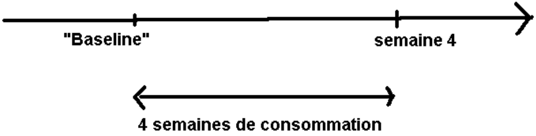
Development of a methodology for analyzing nutritional data

Myriam Tami

20th march 2013

1. Context
2. Establishment of a target zone
3. Target zone and results obtained
4. Alternative : development of a new method
5. Method principle
6. Results
7. Conclusion

Context



- Clinical study : collect food intake. (Nutritional data).
- 308 statistical units (observations) ($i \in \{1,...,308\}$) : first visit → test/control.
- Nutritional recommendations (12 nutrients) : cholesterol, iron, calcium, protein, fat, carbohydrates, phosphorus, potassium, sodium...
- Goal : Establish a target zone based on nutritional recommendations (position).

Target zone : based on nutritional Recommendations

- Nutritional recommendations :

  Examples :

  → Iron ⇝ [12mg, 28mg]

  → Percentage of daily energy intake

Lip	[30%, 40%]
GluT	[45%, 60%]

→ gram body weight

ProtT : [0.66 g/Kg, 2.2 g/Kg]

Target zone : Distances

- Distance for each nutrient $k \in \{1, ..., 12\}$:
  $\longrightarrow$ Measurement of the excess relative to the zone : $\bar{d}_k$.
  $\longrightarrow$ Measurement of deficiency to the zone : $\underline{d}_k$.

$$\begin{aligned} \bar{d}_k(x, R_k) &= 0 \qquad si\ x \leq b_k \\ &= \frac{|x - b_k|}{b_k} \quad si\ x > b_k \end{aligned}$$

$$\begin{aligned} \underline{d}_k(x, R_k) &= \frac{|x - a_k|}{a_k} \quad si\ x < a_k \\ &= 0 \qquad si\ x \geq a_k \end{aligned}$$

Such that,
$R_k = [a_k, b_k]$ the recommendation for each nutrient k.
x the nutrient consumption by a statistical unit $\underline{x}$.

Target zone : Distances

- A global distance :

$\longrightarrow$ Measurement of the global excess to the target zone : $\bar{\delta}$.

$\longrightarrow$ Measurement of global deficiency to the target zone : $\underline{\delta}$.

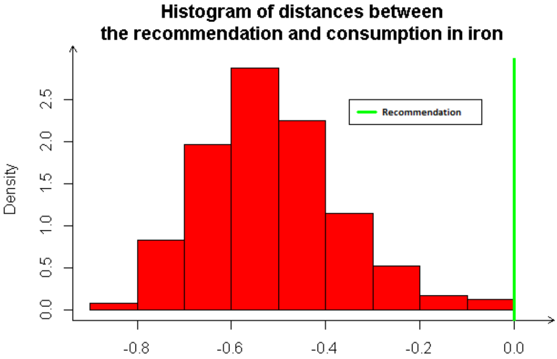
$$\bar{\delta}(\underline{x}, \underline{R}) = [\quad \sum_{k} (\bar{d}_k(\underline{x}, R_k))^p \quad]^{1/p}$$

$$\underline{\delta}(\underline{x}, \underline{R}) = [\quad \sum_{k} (\underline{d}_k(\underline{x}, R_k))^p \quad]^{1/p}$$

$\Longrightarrow$ Matrix of distances 308*12

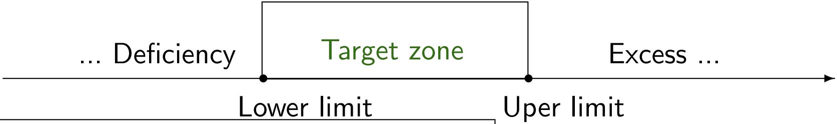
$\hookrightarrow$ new data : $\bar{\delta} - \underline{\delta}$

Iron



1.14% of people in the target zone

Schema : global distance



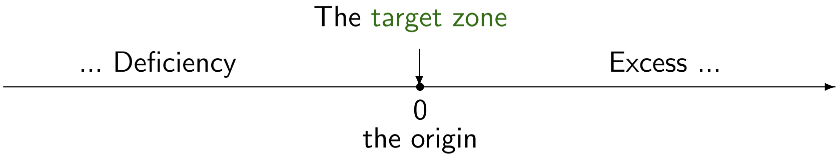
If $\bar{\delta}(\underline{x}, \underline{R})$=0, $\underline{x}$ is here

If $\underline{\delta}(\underline{x}, \underline{R})$=0, $\underline{x}$ is here

If $\bar{\delta}(\underline{x}, \underline{R})$=0
<u>and</u> $\underline{\delta}(\underline{x}, \underline{R})$=0,
$\underline{x}$ is here

Conclusion :
If $\bar{\delta}(\underline{x}, \underline{R}) = 0$ and $\underline{\delta}(\underline{x}, \underline{R}) = 0$, $\underline{x} \in \underline{R}$ the target zone.

