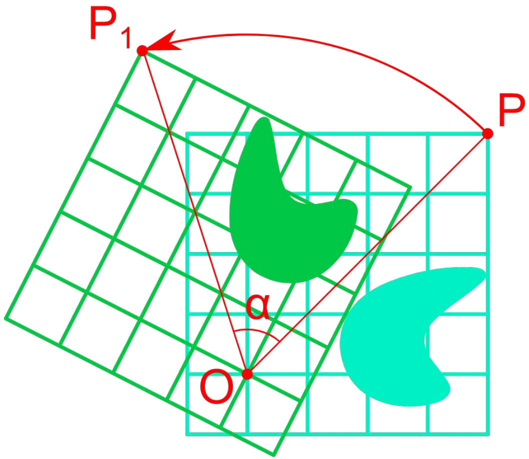
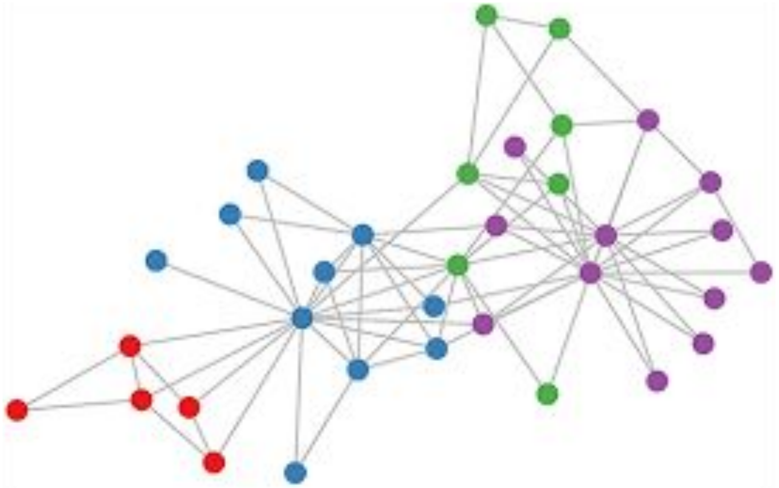




Proyecciones Multidimensionales

una mirada hacia el descubrimiento de patrones y análisis de datos en alta dimensión

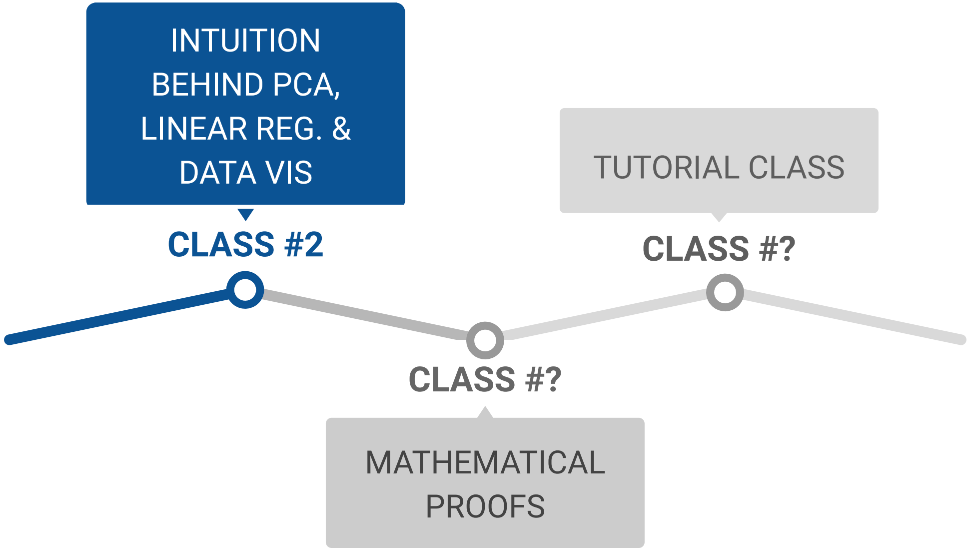
SECTION

2

Prof. Dr. Diego Nascimento

UNIVERSIDAD DE ATACAMA 1857

DPTO. MATEMÁTICA | FACULTAD DE INGENIERÍA



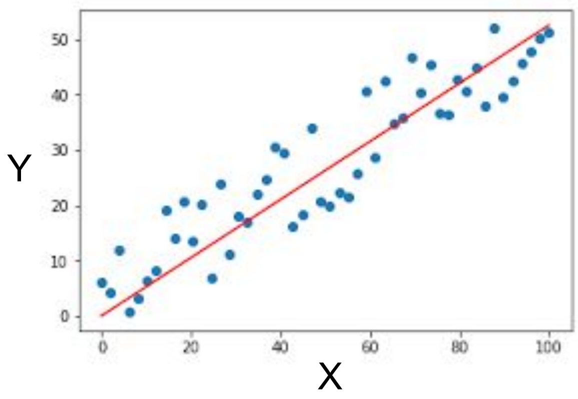
TODAY'S CLASS AGENDA

(INTRO CLASS)

- LINEAR STATISTICAL MODEL
- DATA TRANSFORMATION/ROTATION
  - PCA
  - ICA
  - FA
- REAL-WORLD EXEMPLIFICATION
  - RELIABILITY-CENTERED MAINTENANCE
- SOME REFERENCES

EXPLICABILITY OF THE PHENOMENON …VARIABLES ASSOCIATION

Usually one adopts the COVARIANCE or CORRELATION (pearson) to summarize the relationship between two events in just one number, that is,



$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\rho = \frac{COV(Y,X)}{\sqrt{VAR(Y)VAR(X)}}$$

$$-1 \leq \rho \leq 1$$

or, explain the dynamic association between them, LINEAR REGRESSION how the X unit changes the variable Y.

MULTIPLE EXPLANATORY VARIABLES

Extending the concept of linear relationship of two variables into multiple <u>INDEPENDENT</u> explanatory variables (Xs) that impact the response variable (Y).

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \text{erro}$$

Intercept

Angular Coefficient

Angular Coefficient

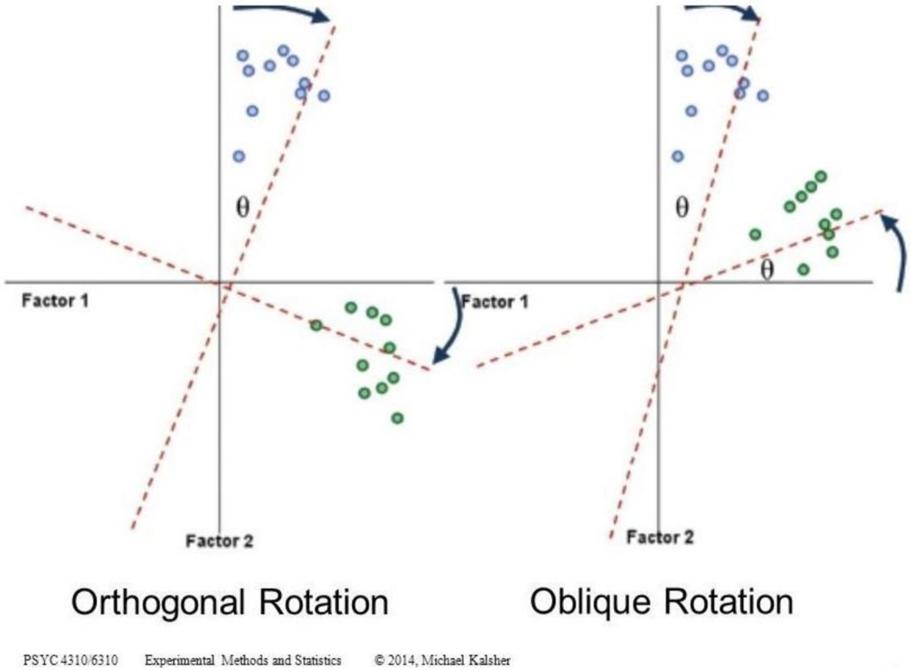
MATRIX FORM **Y** = **Xβ**

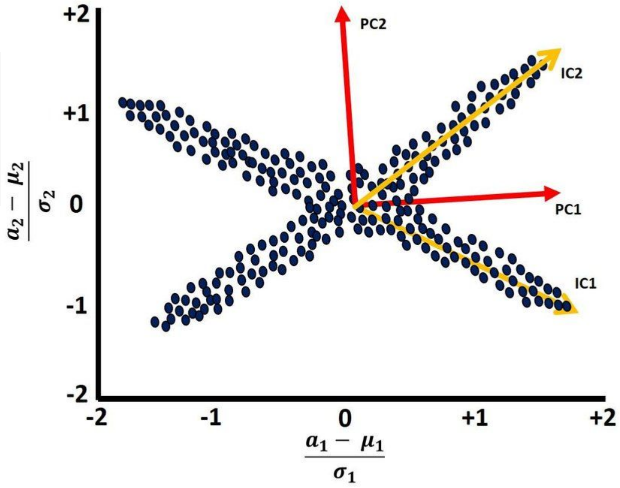
QUESTIONS TO BE ASKED…

- What if those explanatory variables are related?
- How could the combination of these relationships summarize such association?
- Could these summarizations describe some pattern?
- Or even, could some pattern be visually extracted (embedding in dimension less than 4)?

