enbis European Network for Business and Industrial Statistics

Consumers' satisfaction with a product analysed through the lens of fuzzy theory

ENBIS 2021 Spring Meeting

- Biasetton Nicolò, Università degli studi di Padova
- Disegna Marta, Bournemouth University
- Salmaso Luigi, Università degli studi di Padova

dtg

Università degli Studi di Padova

Introduction

Aim

Cluster consumers based on their satisfaction with an Electric engine Pressure Washer with a focus on more dissatisfied consumers.

Problem

Satisfaction on different product characteristics is collected using Likert-type scales questions

Method

Fuzzy *C*-Medoid for fuzzy data.

Case study

Product: Electric engine Pressure Washer

Sample size: 1125 consumers

Variables:

KPI

- Overall rating
- Overall cleaning
- Handling easiness
- Propensity for buying
- Satisfaction with customer service

Drivers groups

- Demographic characteristics
- Usage characteristics
- Issues
- Stains cleaned
- Other characteristics

Note: **KPIs** are used to cluster consumers and **Drivers** are used to profile clusters

Cluster analysis - steps

Step 1	Step 2	Step 3	Step 4	Step 5
Transform the set of segmentation variables (if required)	Choice of **dissimilarity measure**	Choice of **clustering algorithm**	**Cluster validation** Select the best **partition**, i.e. the optimal number of clusters	**Labelling** and **profiling**

Step 1: Data Re-scaling

Rescale

1. Due to the low proportion of negative score, we decided to re-scale into a **3-point scale** the original 5-point scale to better higlight unsatisfied people and their commonalities.

Original scale	negative	1	2	3	4	5	positive

New scale	negative	1	2	3	positive

Filter

2. Customers who gave **maximum score** (5) to **all KPI** have been **excluded**. A total of 260 customers have been removed from the dataset and the final sample size is 865 observations.

Step 1: Fuzzification Likert type scale

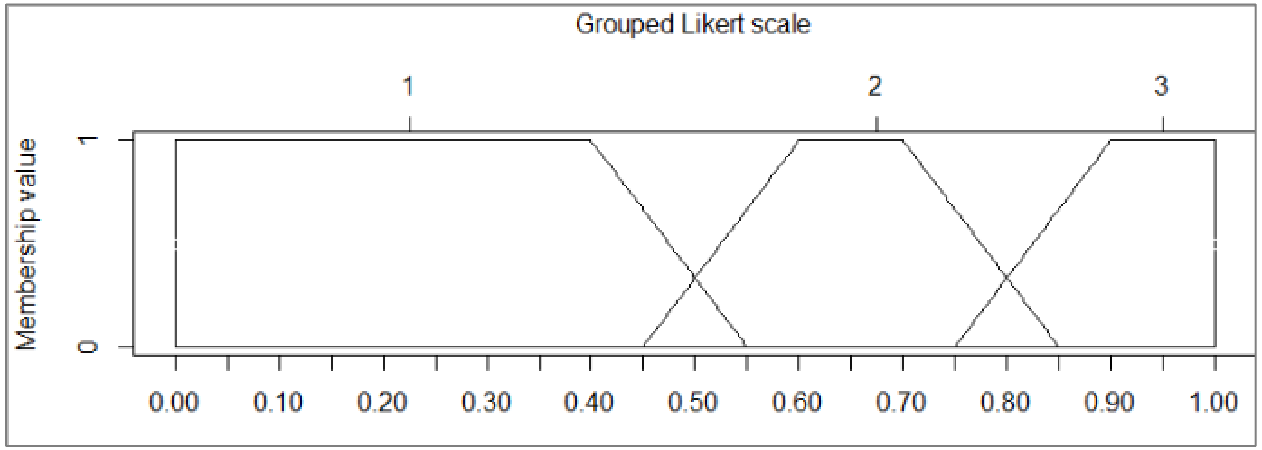
Fuzzify

3. Since:

- KPI have been collected using different scales not fully comparable
- Likert-type scale entiles a certain degree of **uncertainty/vagueness** (Davidov et al. 2014; D'Urso 2007; D'Urso et al. 2013; Disegna et al. 2018):

KPI variables have been recoded into **fuzzy data**, i.e. each individual score or expression is recoded into a range of possible values (Dubois and Prade 1988).