Description zone



↪ No individuals do the following sets of recommendations for 12 nutrients.

$\Longrightarrow$ Study of statistical units outside the target zone :

↪ CAH, K-"median", ACP.

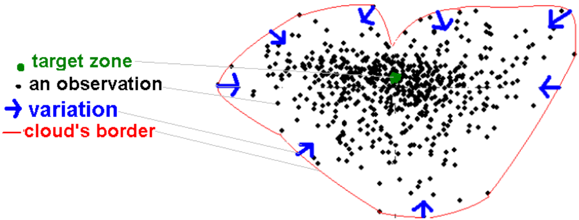
Target zone and results obtained

- The Target zone is empty.
  $\Longrightarrow$ No observation follows all (12) nutritional recommendations.
- Study of observations which are outside the target zone.
  $\rightarrow$Homogeneous : no particular group emerges.
  $\Longrightarrow$ The target zone is restrictive.

Alternative : development of a new method

- Idea : be able to TARGET a less restrictive zone.
- Make a "scanner" (in 12 dimensions) of the observations cloud.

  Variation of the cloud's border and respect of the shape of the cloud :



- Target a zone = choose a border.
- Choose a border = choose a parameter : α (exigency).

What is the α parameter ?

- $\alpha \in$ [0,1].
- α is a percentage.
- Definition :
  $\forall$ i $\in$ {1,...,308}
  x_i i-th observation of 308
  $\longrightarrow \alpha =$ order quantiles q^i_α such that $q^i_\alpha = ||x_i||$.
- To each observation x_i we denote α_i the value corresponding to α.

What is the utility of α parameter ?

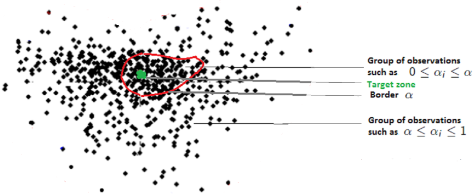
- Choose a border = choose a value α.
- Utility : isolate two groups :

  A group of observations x_i such that

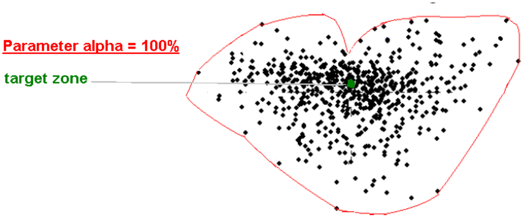
  $0 \leq \alpha_i \leq \alpha$

  The group of remaining observations x_i such that

  $\alpha < \alpha_i \leq 1$

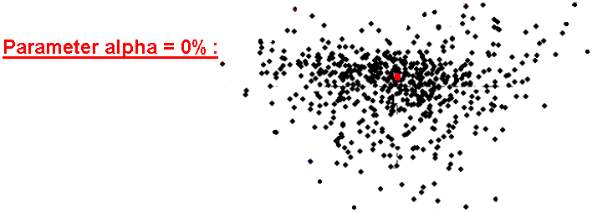

Group of observations
such as $0 \leq \alpha_i \leq \alpha$
Target zone
Border α
Group of observations
such as $\alpha \leq \alpha_i \leq 1$

Example : $\alpha = 1$



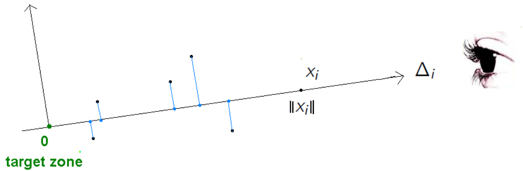
- $\alpha = 1.$
- The target zone contains all the individuals. $(\alpha = 1 \Leftrightarrow \alpha = 100\ \%)$. → No requirement about the consumption quality of individuals.

