Facultat d'Economia i Empresa

Facultat de Matemàtiques i Estadística

Treball Final de Grau

Desenvolupament d'un sofware per l'explotació de bases de dades periòdiques. Dades del CIS.

Grau d'Estadística

Autor: Ernest Orriols Bertran

Directors: Dr. Josep M Oller i Dr. Esteban Vegas

Convocatòria: Setembre 2018

Resum

Cada any el CIS elabora una enquesta periòdica amb el títol *Opinión pública y política fiscal* on demana als entrevistats una sèrie de preguntes sobre la seva persona, com per exemple: edat, nivell socioeconòmic, estatus, etc. També demana l'opinió que té l'entrevistat sobre els impostos i com aquests els administra l'Estat. Aquesta enquesta s'ha fet durant els últims vint-i-un anys, és a dir, tenim una base de dades que canvia en el temps, una sèrie temporal. L'objectiu principal d'aquest treball és implementar un sistema informàtic interactiu que faciliti l'explotació d'aquesta base de dades mitjançant les eines estadístiques més elementals i altres més avançades utilitzant tècniques pròpies d'anàlisi multivariant.

Paraules clau: *CIS*, *ACP*, *Sèries Temporals*, *Dashboard*, *Shiny*.

Abstract

Each year the CIS draws up a periodic survey under the title *Opinión pública y política fiscal* where he asks the interviewees a series of questions about his person, such as for example: age, socio-economic status, status, etc. He also asks for his opinion the interviewee about taxes and how these are administered by the State. This survey It has been done for the last twenty-one years, that is, we have a database that changes in time, a temporary series. The main objective of this work is to implement a Interactive computer system that facilitates the exploitation of this database through the most basic statistical tools and other more advanced using own techniques of multivariate analysis..

Keywords: *CIS*, *PCA*, *Time Series*, *Dashboard*, *Shiny.*

Índex

Índex de figures

Notacions

- *ACP* Anàlisi de Components Principlas
- *CIS* Centro de Investigaciones Sociológicas
- *N.C.N.S* No contesta, no sabe
- ψ Component principal

Capítol 1

INTRODUCCIÓ

El present document constitueix la Memòria del Treball de Final del Grau impartit per la Universitat de Barcelona i la Universitat Politècnica de Catalunya. En aquesta primera part introductòria, s'exposarà la justificació, els objectius i la metodologia emprada al llarg d'aquest treball, juntament amb la seva estructura.

En els últims anys el CIS, *Centro de Investigaciones Sociológicas*, ha elaborat molts tipus d'estudis mitjançant enquestes, un d'aquests estudis és *Opinion pública y política fiscal*. Aquest estudi és anual, és a dir, es repeteix cada any. Un cop depurades i estandarditzades aquestes bases de dades associades a cada estudi obtindríem una gran base de dades periòdica. A partir d'aquesta i mitjançant el sofware R obtindríem un *Dashboard* on l'usuari tindria accés a un programa interactiu per explotar la base de dades.

1.1 Justificació

Saber l'opinió de la societat sempre és útil per encarar els diferents problemes de la millor manera possible, en aquest cas en concret en la recaptació i gestió dels diferents impostos. Com he dit abans la característica principal d'aquest estudi radica en treballar amb una gran base de dades periòdica, bàsicament podríem dir que és una sèrie temporal. Aquest fet obre la porta per veure l'evolució en el temps de les diferents variables.

Cal afegir que les eines que ofereix el CIS en la seva pàgina web són bastant limitades car només hi ha taules de freqüència d'una o dues variables i cap gràfic, a més no es poden veure les variables com canvien a través del temps.

1.2 Objectius

El principal objectiu d'aquest treball és crear una aplicació que permeti l'explotació de la base de dades perquè l'usuari final la utilitzi de la forma que li convingui més, usant

tècniques estadístiques bàsiques fins altres més elaborades com l'anàlisi de components principals (*ACP*). Per arribar aquest objectiu s'han de complir certs requisits intermedis.

Els citem a continuació:

- Depurar i estandarditzar tota la base de dades resultant.
- Elaborar gràfics interactius com gràfics mosaic, amb la seva respectiva taula de freqüència, i diagrames de sectors.
- Utilitzar l'anàlisi de components principals a través del temps.
- Usar sèries temporals per veure com les variables canvien amb el temps.

1.3 Metodologia

La primera part del treball s'ha dedicat a la depuració i normalització de la base de dades periòdica. Recalcar que totes les enquestes amb la seva respectiva base de dades s'han obtingut del CIS [1].

Un cop obtinguda aquesta gran base de dades periòdica s'ha procedit a crear l'eina interactiva a través de la construcció d'un *dashboard*. Aquest està operatiu en el lloc web següent `https://orri.shinyapps.io/DASHBOARDMILLORA/` on podem veure el *Dashboard* final. Cal tenir en compte que l'ús de les eines interactives com *dashboard* o *shiny* no s'estudien de forma extensiva al Grau d'Estadística, per tant ha calgut una gran recerca i temps per aprendre a utilitzar aquestes eines. A l'hora d'aplicar l'ACP a través del temps he utilitzat la diferent documentació que els meus tutors m'han facilitat.

1.4 Estructura

La memòria següent d'aquest treball consta de cinc capítols, incloent-hi aquest, la introducció.

En el segon capítol s'explica els passos seguits per la depuració i estandarització de la base de dades resultant.

En el tercer capítol s'exposa el *Dashboard* final, amb el seu corresponent manual d'utilització i exemples.

En el quart capítol, es presenta l'ús pràctic de la aplicació.

I finalment en el cinquè capítol presento les conclusions finals del treball.

Capítol 2

PREPROCESSAMENT I DESCRIPCIÓ DE LA BASE DE DADES

2.1 Programa utilitzat

Per poder dur a terme aquest treball he utilitzat només un sol *Software*, el *Software* en qüestió ha estat R amb el seu respectiu llenguatge per la importació de dades i la seva posterior normalització. També he utilitzat els paquets *flexdashboard* i *Shiny* per crear el *dashboard* final.

2.2 CIS

Totes les bases de dades utilitzades les he extret del CIS, *Centro de Investigaciones Sociológicas*, el CIS és una institució creada el 1962, el seu objectiu principal és l'estudi científic de la societat espanyola mitjançant la realització d'enquestes. En el nostre cas l'enquesta estudiada és *Opinión pública y política fiscal*, on l'entrevistador pregunta un seguit de qüestions sobre la gestió, administració i la recaptació d'impostos, aquesta enquesta s'ha fet periòdicament des del 1997 on l'univers d'aquesta enquesta és la població espanyola de 18 o més anys. Tanmateix afegir que cada enquesta té aproximadament 2500 individus.

2.3 Depuració i estandardització

Preparar la base de dades periòdica per fer el posterior *dashboard* ha sigut la part més llarga de tot el treball, ja que, havia d'estandarditzar cada base de dades corresponent a cada enquesta una per una. Aquesta normalització consistia en eliminar variables redundants o que he considerat que no eren útils, canviar o simplificar els nivells de les variables, com per exemple el nivell *N.C.N.S*, aquesta és una categoria que agrupa les

categories inicials de no contesten (N.C.) i no saben (N.S.) degut a que són categories minoritaries. Finalment he canviat els noms de les variables. A més, a mesura que les enquestes eren més recents s'introduïen preguntes noves, aquestes preguntes no estaven en les anteriors enquestes i, per tant les vaig haver d'esborrar, perdent així una variable. Mitjançant la funció *rbind* de R he ajuntat totes les bases de dades per formar la base de dades periòdica, aquesta base de dades final conté 51633 individus i 49 variables, el període temporal de l'estudi comprèn des de 1997 fins al 2017

2.4 Definició de les variables

- **ESTU** Estudi en qüestió, l'any en que es va fer el estudi.
- **CCAA** Comunitat autònoma de l'entrevistat.
- **TAMUNI** Tamany del municipi de l'entrevistat.
- **FREQ** Amb quina freqüència parla l'entrevsitat sobre els serveis públics: "Mucho", "Bastante", "Poco", "Ninguno"o "N.C".
- **ENSENYAMENT** En quina mesura molt, bastant, poc o gens funciona l'ensenyament.
- **SANITAT** En quina mesura molt, bastant, poc o gens funciona la sanitat.
- **JUSTICIA** En quina mesura molt, bastant, poc o gens funciona la justicia.
- **SERVEISSSOCIALS** En quina mesura molt, bastant, poc o gens funcionen els serveis socials.
- **TRANSPORT** En quina mesura molt, bastant, poc o gens funciona el transport públic.
- **FRASE** Variable categòrica la qual reflexa l'opinio de l'entrevistat sobre els impostos. pot prendre tres valors:
  - 1->Los impuestos son un medio para redistribuir mejor la riqueza en la sociedad.
  - 2->Los impuestos son algo que el Estado nos obliga a pagar sin saber muy bien a cambio de qué.
  - 3->Los impuestos son necesarios para que el Estado pueda prestar servicios públicos.
- **RESPIMP** Segons l'entrevistat els españols som responsables i consients amb els impostos pot prendre aquests valors:"Muy","Bastante","Poco","MuyPoco"o "N.C.N.S.".

