# FROM STABLE TO…”THE TABLE”
or
# how to do customer “in house” trials

Dr. Alberto Morillo Alujas
Dr. Emilio López Cano
Dr. Daniel Villalba Mata

- Dr. Alberto Morillo Alujas
  - Ph.D. in Veterinary, degree in Statistics and specialist in animal production, consultant of Tests and Trials, S.L.U. in Monzón, Spain
- Dr. Emilio López Cano
  - Ph.D. in Statistics, specialist in quality systems, professor in URJC, Madrid, Spain
- Dr. Daniel Villalba Mata
  - Ph.D. in Agricultural Engineer, specialist in animal production, professor in UdL, Lleida, Spain

# How to change the mind

"There is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things because innovation runs into the hostility of all those the former situation benefits and only meet lukewarm defenders in whom wait for benefits of the new one."

*Nicolas Maquiavelo, 1469-1527.*

PARTNERS IN PROGRESS

# Validity

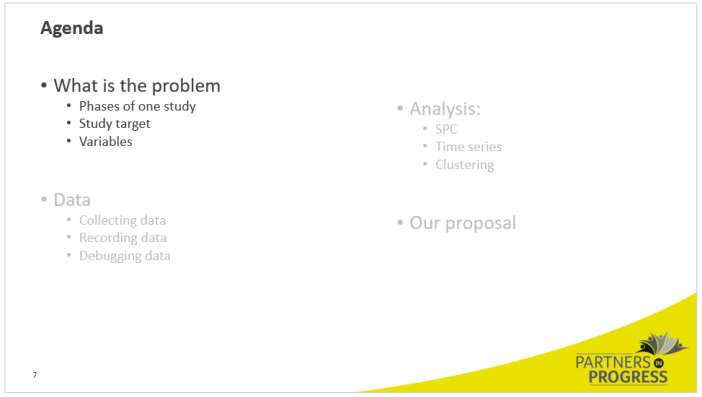
## Internal

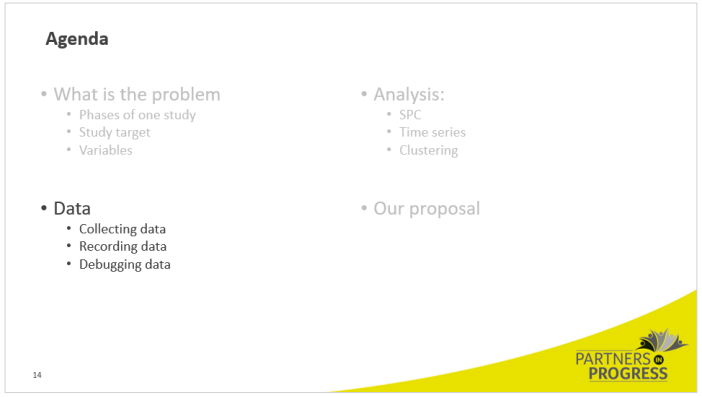
- To know how “something” works or performs
- Controlled studies
- Pen trials
- All factors under control
- Results tell you how your “something” works

## External

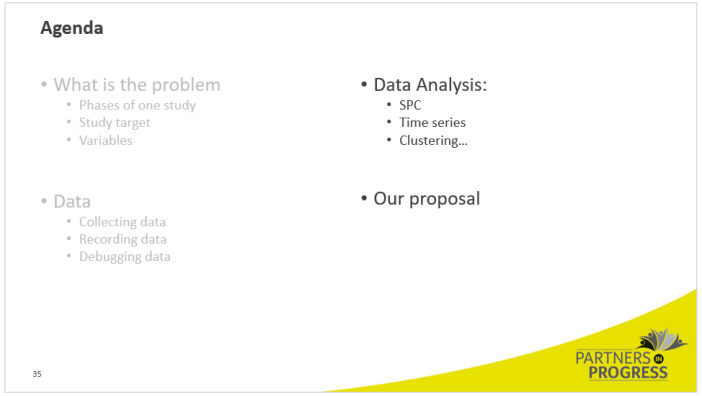
- You know “something” works in pen trials
- You want to try it in all circumstances
- Few factors under control
- Results tell your customer how “something” works in its own circumstances

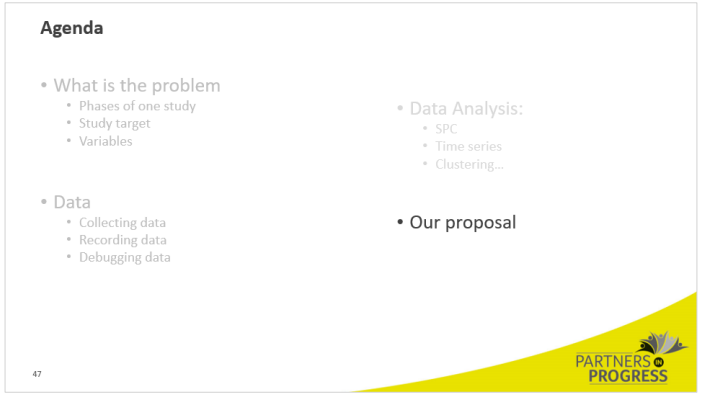
# How to do customer “in house” trials








PARTNERS IN PROGRESS

# Agenda

- What is the problem
  - Phases of one study
  - Study target
  - Variables
- Data
  - Collecting data
  - Recording data
  - Debugging data
- Analysis:
  - SPC
  - Time series
  - Clustering
- Our proposal

# What is the problem?

*“A PROBLEM WELL STATED IS A PROBLEM HALF SOLVED”*

*CHARLES FRANKLIN KETTERING*

## Where is the problem?

- To define what is an experimental unit
- To understand what is a relational DB
- Fear to manage numbers
- “Feelings” instead “Realities”
- Lack of statistical education

# Study plan: Stages of organizational work

- Data debugging
- Original dataset or matrix
- Calculated or creation of new variables
- Data analysis
- Discussion and interpretation
- Conclusions

- Study target: Hypothesis
- Study design: Protocol
- Population: Individuals, batches or farms
- Data form
  - Capture data from registers
  - Capture data from the application form
- Data introduction

# Objective of the study

- A good definition of the study's objective will allow us to define the study variables correctly.
- Main objective and main response variable
- Secondary objectives and response variables

# Phases of the study

| | |
|---|---|
| Plan | Protocol's components |
| Concept | What is the question to be answered<br>Bibliographic review<br>Previous knowledge<br>**Write the hypothesis and targets** |
| Design | Study type |

# Phases of the study

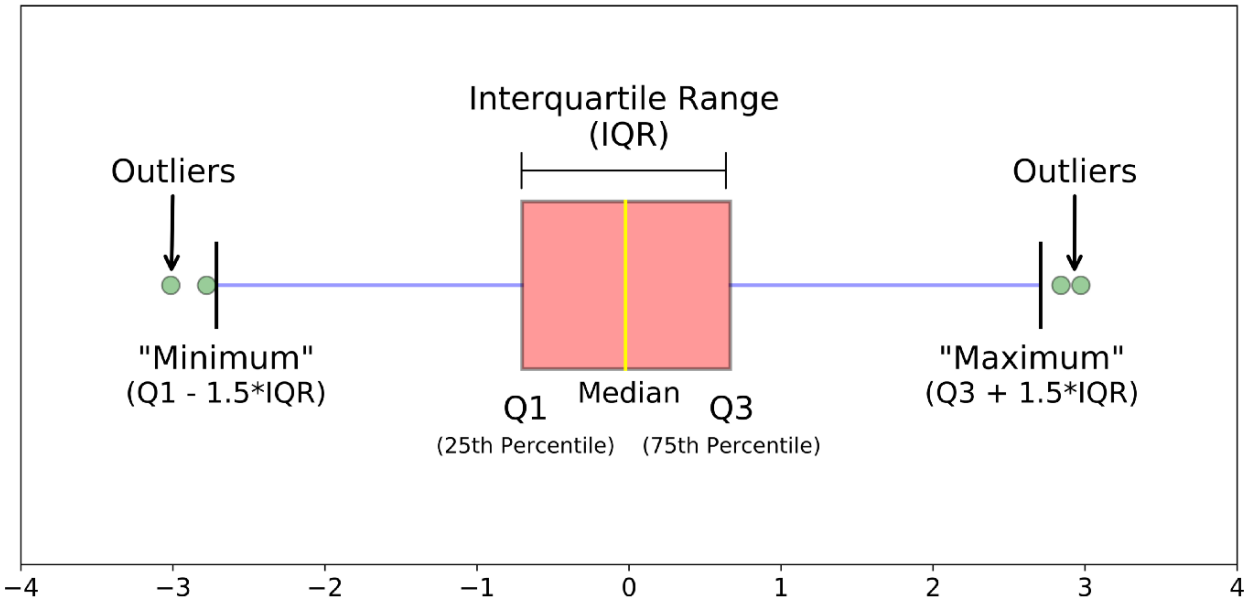
| | |
|---|---|
| Variables in the study | **Variables selection**<br>Metrics of the variables |
| Study population | Selection criteria<br>Type of sampling<br>**Sample size calculation** |
| Collecting data | Information sources<br>Questionnaire. Validation |
| Analysis strategy | Data processing |

# Agenda

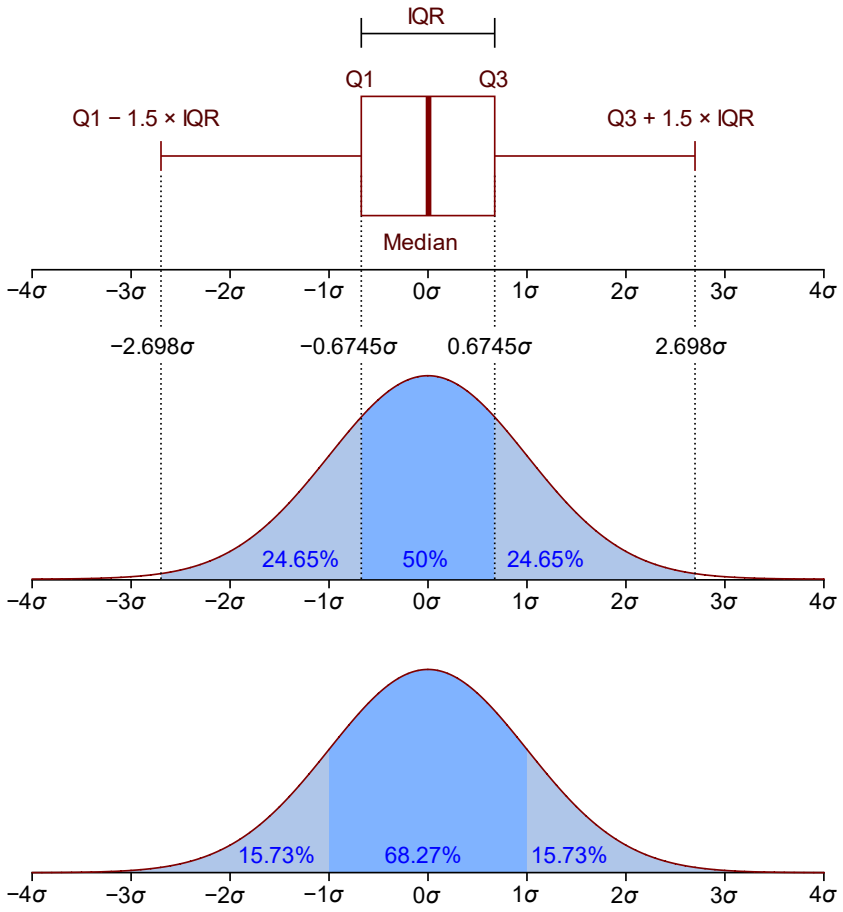
- What is the problem
  - Phases of one study
  - Study target
  - Variables
- Data
  - Collecting data
  - Recording data
  - Debugging data
- Analysis:
  - SPC
  - Time series
  - Clustering
- Our proposal