Following Sii et al. (2001), the **trapezoidal fuzzy number** has been used in this study to recode the 3-point Likert-type scale satisfaction variables:



Step 2: Dissimilarity measure

Since segmentation variables are trapezoidal fuzzy numbers, the dissimilarity between two units is measured by comparing the fuzzy data observed on each unit.

Therefore, the **trapezoidal fuzzy distance** between the i-th and j-th units (with $i \neq j$) suggested by Coppi et al. (2012) has been computed as follows:

$$d_F^2(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) = \left[\omega_M^2 \left(\left\|\boldsymbol{m}_{1i} - \boldsymbol{m}_{1j}\right\|^2 + \left\|\boldsymbol{m}_{2i} - \boldsymbol{m}_{2j}\right\|^2\right) + \omega_S^2 \left(\left\|\boldsymbol{l}_i - \boldsymbol{l}_j\right\|^2 + \left\|\boldsymbol{r}_i - \boldsymbol{r}_j\right\|^2\right)\right]$$

where $\tilde{\boldsymbol{x}}_i \equiv \left\{\widetilde{x_{iy}} = (m_{1iy}, m_{2iy}, l_{iy}, r_{iy})_{LR}\right\}$ denote the fuzzy data vector for the ith object;

$\boldsymbol{m}_{1i} \equiv (m_{1i1}, \ldots, m_{1iy}, \ldots, m_{1iY})'$, $\boldsymbol{m}_{2i} \equiv (m_{2i1}, \ldots, m_{2iy}, \ldots, m_{2iY})'$, $\boldsymbol{l}_{1i} \equiv (l_{i1}, \ldots, l_{iy}, \ldots, l_{iY})'$, $\boldsymbol{r}_{1i} \equiv (r_{i1}, \ldots, r_{iy}, \ldots, r_{iY})'$, the operator $\|\ldots\|^2$ denote the squared Euclidean distance and ω_M, ω_S are suitable weights for the center component and the spread component constrained by the following conditions: $\omega_M + \omega_S = 1$ (normalization condition) and $\omega_M \geq \omega_S \geq 0$ (coherence condition)

Step 3: Fuzzy *C*-medoid clustering algorithm

The fuzzy *C*-medoid algorithm has the advantages of both **Fuzzy** and **Partition Around Medoids** (PAM) algorithms:

- **Fuzzy clustering algorithm**
  - ✓ Allow to represent the real world, often characterised by unclear boundary among clusters, in a more realistic way
  - ✓ More computationally efficient because dramatic changes in the cluster membership are less likely to occurr
  - ✓ Less affected by local optima
  - ✓ Allow to cope with uncertainty in the assignement of each unit to a cluster
- **PAM**
  - ✓ Define the prototype of each cluster as an actually observed unit. This prototype is called **Medoid**

The algorithm can be formalized as the following minimisation problem:

$$\min_{u_{ic}} \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^{p} d_{ic}^{2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_c)$$

$$\text{Subject to } \sum_{c=1}^{C} u_{ic} = 1 \text{ and } u_{ic} \geq 1$$

Where:

- u_{ic} is the membership degree of the *i*-th unit to the *c*-th cluster ($c = 1,\dots, C$);
- $p > 1$ is a weighting exponent that controls the fuzziness of the obtained partition (the more p is near to 1 the more the partition is closer to a crisp one)
- $d_{ic}^{2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_c)$: trapezoidal fuzzy distance between *i*-th unit and *c*-th medoid