- **RESPIMPPER** Com es considera el propi entrevistat en termes de responsabilitat i consciència respecte els impostos, pot prendre aquests valors:"Muy","Bastante", "Poco","MuyPoco"o "N.C.N.S.".
- **SOCIETAT** Variable que indica si la societat es beneficia molt, bastant, poc o gens dels impostos que retornen en forma de serveis.
- **PERSONAL** Variable que indica si la persona entrevistada paga més impostos dels que rep en forma de servei, pot prendre aquests valors: "Más", "MásMenos","Menos"o "N.C.N.S.".
- **ENSENYAMENTvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a l'ensenyament segons l'entrevistat.
- **OBRESvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a les obres públiques segons l'entrevistat.
- **DESEMPLEATvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la protecció dels desempleats segons l'entrevistat.
- **DEFENSAvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la defensa segons l'entrevistat.
- **SEGURETATvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la seguretat segons l'entrevistat.
- **SANITATvi**Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la sanitat segons l'entrevistat.
- **CULTURAvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la cultura segons l'entrevistat.
- **VIVENDAvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la vivenda segons l'entrevistat.
- **JUSTICIAvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a la justicia segons l'entrevistat.
- **PENSIONSvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos a les pensions segons l'entrevistat.
- **TRANSPORTvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos al transport públic segons l'entrevistat.
- **MEDIAMBIENTvi** Variable que indica si es dedica masses, els necessaris o molt pocs recursos al mediambient segons l'entrevistat.

- **GENERAL** Variable que indica si segons l'entrevistat els espanyols, en general paguem molts impostos.
- **COMPARACIÓ** Variable que indica si segons el entrevistat els espanyols, en general paguem més,igual o menys impostos comparat amb altres països europeus.
- **JUSTICIAIMP** Segons l'entrevistat es fa justícia amb els impostos.
- **FRAUDEFISCAL** Segons l'entrevistat existeix frau fiscal a Espanya i amb quina quantitat.
- **LUCHAFRAUDE** L'administració fa esforços per lluitar contra el frau fiscal, segons l'entrevistat.
- **DECLARA** Amb les persones que coneix l'entrevistat, aquestes declaren tots els impostos segons ell. Aquesta variable pot prendre diferents valors: "Toda", "Bastante","Poca","Ninguna"o "N.C.N.S.".
- **DECLARAIVA** Amb les persones que coneix l'entrevistat, aquestes declaren l'IVA segons ell.
- **ACORD1** L'entrevistat està d'acord o en desacord amb aquesta afirmació:
  - "si la gente no engaña más a Hacienda, es por miedo a una revisión"
- **ACORD2** L'entrevistat està d'acord o en desacord amb aquesta afirmació:
  - "todo el mundo engaña algo al pagar sus impuestos, y la Administración ya cuenta con ello"
- **ACORD3** L'entrevistat està d'acord o en desacord amb aquesta afirmació:
  - "en realidad no está mal ocultar parte de la renta, porque eso no perjudica a nadie"
- **ACORD4** L'entrevistat està d'acord o en desacord amb aquesta afirmació:
  - "engañar a Hacienda es engañar al resto de los/as ciudadanos/as"
- **TIPOIMP** Com s'hauria de recaudar els impostos segons l'entrevistat, amb impostos indirectes o directes.
- **DECLARACIÓN** Ha fet l'entrevistat la declaració de la renda.
- **SIGNO** Ha sigut positiva (a pagar) o negativa (a tornar).

- **DRCHAIZDA** Com es definiria l'entrevistat en termes de dreta o esquerra on 1 és totalment a l'esquerra i 10 totalment a la dreta. Es poden prendre valors intermitjos entre 1 i 10.
- **SEXE** Sexe de la persona.
- **EDAT** De la persona
- **ESTADOCIVIL** Estat civil de l'entrevistat.
- **SABELEER** L'entrevistat sap llegir.
- **MASINGRESOS** Qui té els millors ingressos? "el entrevistado", "Otra persona", ïguales"o "N.C.".
- **ESTUDIOS** Estudis de l'entrevistat.
- **ESTATUS** Estatus socioeconòmic de la persona entrevistada.
- **ANYS** Any de l'estudi.

Un altre punt important és que en les enquestes del 2000, 2001, 2002, 2003 i 2004 les següents variables no estan presents, però en les enquestes restants sí.

FREQ	ENSENYAMENT	SANITAT	JUSTICIA
SERVEISSSOCIALS	TRANSPORT	FRASE	SOCIETAT
PERSONAL	ENSENYAMENTvi	OBRESvi	DESEMPLEATvi
DEFENSAvi	SEGURETATvi	SANITATvi	CULTURAvi
VIVENDAvi	JUSTICIAvi	PENSIONSvi	TRANSPORTvi
MEDIAMBIENTvi			

De la mateixa manera les variables: DECLARA, DECLARAIVA i TIPOIMP no estan presents en les enquestes 2011 i 2012. Finalment, la variable ESTATUS no apareix en l'enquesta del 2006. Excepte la variable EDAT totes les altres variables són categòriques.

Capítol 3

DASHBOARD

Un cop acabada la part més llarga del treball, la depuració i neteja de la base de dades periòdica procediré a explicar les diferents parts del dasboard.

3.1 Què és un dasboard?

Dasboard[4] és un paquet d'R que facilita la creació d'aplicacions interactives orientades a l'anàlisi i visualització de dades a través d'R. Crear aquest aplicatiu és l'objectiu final d'aquest treball,la intenció és facilitar la tasca a l'usuari final.

3.2 Manual d'ús

Primer de tot explicaré certes qüestions que s'han de tenir en compte a l'hora d'utilitzar l'aplicació, el primer que veiem de l'aplicatiu és la següent imatge.

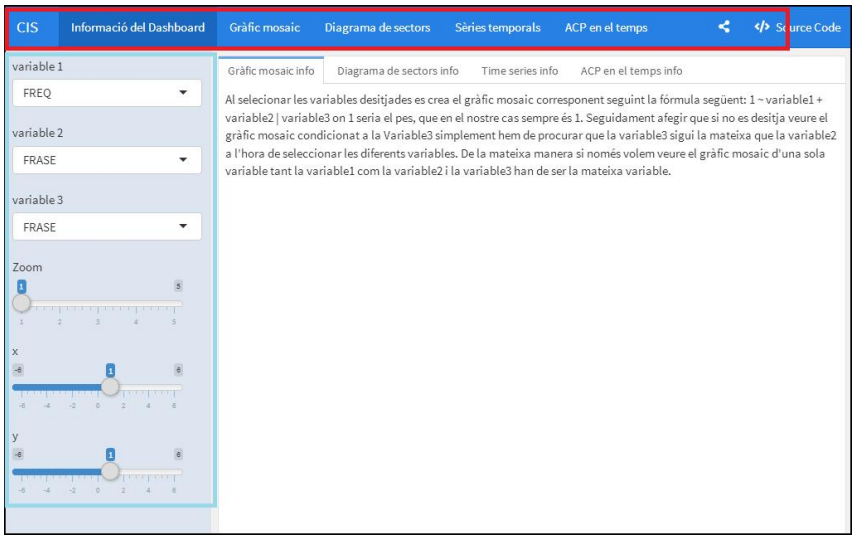

Figura 3.1: Dashboard

Remarcada amb el color blau hi ha la barra lateral, aquí trobem les diferents variables que triarem per mostrar els diferents gràfics que ens interessen, també tenim una barra per poder fer zoom al gràfic, les altres dues són per desplaçar aquest cap als costats o cap a dalt, no obstant aquestes funcions de zoom i/o desplaçament del gràfic només estan disponibles per l'ACP en el temps. En aquesta primera pàgina trobem informació de com funciona el *dashboard*, si ens fixem podem veure remarcat de color vermell les diferents pestanyes de l'aplicatiu, si cliquem sobre una d'aquestes pestanyes obtindrem el gràfic d'acord amb les variables elegides i la pestanya seleccionada.

