Shangai Data: Case study
Deliverable 1

Toni Ballester
Anna Berruezo
ADEI. 2014-2015 Q1

Outline

En aquest estudi tractarem la mostra "Shangai Data", una mostra de 489 individus que han valorat diferents aspectes de les instal·lacions d'un allotjament. Aquests aspectes es poden valorar mitjançant diferents variables quantitatives o qualitatives.

L'estudi es divideix en dues parts: un PCA i un MCA, tots dos enriquits per clustering. Tots els càlculs i els gràfics que es presentaran a continuació s'han realitzat amb l'ajuda del programari R.

Aplicació del PCA enriquit per clustering

Inèrcia dels eixos. Quins factors retenim?

	eigenvalue
comp 1	6.58
comp 2	1.38
comp 3	0.93
comp 4	0.71
comp 5	0.62
comp 6	0.60
comp 7	0.47
comp 8	0.43
comp 9	0.38
comp 10	0.32
comp 11	0.24
comp 12	0.20
comp 13	0.15

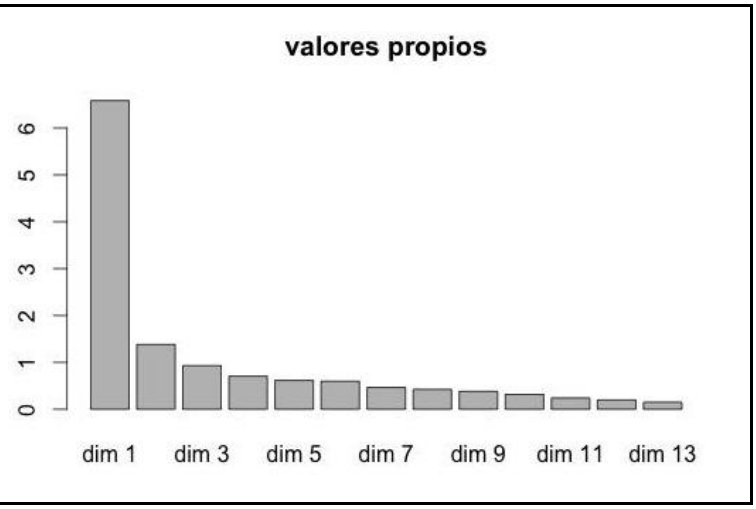

Figura 1. Eigenvalues: valors propis

En un PCA estandarditzat la inèrcia mitjana dels valors propis és 1, per tant només hem de mantenir aquells valors propis superiors a 1. En conseqüència, si observem la Figura 1, podem deduir que centrarem aquest estudi en les dues primeres dimensions. Ara bé, la dimensió 3 té una valor molt proper a 1, les seves dades podrien arribar a ser significatives i aportar informació útil. No descartem utilitzar aquesta tercera dimensió en algun moment.