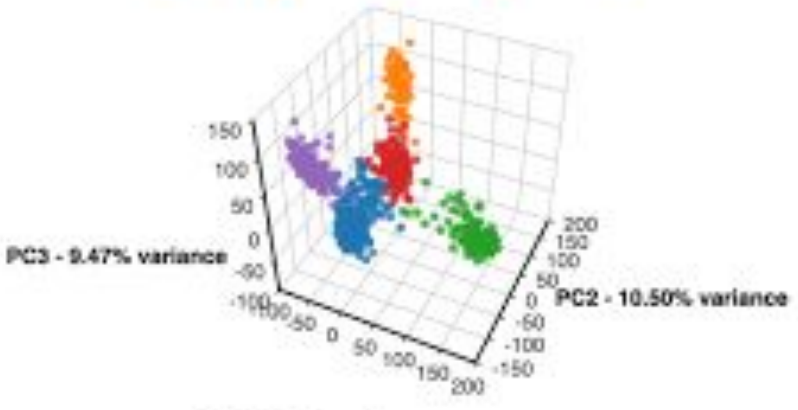
THREE HELPFUL ALGORITHMS

- FACTOR ANALYSIS (FA)





- INDEPENDENT COMPONENT ANALYSIS (ICA)
- PRINCIPAL COMPONENT ANALYSIS (PCA)

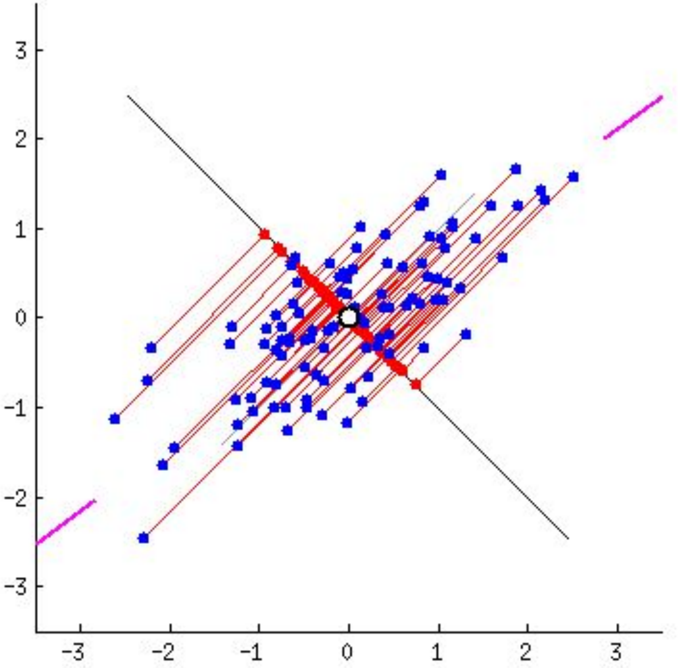


1) LET'S COMBINE RELATED VARIABLES

One projection is the **space rotation** that guarantees the best linear combination which **maximizes the explainability** (variability explanation), equivalent to the eigenvalues and eigenvectors decomposition.

- Reasoning the source of variations in data
- Understand pairwise correlation between attributes of data
- Reduce dimensions with little 'distortion'
- Low dimensional data visualization

PRINCIPAL COMPONENT ANALYSIS (PCA)



MATHEMATICAL RESTRICTIONs & THEIR IMPLICATIONS

What PCA does is decomposition of the total explanation (variability) into “best vectors” for projections, so-called PRINCIPAL COMPONENTS.

- Guarantee they are order by importance;
- Orthogonal from each other (Statistically = Independence).

Direction with Min Reconstruction Error := Direction with Max Variance

OBS: In PCA, observations are considered to be independent (maybe a strong supposition!).

OBS2: Since all calculations are based on the decomposition of the variance-covariance matrix, the best rotation obtained from PCA is only guaranteed for continuous variables!

OBS3: *PCA reasoning is to find the spectral decomposition (eigensystem) of the Covariance matrix originated from the normalized Xs variables.

LINEAR COMBINATION ACROSS VARIABLES (PCA)

The Principal Components (PCs) are a linear combination of all variables, ranked by importance decreasily and obtained based on the total variance explanation decomposition.

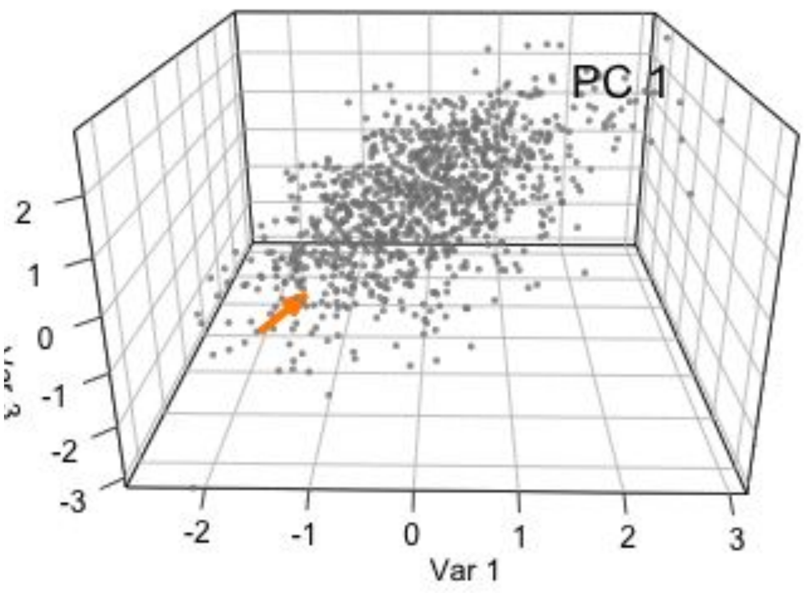
$T : \mathbb{R}^p \longrightarrow \mathbb{R}^p$

$Ax=\lambda x$

$\det(A-\lambda I)=0$

$$\begin{cases} X_1^* = a_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ X_2^* = a_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ \vdots \\ X_p^* = a_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p, \end{cases}$$

Where $a'_1=[a_{11}, a_{12}, \ldots, a_{1p}]$ is a line vector (1 x p) and $\mathbf{X}=[X_1, X_2, \ldots, X_p]'$ is a column vector (p x 1), resulting in a scalar number $X^*_1\ldots$ then, A is a matrix [a1,...,ap] and λ is a diagonal matrix.



$\mathbb{R}^3$ rotation

STEP 1) Position the origin of the system based on the average of each dimension

STEP 2) Calculate the direction of the largest variation (PC 1).

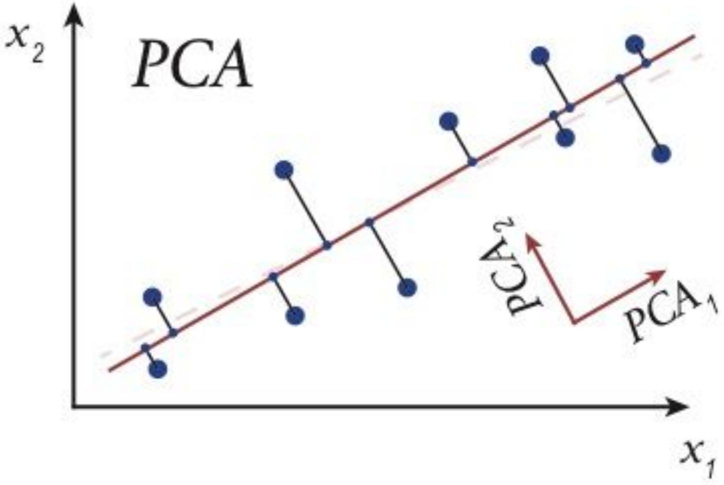
STEP 3) Orthogonal to the dimension obtained in STEP 2, a largest variation is calculated (PC 2).