# Data debugging examples

## Boxplots: help(boxplot). Outliers




https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51




By Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285



PARTNERS IN PROGRESS

# Agenda

- What is the problem
  - Phases of one study
  - Study target
  - Variables
- Data
  - Collecting data
  - Recording data
  - Debugging data
- Data Analysis:
  - SPC
  - Time series
  - Clustering…
- Our proposal

# Types of data analysis

## DESIGNED ANALYSIS

Anova Table

| Sources | df | SS | MSS | F-value | Table Value | | |
|---|---|---|---|---|---|---|---|
| | | | | | 5% | 1% | 0.1% |
| Variety | $3(n_1)$ | 166.19 | 55.40 | 71.03*** | 3.9 | 7.0 | 13.9 |
| Replication | $3(n_1)$ | 3.19 | 1.06 | 1.36 | | | |
| Error | $9(n_2)$ | 7.06 | 0.78 | | | | |
| Total | 15 | | | | | | |

$$\text{MSS} = \text{mean sum of squares} = \frac{\text{SS}}{\text{df}}$$

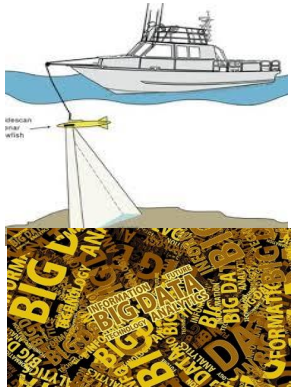
$$\text{Calculated F-value} = \frac{\text{MSS of source}}{\text{MSS of error}}.$$

- Requires trial design
- Reduced numbe[illegible] experimental [illegible]ts
- Expensive
- If we [illegible]des[illegible]ed → Conclusive resu[illegible]

## FISHING ON DATA

- Ofte[illegible]ithou[illegible]gn
- A lo[illegible]ot always a lot of [illegible]ental units
- Chea[illegible]
- Ris[illegible]s!

## BIG DATA ANALYSIS




- Big dataset
- New statistical [illegible]proaches to avoid bia[illegible]
- Mand[illegible]tor[illegible]teraction betw[illegible]tisticians and people[illegible] know the origin of data

PARTNERS IN PROGRESS

# Fishing on data

**Real example of Poultry data**

Experimental unit: Farm

Data available: **EPEF** (European Production Efficiency Factor)

Two treatments applied in different farms

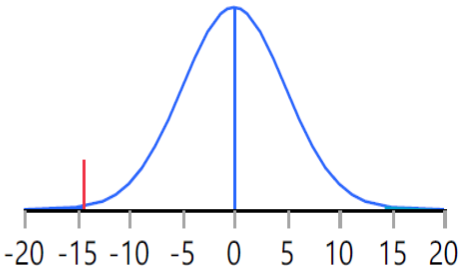
n: around 100 to 200 farms per treatment

| LEVEL | N farms | EPEF | EE | LCL 95% | HCL 95% |
|---|---|---|---|---|---|
| NEW PRODUCT | 190 | **348.7** | 2.7 | 343.4 | 354.0 |
| OLD PRODUCT | 105 | 334.3 | 4.1 | 326.1 | 342.5 |

**Prueba t**

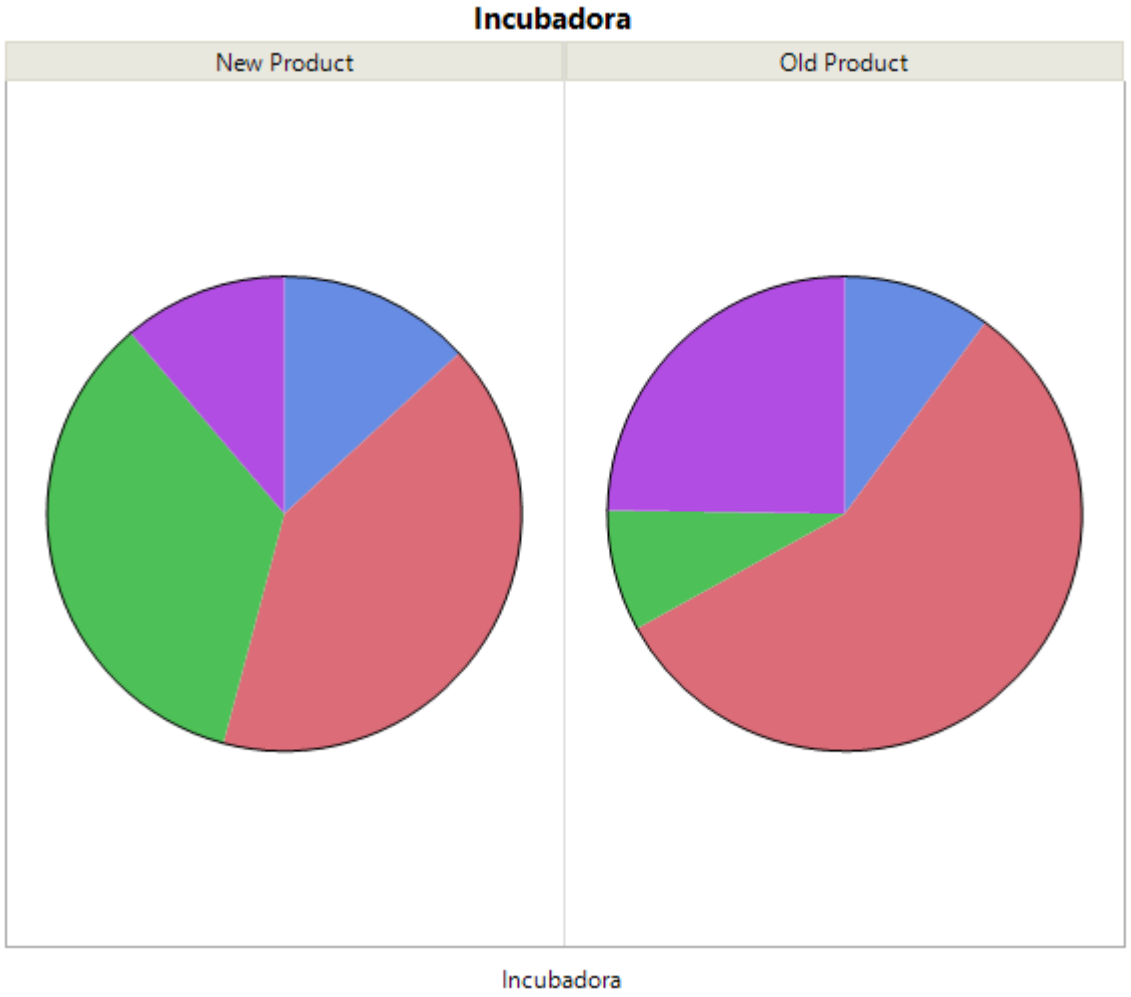
Asumiendo varianzas desiguales

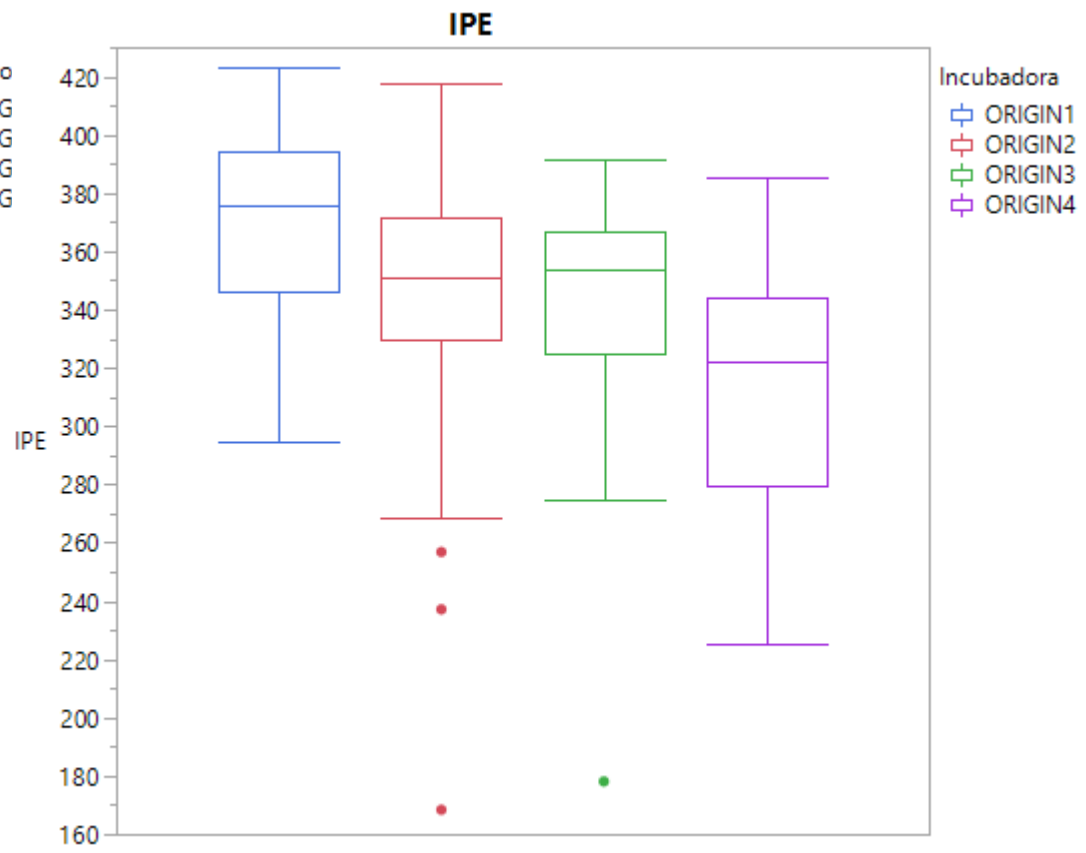
| | | | |
|---|---|---|---|
| Diferencia | -14.384 | Razón t | -2.9265 |
| Error estándar de la diferencia | 4.915 | Grados de libertad | 192.1443 |
| Diferencia del límite de control superior | -4.690 | Prob > \|t\| | 0.0038* |
| Diferencia del límite de control inferior | -24.078 | Prob > t | 0.9981 |
| Confianza | 0.95 | Prob < t | 0.0019* |




PARTNERS IN PROGRESS

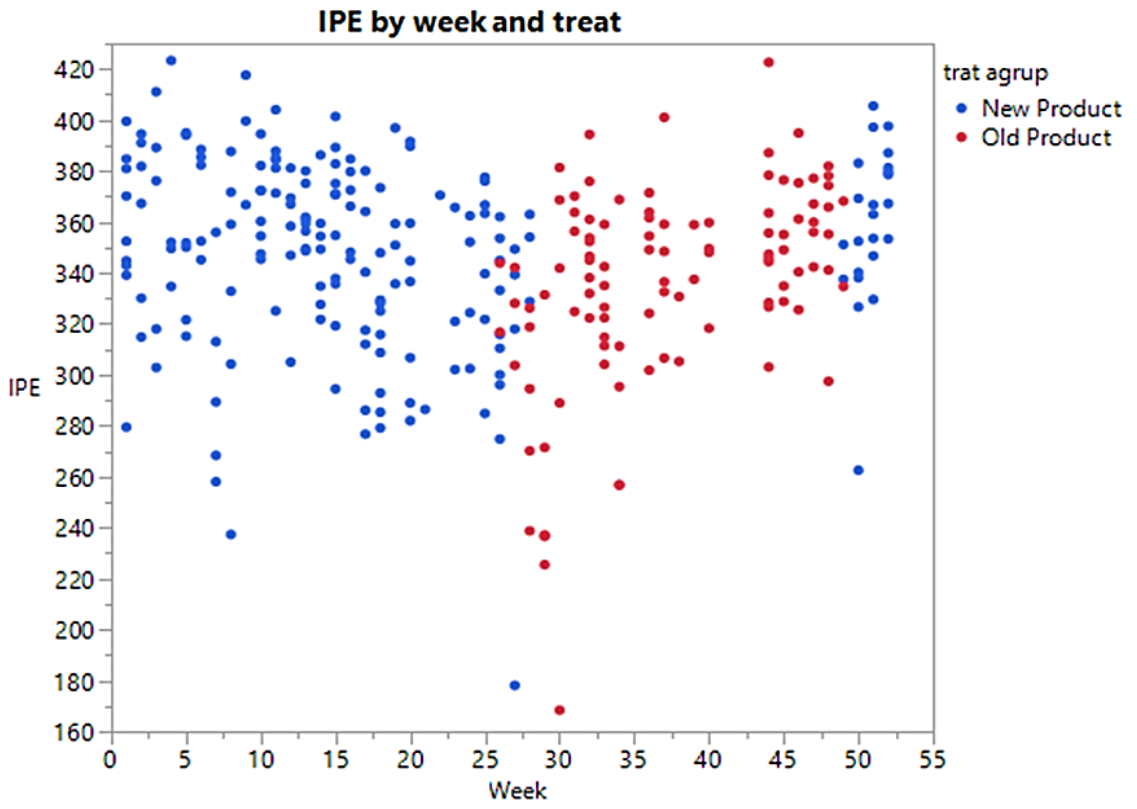
# Bias: chicken's source (hatchery)!