Interpretació dels factors i dels eixos

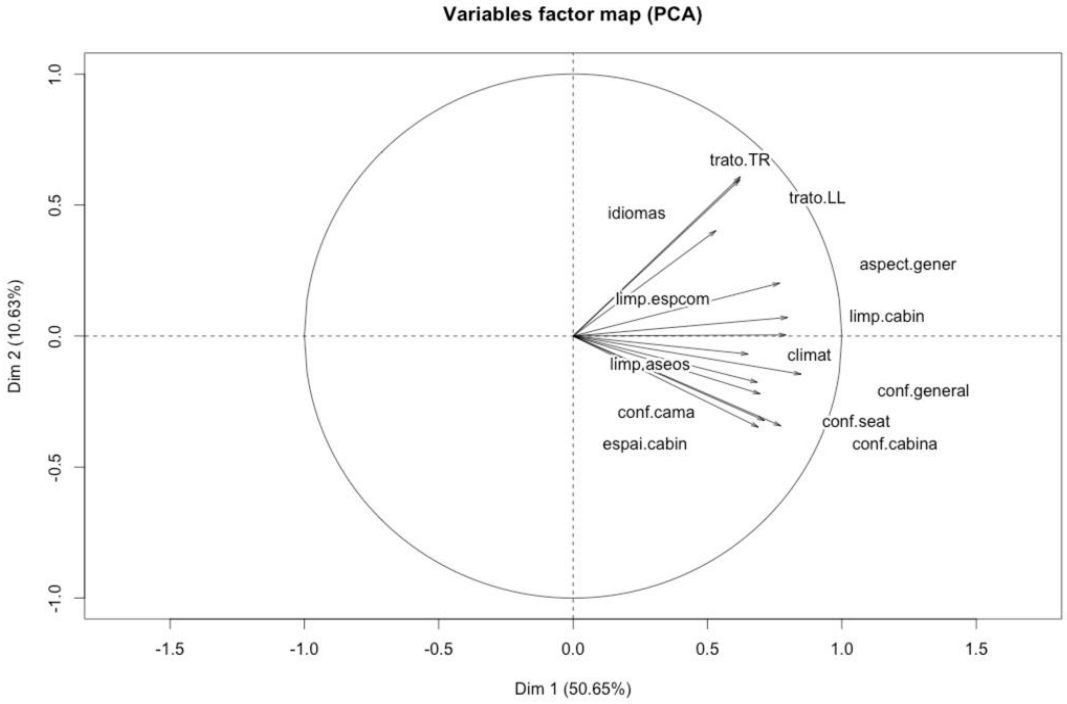

Figura 2. PCA. Mapa factorial de les variables en funció de les dimensions 1 i 2.

Les dimensions són tendències de puntuacions que han fet els diferents individus.
A partir de les figures 2 i 3 podem afirmar que la dimensió 1 en general té bones puntuacions, sobretot respecte el confort i la neteja. En canvi, a les altres dues dimensions podem observar puntuacions negatives entre algunes de les variables. A la dimensió 2 es puntua positivament el tracte però negativament l'espai i el confort a cabina. A la dimensió 3 es puntua positivament el confort i l'espai a cabina i negativament la neteja.
A més, a la Figura 2 podem observar que les variables conf.seat i conf.cabina presenten una correlació alta. Les variables trato.TR i idiomas també.

```
$Dim.1$quanti
             correlation
conf.general   0.8481471
limp.espcom    0.7985376
limp.cabin     0.7920013
conf.cabina    0.7726858
aspect.gener   0.7683819
conf.seat      0.7114754
conf.cama      0.6958620
espai.cabin    0.6888139
limp.aseos     0.6856033
climat         0.6511372
trato.TR       0.6219750
trato.LL       0.6204638
idiomas        0.5312163
```

```
$Dim.2$quanti
             correlation
trato.TR       0.6076557
trato.LL       0.5966692
idiomas        0.4026411
aspect.gener   0.2021238
conf.general  -0.1458578
limp.aseos    -0.1766384
conf.cama     -0.2208818
conf.seat     -0.3225243
conf.cabina   -0.3429921
espai.cabin   -0.3477319
```

```
$Dim.3$quanti
             correlation
espai.cabin    0.4112225
conf.cabina    0.2820723
trato.LL       0.2484574
trato.TR       0.2363126
conf.cama      0.1639467
climat         0.1075416
aspect.gener  -0.1481324
limp.aseos    -0.3400374
limp.cabin    -0.4291976
limp.espcom   -0.4473558
```

Figura 3. Taula amb els valors de correlació de les dimensions 1, 2 i 3.

Tal i com hem havíem dit a l'inici d'aquest estudi, la dimensió 3 ens pot aportar informació rellevant. En aquest cas en relació a les variables de neteja. A la Figura 6 podem observar que la dimensió 3 té una tendència cap a la insatisfacció respecte la neteja en general.