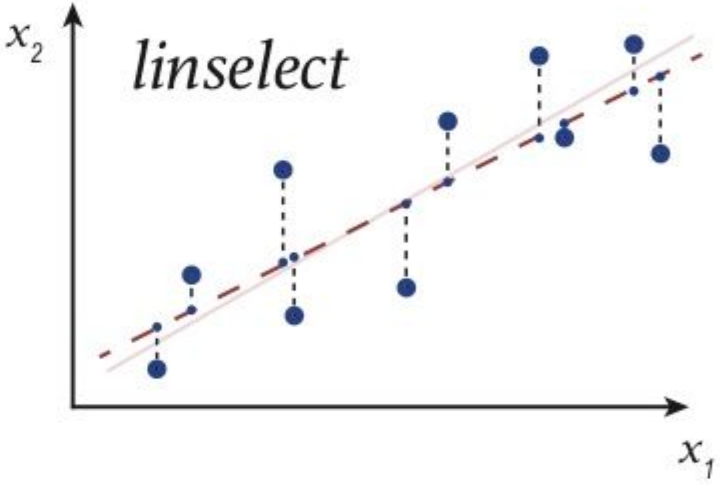
STEP 4) Orthogonal to the dimensions obtained previously, a new direction of the last dimension is obtained (PC 3).

$$PC_1 = w1\ X_1 + w2\ X_2$$

$$PC_2 = w3\ X_1 + w4\ X_2$$

$$X_2 \sim X_1$$





In Python

```
> from sklearn.decomposition import PCA
> model = PCA(n_components=p)
> principal_components = model.fit_transform(DATA)
```

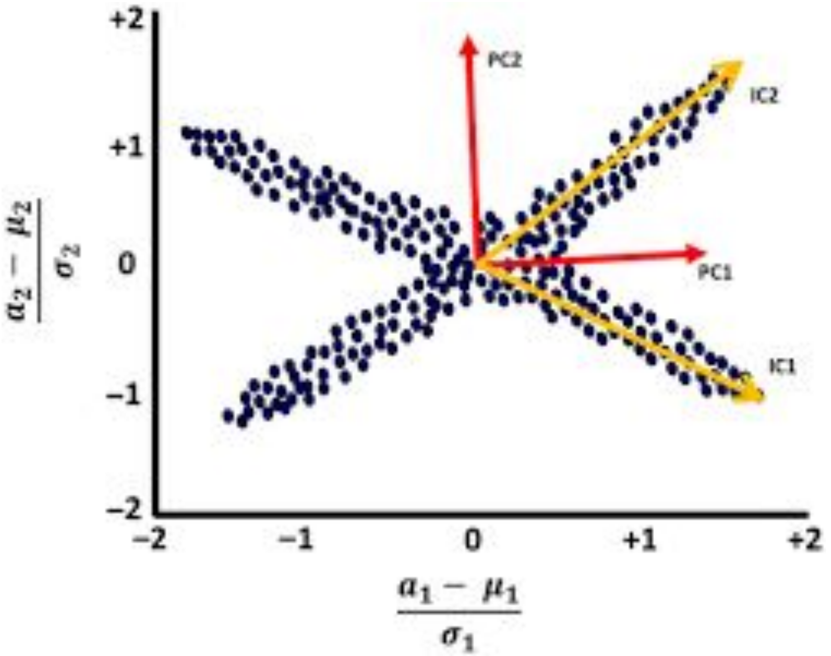
In R

```
> prcomp(DATA, scale = FALSE)   # or
> princomp(DATA, cor = FALSE, scores = TRUE)
```

2) LET'S COMBINE DIFFERENTLY CHARACTERISTICS

Second projection another **space rotation** that guarantees the best linear combination which **maximizes the independence across variables** (separates information) through the MUTUAL INFORMATION metric.

- Mutual Information across the created components are ZERO, $I(Y_i, Y_j) = 0$, that is Statistical Independent.
- Mutual Information across the created components and the original features are MAXIMUM.

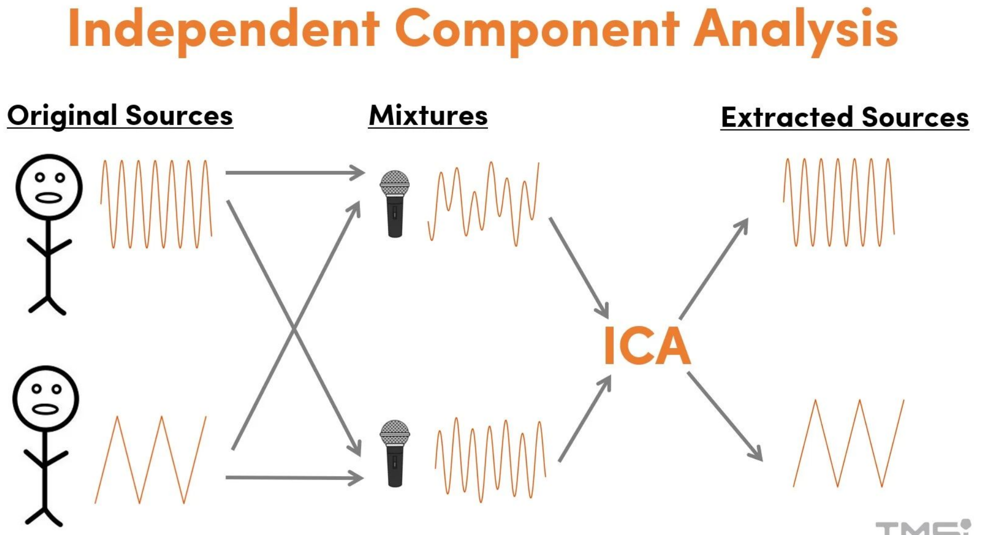


Comon, P. (1994). Independent component analysis, a new concept?. *Signal processing*, *36*(3), 287-314.

COCKTAIL PARTY –EXAMPLE–

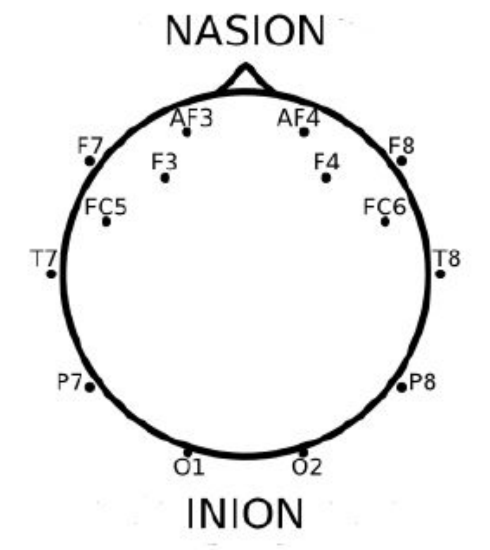
The scenario is based on a real-world situation where, in a noisy environment like a cocktail party, multiple conversations occur simultaneously. Despite the overlapping sounds, a person is able to focus on a single speaker while filtering out other voices and background noise.

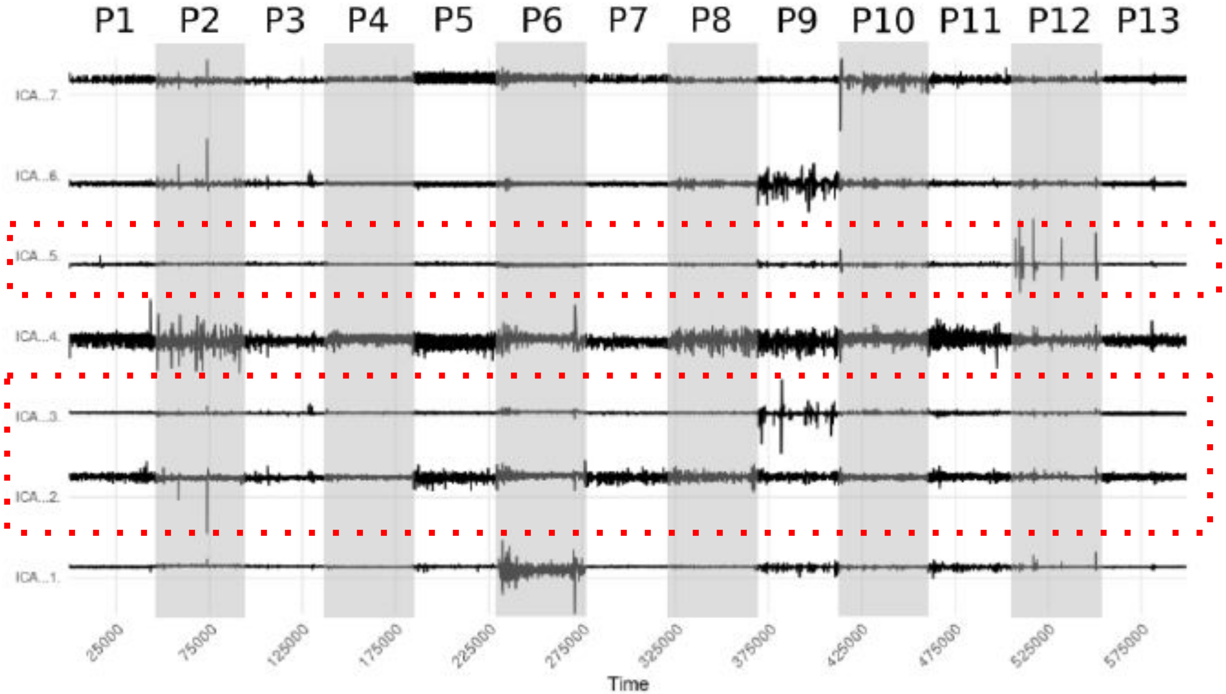
The microphones will capture different mixtures of the same voices. Using ICA, we can process these recordings and separate them back into the individual voices, allowing us to "isolate" specific conversations from the mixture.