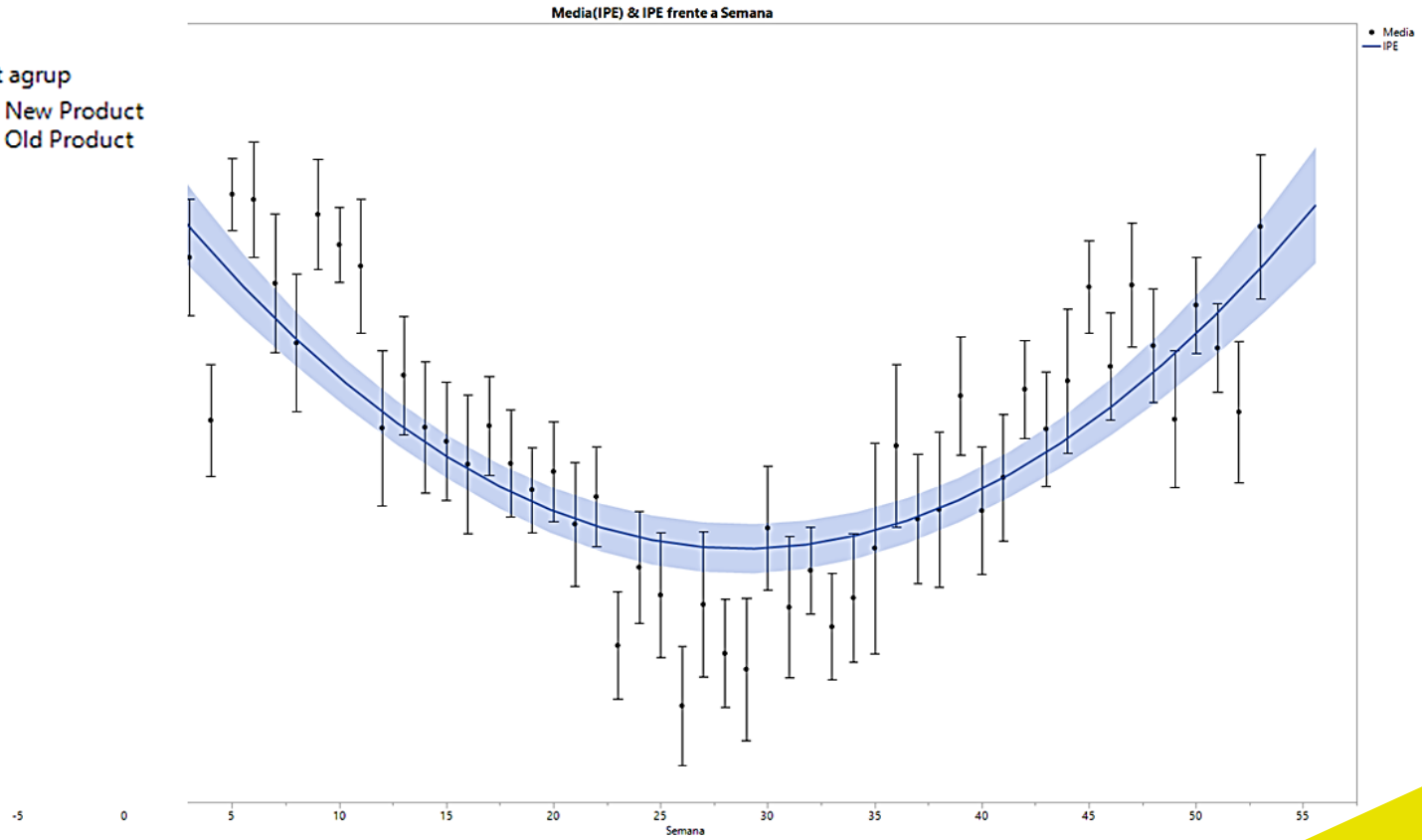



PARTNERS IN PROGRESS

# Bias: Week of year



# Analysis including origin and week

**Pruebas de los efectos**

| Fuente | N parámetros | Grados de libertad | Suma de cuadrados | Razón F | Prob > F |
|---|---|---|---|---|---|
| Origin | 3 | 3 | 62978.464 | 17.7168 | <.0001* |
| week | 1 | 1 | 2765.896 | 2.3343 | 0.1277 |
| week*week | 1 | 1 | 33399.181 | 28.1871 | <.0001* |
| treatment | 1 | 1 | 55.892 | 0.0472 | 0.8282 |

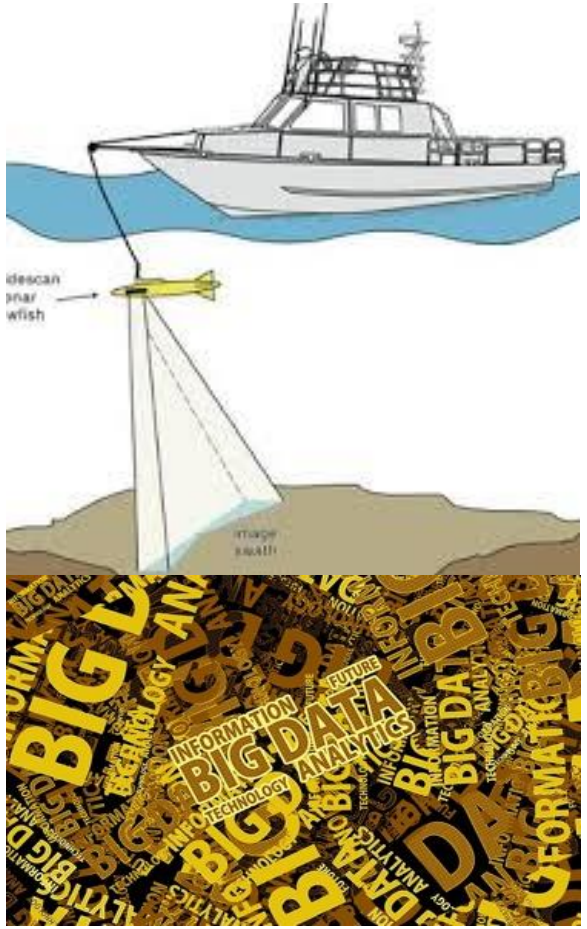
**Tabla de medias de mínimos cuadrados**

| Nivel | Media de mínimos cuadrados | Error estándar | Media |
|---|---|---|---|
| New Product | 329.69444 | 4.4929381 | 348.673 |
| Old Product | 330.99741 | 4.3062538 | 334.289 |

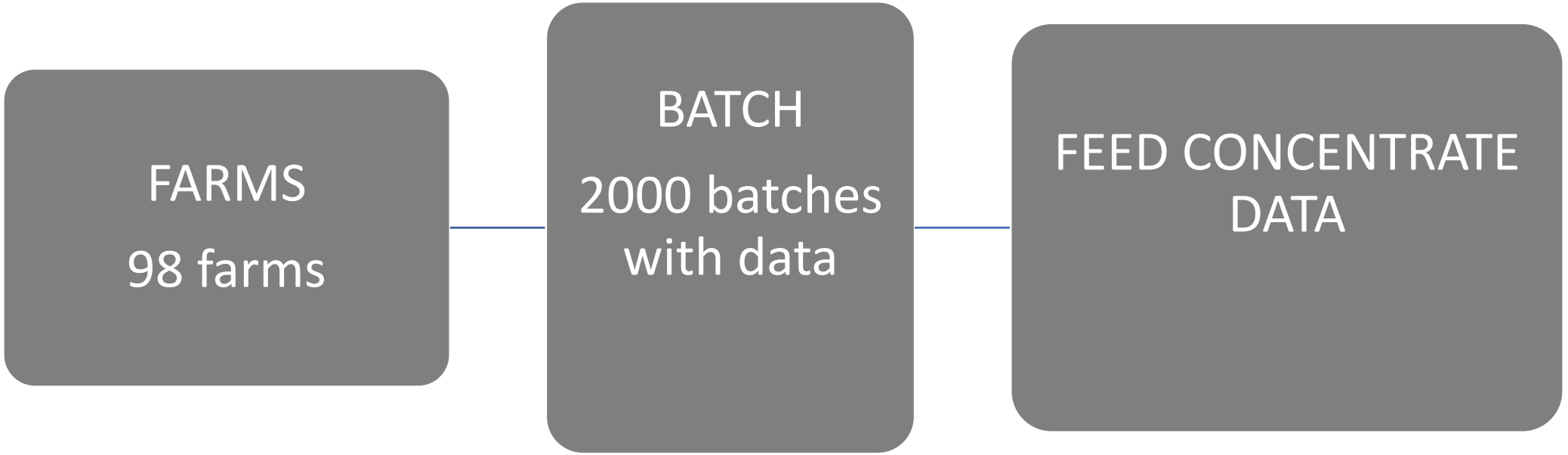
PARTNERS IN PROGRESS

# Big (and not so big) data analysis

- Find the **effects** that explain the **variability** of data
- **Summarize** the effects in order to **simplify** the interpretation
- Summarize the **response** or **performance** of the system
- Find the **association** between effects and responses



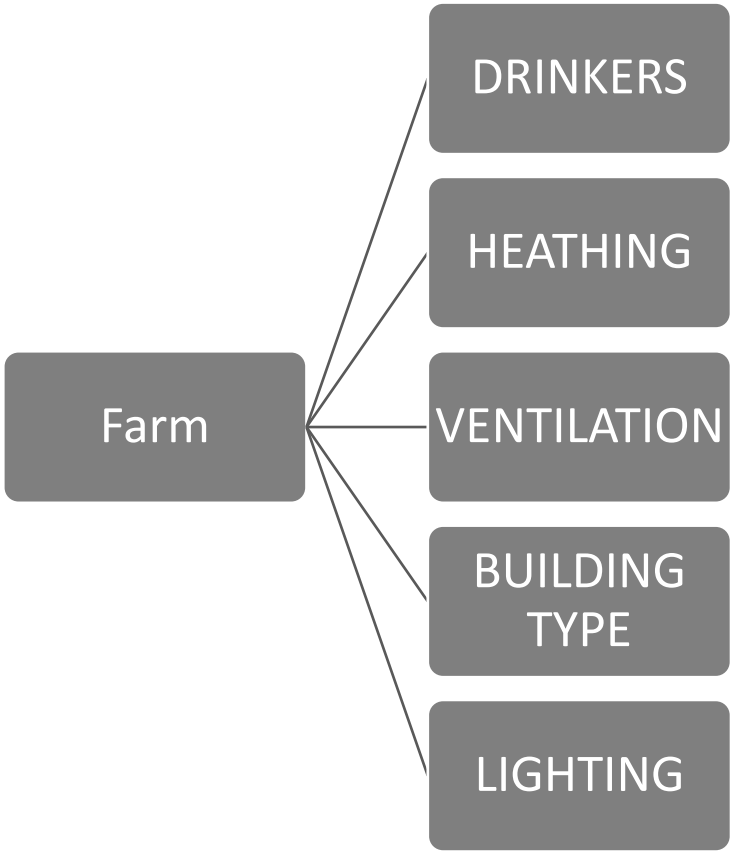
# Merging different databases




PARTNERS IN PROGRESS

# Find EFFECTS that explain variability

## I. Type of farm/ Structural characteristics




More than 50 variables defining each farm!