3.3 Gràfic mosaic

En la segona pestanya tenim el gràfic mosaic, aquest tipus de gràfic no deixa de ser una representació d'una taula de freqüències on cada cel·la de la taula de freqüències queda representada en una regió rectangular proporcional a la cel·la en qüestió. En seleccionar les variables desitjades es crea el gràfic mosaic corresponent, seguint la fórmula següent: $1 \sim$ variable1 + variable2 | variable3 on 1 seria el pes, que en el nostre cas sempre és 1. Seguidament afegir, que si no es desitja veure el gràfic mosaic condicionat a la variable3, simplement hem de procurar que la variable3 sigui la mateixa que la variable2 a l'hora de seleccionar les diferents variables. De la mateixa manera si només volem veure el gràfic mosaic d'una sola variable tant la variable1 com la variable2 i la variable3 han de ser la mateixa variable. En el programa podem veure tant el gràfic com la taula de freqüència.

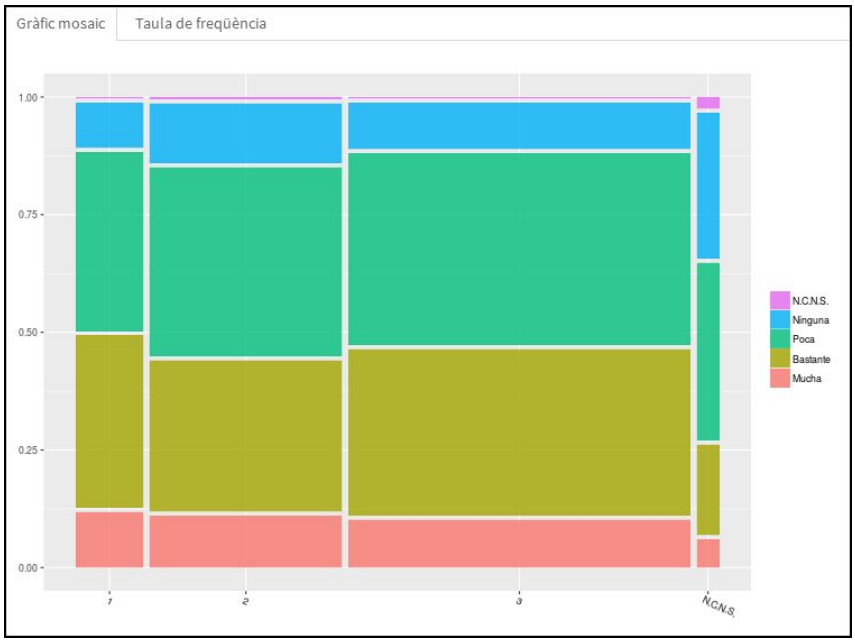

Figura 3.2: Representació de les variables FREQ i FRASE

També si volem, podem veure la taula de freqüència associada al gràfic.

Gràfic mosaic | Taula de freqüència

	1	2	3	N.C.N.S.
Mucha	1.32	3.51	5.77	0.22
Bastante	4.10	10.30	20.09	0.72
Poca	4.25	12.84	23.17	1.41
Ninguna	1.07	4.06	5.56	1.16
N.C.N.S.	0.04	0.14	0.18	0.09

Figura 3.3: Taula de freqüència de les variables FREQ i FRASE

Per fer aquests gràfics, he utilitzat la funció *ggmozaic*[3]. Tal com he dit abans si utilitzem la tercera variable obtenim el gràfic mosaic condicionat a aquesta amb la seva respectiva taula de freqüència.

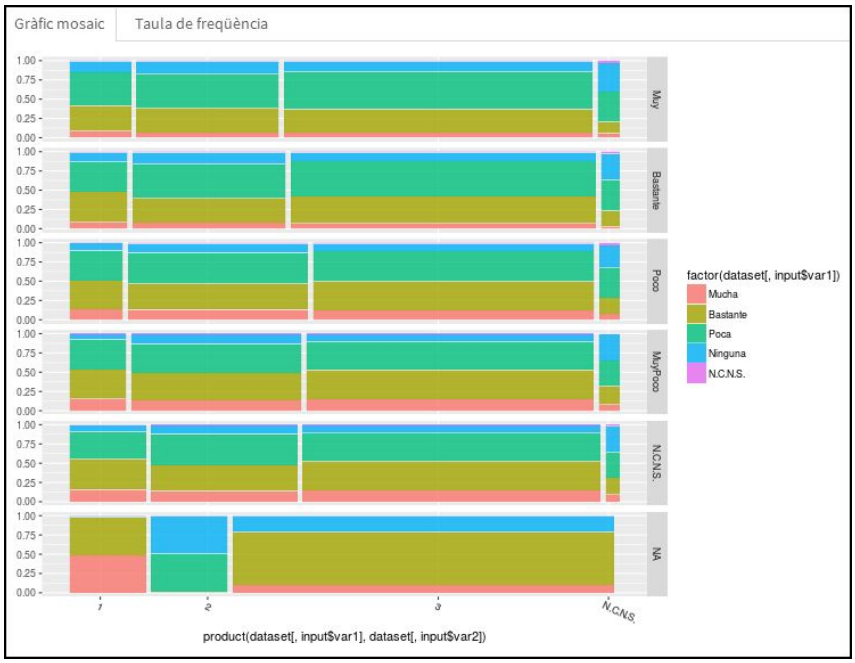

Figura 3.4: Gràfic mosaic de les variables FREQ i FRASE condicionat a la variable RESPIMP

Podem veure també la taula de freqüència associada al gràfic. En la següent imatge només es veu un tros de la taula de freqüència.

3	Poca	1	Muy	159
4	Ninguna	1	Muy	45
5	N.C.N.S.	1	Muy	1
6	Mucha	2	Muy	171
7	Bastante	2	Muy	347
8	Poca	2	Muy	358
9	Ninguna	2	Muy	151
10	N.C.N.S.	2	Muy	7
11	Mucha	3	Muy	190
12	Bastante	3	Muy	554
13	Poca	3	Muy	599
14	Ninguna	3	Muy	222
15	N.C.N.S.	3	Muy	4
16	Mucha	N.C.N.S.	Muy	9
17	Bastante	N.C.N.S.	Muy	26

Figura 3.5: Taula de freqüència de les variables FREQ i FRASE condicionat a la variable RESPIMP

3.4 Diagrama de sectors

He fet servir uns diagrames de sectors per veure com evolucionen les variables categòriques en el temps, tenim tants diagrames com nombre d'estudis. Seguint amb la variable FRASE veiem com aquesta canvia, és curiós veure els canvis que hi ha hagut durant la crisi. Es pot veure una reducció de la frase tres i un augment de la frase dos, la frase tres sempre és la més elegida. La variable interactiva en aquest cas és la variable1.

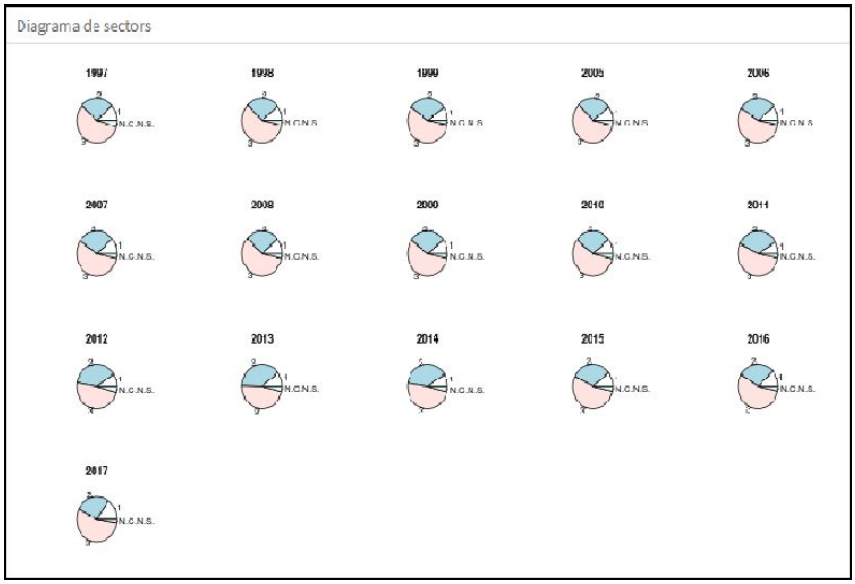

Figura 3.6: Diagrama de sectors de la variable FRASE, en el temps

En aquesta imatge es pot apreciar el que s'ha comentat anteriorment, podem veure que els diagrames de sectors que corresponent als anys 2000, 2001, 2002, 2003 i 2004 no apareixen, ja que en les enquestes d'aquests anys la variable FRASE no existia.

