study carried out by Marcos Malosetti, who evaluated the performance of genotypes in different environments using Linear Mixed Model.





**Picture 04b:** Plot of the correlation between yield (ton ha<sup>-1</sup>) and heading date (days after 1st January) in each of 10 environments.

Picture 04b illustrates the correlation between yield (ton ha<sup>-1</sup>) and heading date (days after 1st January) in each of 10 environments. This can be very useful for you for the analysis of **crop cultivar breeding and evaluation trials**.

Through this example, you understand that **Linear Mixed model is decision tool for the management of risks or constraints linked to the environment and to the Agriculture**, either biotic (insect pests, diseases) or abiotic (drought, flooding, iron toxicity of the soil).

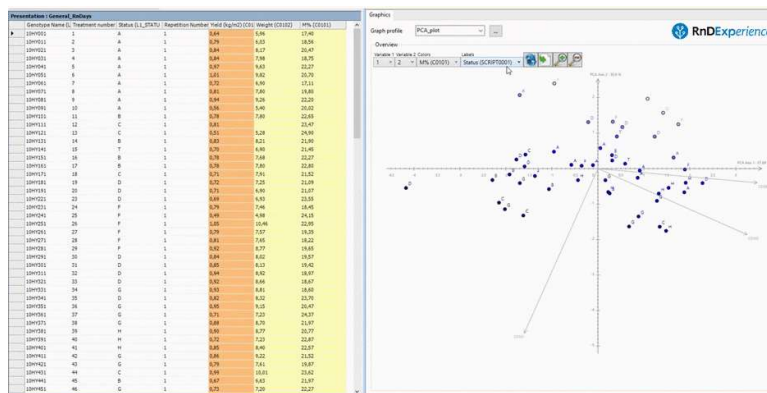
## Usefulness of your agronomy software's statistical wizard

Accurate statistics is the key to make sure your results are reliable all along your experimentation. You know the idiom "correlation is not causality": It can be tempting to draw hasty conclusions based on biased samples, inadequate analysis method or erroneous data... but **the statistical tools of your agro-research software help you trust your breeding and testing results** at the key steps of your plant research processes:

- Choose the most appropriate design
- Identify your best performing lines with GCA
- Find the best parents for future crosses with SCA
- Evaluate environment response with GxE matrix
- Compare the performances of your treatments with a valid ANOVA and a complete set of statistic calculations
- Explore your data with interactive graphs for ACP, scatter plots, box plots and histograms
- Characterize traits transmissions with pedigree visualization







**Picture 05:** Example of dynamic PCA graph in RnDExperience® software.

Combining analytical tools with performant data management tool and R&D resources planning is the key of success of your agronomy campaigns. Doriane through our products and services is leading researchers to their goals!

#### References:

- 1- Prof. Dr. Ir. Romain Lucas GLELE KAKAÏ is Lab Director of the Laboratory of Biomathematics and forestry estimations of University of Abomey-Calavi.
- 2- Erwann Lagabriele (2007). Planning of biodiversity conservation and territorial modeling on Reunion Island. Geography. University of La Réunion, France.
- 3- Lande R. Quantitative genetic analysis of multivariate evolution, applied to brain-body size allometry. Evolution. 1979;33:402–416
- 4- Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sinauer Associates; Sunderland, MA: 1998
- 5- Alaye H. Magloire Firmin OTEYAMI, Agronomist, Geneticist-breeder in Benin, author of this article
- 6- Abdi, Hervé & Williams, Lynne. (2010). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2. 433 - 459. 10.1002/wics.101.

Figure 1: Kouamé, Didier & Pene, Crépin & Zouzou, Michel. (2018). Evaluating varietal resistance of Sugarcane to the Tropical African Cane borer (*Eldana saccharina* Walker) in Ivory Coast.