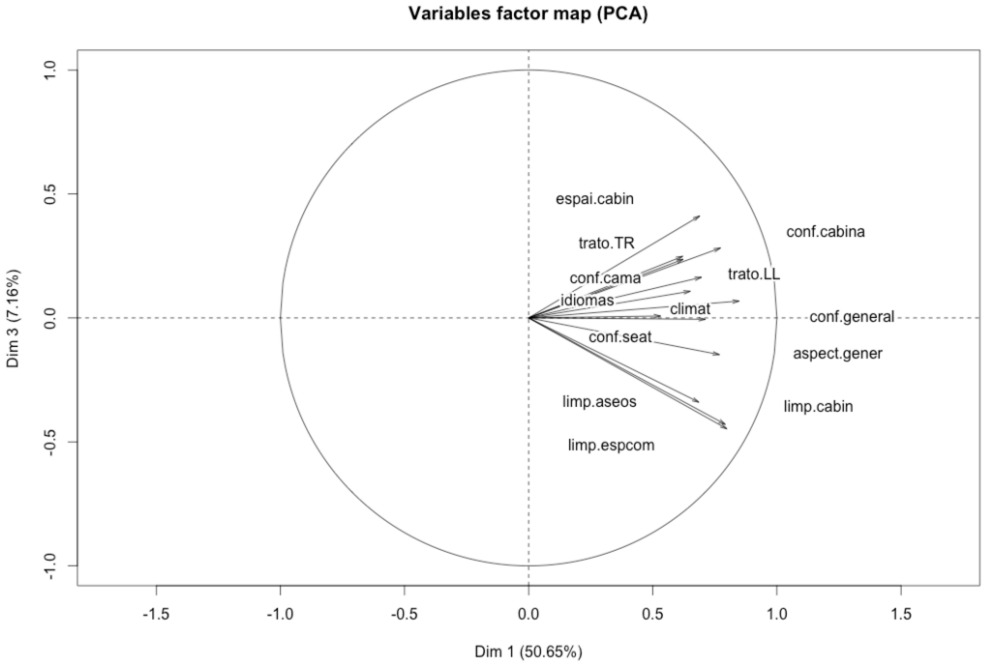

Figura 4. PCA. Mapa factorial de les variables a les dimensions 1 i 3.

Punts de vista dels individus

A partir de la Figura 5 podem observar que la gran majoria d'individus de la mostra estan més o menys centrats amb una lleugera tendència cap a satisfets. Els que estan concentrats a la part dreta del gràfic són els més satisfets en tots els aspectes. En canvi, els que estan dispersos per la part esquerra del gràfic són els menys satisfets, moltes vegades per aspectes diferents, és a dir, que han puntuat malament determinats aspectes.

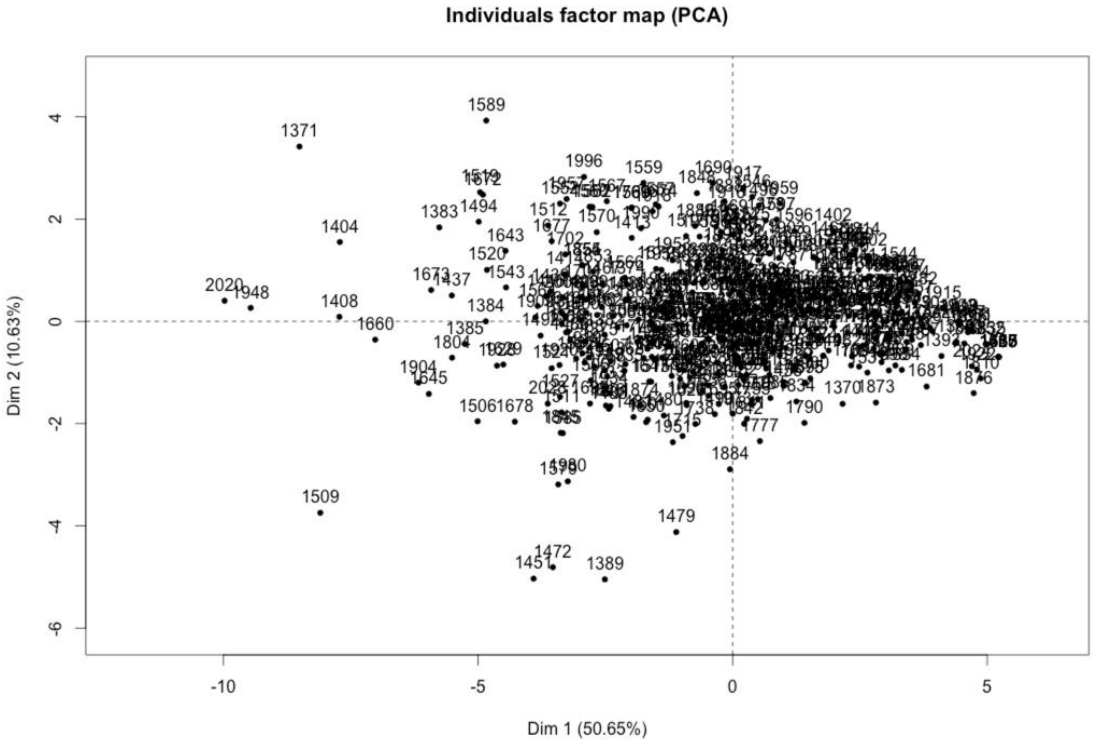

Figura 5. PCA. Mapa factorial dels individus a les dimensions 1 i 2.