3.5 Sèrie temporal

Podem veure aquest fenòmen encara de forma més evident, la reducció del nivell 3 i l'augment del nivell 2 de la variable FRASE, utilitzant la sèrie temporal associada. La variable interactiva en aquest cas és la variable2.

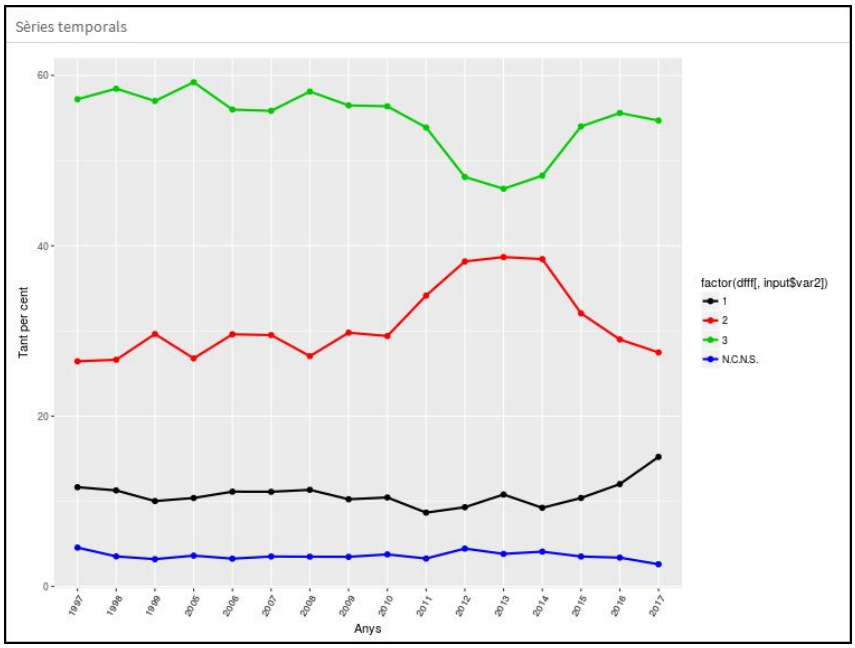

Figura 3.7: Sèrie temporal de la variable FRASE

3.6 ACP en el temps

Primerament definir que és l'ACP (Anàlisi de Components Principals). És una tècnica exploratòria de la base de dades, la principal característica que té l'ACP és que pots reduir la dimensionalitat de la base de dades perdent la mínima informació possible, d'aquesta manera pots representar les dades en un pla.

Un component principal és una combinació lineal de les variables a estudiar. Tenim tants components principals, ψ, com variables a estudiar, ψ =XU. Resolent el problema de programació lineal següent maximitzem la variància del component principal, d'aquesta manera conservem la màxima informació possible.

$$
\begin{array}{lll}
\text{Max} & Var(\psi) = U'X'NXU = \lambda U & \\
\text{subjecte a} & U_i'U_i = 1, & i = 1, ..., n \\
(1) & U_i'U_j = 0, & j = 1, ..., n \\
(2) & i \neq j &
\end{array}
$$

La primera equació (1) és perquè el problema tingui solució i la segona (2) és perquè d'aquesta manera els components principals estiguin incorrelacionats proporcionant així una gran propietat. Afegir també que les unitats de mesura influeixen en l'ACP, ja que la variància depén de les unitats de mesura, si no vols que influeixin hem de standaritzar la base de dades, per defecte la funció *prcomp* de R ja ho fa així, a no ser que indiquis el contrari. Amb aquesta idea general de l'ACP explicada ara podrem aplicar-la en el nostre cas. Com he dit abans la nostra base de dades és periòdica, és a dir, és una base de dades temporal, amb l'ús d'un exemple explicaré pas per pas l'ACP a través del temps.

3.6.1 Pas 1

Primer de tot s'ha de calcular la taula de freqüències d'aquesta manera obtindrem els percentatges.

	Muy	Bastante	Poco	MuyPoco	N.C.N.S.
Clase alta	0.82918740	5.4311774	5.8457711	1.16086235	0.33167496
Clase alta	1.05263158	6.1473684	6.4842105	1.34736842	0.58947368
Clase alta	0.50146260	6.5608023	6.1429168	1.04471375	0.54325115
Clase alta	1.25944584	9.1519731	4.4080605	0.83963056	0.79764903
Clase alta	1.20182346	6.9208454	4.3928719	0.45586407	0.82884376
Clase alta	1.12640801	8.1351690	4.1301627	0.83437630	1.08468919
Clase alta	1.21389703	8.3717036	4.3114274	0.54416074	0.83717036
Clase alta	1.17105813	8.2392304	4.5169385	0.71099958	0.71099958
Clase alta	1.04690117	6.8676717	5.7370184	1.38190955	0.71189280
Clase alta	1.37484836	7.6021027	6.3485645	1.90052568	0.48524060
Clase alta	1.22000813	7.2387149	6.4253762	1.42334282	0.69133794
Clase alta	1.56636439	8.1615829	6.9661995	2.06100577	0.57708162
Clase alta	0.98887515	6.7161104	6.3864854	1.77173465	0.57684384
Clase alta	0.74657818	4.8112816	7.3828287	2.32268768	0.49771879
Clase alta	1.13314448	5.6252529	8.3367058	2.91380008	0.60704168
Clase alta	0.84609186	5.5197421	8.1385979	2.98146656	0.32232071
Clase alta	1.21506683	6.1563386	9.0319968	2.87565816	0.44552450
Clase alta	1.01091791	4.3671654	8.2895269	3.72017792	0.36393045
Clase alta	0.96813231	4.8406616	10.4074223	3.20778540	0.68576029

Figura 3.8: Taula de freqüències en percentatge de les variables ESTATUS i RESPIMP en el temps

En la imatge podem veure només el nivell *Clase alta* i com aquest canvia els seus percentages en el temps categoritzada amb els nivells de RESPIMP. Tot i que, en la imatge no es veu aquest *data frame* conté tots els nivells de la variable ESTATUS.

3.6.2 Pas 2

D'aquest nou *data frame* creat tranformem els percentatges usant per això la distància Helinger.

3.6.3 Pas 3

En aquesta matriu de distàncies apliquem la funció, *prcomp*, d'aquesta manera obtenim els components principals. En aquest cas aconseguim reduir la dimensió de cinc a dos, ja que agafem només els dos primers components.

3.6.4 Pas 4

En el gràfic següent podem veure el resultat final, cal observar que els punts estan units segons l'estatus per un color diferent mitjançant fletxes. Si ens fixem, podem veure dos tipus de lletra P de principi i F de final, representant respectivament el punt inicial i el punt final, no deixa de ser una sèrie temporal. Indicar també que es representen les categories com fletxes centrades a l'origen. Per últim afegir que les variables interactives són la variable1 i la variable2.