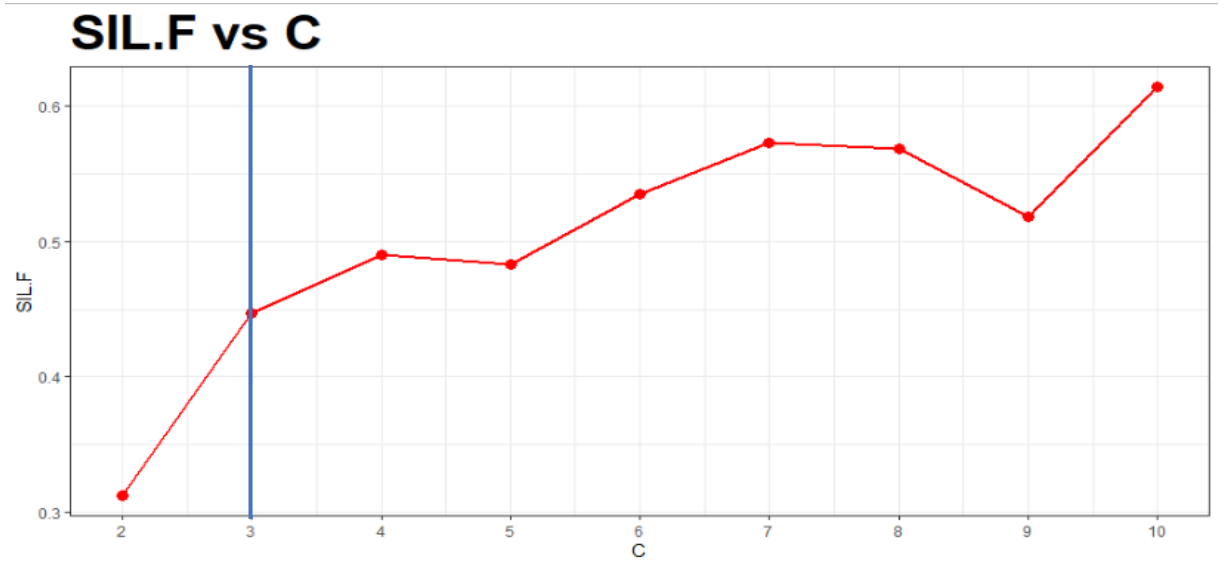
Step 4: Cluster validation

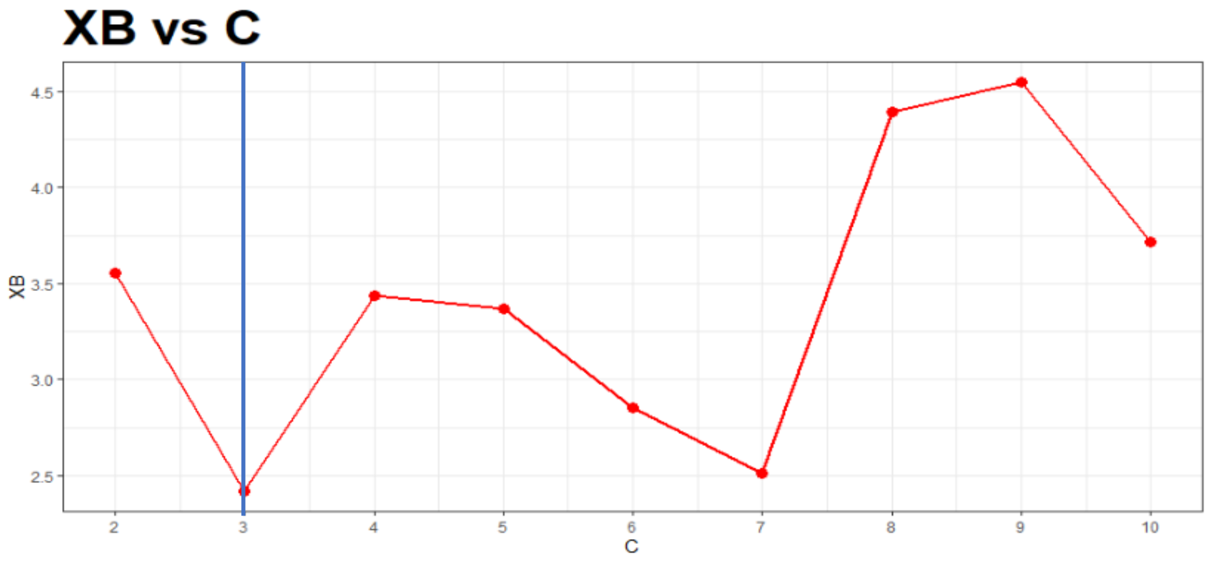
To select the best partition, (the optimal number of clusters) two complex indicators have been adopted:

- **Fuzzy Silhouette (SILF)** – to be maximized (at least locally)
- **Xie-Beni (XB)** – to be minimized (at least locally)

Both SILF and XB measure the **separability** between the clusters and the **homogeneity** of units inside the same cluster:

The SILF and XB for $C \in [2, 10]$ have been computed. From the visual inspection of the graphs, the optimal partition is 3 clusters.