Example : $\alpha = 0$



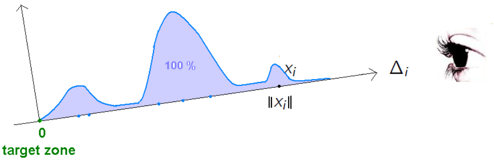
- $\alpha = 0$.
- The target zone is empty.
  $\rightarrow$ High exigency about the alimentary consumption quality of the individuals.

Method principle : step 1



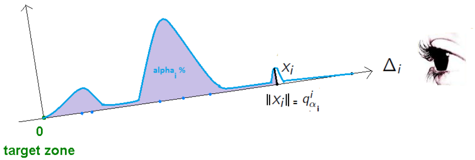
- From each observation x_i, we study the other observations positioning according to two points following : the target zone and x_i.

Method principle : step 2



- A curve (density) is obtained.
  $\longrightarrow$ Illustration of the cloud observed from x_i's position.

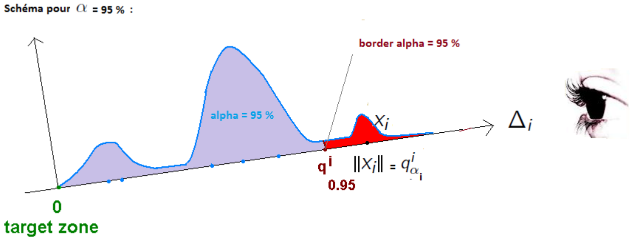
Method principle : step 3



- Evaluation of α_i values associated to each x_i, i $\in$ {1,...,308}.

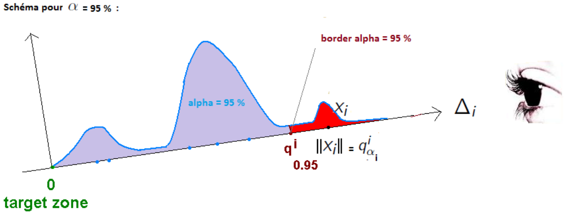
Method principle : step 4

- We choose a value of α : the border.
- For example $\alpha = 95\%$.



- We compare the value α chosen and α_i, i fixed.

Method principle : step 5

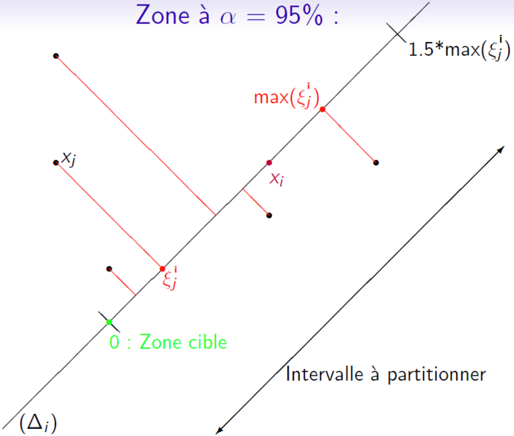


- Decision :
  $\longrightarrow$ If $\alpha_i > \alpha$ then x_i is outside of border chosen : x_i is outlier of the α zone chosen.
  $\longrightarrow$ If $\alpha_i \leq \alpha$ then x_i is inside of the border α chosen, i.e : inside the α zone.

Especially for statisticians : the principle construction

Especially for statisticians : the principle construction

Zone à $\alpha = 95\%$:



Especially for statisticians : the principle construction

- On each Δ_i, we partition the interval $[0, 1.5 * max_j(\xi_j^i)]$ so as to obtain 100 coordinates y_k^i, $k \in \{1,...,100\}$.
- In each y_k^i, we estimate value of the density.
- Multivariate and non-parametric kernel density estimatior :

$$\bar{f}_i(y_k^i) = \frac{1}{308 * h_1 * ... * h_{12}} \sum_{j=1}^{308} \prod_{d=1}^{12} K(\frac{y_k^{i,d} - x_j^d}{h_d})$$

$\longrightarrow$ d=12, the dimension.

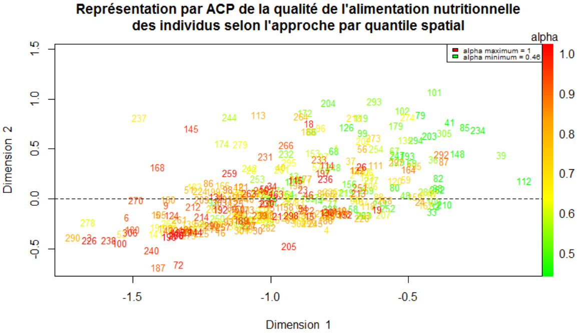
$\longrightarrow$ h_d diagonal elements of H (the bandwidth, smoothing).