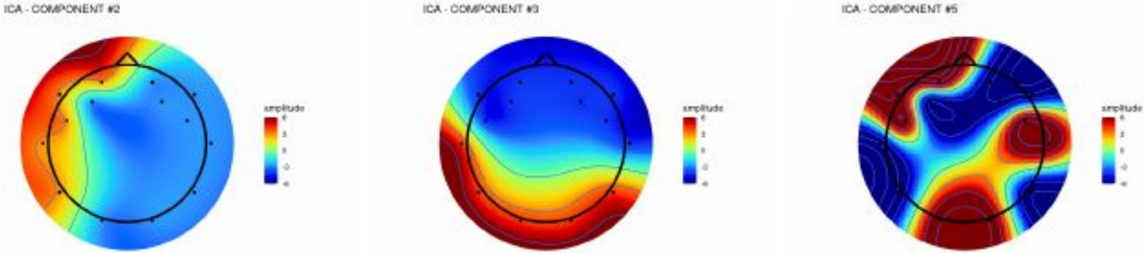


EEG ICA –Example–

$$T: R^{14} \rightarrow R^{7}$$



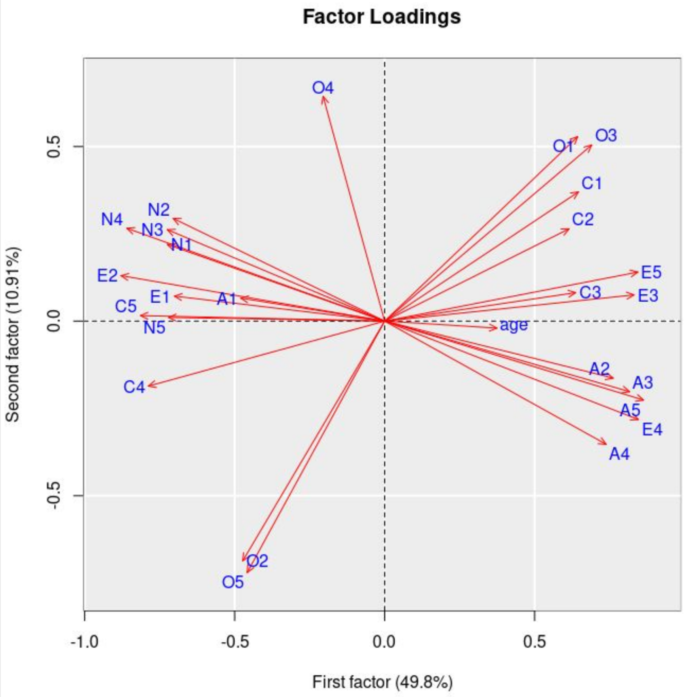




3) ANOTHER LOOK TO THE CHARACTERISTICS

Third projection is towards the **space rotation** that guarantees the best linear combination using **correlational structure on the observed variables** to maximizes the explicability.

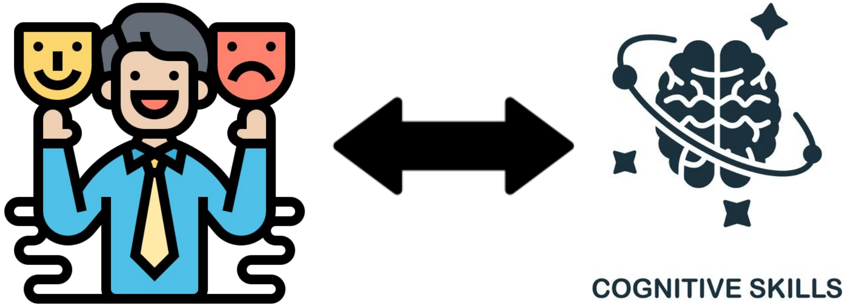
- Incorporates potentially less amount of features towards explaining (with less dimension) latent variables/factors.



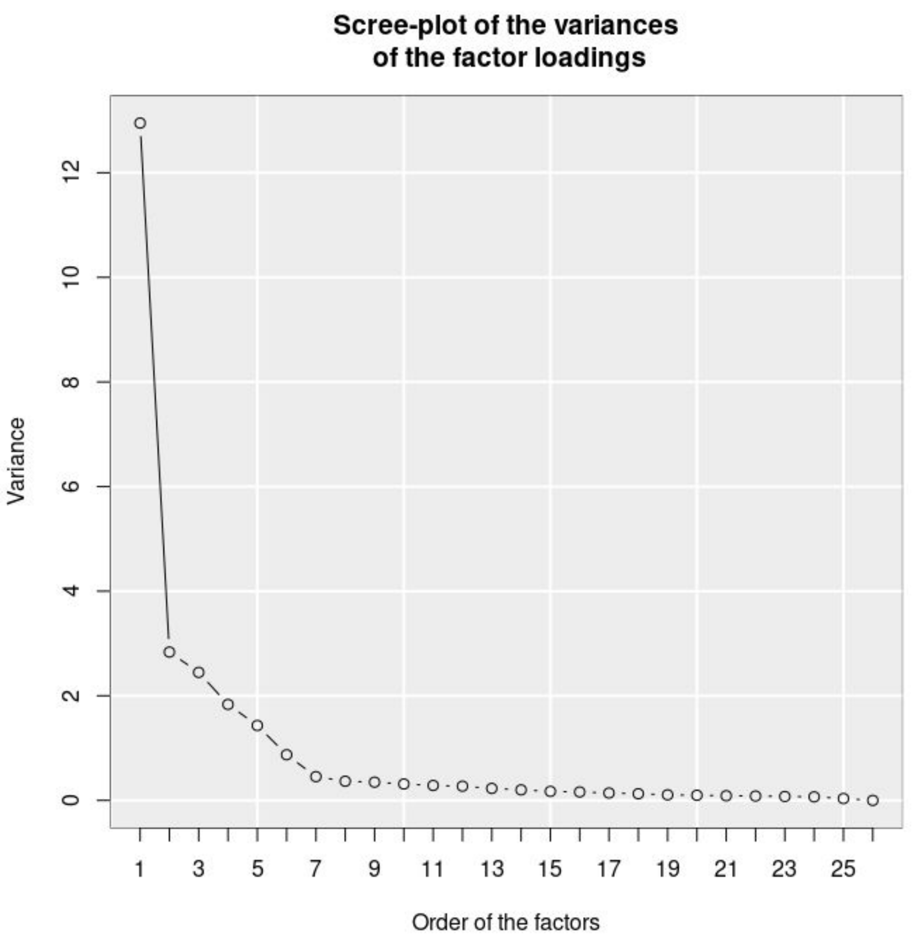
Comrey, A. L., & Lee, H. B. (2013). A first course in factor analysis. Psychology press.

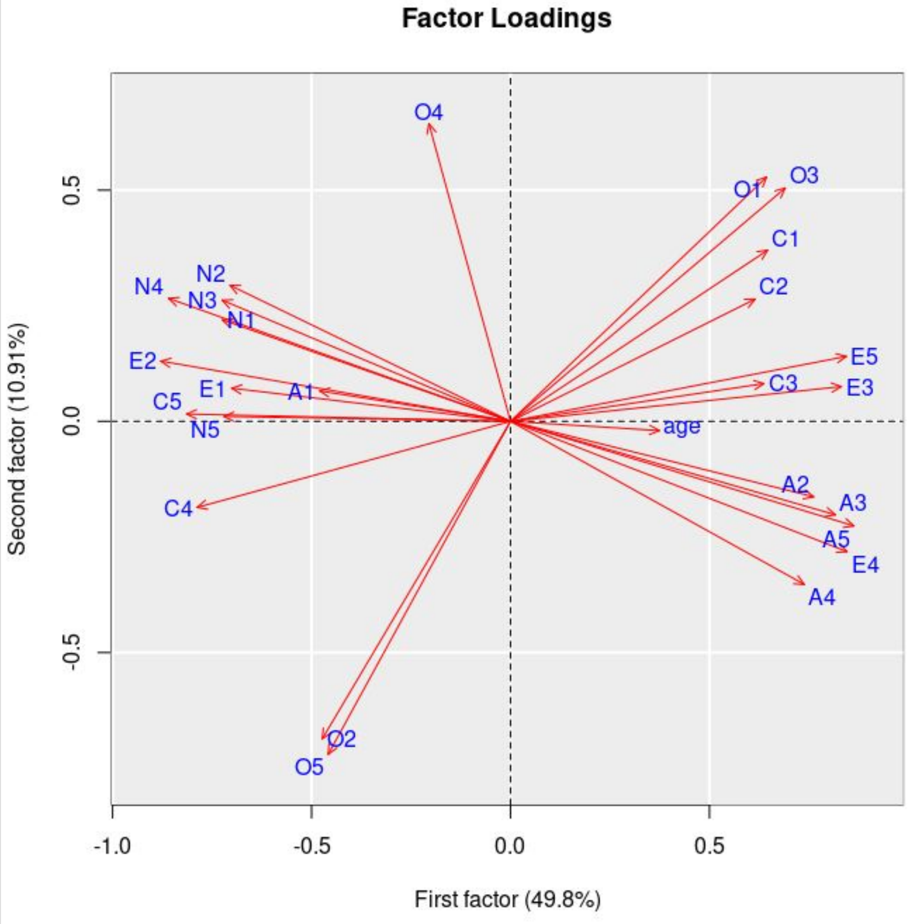
Psychological Test –EXAMPLE–

One experiment aimed to examine the relationship between Personality & Cognition. The dataset contains 2,800 observations and includes **28 variables** (gender, age, education, and 25 self-report personality items). The dataset that can be found in R under package *psych*, file name is *bfi*.