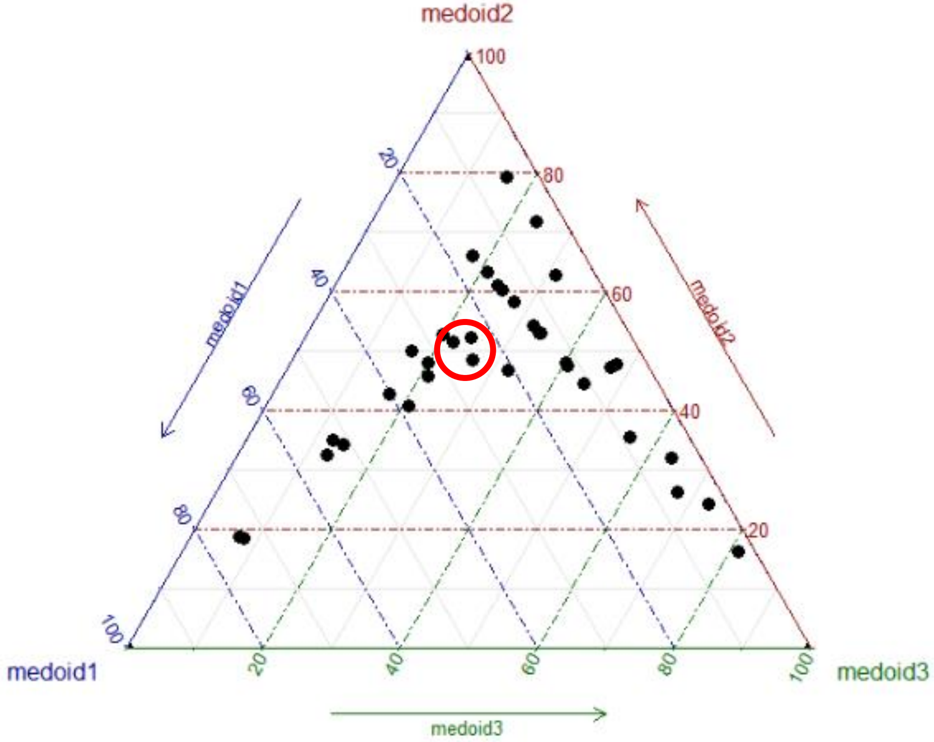
Step 5: Clustering results

The final number of clusters is 3

The distribution of each customer in the three groups is represented in the ternary plot graph.

Note that:

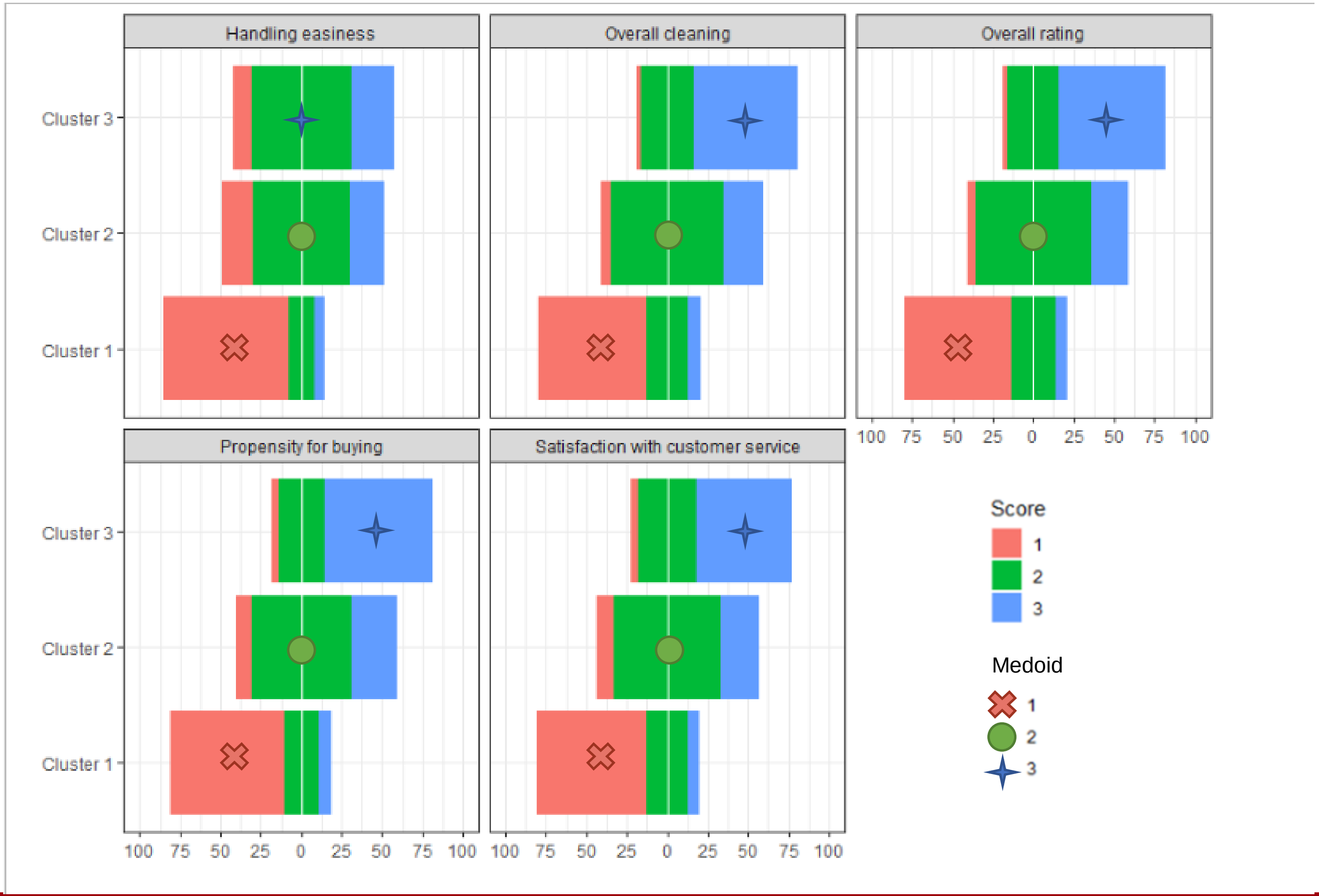
- each consumer is represented by three values (membership degrees) which represent how much is/she belongs to the three clusters.
- the medoids (representatives of each cluster) are visualised at the vertices of the triangle.