$\longrightarrow$ K is the gaussian kernel.

- For each estimated density $\hat{\phi}_i = \hat{f}_i \|x_i\|^{12-1}$, we estimate the quantile α % : q_α^i.
- Comparaison of $\|x_i\|$ et q_α^i (α chosen).

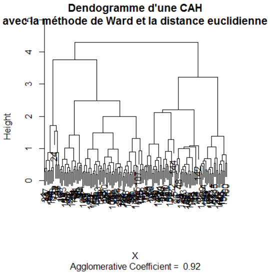
$\Longrightarrow$ Variation of α : "scanner" of the cloud.

Method's graphic results



Study of the group "outlier" : CAH

- Dendogram obtained by aggregation method of Ward and the Euclidean metric.



$\hookrightarrow$ k = 2 ou 3 clusters.

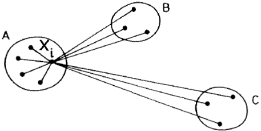
- Quality of clustering : silhouette coefficient.

Study of the group "outlier" : "K-median"

- **Identification of the number of cluster** based on median (robust to outliers).
- Definition :

$$s(X_i) = \frac{b(X_i) - a(X_i)}{max\{a(X_i), b(X_i)\}}, \text{ for each } x_i.$$

- $a(X_i)$ is the average of dissimilarities between X_i and all other observations in the group to which is belong.
- $b(X_i)$ is the average of dissimilarities between x_i and the observations in the group closest to the group of x_i.

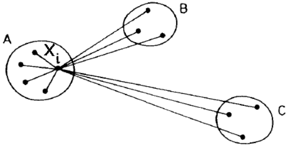


Study of the group "outlier" : "K-median"

- **For each group** of a size k partition of all observations, **there is a silhouette coefficient** which is the average of $s(x_i)$'s group.
- **The global silhouette coefficient s(x)** (which is indicated in a graphic output) is the **average of all silhouette coefficients of groups**.
  $\hookrightarrow$ -1 $\leq$ s(x) $\leq$ 1
  $\hookrightarrow$ s(x) is calculated for a number of clusters k $>$ 1.
  $\hookrightarrow$ s(x) measures the quality of the clustering (the lower s(x), the less well x is classified).

Study of the group "outlier" : "K-median"

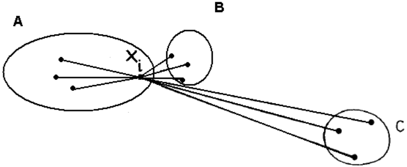
Example k = 3 :
s(X_i) $\simeq$ 1 $\Leftrightarrow$ X_i is in the right group.



$\hookrightarrow$ b(X_i) corresponds to the average of dissimilarities associated with group B since C is more distant.
$\hookrightarrow$ $a(X_i)$ is the average of dissimilarities between X_i and all other observations in group A.

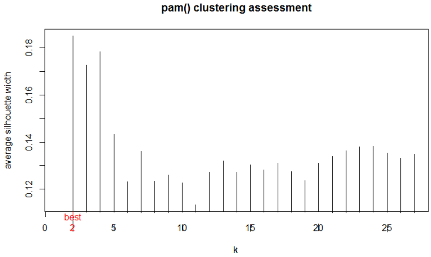
Study of the group "outlier" : "K-median"

Example k = 3 :
s(X_i) ≃ -1 ⇔ X_i is in the wrong group.



↪ b(X_i) corresponds to the average of dissimilarities associated with group B since C is more distant.
↪ $a(X_i)$ is the average of dissimilarities between X_i and all other observations in group A.

Study of the group "outlier" : K-means ou "K-median" : result

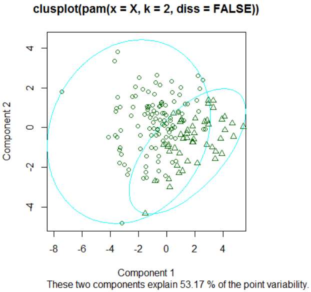


- k = 2 clusters.
- The average silhouette is 0.19.

  ⟶ Quality of clustering is low.