Hem seleccionat els individus situats als extrems dels eixos vertical i horitzontal.

D'una banda tenim els més allunyats de l'eix vertical: el més alt, 1589, i el més baix, 1451. El primer ha puntuat molt baix les variables de confort i comoditat i molt alt les variables de tracte, aspecte general i idiomes. El segon ha puntuat contràriament al primer.

D'altra banda tenim els més allunyats de l'eix horitzontal: el més a la dreta, el 1810, i el més a l'esquerra, el 2020. El primer ha puntuat màximament totes les variables. El segon ha puntuat negativament la majoria de les variables.

Individu: 1451		Individu: 1589		Individu: 2020		Individu: 1810	
confort.general	5	confort.general	2	confort.general	0	confort.general	10
confort.cabina	8	confort.cabina	2	confort.cabina	0	confort.cabina	10
espacio.cabina	6	espacio.cabina	2	espacio.cabina	0	espacio.cabina	10
comodidad.cama	8	comodidad.cama	1	comodidad.cama	1	comodidad.cama	10
comodidad.asie	8	comodidad.asie	1	comodidad.asie	0	comodidad.asie	10
climatizacion	5	climatizacion	1	climatizacion	0	climatizacion	10
conser.limp.es	2	conser.limp.es	5	conser.limp.es	3	conser.limp.es	10
conser.lim.cab	6	conser.lim.cab	5	conser.lim.cab	2	conser.lim.cab	10
conser.lim.ase	4	conser.lim.ase	5	conser.lim.ase	3	conser.lim.ase	10
trato.acogida	3	trato.acogida	10	trato.acogida	5	trato.acogida	10
trato.trayecto	2	trato.trayecto	10	trato.trayecto	2	trato.trayecto	9
aspectos.gener	7	aspectos.gener	8	aspectos.gener	3	aspectos.gener	9
conoc.idiomas	2	conoc.idiomas	8	conoc.idiomas	8	conoc.idiomas	10

Figura 6. Taula amb les puntuacions d'individus als extrems dels eixos.

Distància entre els nivells de satisfacció

La figura 7 ens mostra que les diferents categories de la variable satisfacció formen una línia gairebé recta seguint l'eix horitzontal. A més, estan ordenades de menys a més satisfeta. Per tant, podem deduir que les respostes han estat coherents.

Podem observar una gran diferència entre la variable "nada" i la variable "poco". Aquest fet ens indica que són variables molt diferenciades i que els individus que es troben a "nada" realment estan molt insatisfets. Ara bé, per contra trobem que les categories "poco" i "regular" estan una mica més juntes. Si estiguessin una mica més unides podríem arribar a dir que descriuen el mateix nivell satisfacció tot i ser categories diferents.