These customers almost equally belong to all three clusters, meaning that they are not well classified.

Step 5: Labeling

Per each KPI, the weighted proportion of each score (1-3) by cluster is visualised. Note that the membership degree is used as weight.



Cluster 1 groups consumers with the highest proportion of 1 = «Dissatisfied»

Cluster 2 groups consumers with the highest proportion of 2 = «Satisfied»

Cluster 3 groups consumers with the highest proportion of 3 = «Completely satisfied»

Step 5: Profiling - FML model

Fuzzy C-medoid clustering algorithm

OUTPUT

u_{ic} : membership degree for the i-th unit to the c-th cluster constrained to:

- $0 \leq u_{ic} \leq 1$
- $\sum_{c=1}^{C} u_{ic} = 1$

Aim:
predict the probability to belong to each cluster based on the Drivers

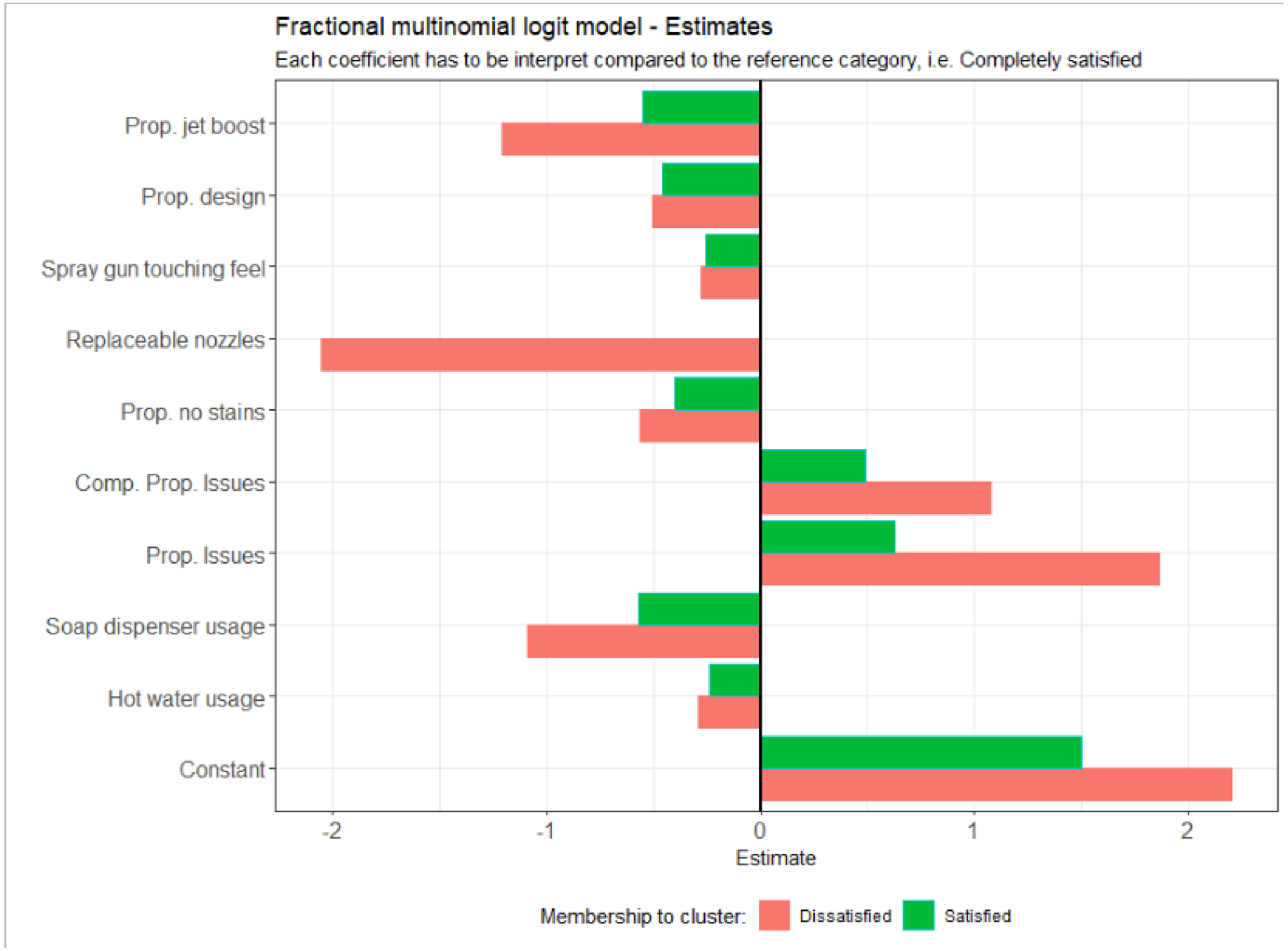
Method:
Since the dependent variables are the memberships obtained through the cluster analysis, i.e. variables define in the interval [0-1], the **Fractional multinomial logit** (FML) model has been adopted:

$$P(u_{ic}|X_i) = G(X_i, \beta_c) = \frac{e^{X_i \beta_c}}{\sum_{c=1}^{C} e^{X_i \beta_c}} \quad , \qquad c = 1, \dots, C$$

where

- X_i represents the vector of independent variables observed for the i-th unit;
- β_c represents the estimated vector of model parameters for the c-th cluster.

Step 5: FML results



Higher the **proportion of issues encountered more often compared to usual product**, higher the probability to belong to the unsatisfied group compared to the completely satisfied group.

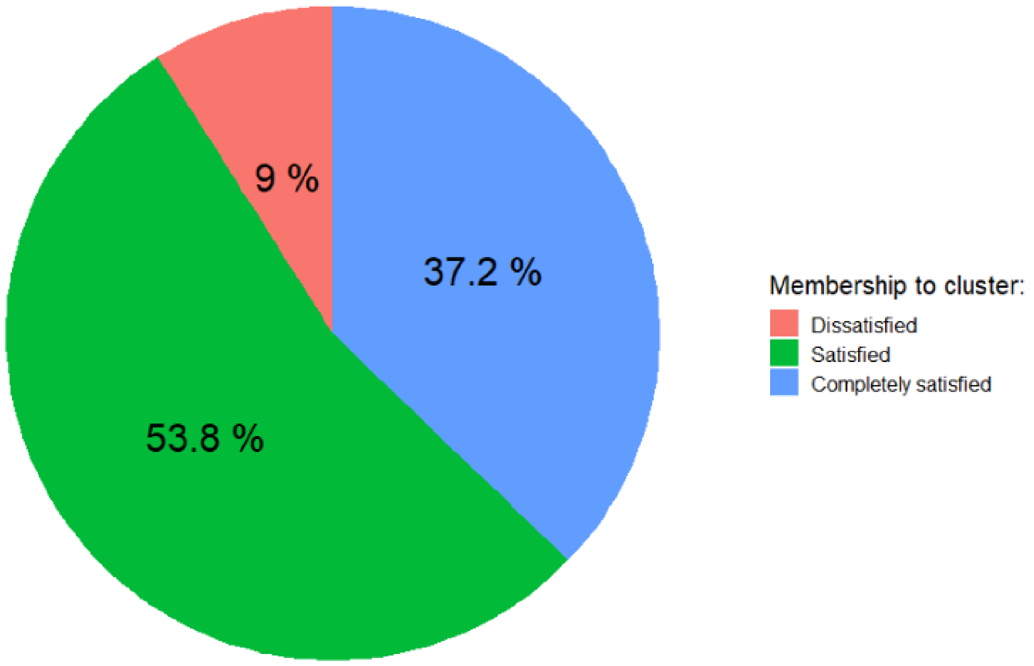
The presence of **replaceable nozzles** (instead of adjustable), lower the probability to belong to the unsatisfied group compared to the completely satisfied group.