Study of the group "outlier" : K- median

- "K-median" ⇒ 2 clusters.



- cloud "homogeneous" : No clear group structure.

Conclusion

- Measurement criterion : distance.
- No individuals in the target zone.
- Alternative approach : "scanner" of the cloud (variation of the cloud's border i.e : α).
- Groups of individuals (outliers compared to the cloud of points observed) are formed depending on the α value chosen. $\hookrightarrow$ A classical study of these groups "oulier" is possible (CAH, K-means/K-median, AFD...).
- No interest to cluster the data group "outlier" by classes.

THANK YOU

THANK YOU FOR YOUR ATTENTION

Bibliography

[1] To Meet Nutrient Recommendations, Most French Adults Need to Expand Their Habitual Food Repertoire, , The Journal of Nutrition, 2009.

[2] Development of the Healty Eating Index-2005, Journal of the AMERICAN DIETETIC ASSOCIATION, 2008.

[3] Evaluation of the Healthy Eating Index-2005, Journal of the AMERICAN DIETETIC ASSOCIATION, 2008.

[4] Silhouettes : a graphical aid to the interpretation and validation of cluster analysis, Elsevier Science Publishers B.V. (North-Holland), 1987.

Bibliography

[5] Clustering in an Object-Oriented Environment, Departement of Mathematics and Computer Science, U.I.A, Universiteitsplein 1, B-2610 Antwerp, Belgium.

[6] Central region for multidimensionnal data : the spatial quantile approach, messieurs Gannoun et Bry, en préparation.

[7] Le logiciel R : Maîtriser le langage, Effectuer des analyses statistiques, Rémy Drouilhet, Pierre Lafaye de Micheaux, Benoît Liquet, 2010.

[8] Initiation à la statistique avec R, Frédéric Bertrand, Myriam Maumy-Bertrand, 2010.

Bibliography

[9] Classification et Prévision des Données Hétérogènes : Application aux Trajectoires et Séjours Hospitaliers, Thèse de Haytham ELGAZEL, Université Claude Bernard Lyon, 2007.

[10] Économétrie non paramétrique, Stéphane Adjemian, Université d'Evry, 2004.

[11] Convergence de l'estimateur multivarié à noyau de la régression, Karima Lagha et Smail Adjabi, Laboratoire LAMOS, Université de Béjaia, Algérie.

[12] Study NU325, Data review meeting, Nutritional data, Report, version 1.4-30 March 2011

Recommendations

- EAR : "Estimated Average Requirement"*
- RDA : "Recommended dietary allowance"*

∗ According to : *To Meet Nutrient Recommend tions, Most French Adults Need to Expand Their Habitual Food Repertoire*.

Recommendations

TABLE : Recommendations form of simple intervals

[EAR(Chol), Upperlim]	[RDA(Chol), Upperlim]
[0mg, 300mg]	[0mg, 300mg]
[EAR(FibT), Upperlim]	[RDA(FibT), Upperlim]
[19g, ∞]	[25g, ∞]
[EAR(Fe), Upperlim]	[RDA(Fe), Upperlim]
[12mg, 28mg]	[16mg, 28mg]
[EAR(Iode), Upperlim]	[RDA(Iode), Upperlim]
[116μg,600μg]	[150μg, 600μg]
[EAR(Mg), Upperlim]	[RDA(Mg), Upperlim]
[277mg, 700mg]	[360mg, 700mg]

Recommendations

TABLE : Recommendations form of simple intervals

[EAR(P), Upperlim]	[RDA(P), Upperlim]
[578mg,2500mg]	[750mg,2500mg]
[EAR(K), Upperlim]	[RDA(K), Upperlim]
[2387mg, ∞]	[3100mg, ∞]
[EAR(Na), Upperlim]	[RDA(Na), Upperlim]
[1500mg, 2365mg]	[1500mg, 2365mg]
[EAR(Cu), Upperlim]	[RDA(Cu), Upperlim]
[1,2mg,5mg]	[1,5mg, 5mg]
[EAR(VitC), Upperlim]	[RDA(VitC), Upperlim]
[85mg, 500mg]	[110mg, 500mg]

Recemmandations

TABLE : Recommendations If age $\leq$ 55 years old

[EAR(Ca), Upperlim]	[RDA(Ca), Upperlim]
[693mg, 2500mg]	[900mg, 2500mg]
[EAR(Zn), Upperlim]	[RDA(Zn), Upperlim]
[7.7mg,25mg]	[10mg, 25mg]
[EAR(Se), Upperlim]	[RDA(Se), Upperlim]
[39μg,350μg]	[50μg,350μg]