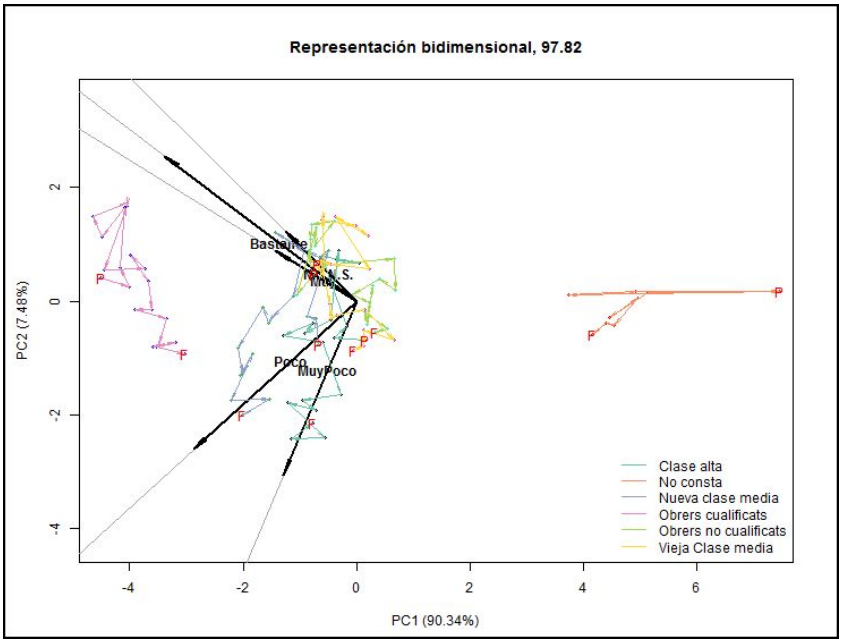

Figura 3.9: ACP en el temps de les variables ESTATUS i RESPIMP

També mitjançant les barres del Dasboard podem fer zoom al gràfic.

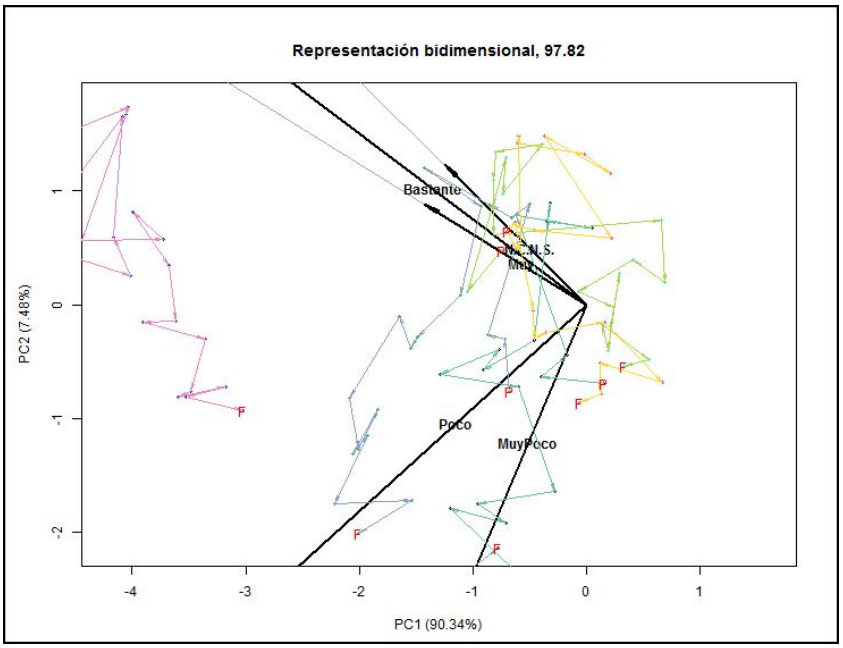

Figura 3.10: Zoom de l'ACP en el temps de les variables ESTATUS i RESPIMP

D'acord amb el gràfic anterior es pot observar que la seva posició en favor de la responsabilitat respecte els impostos ha empitjorat en el temps de forma generalitzada, independentment del nivell de la variable ESTATUS.

Capítol 4

EXEMPLES D'APLICACIÓ

Explicat el funcionament del dashboard només queda que l'usuari utilitzi l'aplicació de la forma que li resulti més útil. Tot seguit, faré alguns exemples d'aplicació per saber les tendències i opinions que tenen els ciutadans d'Espanya sobre els impostos i com aquests els administra l'Estat.

4.1 Impostos en forma de serveis

Primerament dues variables molt interssants són les variables PERSONAL i SOCIETAT. Recordem la seva descripció:

- **SOCIETAT** Variable que indica si la societat es beneficia molt, bastant, poc o gens dels impostos que retornen en forma de serveis.
- **PERSONAL** Variable que indica si la persona entrevistada paga més impostos dels que rep en forma de servei. Pot prendre aquests valors: "Más", "MásMenos","Menos"o "N.C.N.S.".
- **RESPIMPPER** Com es considera el propi entrevistat en termes de responsabilitat i consciència respecte els impostos, pot prendre aquests valors:"Muy","Bastante", "Poco","MuyPoco"o "N.C.N.S.".

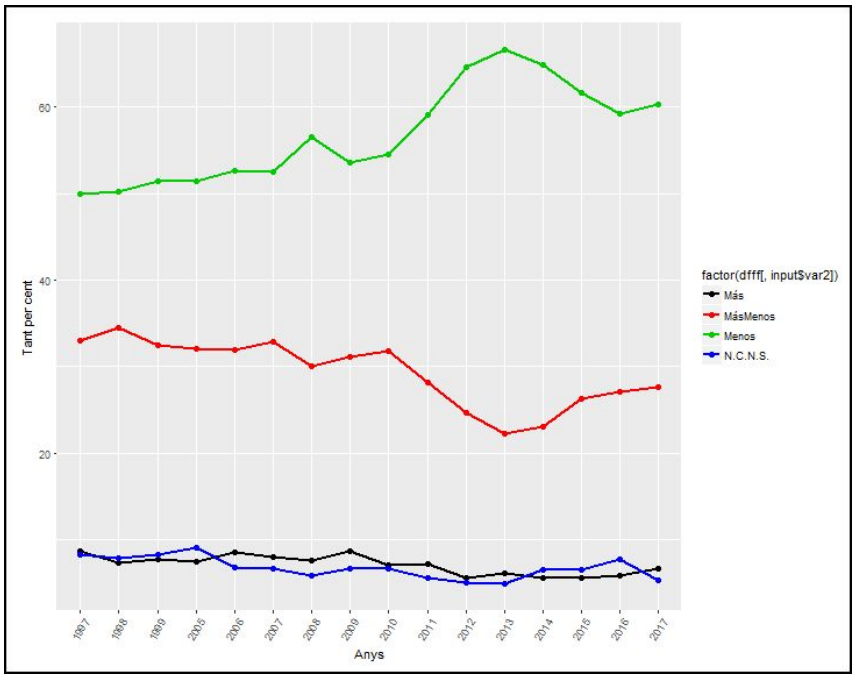

Figura 4.1: Sèrie temporal de la variable PERSONAL

Observant la sèrie temporal de la variable PERSONAL podem veure que com a mínim el cinquanta per cent dels entrevistats creu que paga més impostos dels que rep en forma de servei. Aquesta tendència s'agreuja encara més a partir del 2010. El punt més fatídic és el 2013, en el qual arriba al 70 per cent. A partir d'aquí aquesta tendència negativa s'atenua una mica.

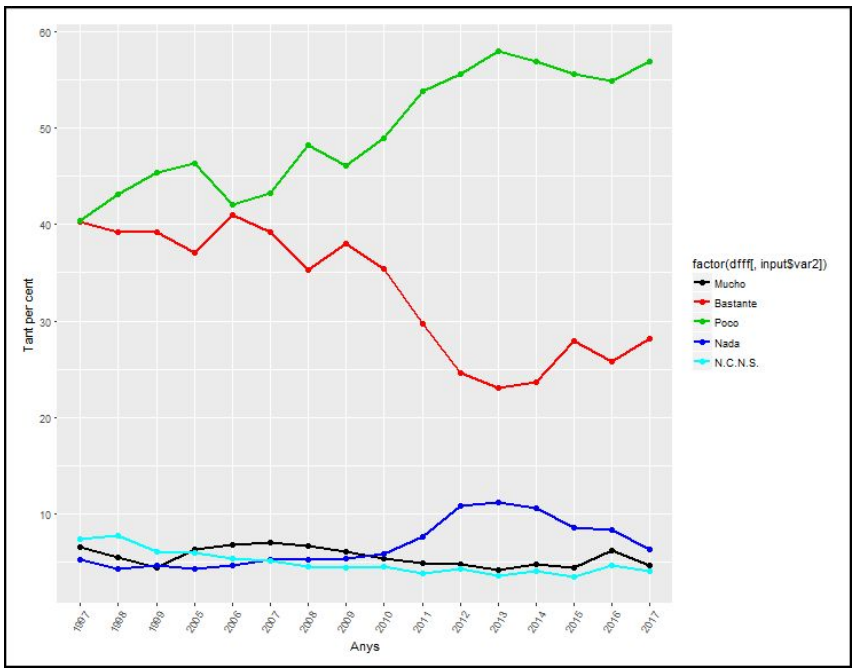

Figura 4.2: Sèrie temporal de la variable SOCIETAT

La variable SOCIETAT també es comporta d'una forma similar, a partir del 2010 el nivell *poco* es dispara. Cal notar que abans del 2010 el nivell *poco* i *bastante* tenien uns nivells en percentatges força semblants.