Figure 4b: Malosetti M, Voltas J, Romagosa I, Ullrich SE, van Eeuwijk FA (2004) Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica 137: 139-145

Figure 5: © Doriane SAS



**Magloire Oteyami**

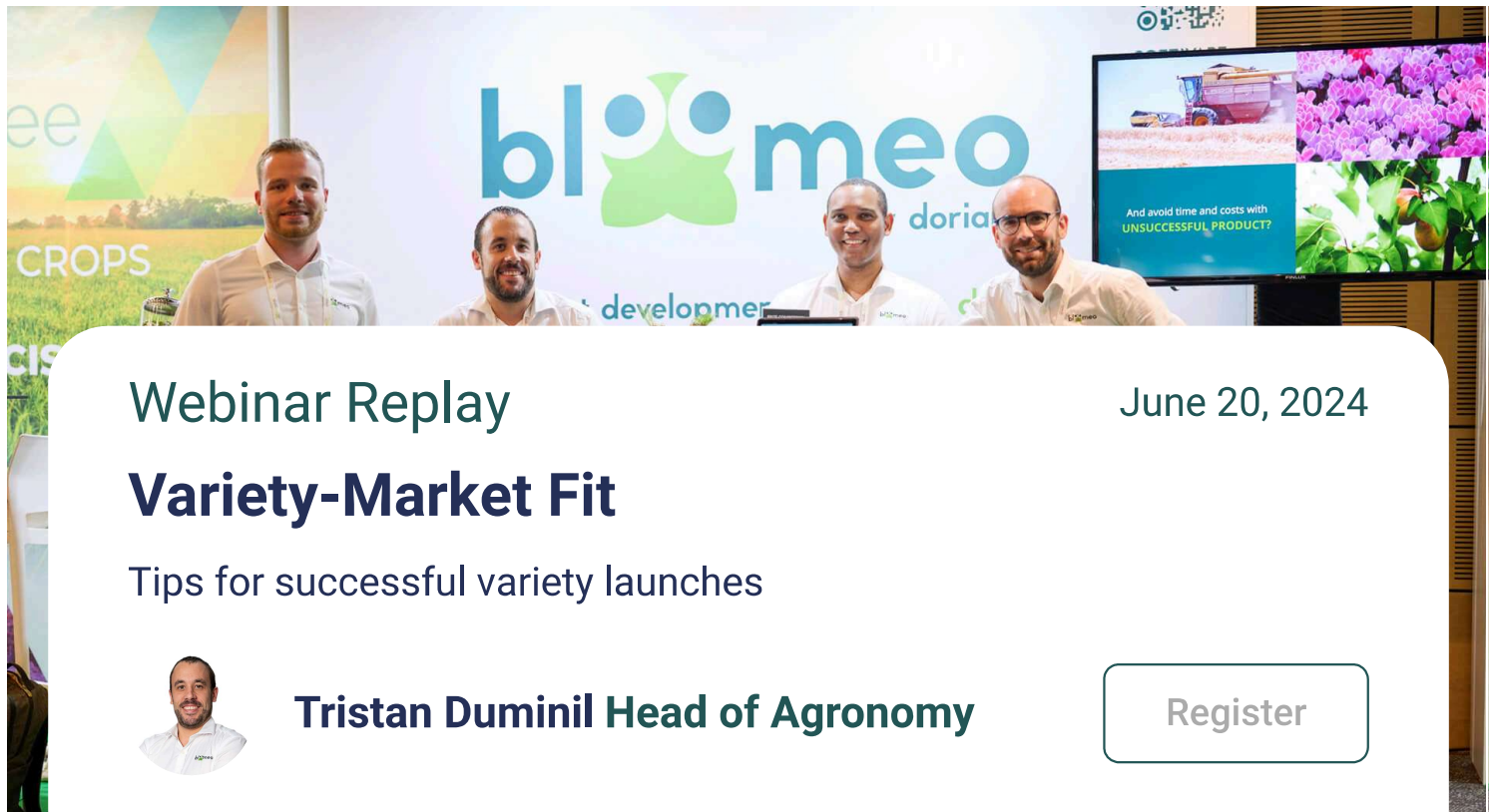
Plant breeder, co-founder of VART-Lab

Did you find this article useful? **Share it!**






# Save the date



**Webinar Replay** June 20, 2024

## Variety-Market Fit

Tips for successful variety launches

 **Tristan Duminil Head of Agronomy**

[Register](#)

## Ask our team about your project



# Clément B.

## Business & Agronomy Engineer

[Book a demo](#)



8 rue de Russie - 06000 Nice -  
France  
contact@doriane.com - +33  
492 478 444

[Legal mentions](#) | [Privacy policy](#) | Doriane ©2024