Recommendations

TABLE : Recommendations If age $\geq$ 55 years old

[EAR(Ca), Upperlim]	[RDA(Ca), Upperlim]
[924mg,2500mg]	[1200mg, 2500mg]
[EAR(Zn), Upperlim]	[RDA(Zn), Upperlim]
[8.5mg,25mg]	[11mg, 25mg]
[EAR(Se), Upperlim]	[RDA(Se), Upperlim]
[46.2μg,,350μg]	[60μg,,350μg]

Recommendations

TABLE : Recommendations for ProtT

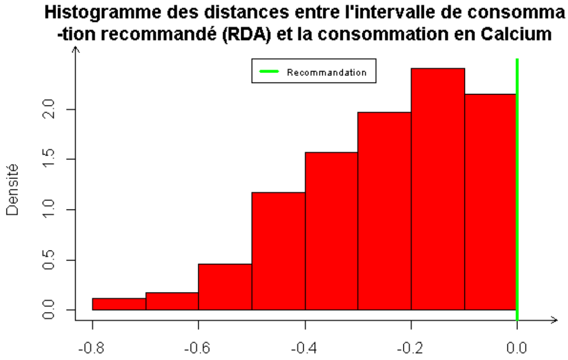
[EAR(ProtT), Upperlim]	[RDA(ProtT), Upperlim]
[0.66 g/Kg, 2.2 g/Kg]	[0.83 g/Kg, 2.2 g/Kg]

Recommendations

TABLE : Recommendations as a percentage of daily energy intake

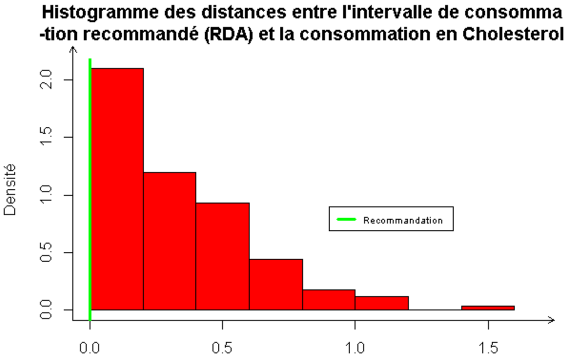
Lip	[30%, 40%]
GluT	[45%, 60%]