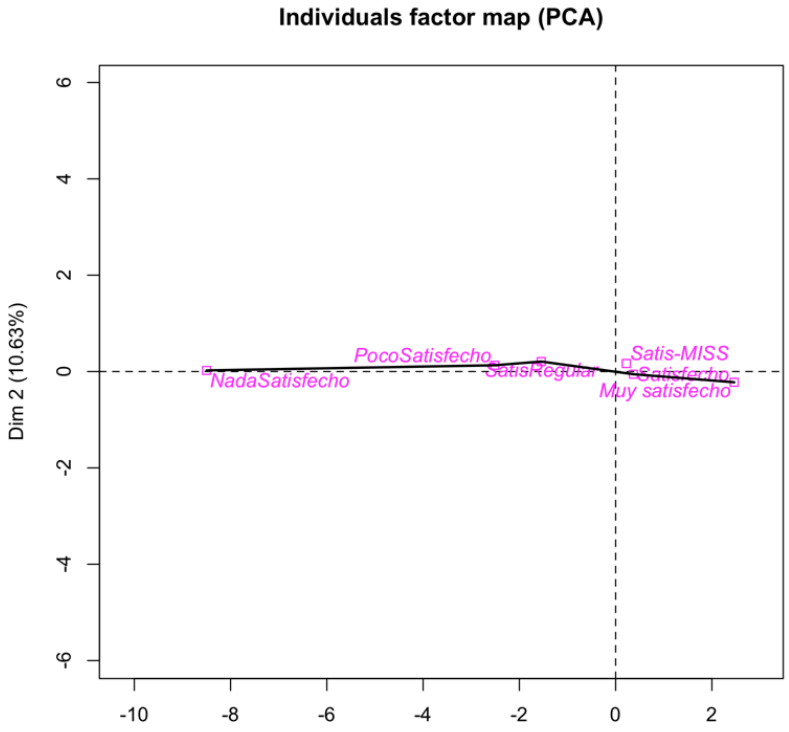

Figura 7. PCA. Mapa factorial de les categories de satisfacció a les dimensions 1 i 2.

Síntesi del PCA a través del clustering

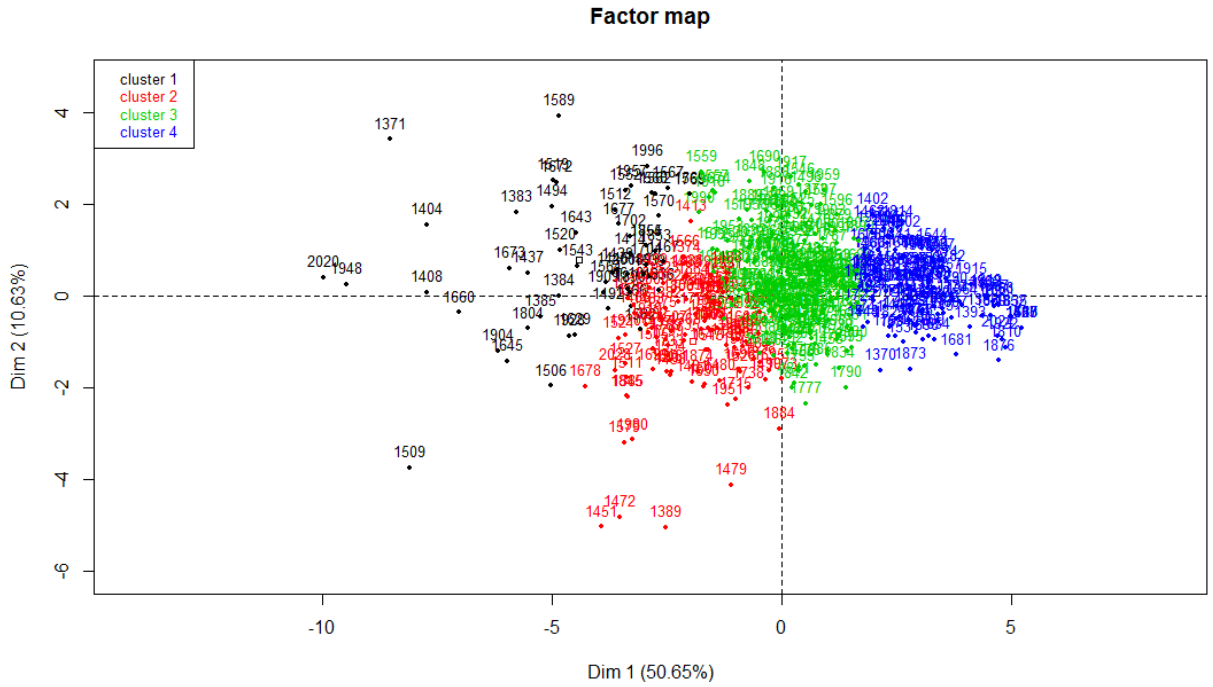

Figura 8. PCA. Mapa factorial dels individus en funció de les dimensions 1 i 2, classificats per clustering.