Not possible to include all the variables→ **CLUSTERING OF HOMOGENEOUS FARMS**

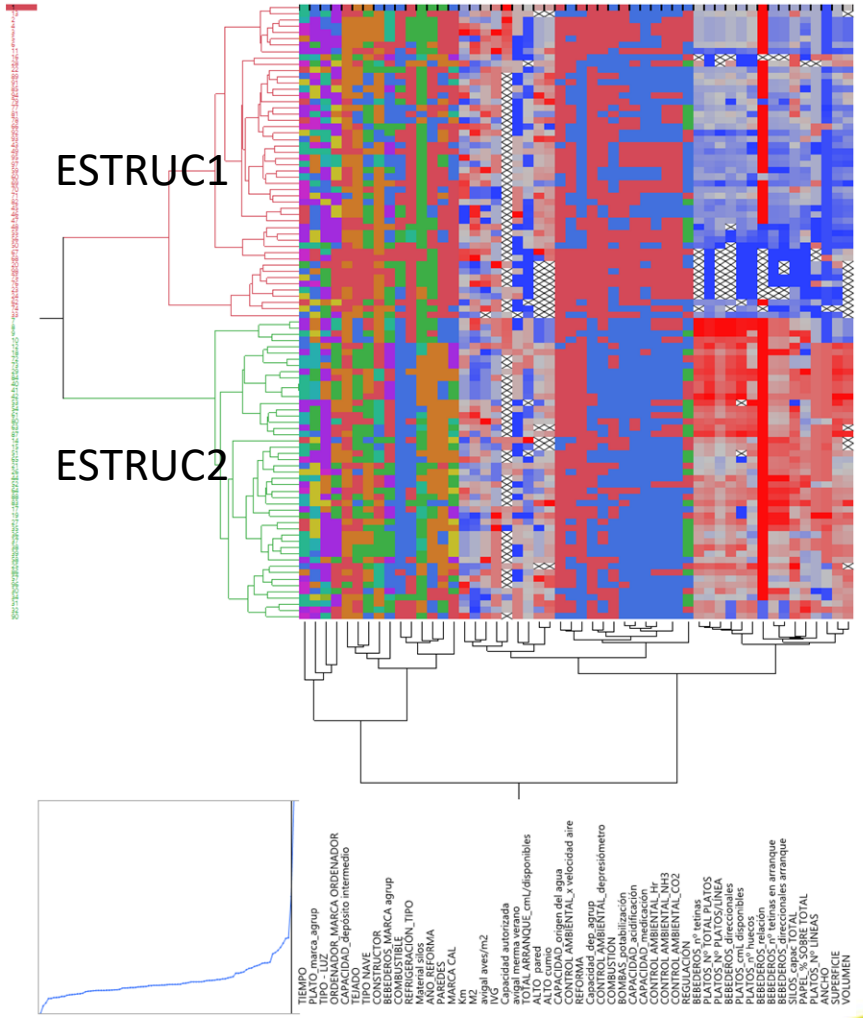
PARTNERS IN PROGRESS

# Clustering farms according to structural variables

**Resumen de columna**

| Columna | R cuadrado | .2 .4 .6 .8 |
|---|---|---|
| PLATOS_cmL disponibles | 0.4466 | |
| PAPEL_% SOBRE TOTAL | 0.4422 | |
| VOLUMEN | 0.4395 | |
| PLATOS_nº huecos | 0.3951 | |
| PAREDES | 0.3784 | |
| AÑO_REFORMA | 0.3613 | |
| Capacidad_dep_agrup | 0.3215 | |
| CAPACIDAD_acidificación | 0.3176 | |
| PLATOS_Nº LÍNEAS | 0.3116 | |
| SILOS_capac TOTAL | 0.2901 | |
| CONTROL AMBIENTAL_NH3 | 0.2801 | |
| CONTROL AMBIENTAL_CO2 | 0.2801 | |
| COMBUSTIBLE | 0.2545 | |
| BOMBAS_potabilización | 0.2521 | |
| MARCA CAL | 0.2349 | |
| Material silos | 0.2112 | |
| REFRIGERACIÓN_TIPO | 0.1958 | |
| CONTROL AMBIENTAL_depresiómetro | 0.1935 | |
| REFORMA | 0.1794 | |
| CAPACIDAD_medicación | 0.1691 | |
| CONTROL AMBIENTAL_Hr | 0.1600 | |
| TEJADO | 0.1348 | |
| BEBEDEROS_relación | 0.1077 | |
| TOTAL ARRANQUE_cmL/disponibles | 0.0918 | |
| CAPACIDAD_depósito intermedio | 0.0845 | |
| ALTO cumio | 0.0649 | |
| CONSTRUCTOR | 0.0632 | |
| M2 | 0.0593 | |
| avigal aves/m2 | 0.0563 | |
| avigal merma verano | 0.0485 | |
| TIPO - LUZ | 0.0466 | |
| ORDENADOR_MARCA ORDENADOR | 0.0402 | |
| REGULACIÓN | 0.0344 | |
| PLATO_marca_agrup | 0.0302 | |
| COMBUSTIÓN | 0.0297 | |
| ALTO pared | 0.0245 | |
| Km | 0.0220 | |
| TIEMPO | 0.0153 | |
| TIPO NAVE | 0.0149 | |
| CONTROL AMBIENTAL_x velocidad aire | 0.0094 | |
| IVG | 0.0006 | |
| Capacidad autorizada | 0.0002 | |
| CAPACIDAD_origen del agua | 0.0001 | |
| BEBEDEROS_MARCA agrup | 0.0001 | |

Porción de la variación total de cada columna absorbida por el conglomerado



# Find EFFECTS that explain variability

## II. Batch characteristics

STRUCTURAL

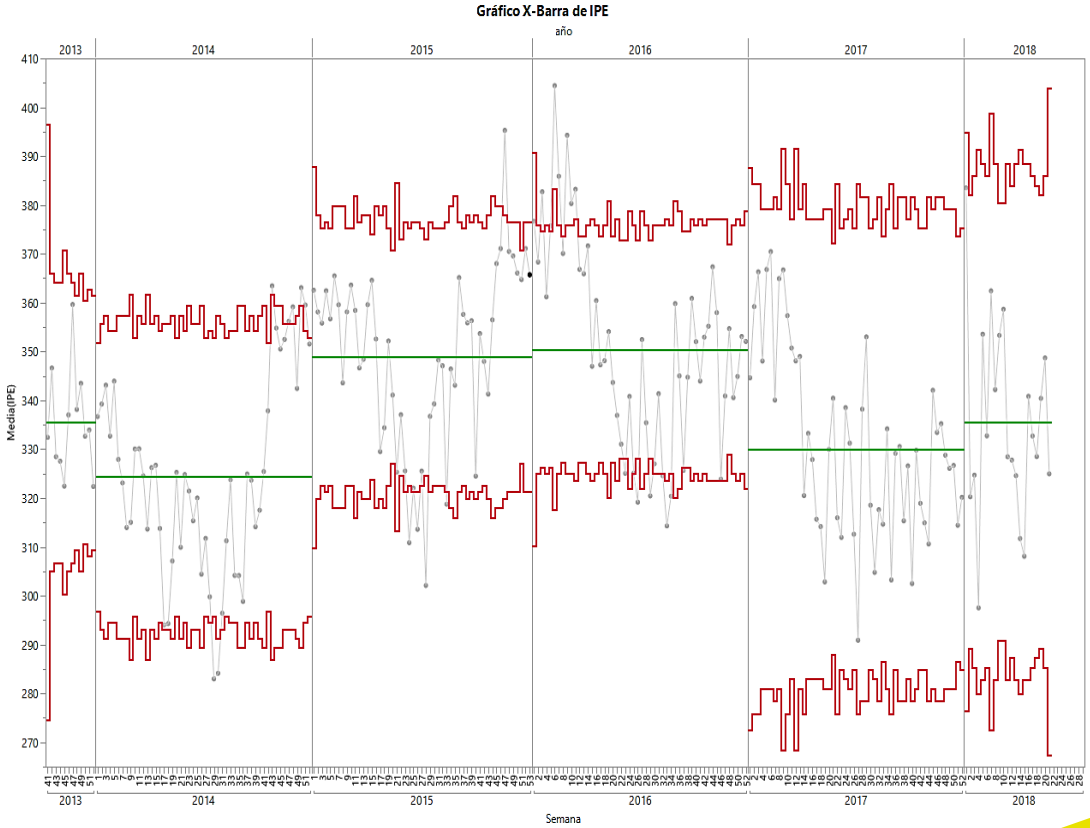
98 farms

2 GROUPS (ESTRUCT1, ESTRUC2)

FIELD DATA (BATCHES)

Explore the variability →

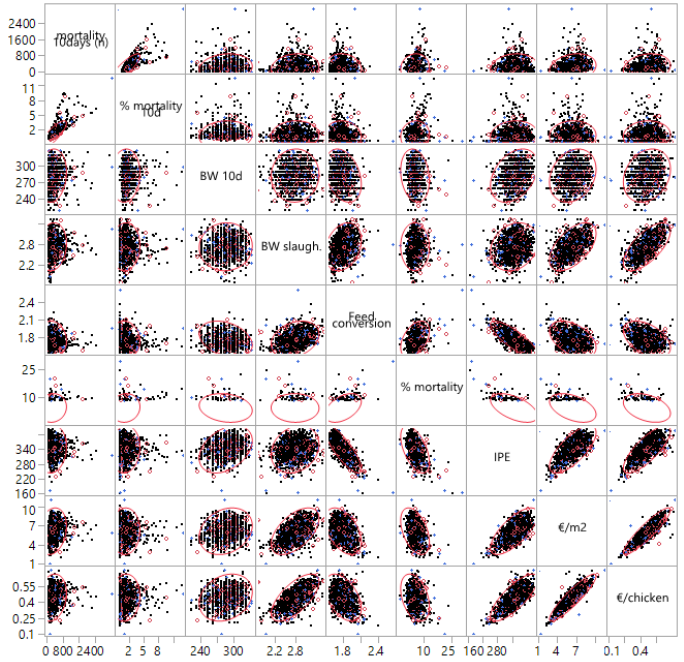
**TIME SERIES ANALYSIS**




PARTNERS IN PROGRESS

# Summarize the response or performance of the system

| Performance variables | Average |
|---|---|
| mortality 10 days (n) | 309 |
| % mortality 10 days | 1 |
| BW 10 days | 285 |
| BW slaughter | 2.8 |
| Feed conversion | 1.8 |
| % mortality | 4.8 |
| EPEF | 343.6 |
| €/chicken | 0.5 |
| €/m² | 6.4 |
| ……. | |