Calcium



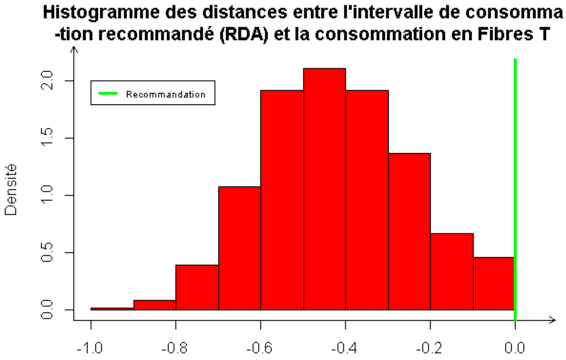
43.18% d'individus dans la zone

Cholesterol



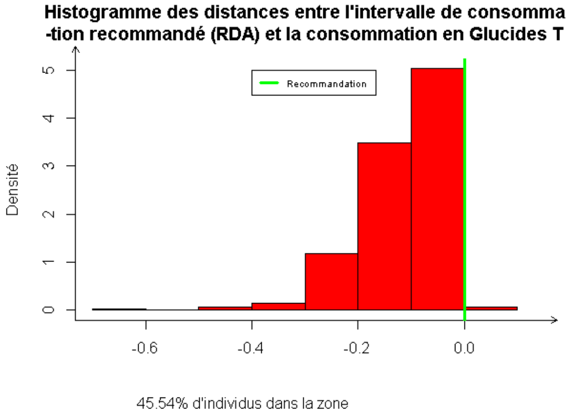
53.89% d'individus dans la zone

Fiber T

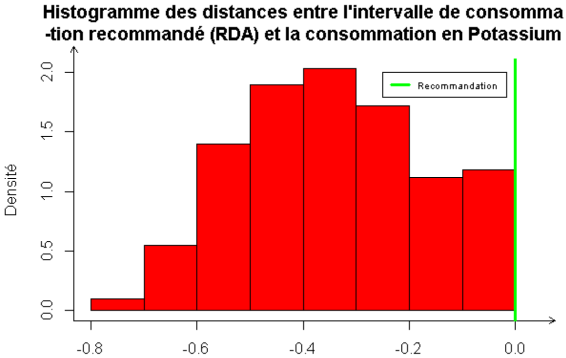


5.03% d'individus dans la zone

Carbohydrate T

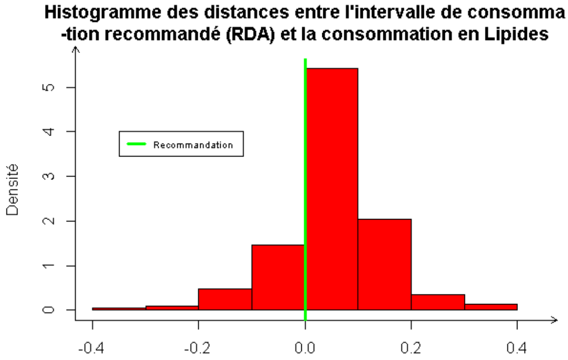


Potassium



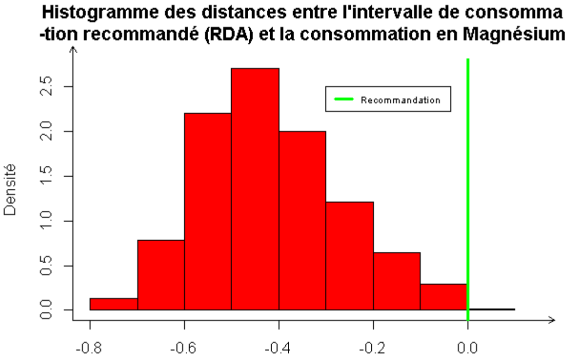
5.52% d'individus dans la zone

Fat



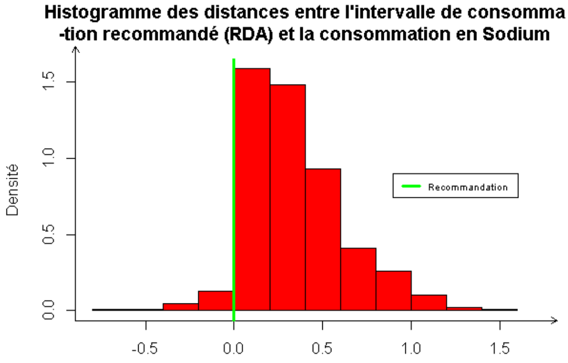
62.5% d'individus dans la zone

Magnesium



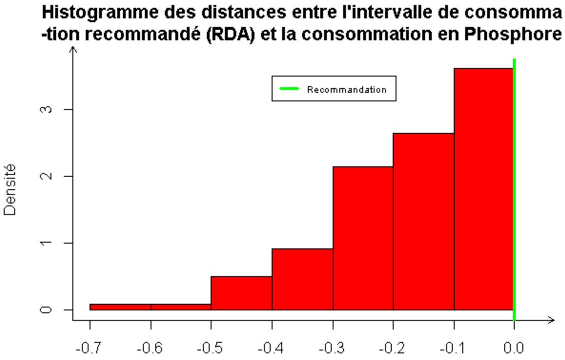
2.11% d'individus dans la zone

Sodium



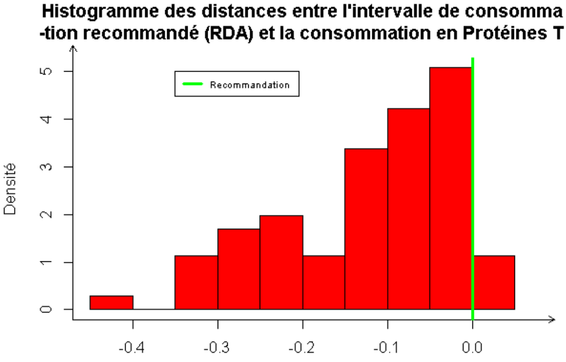
30.84% d'individus dans la zone

Phosphorus



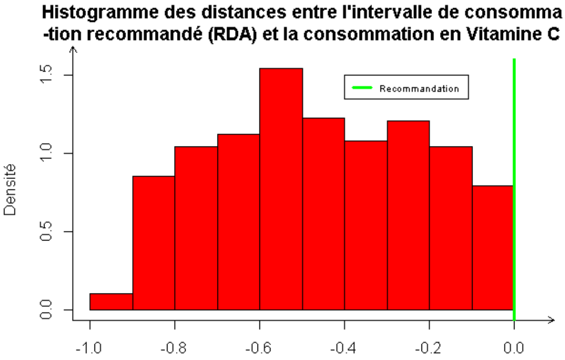
64.45% d'individus dans la zone

Protein T



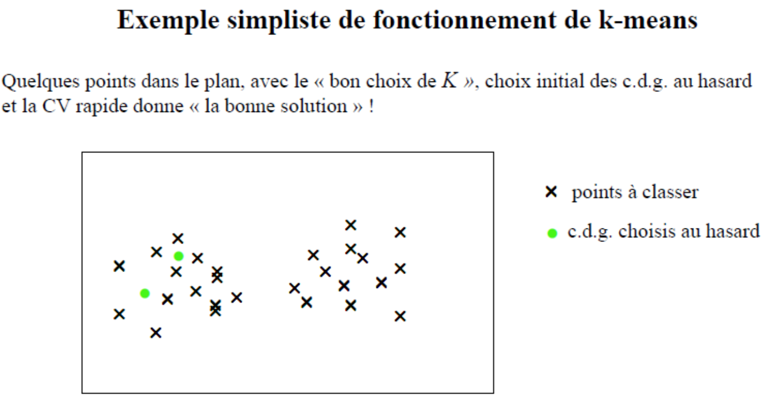
88.47% d'individus dans la zone

Ascorbic acid