Goldberg, L.R. (1999) **A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models.** In Mervielde, I. and Deary, I. and De Fruyt, F. and Ostendorf, F. (eds) Personality psychology in Europe. 7. Tilburg University Press. Tilburg, The Netherlands.


Scree-plot of the variances
of the factor loadings
Variance
Order of the factors
0
2
4
6
8
10
12
1
3
5
7
9
11
13
15
17
19
21
23
25


Factor Loadings
Second factor (10.91%)
First factor (49.8%)
-1.0
-0.5
0.0
0.5
O4
O1
O3
C1
C2
N2
N4
N3
N1
E2
E5
C3
E3
E1
A1
C5
N5
age
A2
A3
A5
E4
A4
C4
O2
O5

MATRIZ DE CORRELAÇÃO

	MR2	MR1	MR3	MR5	MR4	MR6
MR2	1.00	0.24	-0.18	-0.05	-0.01	0.10
MR1	0.24	1.00	-0.23	-0.28	-0.19	-0.15
MR3	-0.18	-0.23	1.00	0.16	0.19	0.04
MR5	-0.05	-0.28	0.16	1.00	0.18	0.17
MR4	-0.01	-0.19	0.19	0.18	1.00	0.05
MR6	0.10	-0.15	0.04	0.17	0.05	1.00

	MR2	MR1	MR3	MR5	MR4	MR6	h2	u2	com
A1	0.11	-0.07	0.07	-0.56	-0.01	0.35	0.379	0.62	1.8
A2	0.03	-0.08	0.09	0.64	0.01	-0.06	0.467	0.53	1.1
A3	-0.04	-0.10	0.04	0.60	0.07	0.16	0.506	0.49	1.3
A4	-0.07	-0.07	0.19	0.41	-0.13	0.13	0.294	0.71	2.0
A5	-0.17	-0.16	0.01	0.47	0.10	0.22	0.470	0.53	2.1
C1	0.05	0.08	0.54	-0.02	0.19	0.05	0.344	0.66	1.3
C2	0.09	0.17	0.66	0.06	0.08	0.16	0.475	0.53	1.4
C3	0.00	0.07	0.56	0.07	-0.04	0.05	0.317	0.68	1.1
C4	0.07	0.10	-0.67	-0.01	0.02	0.25	0.555	0.45	1.3
C5	0.15	0.17	-0.56	0.02	0.10	0.01	0.433	0.57	1.4
E1	-0.14	0.61	0.09	-0.14	-0.08	0.09	0.414	0.59	1.3
E2	0.06	0.68	-0.03	-0.07	-0.08	-0.01	0.559	0.44	1.1
E3	0.02	-0.32	0.01	0.17	0.38	0.28	0.507	0.49	3.3
E4	-0.07	-0.49	0.03	0.25	0.00	0.31	0.565	0.44	2.3
E5	0.16	-0.39	0.27	0.07	0.24	0.04	0.410	0.59	3.0
N1	0.82	-0.09	-0.01	-0.09	-0.03	0.02	0.666	0.33	1.1
N2	0.83	-0.07	0.02	-0.07	0.01	-0.07	0.654	0.35	1.0
N3	0.69	0.13	-0.03	0.09	0.02	0.06	0.549	0.45	1.1
N4	0.44	0.43	-0.14	0.09	0.10	0.01	0.506	0.49	2.4
N5	0.47	0.21	-0.01	0.21	-0.17	0.09	0.376	0.62	2.2
O1	-0.05	-0.01	0.07	-0.04	0.57	0.09	0.357	0.64	1.1
O2	0.12	0.01	-0.09	0.12	-0.43	0.28	0.295	0.70	2.2
O3	0.01	-0.10	0.00	0.05	0.65	0.04	0.485	0.52	1.1
O4	0.10	0.34	-0.05	0.15	0.37	-0.04	0.241	0.76	2.6
O5	0.04	-0.02	-0.04	-0.01	-0.50	0.30	0.330	0.67	1.7
gender	0.20	-0.12	0.09	0.33	-0.21	-0.15	0.184	0.82	3.6
education	-0.03	0.05	0.01	0.11	0.12	-0.22	0.072	0.93	2.2
age	-0.06	-0.02	0.07	0.16	0.03	-0.26	0.098	0.90	2.0

A general Bayesian Blind Separation of Sources model shows that PCA, ICA, and FA are submodels.

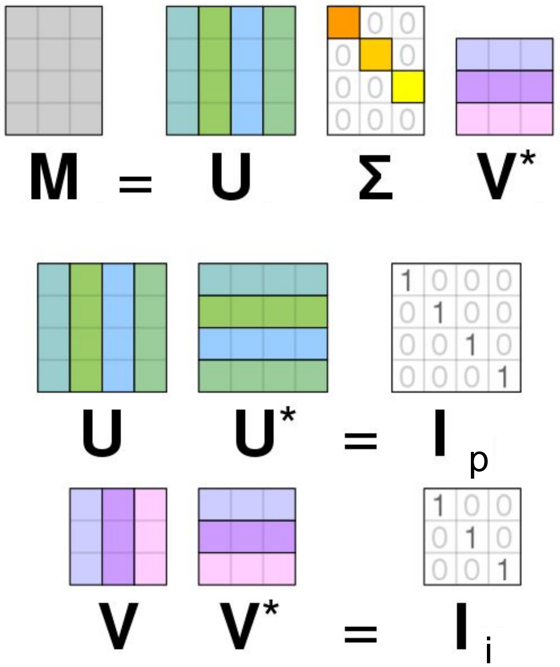
The general blind separation of sources model is

$$\underset{(p\times 1)}{(x_i|s_i,m)} = \underset{(p\times 1)}{f(s_i|m)} + \underset{(p\times 1)}{\epsilon_i,}$$

$$(\epsilon_i|\Psi) \sim N(0,\Psi)$$

$$(x_i|\mu,\Lambda,s_i,m) \sim N(\mu+\Lambda s_i,\Psi),$$

$$p(x_i|\mu,\Psi,m,s_i,\Lambda) =$$
$$(2\pi)^{-\frac{p}{2}}|\Psi|^{-\frac{1}{2}}e^{-\frac{1}{2}(x_i-\mu-\Lambda s_i)'\Psi^{-1}(x_i-\mu-\Lambda s_i)}.$$



DANIEL ROWE (1999). Bayesian blind source separation. IEEE Trans. Signal Processing

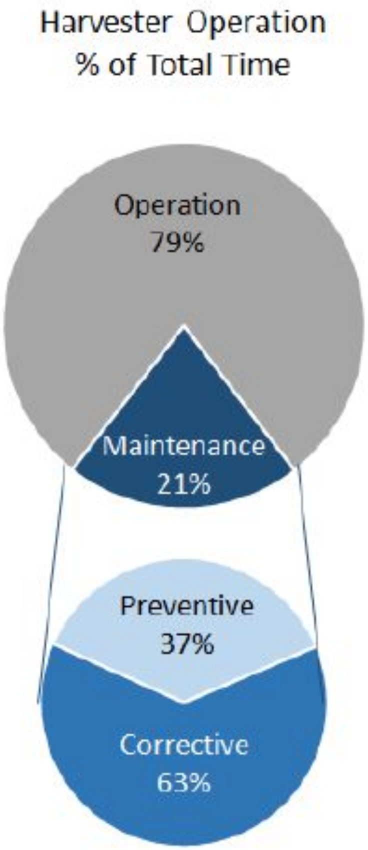
EXEMPLIFICATION:

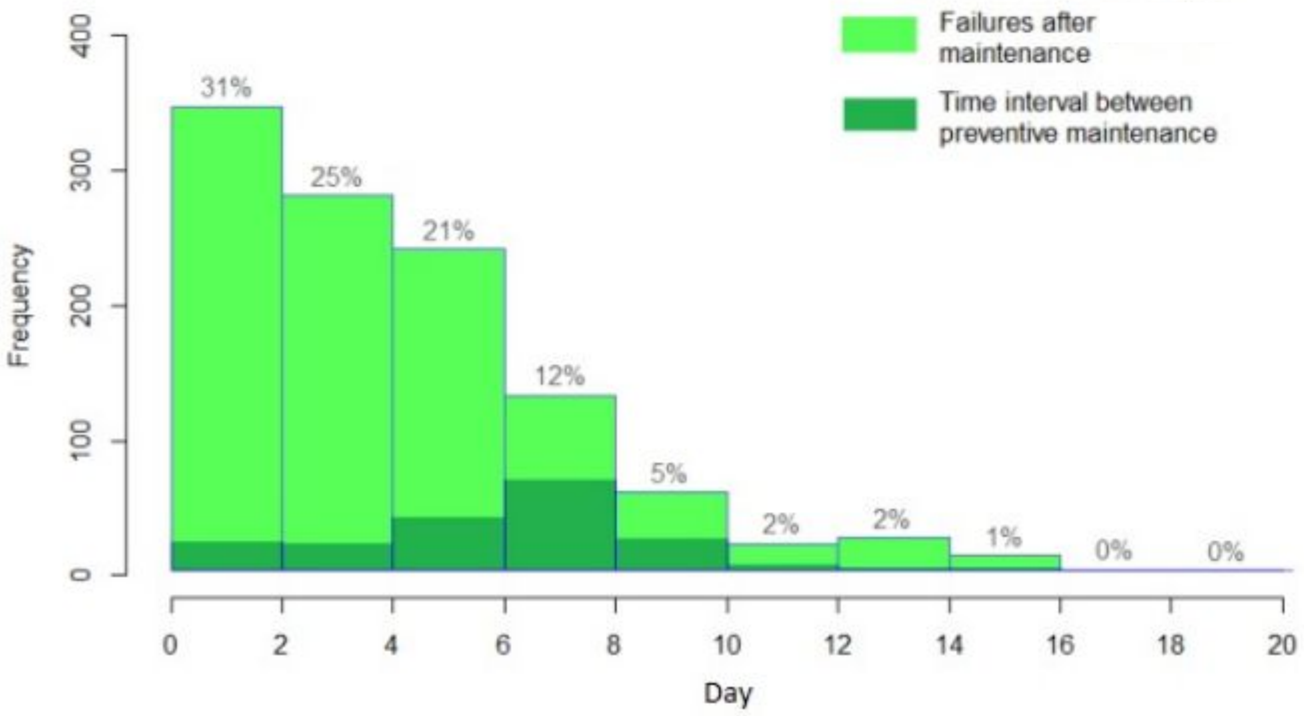
- [Reliability-Centered Maintenance]()

https://github.com/ProfNascimento/ECU

Nascimento D.C., Ramos P.L., Ennes A., Cocolo C., Nicola M.J., Alonso C., Ribeiro L.G., Louzada F. A reliability engineering case study of sugarcane harvesters. *Gestão & Produção*. 2020 Jul 27;27.

- **<u>10 Mechanical Components</u>** related with harvester.


Harvester Operation
% of Total Time
Operation
79%
Maintenance
21%
Preventive
37%
Corrective
63%


Failures after maintenance
Time interval between preventive maintenance
Frequency
Day
0
100
200
300
400
0
2
4
6
8
10
12
14
16
18
20
31%
25%
21%
12%
5%
2%
2%
1%
0%
0%

GOAL: What features/variables are associated with the durability of the equipment (working hours)?

INFERENCIAL MODEL

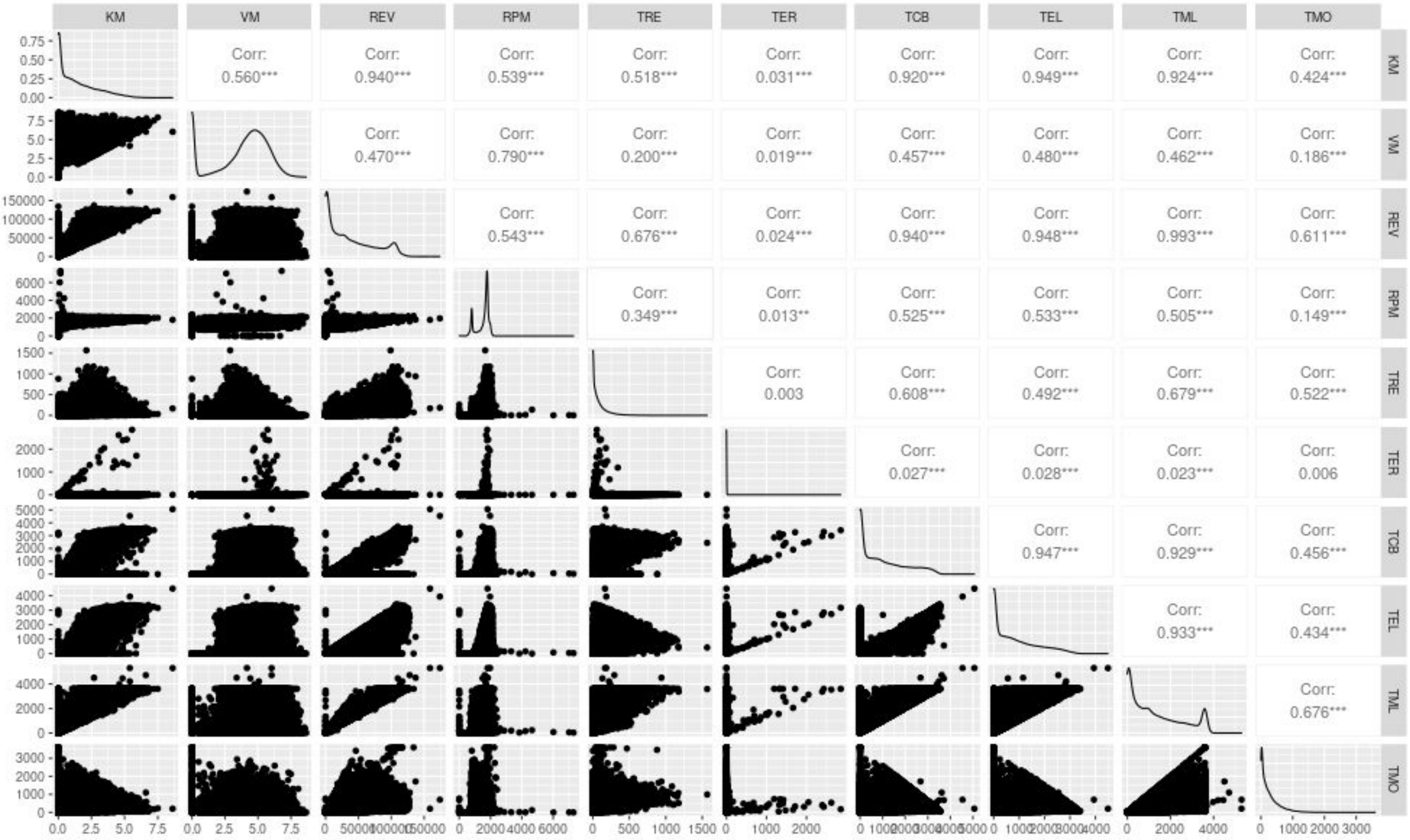
$$Y = g(X1,...,Xp)$$

Dependent Variable (Y) explained by some other factors (Xs, covariables or features)