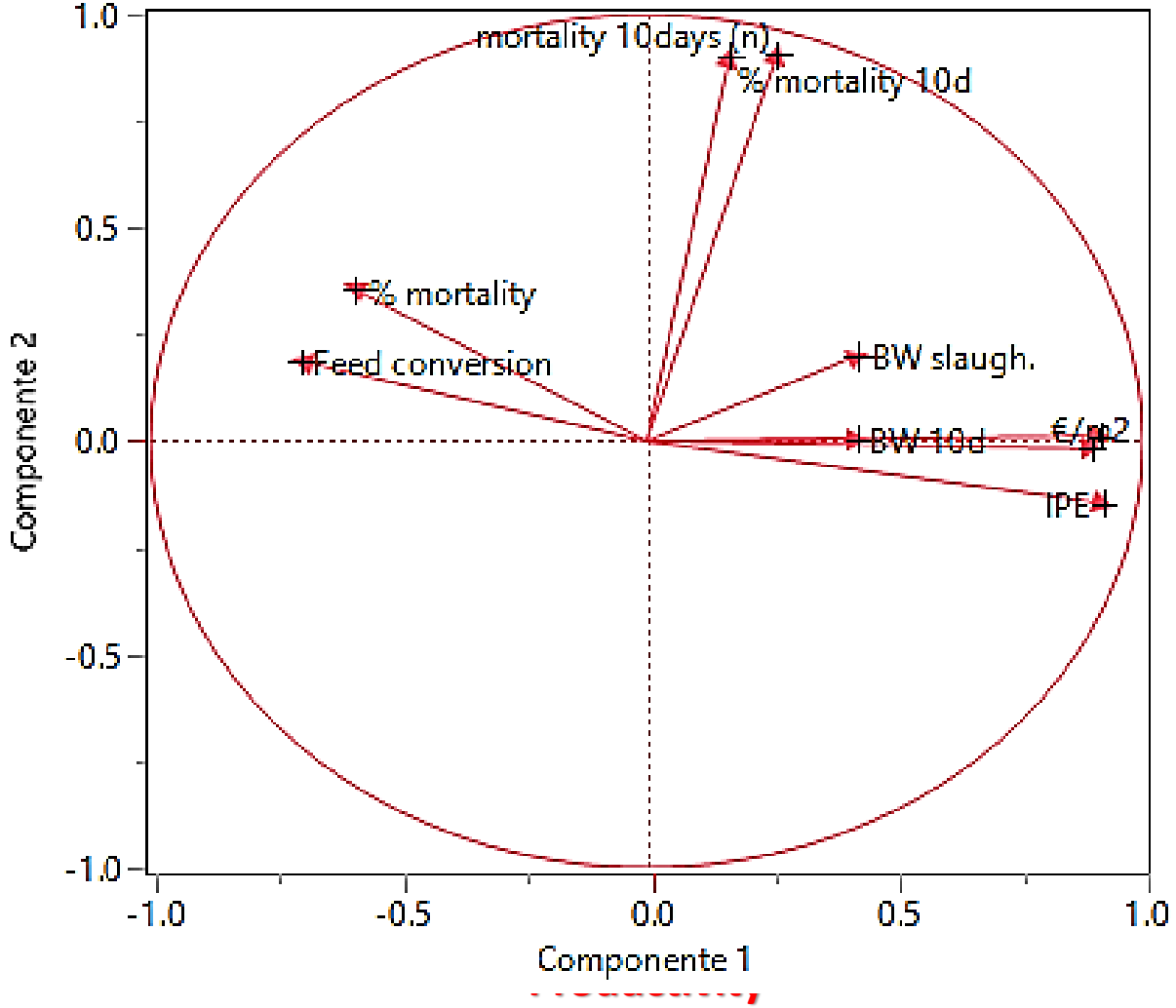

CORRELATED!!

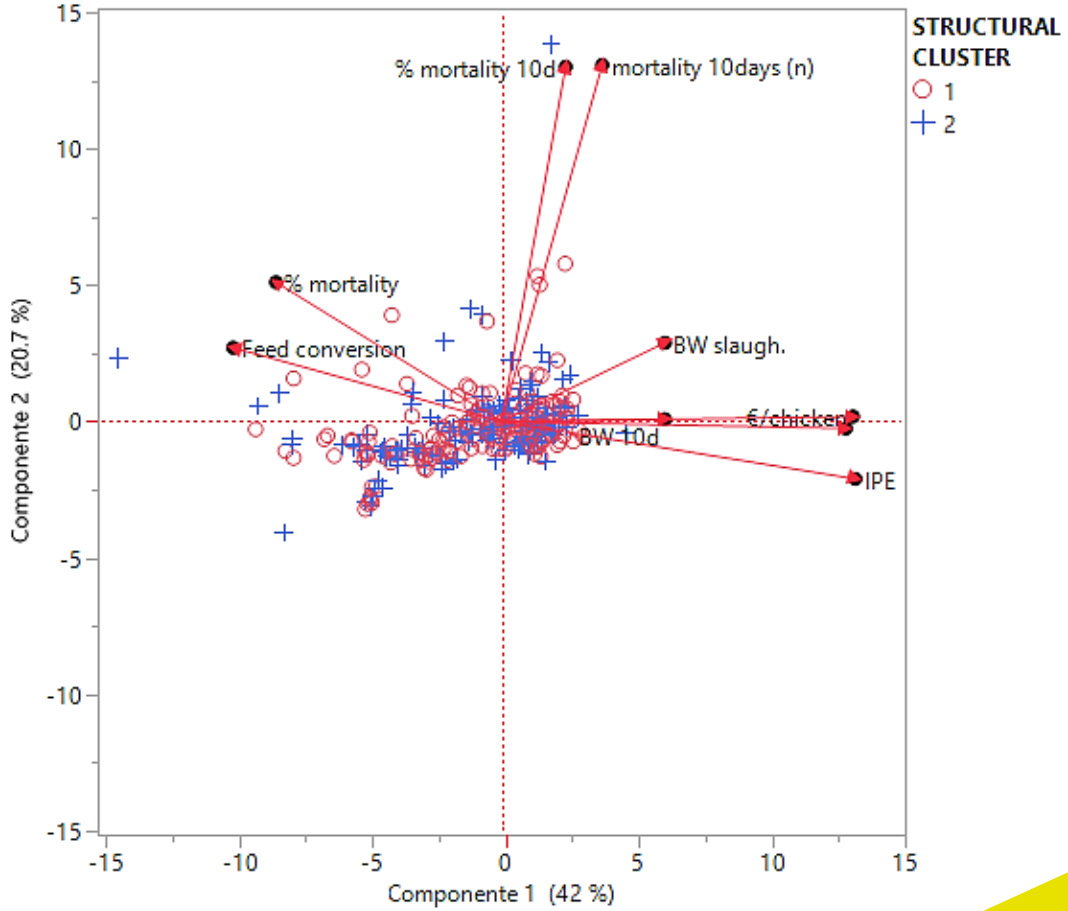
**PRINCIPAL COMPONENT ANALYSIS**

PARTNERS IN PROGRESS

# New variables that summarize the output

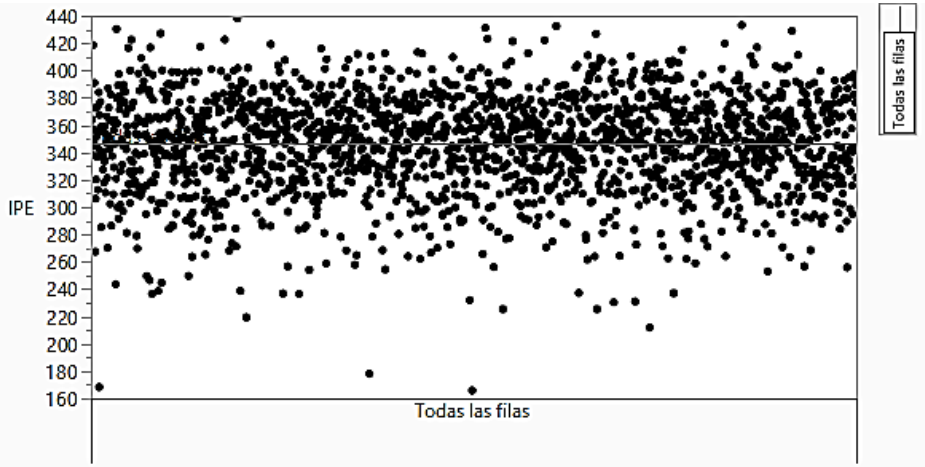
Mortality




PARTNERS IN PROGRESS

# New statistical approaches of *machine learning* : decision trees




Dividir | Podar

| R cuadrado | RMSE | N | Número de divisiones | AICc |
|---|---|---|---|---|
| 0.000 | . | 1945 | 0 | 0 |

**Todas las filas**

| | |
|---|---|
| Conteo | 1945 |
| Media | 346.29964 |
| Desviación estándar | 34.929704 |

**Candidatos**

| Término | SS del candidato | | Log Utilidad | Punto de corte |
|---|---|---|---|---|
| Origin | 123793.5456 | | 25.70548964 | ORIGIN4 |
| color | 109444.0874 | | 22.06635930 | B |
| año | 209094.5932 | * | 47.40623721 | 2014 |
| week | 124808.2313 | | 28.93683789 | 14 |
| treatment | 16011.5572 | | 1.75077293 | Old Product |
| breed | 79728.3539 | | 16.04480872 | otros,cobb |
| STRUCTURAL CLUSTER | 3021.4722 | | 0.89257760 | 2 |

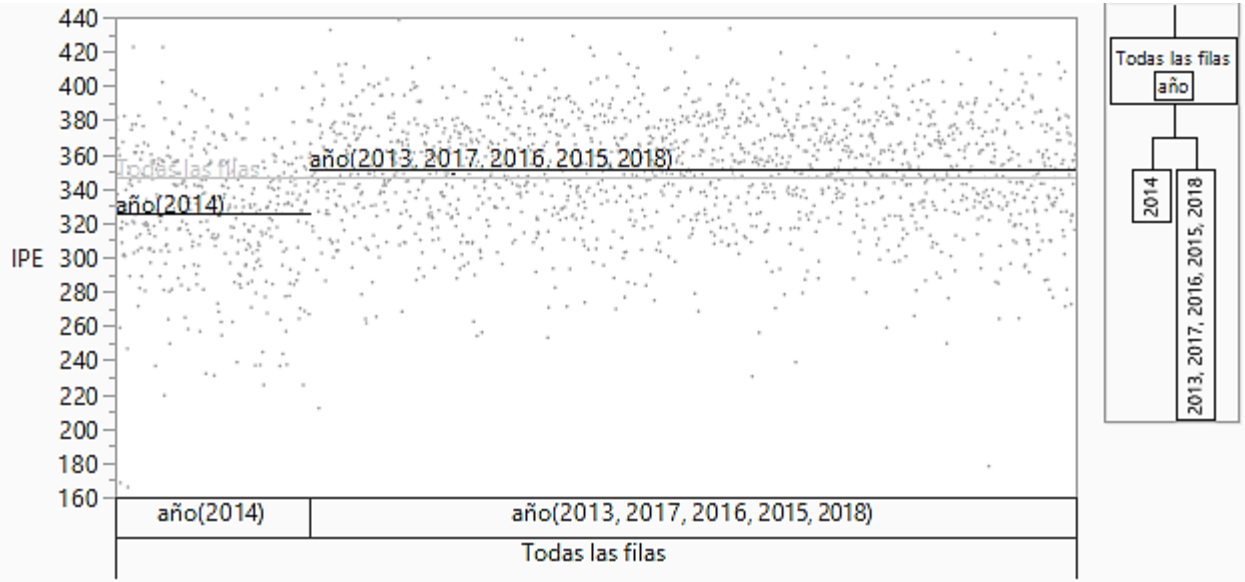
**Mean of EPEF: 346.29. Min 180, Max 440 !!!**

**Which factor explains better this variability??**

| Término |
|---|
| Origin |
| color |
| año |
| week |
| treatment |
| breed |
| STRUCTURAL CLUSTER |

PARTNERS IN PROGRESS

# Year




Dividir Podar

| R cuadrado | RMSE | N | Número de divisiones | AICc |
|---|---|---|---|---|
| 0.088 | 33.345964 | 1945 | 1 | 19167.7 |

**Todas las filas**

| | | Log Utilidad | Diferencia |
|---|---|---|---|
| Conteo | 1945 | | |
| Media | 346.29964 | 47.406237 | 25.822 |
| Desviación estándar | 34.929704 | | |

**año(2014)**

| | |
|---|---|
| Conteo | 393 |
| Media | 325.69518 |
| Desviación estándar | 37.754019 |