21.91% d'individus dans la zone

Kmeans



Exemple simpliste de fonctionnement de k-means

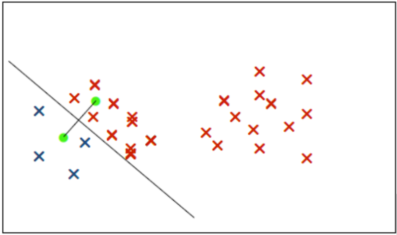
Quelques points dans le plan, avec le « bon choix de K », choix initial des c.d.g. au hasard et la CV rapide donne « la bonne solution » !

✕ points à classer

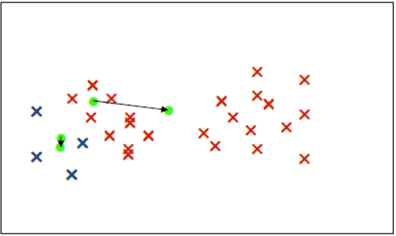
● c.d.g. choisis au hasard

Kmeans

Etape 0-1 : affectation des observations aux c.d.g.

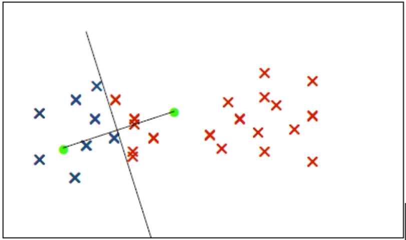


Etape 0-2 : calcul des nouveaux c.d.g.

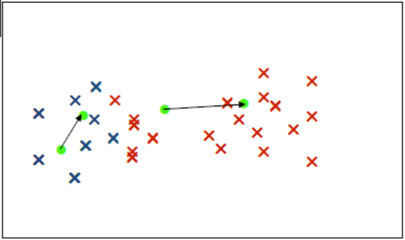


Kmeans

Etape 1-1 : affectation des observations aux c.d.g.

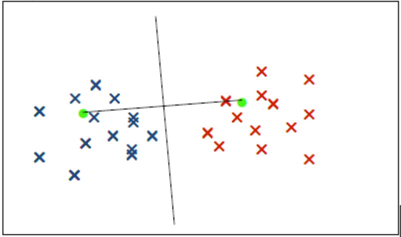


Etape 1-2 : calcul des nouveaux c.d.g.

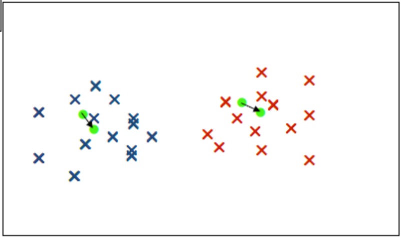


Kmeans

Etape 2-1 : affectation des observations aux c.d.g.



Etape 2-2 : calcul des nouveaux c.d.g.



NONE SEARCH