Tenim 4 clústers amb el següent nombre de components: clúster 1 (97), clúster 2 (141), clúster 3 (121), clúster 4 (130).

```
Cluster 1
          Mean in category  Overall mean
trato.LL          1.998922      8.207039
conf.seat         4.158652      6.194617
trato.TR          6.679624      8.307851
conf.cama         4.882927      6.821978
conf.cabina       4.597938      6.628337
conf.general      4.896907      7.018443
mean.SAT          5.126571      7.166966
```

Al clúster 1 trobem el grup dels insatisfets (categories nada i poco). Ho han puntuat tot per sota de la mitjana. Estan especialment insatisfets amb el tracte i el confort.

```
Cluster 2
            Mean in category Overall mean
espai.cabin     5.801418       5.465164
mean.SAT        6.938419       7.166966
```

El clúster 2 també està format majoritàriament per gent insatisfeta però amb puntuacions no tant per sota de la mitjana com al clúster 1. A més, han puntuat positivament l'espai de la cabina.

```
Cluster 3
             Mean in category Overall mean
trato.TR         9.250478      8.307851
trato.LL         9.133942      8.207039
idiomas          8.687140      7.959233
aspect.gener     8.395084      7.978355
espai.cabin      4.896406      5.465164
conf.seat        5.588385      6.194617
conf.cabina      6.104366      6.628337
```

El clúster 3 està format per gent majoritàriament satisfeta, sobretot amb el tracte, els idiomes i l'aspecte general. Ara bé, han puntuat malament l'espai de la cabina i la comoditat.

```
Cluster 4
            Mean in category  Overall mean
mean.SAT         8.834925      7.166966
```

El clúster 4 està format per gent satisfeta o molt satisfeta. Ho han puntuat tot per sobre de la mitjana.

Aplicació del MCA enriquit per clustering

Inèrcia dels eixos. Quins factors retenim?

	eigenvalue
dim 1	0.56
dim 2	0.35
dim 3	0.26
dim 4	0.19
dim 5	0.14
dim 6	0.13
dim 7	0.12
dim 8	0.11
dim 9	0.11
dim 10	0.11

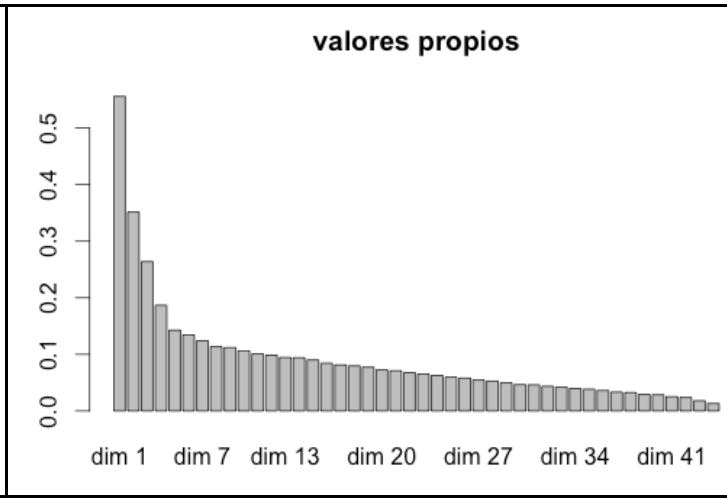

Figura 9. Eigenvalues: valors propis

A partir de la Figura 9 i de la norma del colze, tindrem en compte les 4 primeres dimensions durant aquesta part de l'estudi, el MCA. A partir de la quarta dimensió la diferència entre els diferents valors és molt petita i per tant aportaran informació molt menys rellevant que les anteriors.

Interpretació dels factors

Després d'imputar valors als nuls (valors NA de la mostra) tenint en compte els 2 eixos suggerits per l'estimació de R i obtenir una nova taula de valors, discretitzem les nostres dades a partir dels punts de tall deduïts dels quartils de la mitjana de satisfacció.

Ara podem començar a analitzar el nostre estudi MCA. Per començar, podem observar a la Figura 10 observar que el núvol de punts-individus té una repartició bastant regular sobre el pla factorial. Això vol dir que reflecteix tendències compartides per grups grans d'individus.