▷ Candidatos

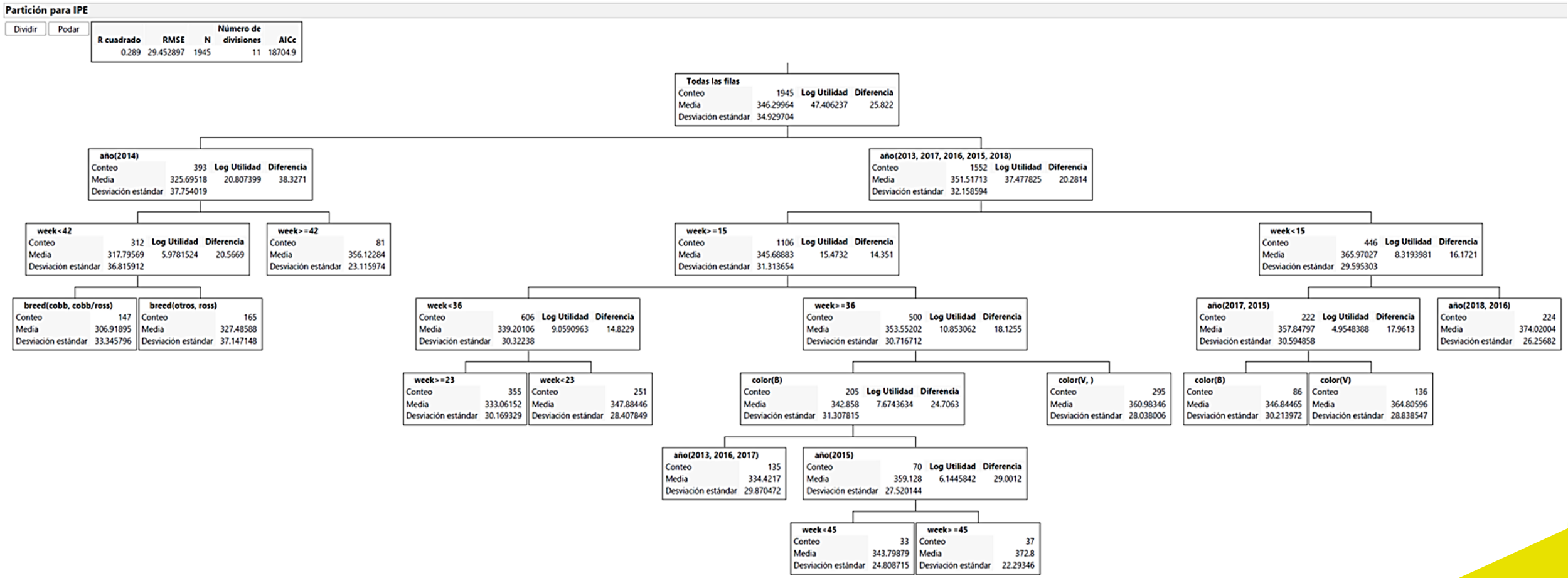
**año(2013, 2017, 2016, 2015, 2018)**

| | |
|---|---|
| Conteo | 1552 |
| Media | 351.51713 |
| Desviación estándar | 32.158594 |

▷ Candidatos

PARTNERS IN PROGRESS

# Complete decision/prediction tree



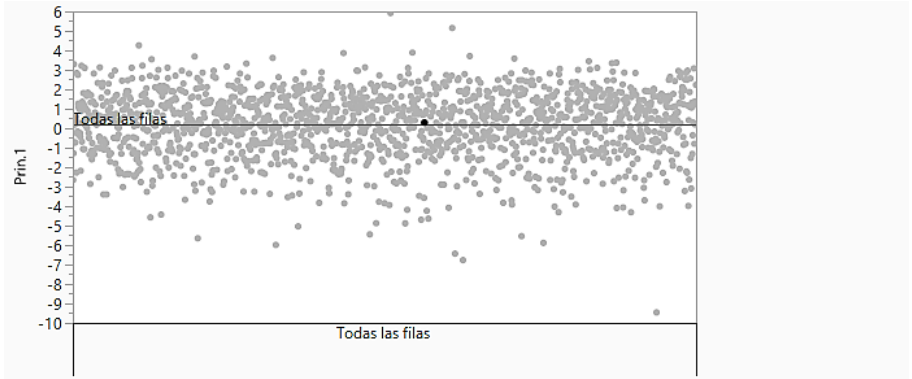

PARTNERS IN PROGRESS

# Even with lots of variables, this nonparametric methods works

## Which variables explain better this variability??

….. 50 variables that define the type of farm

Productivity component




Dividir | Podar

| R cuadrado | RMSE | N | Número de divisiones | AICc |
|---|---|---|---|---|
| 0.000 | . | 1565 | 0 | 0 |

**Todas las filas**

| | |
|---|---|
| Conteo | 1565 |
| Media | 0.1687788 |
| Desviación estándar | 1.6791648 |

**Candidatos**

| Término | SS del candidato | Log Utilidad | Punto de corte |
|---|---|---|---|
| CAPACIDAD_origen del agua | 26.976517 | 2.3504125 | traida,pozo |
| Capacidad_dep_agrup | 42.961345 | 3.7078507 | >10000 |
| CAPACIDAD_depósito intermedio | 54.436114 | 4.0235110 | 2000,no,,<10000 |
| BOMBAS_potabilización | 3.472105 | 0.3332100 | Si, |
| CAPACIDAD_acidificación | 25.384582 | 2.2152485 | si, |
| CAPACIDAD_medicación | 73.306570 | 6.2988258 | ,si |
| BEBEDEROS_MARCA | 83.334794 | 5.3641221 | lubbing plástico,Favega,tigsa,bigduchtman,corti |
| BEBEDEROS_MARCA agrup | 63.658238 | 4.1897043 | tigsa,bigduchtman,corti |
| BEBEDEROS_nº tetinas | 124.874663 | 11.1778992 | 636 |
| BEBEDEROS_direccionales | 59.761839 | 4.5925096 | 500 |
| BEBEDEROS_relación | 21.702771 | 1.4457033 | 25 |
| BEBEDEROS_nº tetinas en arranque | 184.772561 | 17.6066243 | 980 |
| BEBEDEROS_direccionales arranque | 178.969433 | 16.9717589 | 1250 |
| Id 3 | 97.543074 | 8.3489500 | 47 |
| CODIGO 3 | 134.766795 | 12.2186648 | 22728 |

PARTNERS IN PROGRESS

# Agenda

- What is the problem
  - Phases of one study
  - Study target
  - Variables
- Data
  - Collecting data
  - Recording data
  - Debugging data
- Data Analysis:
  - SPC
  - Time series
  - Clustering…
- Our proposal

# Statistical Process Control: to detect change

"Let's understand the variation, as this is the key to understanding and managing numerical chaos“


*Donald J. Wheeler*

*Understanding variation: The Key to Managing Chaos*

http://www.spcpress.com/

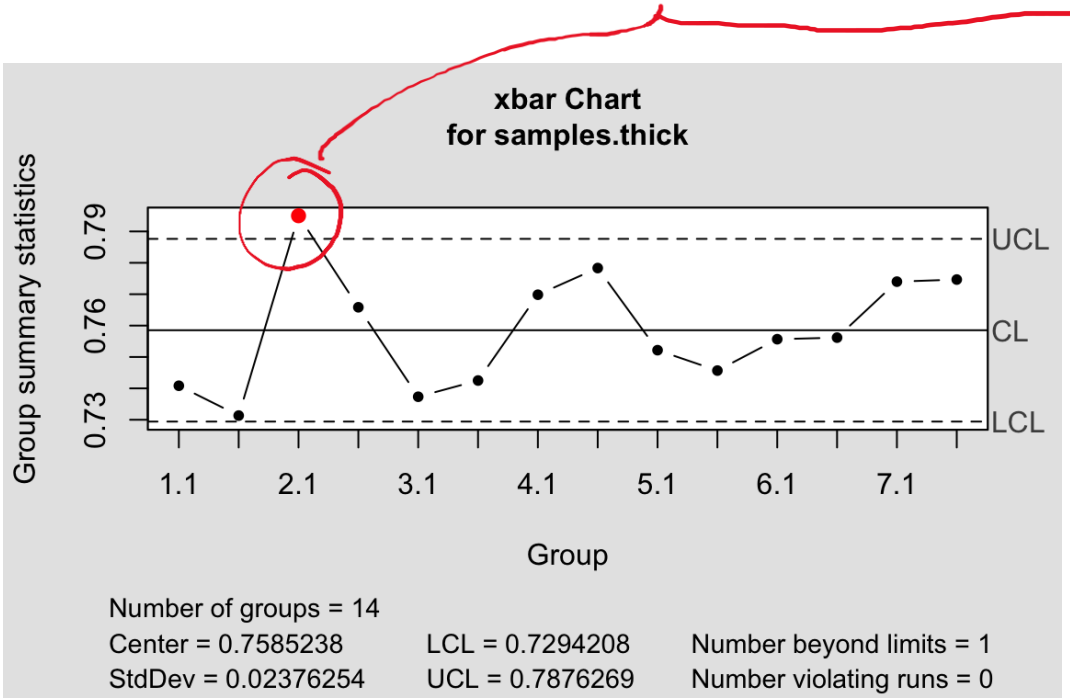
*Activities focused on the use of statistical techniques to reduce variation, increase knowledge about the process and steer the process in the desired way*

ISO 3534-2:2006

PARTNERS IN PROGRESS

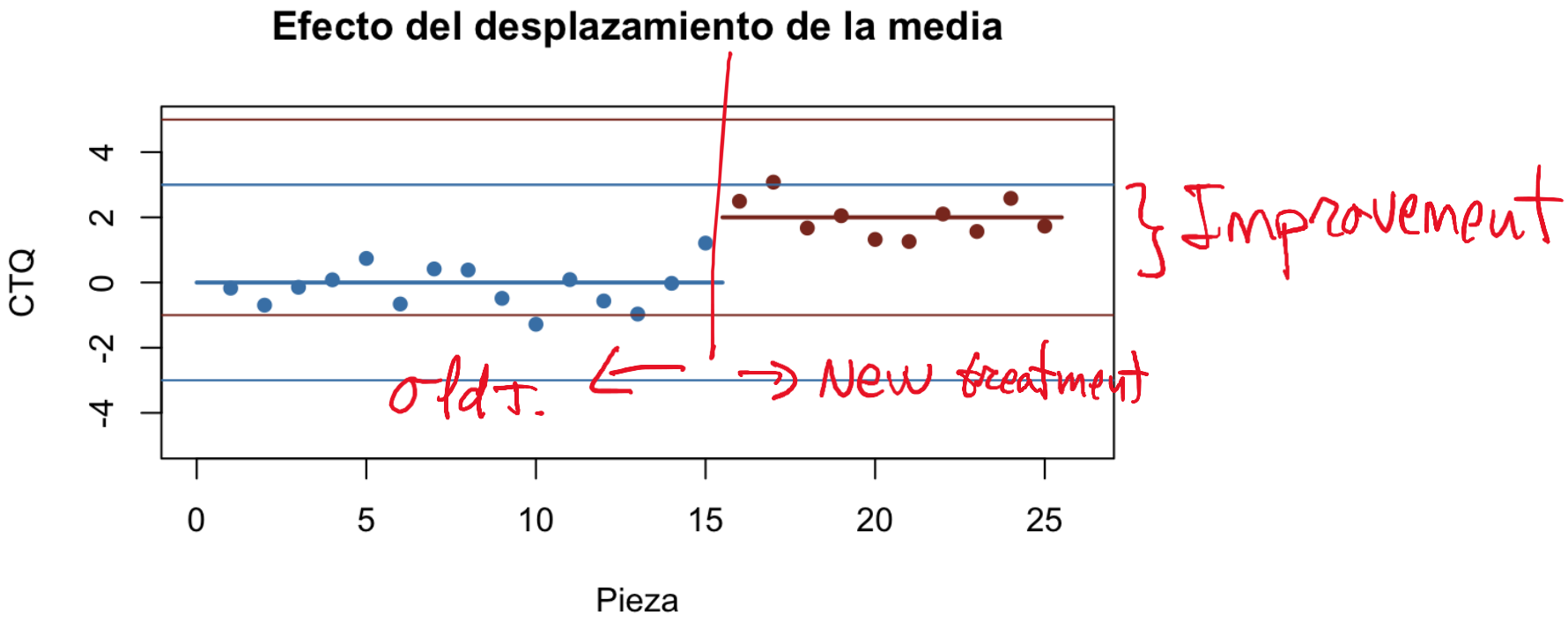
# Control charts

- Detect out-of-control situations
- Usually with the aim of removing special causes of variation