Note: Only variables with estimated coefficients significantly different from zero are reported.

Prediction

What is the estimated probability to belong to each cluster?

Variable	Mean value
Hot water usage	0
Soap dispenser usage	1
Prop. Issue	0.247
Comp. Prop. Issue	0.503
Prop. no stains	0.573
Replaceable nozzles	1
Spray gun touching feel	0
Prop. design	0.830
Prop. jet boost	0.611



Note: These results are obtained considering a fictituous person having mean values for each driver.

References

- Coppi, R., D'Urso, P., Giordani, P., (2012) "Fuzzy and possibilistic clustering for fuzzy data". *Computational Statistics & Data Analysis* 56 (4) 915–927.
- D'Urso, P. (2007). Clustering of fuzzy data. In J. V. De Oliveira, & W. Pedrycz (Eds.). *Advances in fuzzy clustering and its applications* (pp. 155–192). J. Wiley and Sons.
- D'Urso, P., De Giovanni, L., Disegna, M., & Massari, R. (2013). "Bagged Clustering and its application to tourism market segmentation". *Expert Systems with Applications*, 40, 4944–4956.
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). "Measurement equivalence in cross-national research". *Annual Review of Sociology*, 40, 55–75.
- Disegna M., D'Urso P., Massari R. (2018) "Analysing cluster evolution using repeated cross-sectional or- dinal data". *Tourism Management*, 69, 524-536.
- Dubois, D., & Prade, H. (1988). *Possibility theory*. New York: Plenum press
- Sii, H. S., Ruxton, T., Wang, J., (2011) "A fuzzy-logic-based approach to qualitative safety modelling for marine system". *Reliability Engineering and System Safety*, 73, 19-34.

Questions? Comments?

Thanks for your attention!

Nicolò Biasetton
nicolo.biasetton@phd.unipd.it

Marta Disegna
disegnam@bournemouth.ac.uk

Luigi Salmaso
luigi.salmaso@unipd.it

Other formulas

Fuzzy Silhouette index (FS):

$$FS = \frac{\sum_{i=1}^{N}\left(u_{ri} - u_{qi}\right)^{\alpha}\lambda_i}{\sum_{i=1}^{N}\left(u_{ri} - u_{qi}\right)^{\alpha}}, \qquad \lambda_i = \frac{(b_i - a_i)}{max\{b_i, a_i\}}$$

Where:

- a_i is the average distance between the *i*-th unit and the units belonging to the *r* cluster (with which *i* is associated with the highest membership degreee)
- b_i is the minimum (over cluster) average distance of the *i*-th unit to all units belonging to the cluster *q* with $q \neq r$
- r, q are respectively the first- and second-best cluster (accordignly to membership degree) to which *i*.th unit is associated
- $\left(u_{ri} - u_{qi}\right)^{\alpha}$ is the weoght of each λ_i, calculated upon the fuzzy matrix **U** (u_{ri}, u_{qi} are the first and second largest element of the i-th column)
- α is an optional user defined weighting coefficient

The higher the value of FS, the better the assignment og the units to the clusters simultaneously obtaining the minimisation of the intra-cluster distance and the maximisation of the inter-cluster distance

Xie-Beni index (XB)

$$XB = \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ic}^{p} d_{ic}^{2}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_c)}{N {d^2}_{min}}$$

Where:

${d^2}_{min} = min_{c=i,j}\left\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\right\|$ is the minimum distance between cluster centroids.

The more separate the clusters, the larger the ${d^2}_{min}$ and the smaller the *XB* index.

The lower the value of XB, the better the assignment og the units to the clusters simultaneously obtaining the minimisation of the intra-cluster distance and the maximisation of the inter-cluster distance