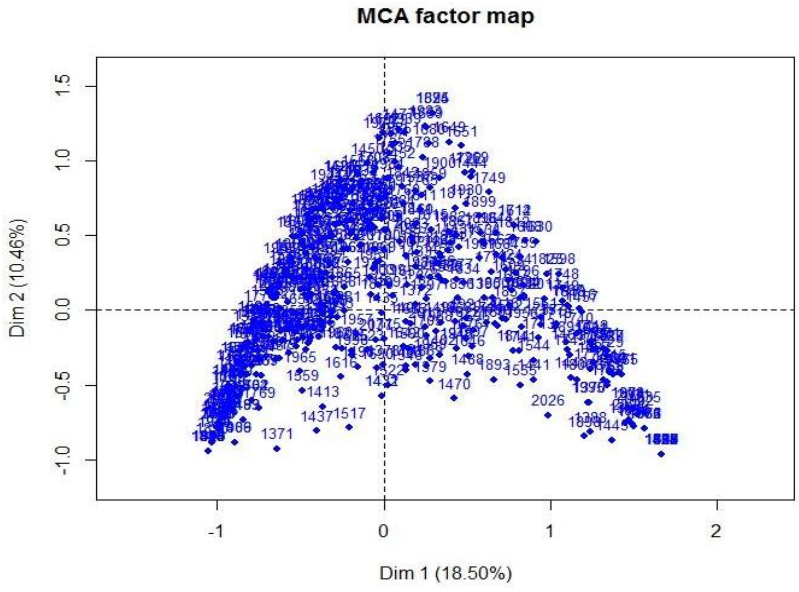

Figura 10. MCA. Mapa factorial dels individus en funció de les dimensions 1 i 2.

També podem observar que es produeix l'efecte Gutmann. Aquest efecte indica que hi ha un fort factor d'escala entre les categories de les variables. Es manifesta per la forma que pren el núvol de punts, gairebé parabòlica. Això és bo ja que ens diu que reflecteix tendències generals i no específiques.

A la Figura 11 podem observar la distribució de la puntuació de les diferents variables en funció de la satisfacció. Com més a la dreta es troben els punts, més disperses estan les puntuacions d'aquella variable i més associació tenen amb la satisfacció. Com més a l'esquerra es troben els punts, més comprimides estan les puntuacions d'aquella variable i menys associació tenen amb la satisfacció.
No s'observa relació entre les variables categòriques i les suplementàries.

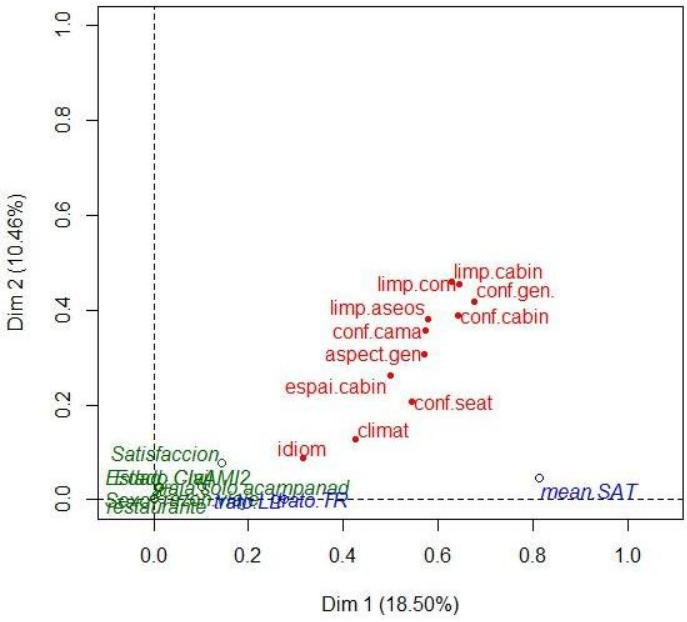

Figura 11. MCA. Mapa factorial de la distribució de les variables quantitatives i qualitatives en funció de les dimensions 1 i 2.