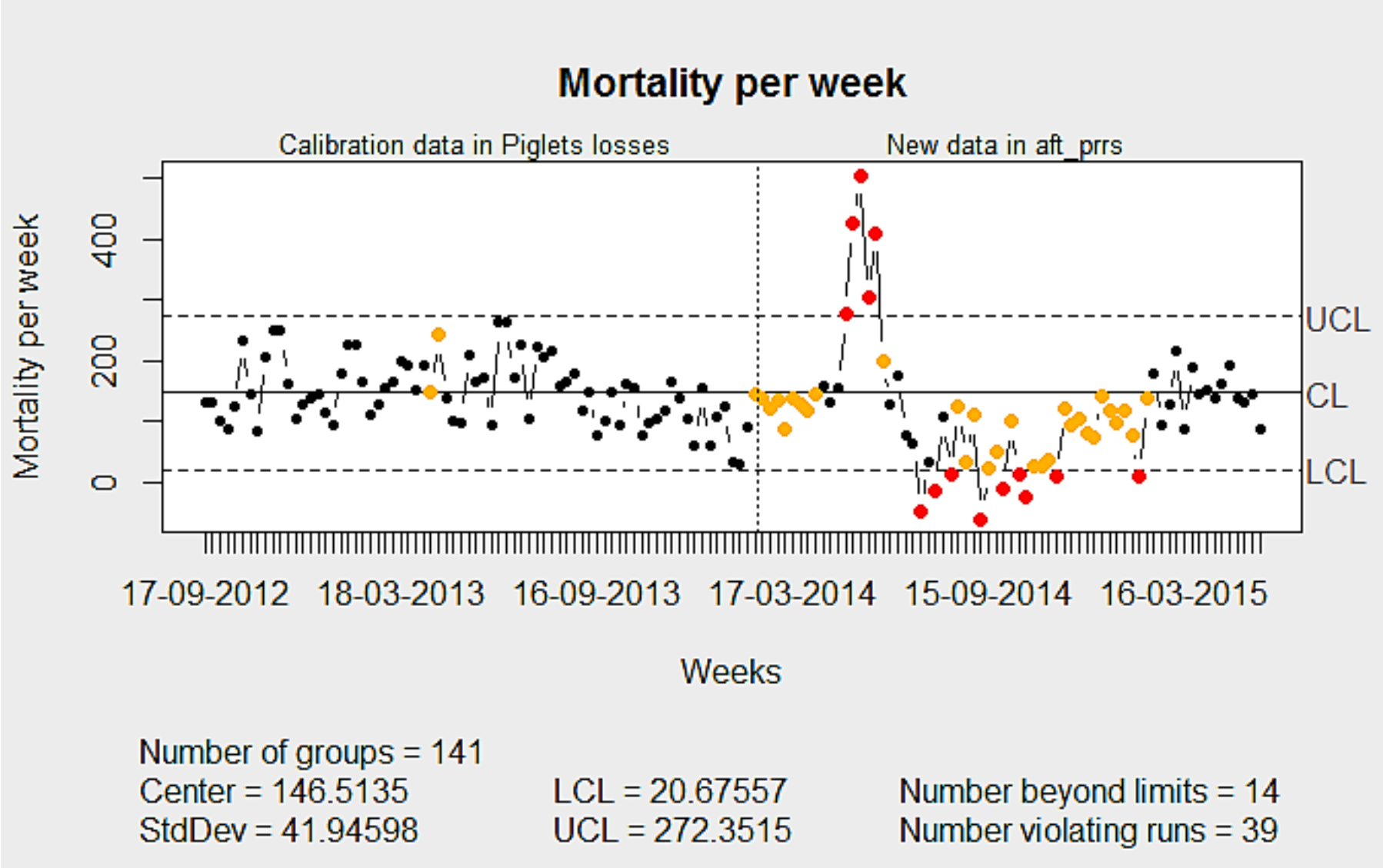


# Control charts to confirm effect

… But useful to detect if a new treatment or method improves performance

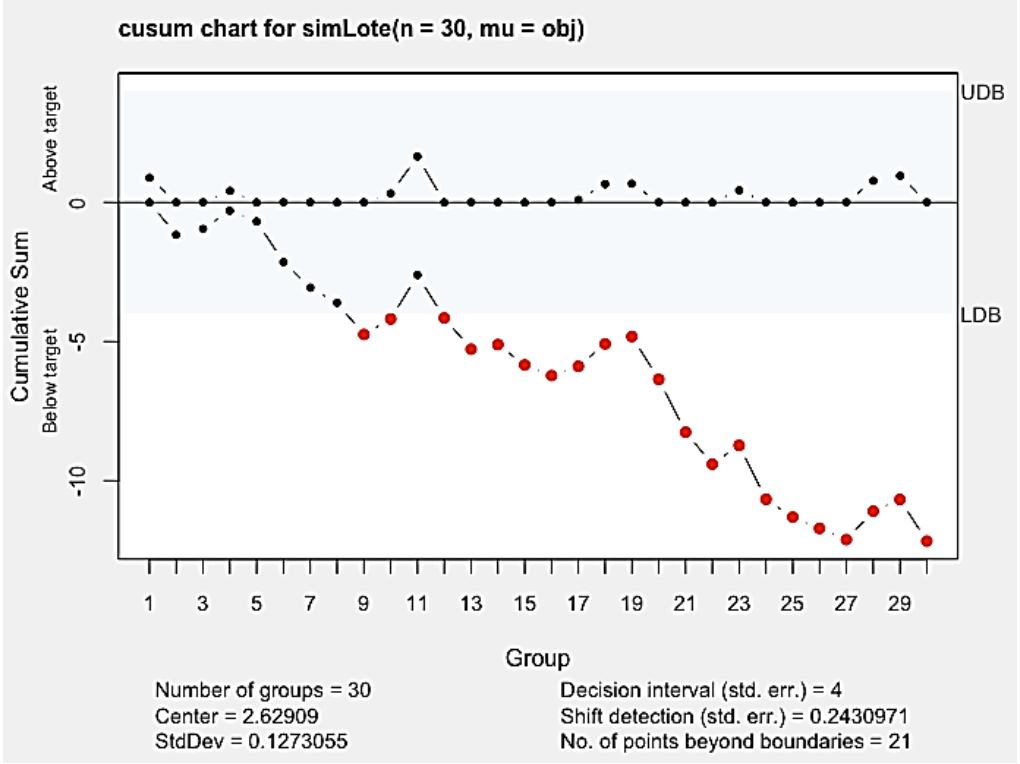


# Xbar.one/Studying data



# Advanced charts

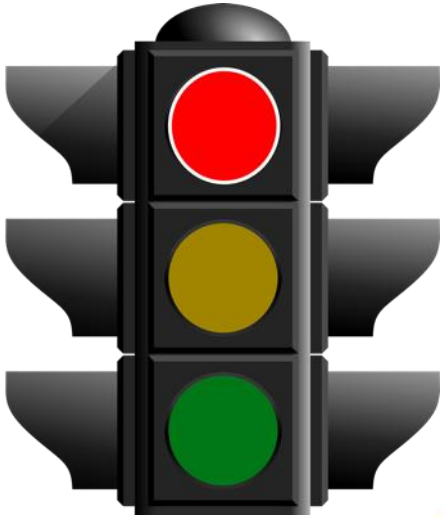
- Shewhart charts (x-bar, individuals) are powerful to detect significant shifts in a process
- But poor at detecting small changes (that could be economically important)
- CUSUM charts detect smaller changes quite fast



# Average Run Length (ARL)

- For a given control chart, the ARL is the number of samples (lot, batch, day, etc.) needed, on average, to detect a given shift in the mean.
- Based on the probability of Type II error.
- Easy to compute for x-bar charts through the standardized normal distribution

Example: the ARL to detect a shift equivalent to 1 standard deviation with an x-bar control chart is 4. For smaller shifts the CUSUM chart is more appropriate



# Simulation of  a “digital twin” farm

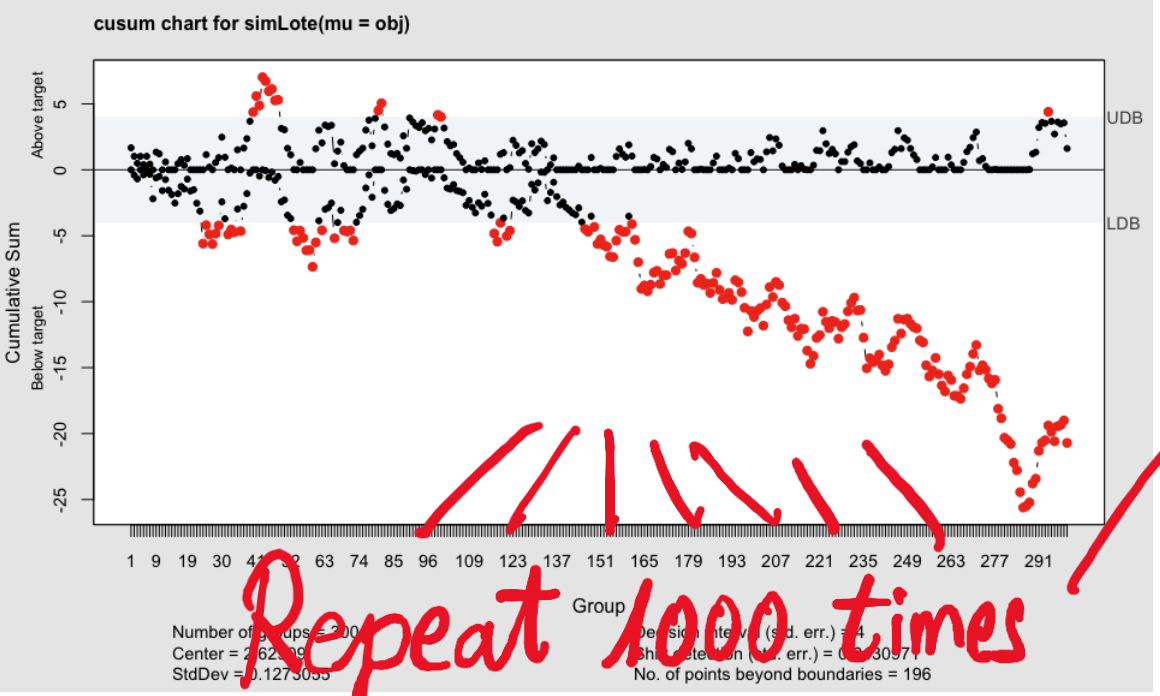
- For CUSUM charts, we can find tables in standards or even program numerical optimization to obtain ARLs
- A more empirical approach is to simulate several lots drawing random variates from the probability distribution of the actual farm