Ara si utilitzem l'ACP en el temps elegint les variables PERSONAL i RESPIMPER podem veure el següent gràfic.

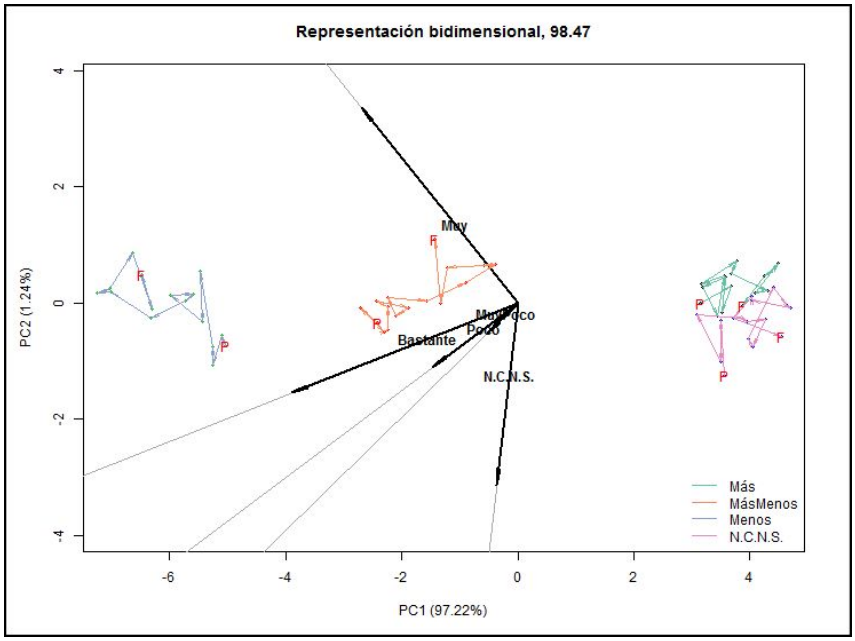

Figura 4.3: ACP en el temps de les variables PERSONAL i RESPIMPER

Veient l'anterior gràfic podem afirmar que tant les persones amb nivell *Menos* i *Más-Menos* de la variable SOCIETAT s'han tornat més responsables.

4.2 Tipus d'impost

Una altra variable molt útil és el tipus d'impost, és a dir, quin tipus d'impost és millor segons l'entrevistat, impostos indirectes com per exemples IVA o impostos directes com l'IRPF. Una variable útil en aquest cas és TIPOIMP recordem la seva definició.

- **TIPOIMP** Com s'hauria de recaptar els impostos segons l'entrevistat, amb impostos indirectes o directes

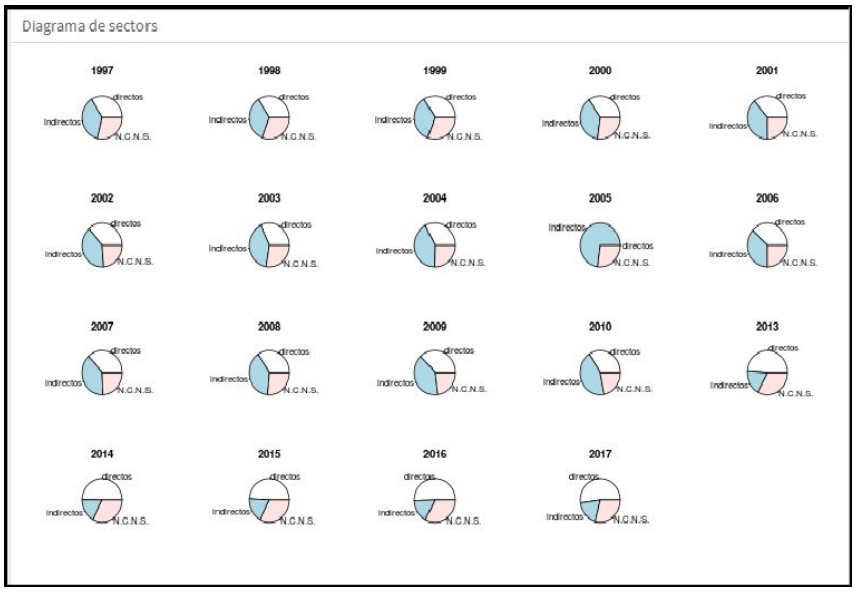

Figura 4.4: Diagrama de sectors en el temps de la variable TIPOIMP

Observant els diferents diagrames de sectors podem veure clarament com al cap dels anys la forma preferida de recaptació són els impostos directes. Si ara utilitzem l'ACP en el temps de les variables TIPOIMP i ESTATUS, observem que tots els grups es desplacen en direcció *directos*, el que fa el canvi més significatiu és el nivell *obrers qualificats* de la variable ESTATUS.

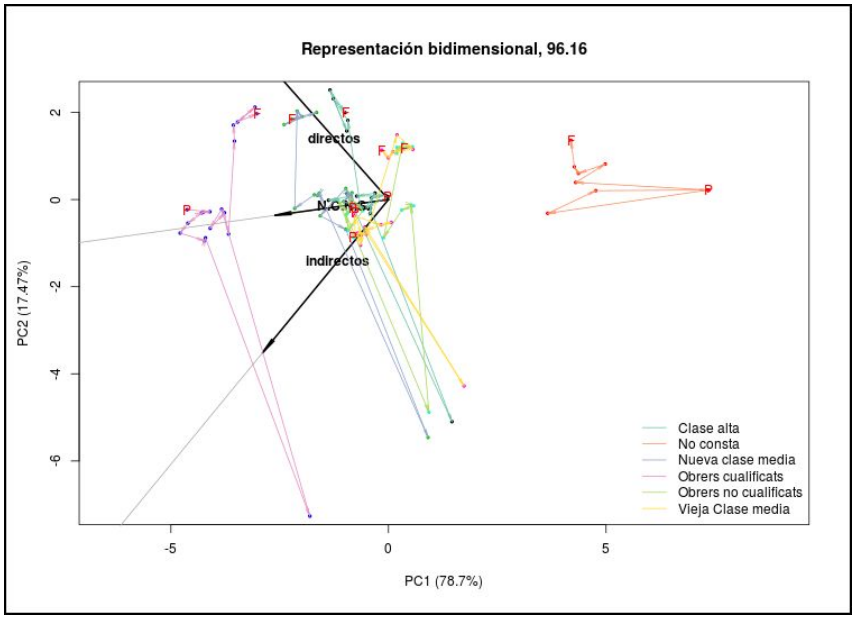

Figura 4.5: ACP en el temps de les variables TIPOIMP i ESTATUS

Si ens fixem bé podem veure que en tots els grups de la variable ESTATUS hi ha un punt extrem en direcció a *indirectos*, això es deu al fet que l'any 2005 molta gent

per alguna raó va preferir de forma generalitzada que els impostos es paguessin de forma indirecta, personalment crec que es tracta d'algun tipus d'error, mirant el diagrama de sectors de l'any 2005 confirmem aquest fet.

4.3 Frau fiscal

Per últim les diferents opinions dels entrevistats respecte el frau fiscal. Dues variables que ens poden ajudar són les següents:

- **ACORD2** L'entrevistat està d'acord o en desacord amb aquesta afirmació:
  - "todo el mundo engaña algo al pagar sus impuestos, y la Administración ya cuenta con ello"
- **FRAUDEFISCAL** Segons l'entrevistat existeix frau fiscal a Espanya i amb quina quantitat.

Observant detingudament el gràfic podem observar que el nivell *Desacord* ha anat augmentat al cap dels anys, la gent s'ha fet més responsable i conscient.