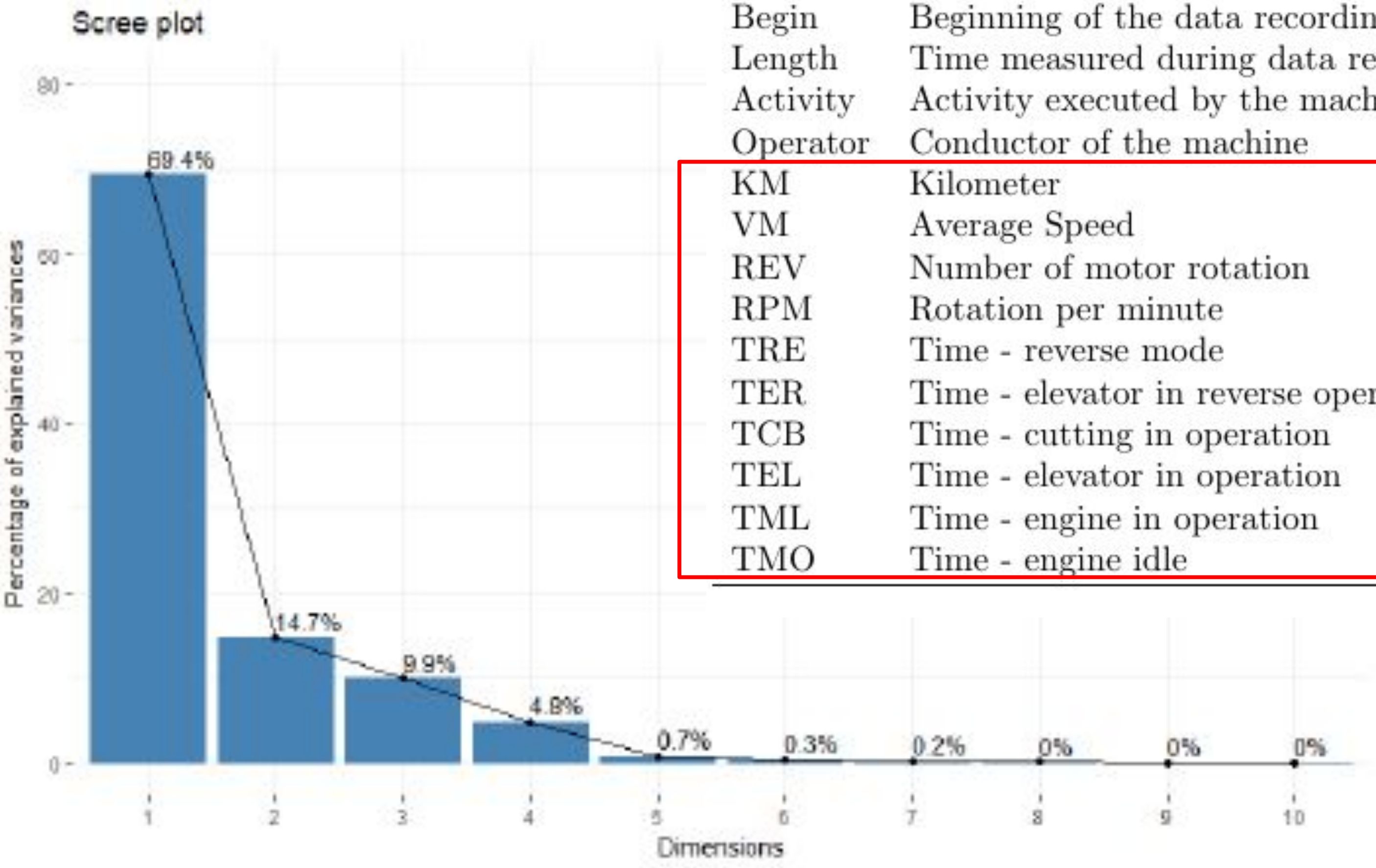
[USUALLY INDEPENDENTS FROM EACH OTHER]

i) ESTIMATE PARAMETERS and;

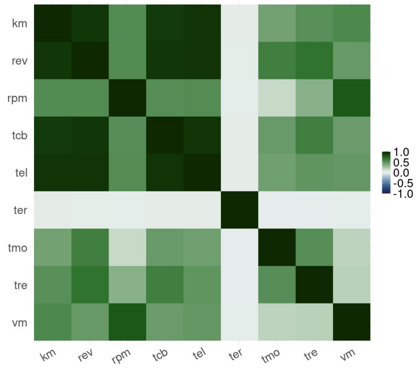
ii) HYPOTHESIS TESTING



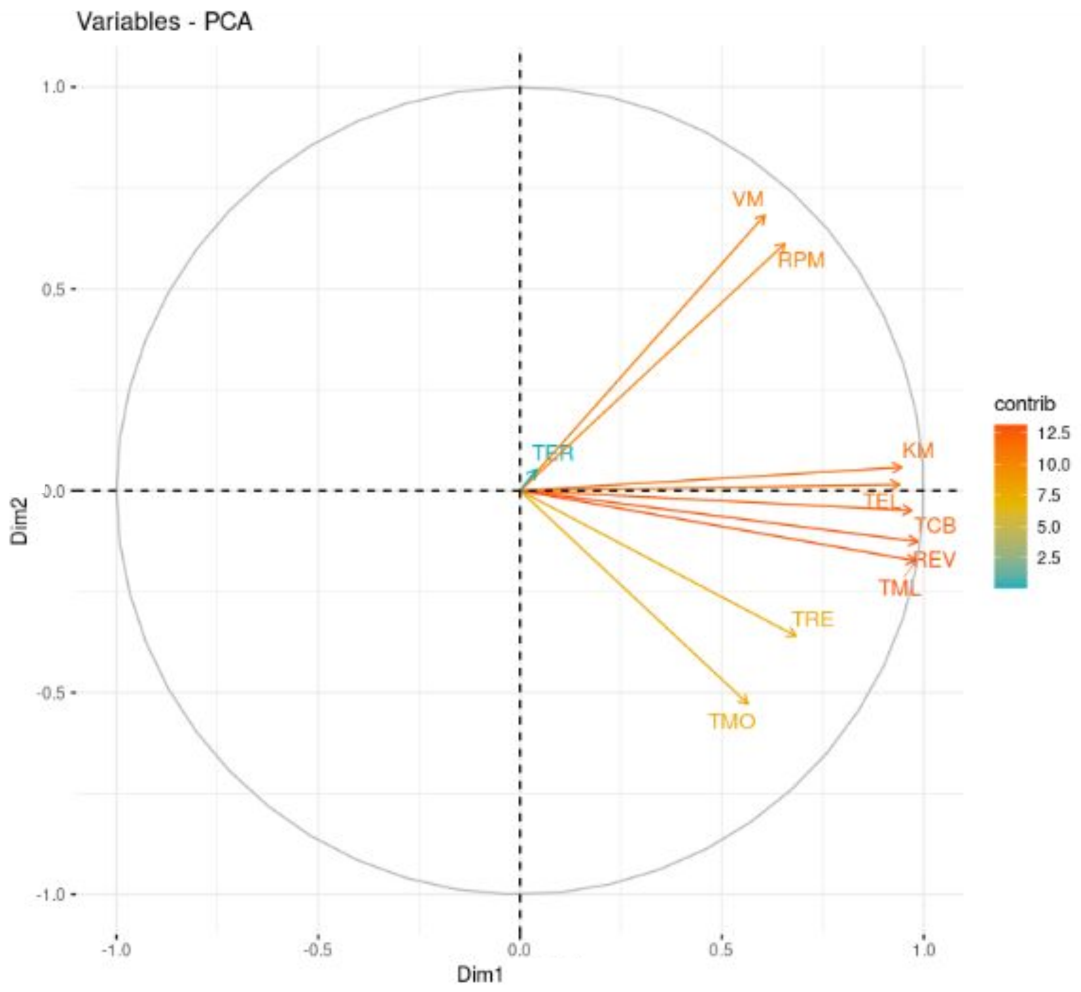
KM
VM
REV
RPM
TRE
TER
TCB
TEL
TML
TMO
Corr: 0.560***
Corr: 0.940***
Corr: 0.539***
Corr: 0.518***
Corr: 0.031***
Corr: 0.920***
Corr: 0.949***
Corr: 0.924***
Corr: 0.424***
Corr: 0.470***
Corr: 0.790***
Corr: 0.200***
Corr: 0.019***
Corr: 0.457***
Corr: 0.480***
Corr: 0.462***
Corr: 0.186***
Corr: 0.543***
Corr: 0.676***
Corr: 0.024***
Corr: 0.940***
Corr: 0.948***
Corr: 0.993***
Corr: 0.611***
Corr: 0.349***
Corr: 0.013**
Corr: 0.525***
Corr: 0.533***
Corr: 0.505***
Corr: 0.149***
Corr: 0.003
Corr: 0.608***
Corr: 0.492***
Corr: 0.679***
Corr: 0.522***
Corr: 0.027***
Corr: 0.028***
Corr: 0.023***
Corr: 0.006
Corr: 0.947***
Corr: 0.929***
Corr: 0.456***
Corr: 0.933***
Corr: 0.434***
Corr: 0.676***



Variable	Explanation
Vehicle	Number of equipments
Begin	Beginning of the data recording
Length	Time measured during data recording
Activity	Activity executed by the machine
Operator	Conductor of the machine
KM	Kilometer
VM	Average Speed
REV	Number of motor rotation
RPM	Rotation per minute
TRE	Time - reverse mode
TER	Time - elevator in reverse operation
TCB	Time - cutting in operation
TEL	Time - elevator in operation
TML	Time - engine in operation
TMO	Time - engine idle



	PC1	PC2
KM	0.37856360	0.05073800
VM	0.24282578	0.59760239
REV	0.39440254	-0.11058193
RPM	0.26292542	0.53610395
TRE	0.27377921	-0.31549596
TER	0.01607808	0.04607790
TCB	0.38847137	-0.04297533
TEL	0.37652152	0.01320167
TML	0.39251238	-0.15236282
TMO	0.22601797	-0.46234702