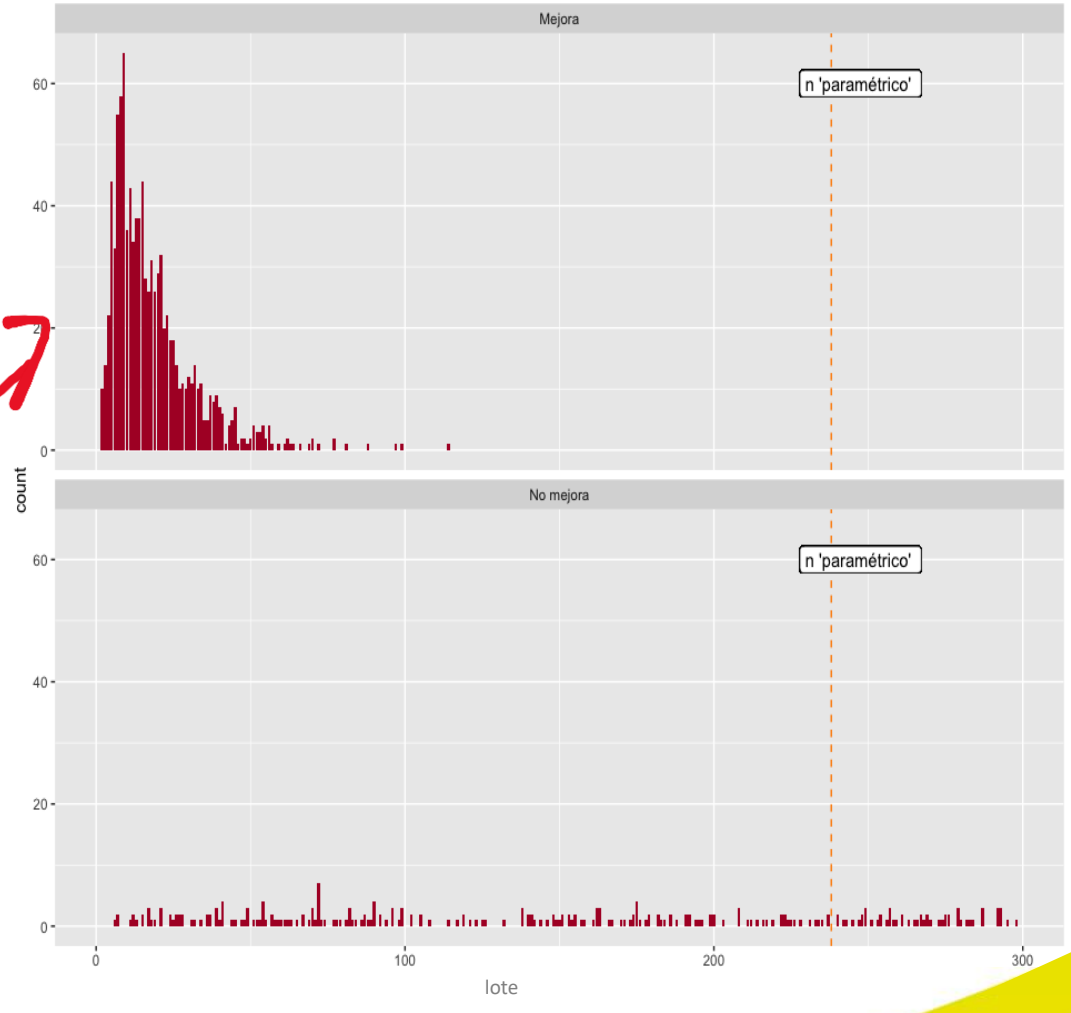
Hence, we find frequency distributions for the ARL and for the false alarms.



PARTNERS IN PROGRESS

# When it is likely to detect the desired change?



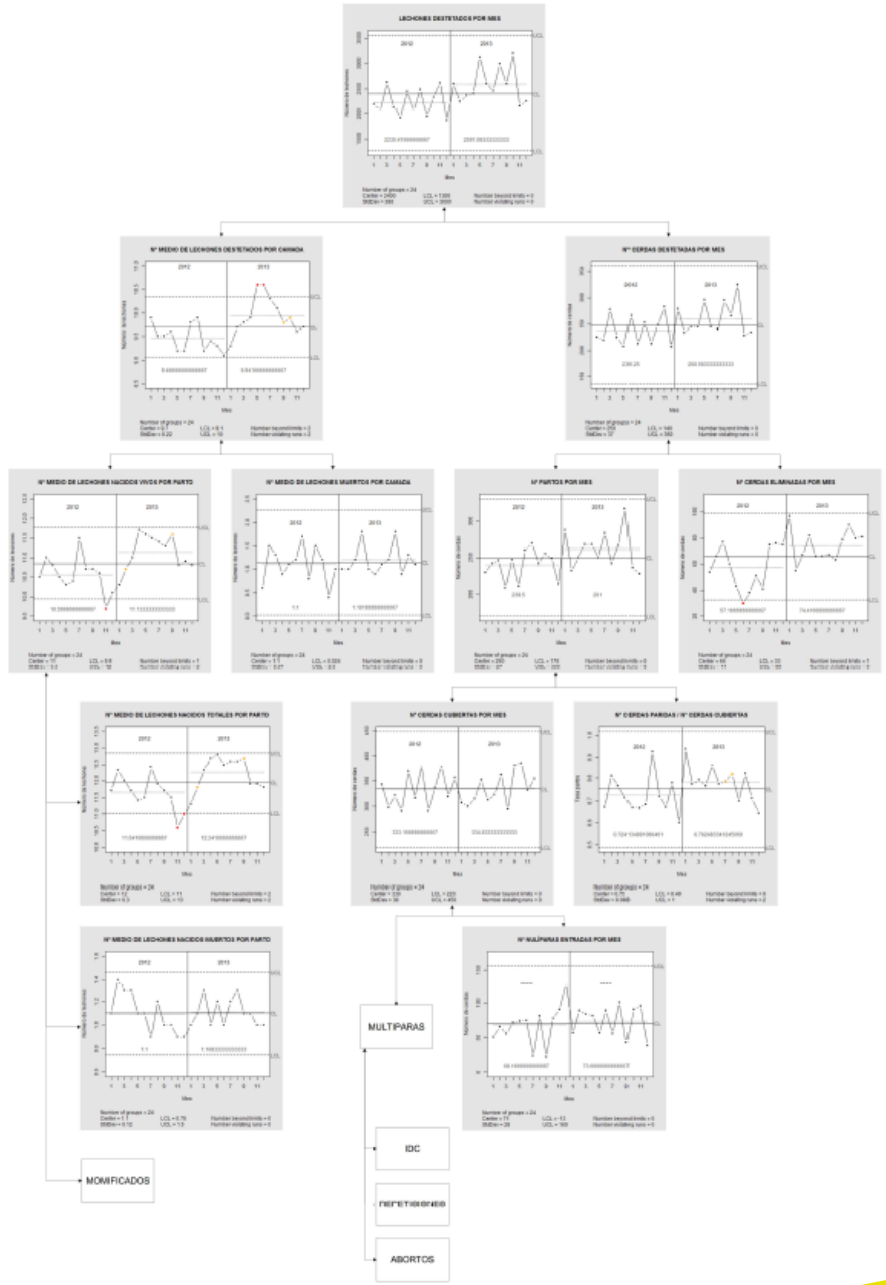



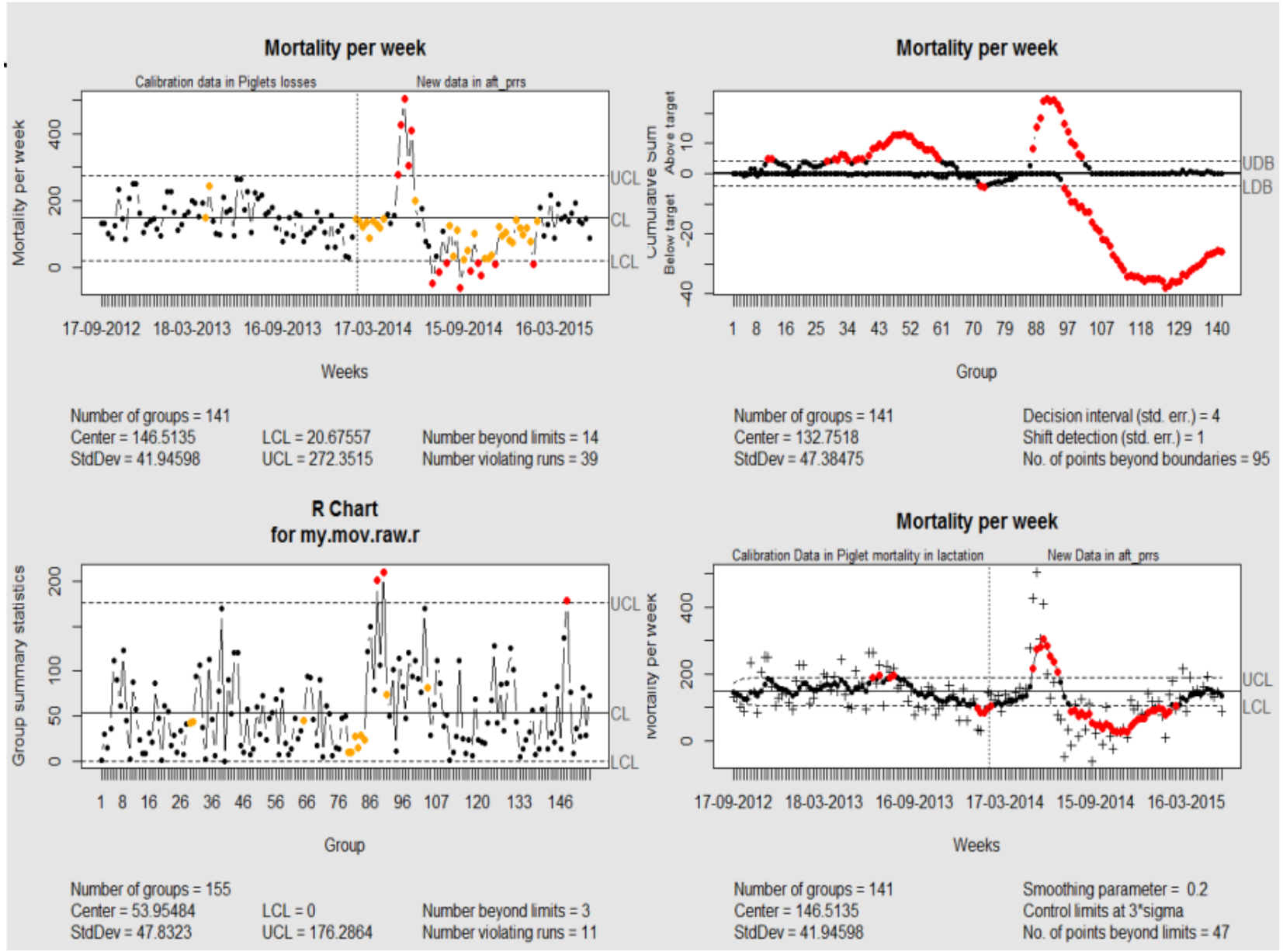

Well before the parametric test "n" estimation, with a low risk of false alarms

# When it is likely to detect the desired change?

- Sample size to detect a change
  - $H_0: \mu = 2.6290903;\ H_0: \mu < 2.6290903; Target, 2.60$; with $\alpha = 0.05$ and $\beta = 0.80$
  - 238 experimental units
- With ARL:
  - 19 experimental units' first signal
  - 45 experimental units in 95% of the simulations
  - 114 experimental units as maximum

TESTS AND TRIALS S.L.
DIAGRAMA PRODUCTIVIDAD 2012-2013








PARTNERS IN PROGRESS

# Agenda

- What is the problem
  - Phases of one study
  - Study target
  - Variables
- Data
  - Collecting data
  - Recording data
  - Debugging data
- Data Analysis:
  - SPC
  - Time series
  - Clustering…
- Our proposal

# Proposal for conducting field trials and product valuation

1. Development of a hypothesis
2. Bibliographical study
3. Protocol development
4. Team training
5. Analysis of customer past performances
6. Choice of farm population and verification of their uniformity
7. Conducting tests
8. Verification
9. SPC and Statistical study and conclusions

# Conclusions

- The money that an animal production company generates is too big
- Business decisions need experiments and studies
- A correct design avoids “outfishing"
- Statistical methods are a tool, *“sine qua non”*
- Statistical knowledge & management is critical in animal production business
- Extensive and critical data analyses will improve your animal business
- Currently, Business Intelligence techniques are “sine qua non” to go from stable to “table”

# Síagro

- The tool you need to take decisions based on your own data:
  - https://www.siagro.es/en/home/

# Questions & Answers

# THANK YOU!

PARTNERS IN PROGRESS