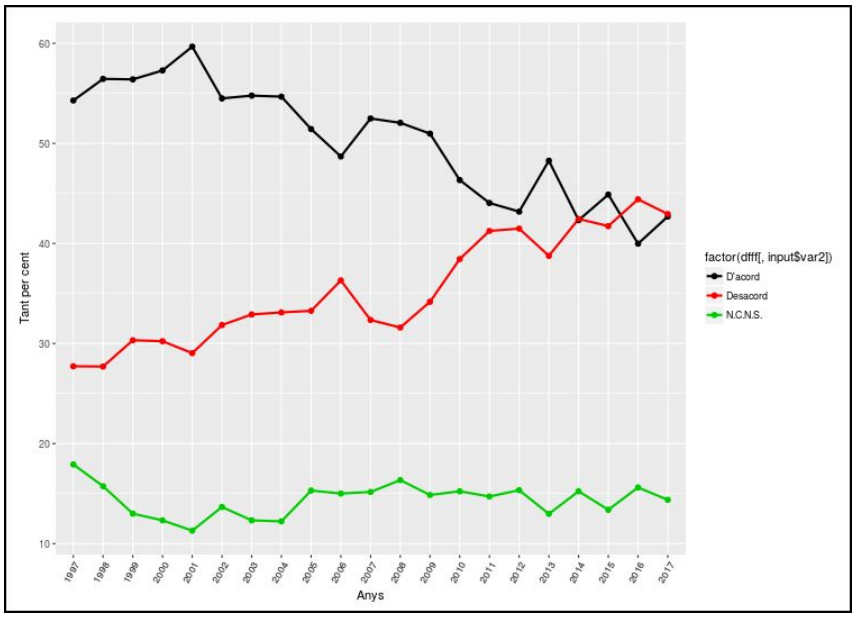

Figura 4.6: Evolució en el temps de la variable ACORD2

Els nivells *muy* i *bastante* han canviat de tendència en direcció a *desacord*, els que opinen que quasi no hi ha frau fiscal tenen una opinió molt dividida entre si estan en *desacord* o *d'acord* amb el que diu la variable ACORD2.

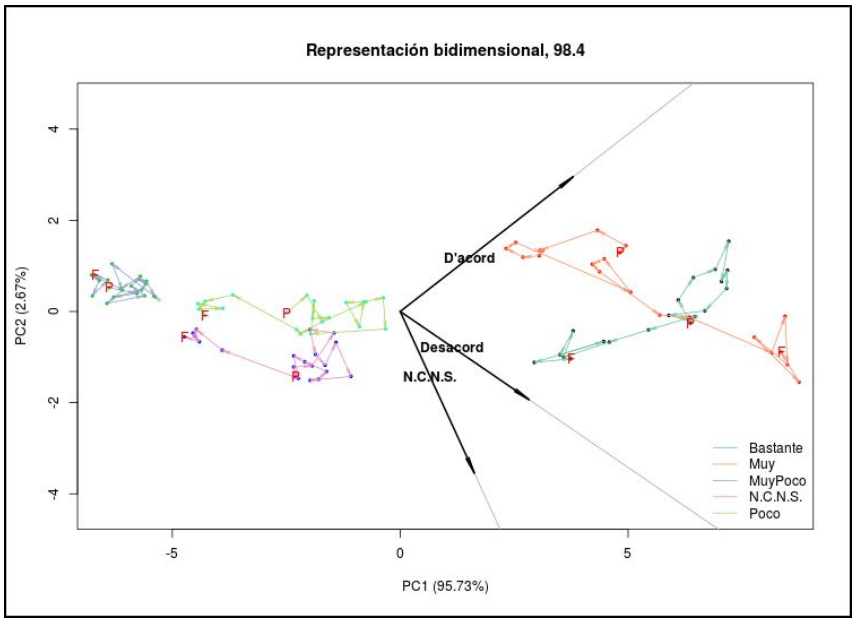

Figura 4.7: ACP en el temps de la variable ACORD2 i FRAUDEFISCAL

Capítol 5

CONCLUSIONS

El present treball s'ha centrat en la creació d'una aplicació web amb la intenció d'oferir una eina gratuïta i eficaç per a l'anàlisi de dades d'una forma fàcil i intuitiva, complint amb l'objectiu principal del treball. El programa serveix per a altres bases de dades semblants, sempre que aquestes siguin periòdiques i tinguin un nombre elevat de variables categòriques, per tant l'aplicació web es força versàtil i pràctica. Amb el programa els professionals que no estan del tot familiaritzats amb l'estadística se'ls facilita feina.

Com a conclusió final de la base de dades periòdica utilitzada, podem veure a través dels gràfics de l'aplicatiu que a partir del 2007-2008, coincidint amb la crisi econòmica, la majoria de variables obren una nova tendència a causa del canvi de mentalitat que va suposar la crisi, aquest fet ens reafirma en la gran utilitat d'aquest programa, ja que podem veure com les variables evolucionen en el temps i ens permet reconèixer noves tendències en les maneres de pensar.

La meva conclusió personal d'aquest treball és que la major part del temps l'he dedicat a estandarditzar la base de dades, que és una feina mecànica i poc gratificant, un cop normalitzada la base crear l'aplicatiu ha sigut més entretingut.

Una possible millora de l'aplicatiu seria la creació d'un sofware que millorés i fes més senzill el preprocessament d'altres bases de dades semblants a la que he utilitzat.

Els estudis que he realitzat d'estadística sempre m'havien semblat molt útils per a la seva aplicació pràctica, fent aquest treball em ratifico en aquesta idea.

Capítol 6

AGRAÏMENTS

Voldria agrair als meus dos professors, tutors d'aquest treball, Dr. Josep M Oller i Dr. Esteban Vegas el seu mestratge, dedicació i disposició per ajudar-me en tot moment.

Bibliografia

[1] Lloc web [Online]. Available: http://www.cis.es/cis/opencms/ES/index.html.

[2] Lloc web [Online]. Available: https://orri.shinyapps.io/DASHBOARDMILLORA/.

[3] Lloc web [Online]. Available: https://cran.r-project.org/web/packages/ggmosaic/vignettes/ggmosaic.html.

[4] Lloc web [Online]. Available: https://rmarkdown.rstudio.com/flexdashboard/.

[5] Lloc web [Online]. Available: https://shiny.rstudio.com/.

[6] Robert L.Kabacoff. *R IN ACTION, Data analysis and graphics with R, Manning Publicacions CO, USA 2015.*