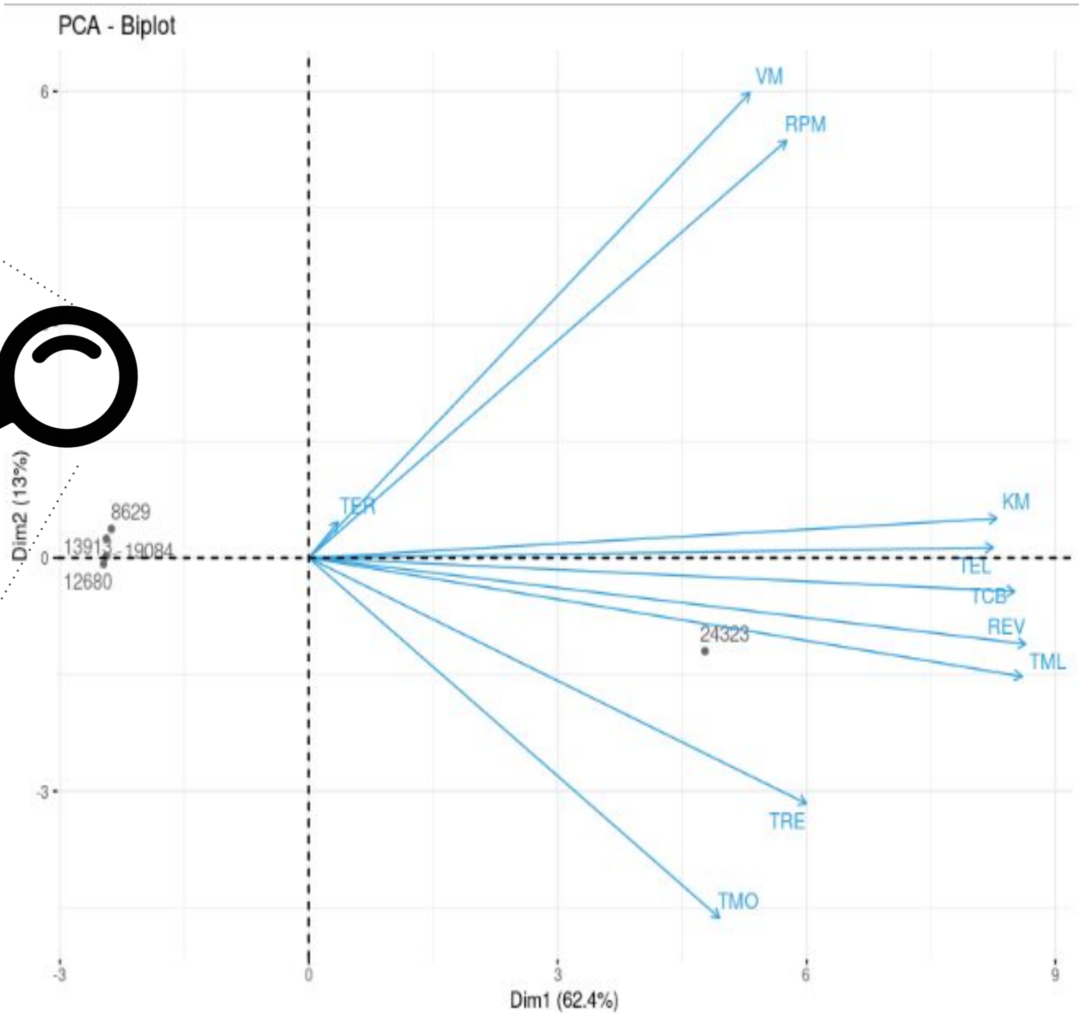


	ID
Xs	**13913**
KM	0.842
VM	5.46
REV	21600
RPM	1680
TRE	12
TER	0
TCB	550
TEL	479
TML	773
TMO	223

Y = 773 hours

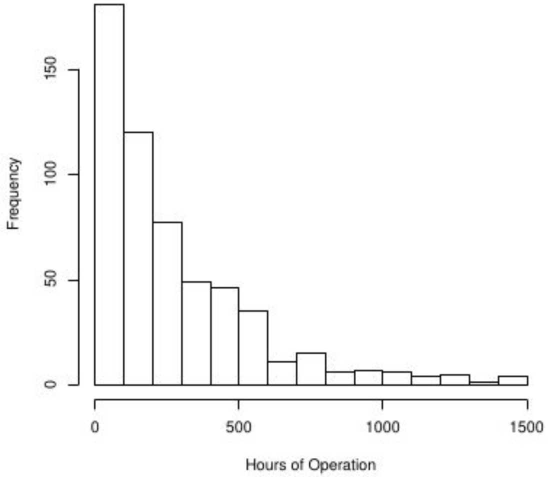
X*s

PC 1	-2.4392
PC 2	0.2469



…THEN MODELING (GAMMA REGRESSION)

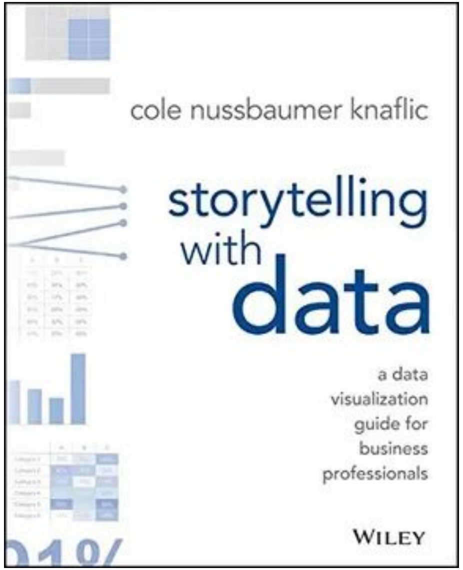
$$\begin{aligned} Y =& g(\beta_0 + \beta_1 \log(\#Failures) + \beta_2 \log(\#Failure\_Cause) + \beta_3 PC1 + \beta_4 PC2 + \beta_5 PC3 + \\ &+ \beta_6 Occurence\_Type + \beta_7 Equipment + \beta_8 Occurrence\_Type * Equipment + \beta_9 Crop + \\ &+ \beta_{10} Occurrence\_Type * Crop + \beta_{11} Occurrence\_Type * \#Failures\_Cause + \beta_{12} Rainfall + \\ &+ \beta_{13} Relative\_Humidity + \beta_1 4 Temperature) \end{aligned}$$

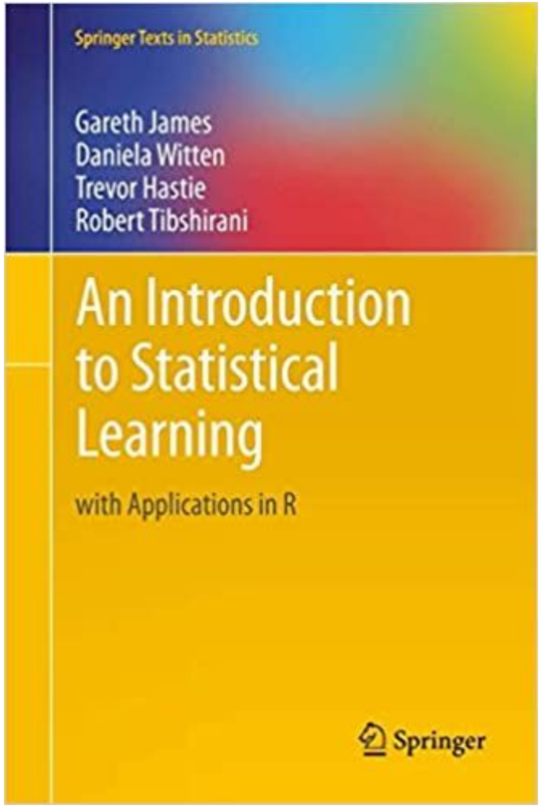


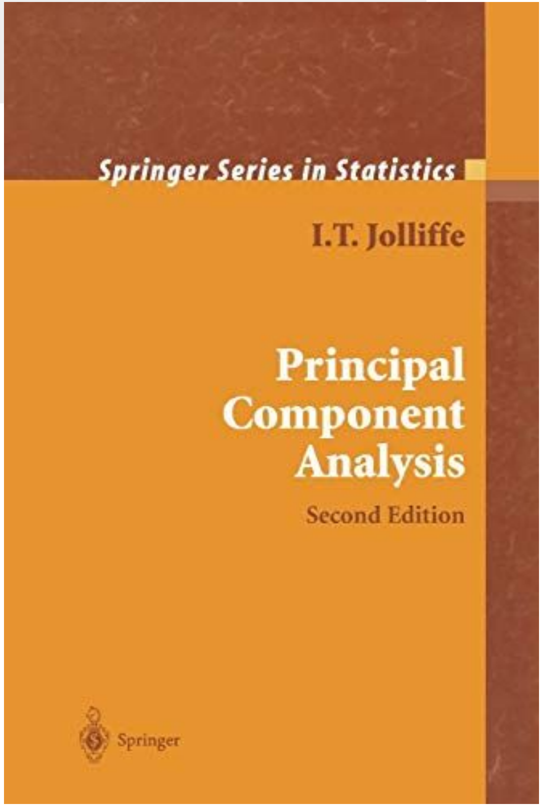
Significant Explainable Variables

- PC1;
- PC2;
- Occurrence Type (Transmission);
- Occurrence Type (Motor & Transmission) in the Crop of 2017;
- Occurrence Type (Transmission) | # Failures;
- Rainfall;
- Max Temperature.

FURTHER INVESTIGATION







UNIVERSIDAD
DE ATACAMA

Diego Nascimento
diego.nascimento@uda.cl
Ph.D. in Statistics (USP/UFSCar)