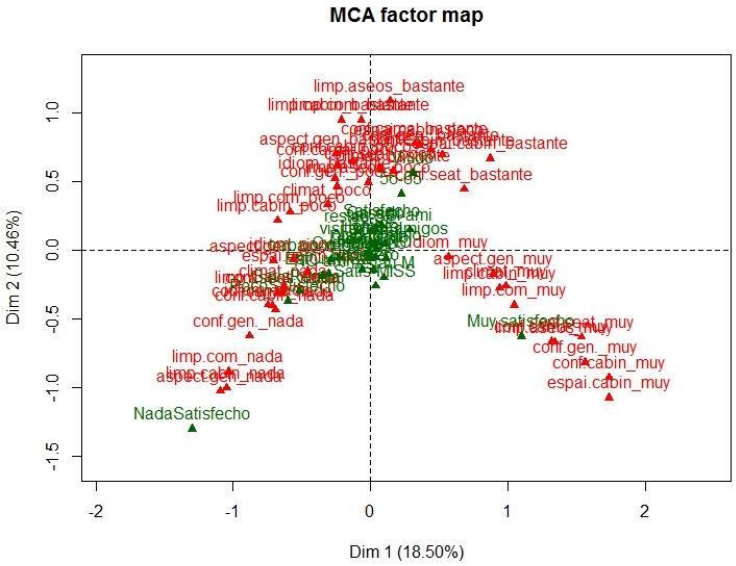

Figura 12. MCA. Mapa factorial de la distribució de les categories de les variables quantitatives i les categories de la variable de satisfacció en funció de les dimensions 1 i 2.

A la Figura 12 podem observar que la distribució de les categories de les variables quantitatives i les categories de la variable de satisfacció tenen formes parabòliques similars. Per tant, tornem a tenir un indicador de que es produeix l'efecte Gutmann.

Síntesi del MCA a través dels clusters

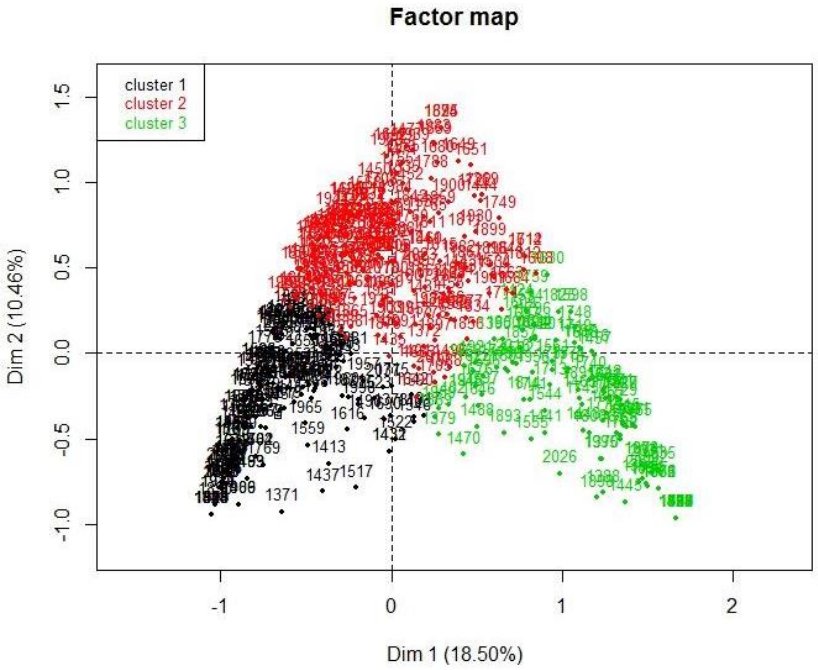

Figura 10. MCA. Mapa factorial dels individus en funció de les dimensions 1 i 2, classificats per clustering.

Tenim 3 clústers amb els següents nombres d'individus: clúster 1 (183), clúster 2 (194) i clúster 3 (112).

Descripció dels clústers

El clúster 1 agrupa els individus insatisfets. Han puntuat negativament per sota de la mitjana global, especialment el tracte.

El clúster 2 agrupa individus que han puntuat per sobre de la mitjana global.

El clúster 3 agrupa individus que han puntuat positivament per sobre de la mitjana global, especialment el tracte.

Conclusions globals

Una primera aproximació a les dades de la mostra fa pensar que tots els individus estan raonablement satisfets amb l'experiència (mitjana de satisfacció de 7.167). Ara bé, un estudi més profund ha revelat que existeixen tendències en les puntuacions. Els individus puntuen determinats aspectes positiva o negativament en funció de les seves característiques personals com l'edat o la raó per la que viatgen. A més, hem pogut observar que si una tendència (dimensió) és molt forta, influeix molt en la resta i moltes vegades les difumina.

Deixant de banda l'estudi en sí, la realització d'aquest treball ens ha permès fer un ús més extens de R del que havíem fet fins ara i aprendre una mica més del seu potencial, per exemple, amb l'ús de paquets addicionals com FactoMineR i MissMDA.