Annex A

CODI R

A.1 Dashboard

```
> #---
> #title: "CIS"
> #output:
> #  flexdashboard::flex_dashboard:
> #    orientation: columns
> #    social: menu
> #    source_code: embed
> #runtime: shiny
> #---
>
> #```{r global, include=FALSE}
> library(flexdashboard)
> library(shiny)
> library(shinydashboard)
> library(rsconnect)
> library(ggmosaic)
> library(crosstalk)
> library(leaflet)
> library(DT)
> library(dplyr)
> library(vcd)
> library(stringr)
> library(forecast)
> library(tidyquant)
> library(timetk)
```

```
> library(sweep)
> library(RColorBrewer)
> library(reshape2)
> library(ggplot2)
> library(zoom)
> library(MASS)
> load("data2.Rdata")
> dataset<-comb
> for (i in 1:length(levels(as.factor(comb$ANYS)))) {
+   talls<-comb[as.factor(comb$ANYS)==levels(as.factor(comb$ANYS))[i],]
+   assign(paste("talle",i,sep=""),talls)
+   }
> dlist<-lapply(str_sort(ls(patt="^talle"), numeric = TRUE,decreasing = F),get)
> timesss<-function(y){
+
+   for (i in 1:length(levels(as.factor((comb$ANYS))))){
+   dlist[[i]]<-as.data.frame(dlist[[i]])
+   dlist[[i]]<-dlist[[i]] %>% group_by(.dots = y) %>% tally()
+   dlist[[i]]$estu<-levels(as.factor(comb$ANYS))[i]
+   dlist[[i]]$n<-((dlist[[i]]$n)/sum((dlist[[i]]$n)))*100
+   }
+   #d44temps<-do.call(rbind, dlist)
+  return(dlist)
+ }
>  #do.call(rbind, timesss("FREQ"))
>
> #noms<-names(sapply(sapply(comb, levels),length)[sapply(sapply(comb, levels),lengt
>
> #```
> #Sidebar {.sidebar}
> #=====================================================================
> #```{r}
> selectInput("var1",label = "variable 1",choices = names(dataset),selected="FREQ")
> selectInput("var2",label = "variable 2",choices = names(comb),selected="FRASE")
> selectInput("var3",label = "variable 3",choices = names(comb),selected="FRASE")
> sliderInput("bins1", "Zoom",
+             min =1, max =5, value = 1,width = "90%")
> sliderInput("x", "x",
+             min =-6, max =6, value = 1,width = "90%")
```

```
> sliderInput("y", "y",
+              min =-6, max =6, value = 1,width = "90%")
> #```
>
> #Informació del Dashboard
> #=====================================================================
> #Row {.tabset .tabset-fade}
> #---------------------------------------------------------------------
>
> ### Gràfic mosaic info
> #```{r}
> renderText({"Al selecionar les variables desitjades es crea el gràfic mosaic corre
> #```
>
> ### Diagrama de sectors info
> #```{r}
> renderText({"En aquest gràfic podem veure un anàlisi descriptiu per anys, la varia
> #```
>
> ### Time series info
> #```{r}
> renderText({"En aquest gràfic podem veure una sèrie temporal, la variable interact
> #```
>
> ### ACP en el temps info
> #```{r}
> renderText({"En aquest gràfic podem veure l'ACP a través del temps, les variables
> #```
>
> #Gràfic mozaic
> #=====================================================================
> #Row {.tabset .tabset-fade}
> #---------------------------------------------------------------------
>
> ### Gràfic mosaic
> #```{r}
> renderPlot({
+
+   if(input$var3!=input$var2){
```

```
+ ggplot(data = dataset) +
+   geom_mosaic(aes(weight = 1, x = product(dataset[,input$var1],dataset[,input$var2
+                   fill=factor(dataset[,input$var1])),na.rm=T)+theme(axis.text.x=el
+ facet_grid(dataset[,input$var3]~.)}else{
+       ggplot(data = dataset) +
+       geom_mosaic(aes(weight = 1, x = product(comb[,input$var2]), fill=factor(data
+       theme(axis.text.x=element_text(angle=-25, hjust= .1)) + labs(x="", title='')
+       guides(fill=guide_legend(title = "", reverse = TRUE))
+
+ }
+   })
> #```
>
>
>
> ### Taula de freqüència
> #```{r}
>
>
> renderTable({
+ if(input$var3==input$var2){
+ matrix(prop.table(table(comb[,input$var1],comb[,input$var2]))*100,ncol = length(le
+
+ }else{
+ xtabs(~comb[,input$var1]+comb[,input$var2]+comb[,input$var3])
+ }
+ },striped = T,rownames = T)
> #```
>
> #Diagrama de sectors
> #=========================================================================
>
> ### Diagrama de sectors
> #```{r}
> renderPlot({
+ par(mar=c(3,3,3,3))
+ par(mfrow=c(5,5))
+ lisDataFrames<- vector(mode = "list", length = length(dlist))
+ for (i in 1:length(dlist)) {
```

```
+ lisDataFrames[[i]]<-(table(dlist[[i]][[input$var1]],droplevels(dlist[[i]][["ESTU"]
+ }
+ #})
+
+ lisDataFrames<-lisDataFrames[lapply(lisDataFrames, sum)!=0]
+ for(i in 1:length(lisDataFrames)){
+   pie(lisDataFrames[[i]],main = colnames(lisDataFrames[[i]]),labels = rownames(lis
+ }
+
+ })
> #```
>
> #Sèries temporals
> #=========================================================================
>
> ### Sèries temporals
> #```{r}
>
> renderPlot({
+   dfff<-do.call(rbind, timesss(input$var2))
+   dfff <- as.data.frame(dfff)
+   dfff$estu<-as.factor(dfff$estu)
+   dfff[,input$var2]<-as.factor(dfff[,input$var2])
+   dfff$n<-as.numeric(dfff$n)
+   dfff<-dfff[complete.cases(dfff),]
+
+ ggplot(dfff, aes(x = dfff$estu, y = dfff$n,group=factor(dfff[,input$var2]))) +
+   geom_line(aes(color = factor(dfff[,input$var2])), size = 1,na.rm = T) +
+   scale_color_manual(values = c(1:length(levels(dfff[,input$var2]))))+
+   geom_point(aes(colour = factor(dfff[,input$var2])), size = 2)+xlab("Anys") + yla
+ })
> #```
>
> #ACP en el temps
> #=========================================================================
>
> ### ACP en el temps
> #```{r}
> renderPlot({
```

```
+ listOfDataFrames <- vector(mode = "list", length = length(dlist))
+
+ for (i in 1:length(listOfDataFrames)) {
+   listOfDataFrames[[i]]<-prop.table(table(dlist[[i]][[input$var1]],dlist[[i]][[inp
+ }
+
+
+ df<-do.call(rbind,listOfDataFrames)
+ fi<-df[order(rownames(df)),]
+ fi<-fi[complete.cases(fi),]
+ fii<-as.data.frame(fi,T)
+ fii$clase<-rownames(fi)
+ fii$clase<-as.factor(fii$clase)
+
+
+ cols <- RColorBrewer::brewer.pal(length(unique(fii$clase)), name = "Set2")
+ if(length(levels(fii$clase))==2){
+ fii$color <- factor(fii$clase, labels = cols[1:2])
+ }else{
+ fii$color <- factor(fii$clase, labels = cols)}
+
+
+
+ Clase <- 2* sqrt(fi)
+
+ #fiic <-with(fii,
+ #         data.frame(clase = levels(clase),
+ #                    color = I(brewer.pal(nlevels(clase), name = 'Dark2'))))
+
+ pc <- princomp(Clase)
+
+ ev <-pc$sdev^2
+ varExp <- ev/sum(ev)*100
+
+
+ #final<-merge(x = fii, y = fiic, by = "clase", all.x = TRUE)
+ final<-fii
+
+
```

```
+ eqscplot(pc$scores[,1:2], pch=20,cex=0.7,asp=1,col=final$color,
+          main=paste0("Representación bidimensional, ", round(varExp[1]+varExp[2],2
+          xlab=paste0("PC1 (" , round(varExp[1],2), "%)"), ylab=paste0("PC2 (",  ro
+
+ lines(for (i in 1:dim(fi)[2]){
+   part <- 100*pc$loadings[i,1:2]
+   arrows(x0=0,y0=0,x1=part[1],y1=part[2],length=0.0, angle=0, code=2,lwd=1,col="da
+ })
+
+ lines(for (i in 1:dim(fi)[2]){
+   part <- 5*pc$loadings[i,1:2]
+   arrows(x0=0,y0=0,x1=part[1],y1=part[2],length=0.2, angle=5, code=2,lwd=2,col="bl
+ })
+
+ lines(for (i  in 1:length(levels(comb[,input$var2]))) {
+   text(x=2*pc$loadings[i,1],y=2*pc$loadings[i,2],labels=levels(comb[,input$var2])[
+ })
+
+ #
+ # representaci?n evoluci?n temporal
+ #
+ a<-c()
+ for (i in 1:length(levels(comb[,input$var1]))) {
+   a<-c(a,which(rownames(pc$score)==levels(comb[,input$var1])[order(levels(comb[,in
+ }
+ #which(rownames(pc$score)==levels(comb$ESTATUS)[order(levels(comb$ESTATUS))][i])[1
+ a<-c(a,dim(pc$score)[1])
+ colors<-levels(as.factor(final$color))
+
+ lines(for (l in 1:(length(a)-1)) {
+   for (i in (a[l]+1):(a[l+1]-1)){
+     k<-i-1; j<-i;
+     x0<- pc$score[k,1];
+     y0<- pc$score[k,2];
+     x1<- pc$score[j,1];
+     y1<- pc$score[j,2];
+     arrows(x0,y0,x1,y1,length=0.1, angle=8, code=2,col=colors[l])}
+ })
+
```

```
+ lines(arrows(pc$score[(nrow(pc$score)-1),1], pc$score[(nrow(pc$score)-1),2] ,pc$sc
+
+
+
+
+ for (t in 1:(length(a)-1)) {
+  text(pc$scores[a[t],1],pc$scores[a[t],2],"P",col = "red")
+ }
+
+
+ for (t in 2:(length(a)-1)) {
+   text(pc$scores[(a[t]-1),1],pc$scores[(a[t]-1),2],"F",col = "red")
+ }
+ text(pc$scores[(dim(Clase)[1]),1],pc$scores[(dim(Clase)[1]),2],"F",col = "red")
+
+ legend(x = 'bottomright',
+        legend = as.character(levels(final$clase)),
+        col = levels(final$color), bty = 'n', xjust = 1, lty=1, cex=1)
+ zoomplot.zoom(fact=input$bins1,x=input$x,y=input$y)
+
+
+ })
> #```
```