Tipologia de despesa a les llars Espanyoles

Autor: Joan Ferrando Ravella
Tutor: Llorenç Badiella

Índex

1 Introducció

L'Institut Nacional d'Estadística (INE) és l'organisme encarregat de la coordinació general dels serveis estadístics de l'Administració General de l'Estat i la vigilància, control i supervisió de procediments tècnics dels mateixos. Fou creada arrel de la Llei del 31 de desembre de 1945, amb l'objectiu d'elaborar i perfeccionar les estadístiques demogràfiques, econòmiques i socials ja existents, la creació d'altres noves i la coordinació amb els serveis estadístics de les àrees provincials i municipals.

Possiblement un dels estudis més rellevants que duu a terme l'INE és l'**"Enquesta de Pressupostos Familiars" (EPF)**, una enquesta de nivell nacional que existeix des de 1997 i que va sorgir arrel d'una necessitat per trobar aproximacions d'habits de consum de les llars Espanyoles per, posteriorment, estudiar la seva evolució en relació amb la fluctuació dels preus dels mateixos productes.
Des dels seus origens, l'EPF ha anat patint canvis metodologics i de disseny mostral que han afectat a tots els aspectes relacionats amb el seu anàlisi. Malgrat això, l'EPF sempre ha mantingut uns principis invariables al llarg dels anys, totes les EPF sempre subministen informació sobre la naturalesa i el destí dels costos del consum i sobre diferents característiques relatives a les condicions de vida de les llars. Els objectius d'aquesta enquesta també han sofert canvis, tot i que van quedar fixats al 2006 i no han canviat des d'aleshores. Aquests objectius són:

- Obtindre estimadors de la despesa agregada anual de les llars, tant a nivell nacional com a nivell de comunitat autònoma, així com la seva classificació segons diverses condicions de la llar.
- Estimar el canvi interanual de la despesa agregada pel conjunt nacional i per comunitat autònoma.
- Estimar el consum en quantitats físiques de determinats productes pel conjunt nacional.

Addicionalment, dins dels objectius principals, destaquen per la seva importància uns altres dos objectius relacionats amb necessitats concretes de diversos usuaris de l'enquesta: l'estimació de la despesa com instrument per l'obtenció del consum privat en la comptabilitat nacional i l'estimació de l'estructura de ponderacions a partir de la despesa necessària pel càlcul de l'IPC.

Per altra banda, donada la magnitud de l'estudi, la informació que l'Enquesta de Pressupostos Familiars pot aportar a diferents àmbits és molt elevada i, ja que aquesta està a disposició de tothom i de forma gratuïta, molts usuaris i empreses de naturalesa molt diversa han fet ús d'aquesta base de dades adaptant-la al seu camp de treball.

Per exemple, els títols d'alguns treballs científics realitzats per centres o usuaris amb major impacte científic són:

- Major fruit and vegetable contributors to the main serum carotenoids in the Spanish diet.
  - Servicio de Nutrición, Clínica Puerta de Hierro, Madrid, Spain.
- Returns to Human Capital in Spain: A Survey of the Evidence.
  - Departament d'Economia Aplicada, Universitat Autònoma de Barcelona, Spain.
- Poverty among children and youth in Spain: The role of parents and youth employment status.
  - Universidad de Vigo, Spain.

Per concloure, l'Enquesta de Pressupostos Familiars és un **estudi molt rellevant** que desperta l'interés analític d'empreses alienes al propi INE i que ha estat analitzat i explotat per diferents objectius i per sectors molt diversos, que generalment han empleat **metodologies típiques del propi camp de coneixement** per extreure les seves conclusions, per exemple, tècniques d'anàlisi economètrica, models de regressió o contrasts d'hipòtesis propis del sector en qüestió.

2 Objectius

Com acabem de comentar, són molts els diferents objectius d'anàlisi que es poden realitzar sobre l'EPF i són molts els treballs realitzats sobre la mateixa, deixant així poques alternatives d'estudi que aportin conclusions innovadores i que no caiguin en redundància.

Com la motivació d'aquest treball és aportar informació a un estudi de molt renom i notablement treballat, s'utilitzarà metodologies poc convencionals que escapen a les clàssicament empleades per diferents camps de treball.

D'aquesta manera, l'objectiu del treball serà ampliar el ventall d'eines estadístiques que poden ser empleades per tal d'atacar un problema des de diferents enfocs i guanyar així perspectiva del mateix. Concretament, s'empraran eines d'anàlisi multivariant utilitzant principalment l'**anàlisi de clústers per tal d'assignar perfils de despesa a cada llar**.

L'estudi és durà a terme emprant l'EPF de 2015.

3 Material i Mètodes

L'Enquesta de Pressupostos Familiars ha anat canviant la seva metodologia adaptant-se a les noves exigències i necessitats dels usuaris per tal d'assegurar la màxima qualitat d'informació. El disseny mostral, de la mateixa manera, també s'ha anat ajustant, actualment conserva la mateixa metodologia que l'aplicada l'any 2011, que és la següent:

L'any 2011 es va establir un mida mostral de 2.392 seccions censals de les aproximadament 36.000 seccions diferenciables a Espanya, seleccionant en cadascuna d'elles 10 llars de les que es recull la informació de tots els membres de la llar que hi resideixen. En aquest sentit, l'EPF es planifica amb una mida mostral de 23.920 llars, tot i que a causa d'abandonaments, la mida final pot ser lleugerament inferior. Cada any es renova la meitat de la mostra, motiu pel qual cada llar col·labora durant un màxim de dos anys.
Amb aquest esquema, la mostra de llars està repartida uniformement en períodes de 14 dies al llarg de l'any, de manera que la mostra anual es divideix en vint-i-sis grups de llars que comencen i acaben la seva col·laboració anual al mateix temps dins de cada grup.

Cada llar romà a la mostra dos anys consecutius incloent en cadascun d'ells un període de catorze dies durant el qual les famílies anoten, en llibretes destinades a aquesta finalitat, tots els béns i serveis adquirits. No obstant, sent dues setmanes un lapse de temps excessivament breu per englobar l'adquisició de tota la gamma de béns i serveis anuals, es sol·licita també, mitjançant una entrevista presencial, informació sobre les compres efectuades amb periodicitat superior a aquest interval, és a dir, adquisicions realitzades durant els dotze mesos anteriors.

Donat que el període de l'estudi s'allarga fins un any, és necessari classificar les despesses segons el que s'anomena **"periodicitat de les despeses"** donat que no tots els béns es poden agrupar dins de les dues setmanes d'anotació directa, és a dir, hi ha despeses de major freqüència que sí s'anotarien a la llibreta i despeses de menor freqüència que s'indicarien amb l'entrevista presencial. L'EPF classifica les despeses segons aquests cinc grups:

- Bisetmanal (despeses de major freqüència o d'imports petits).
- Mensual (despeses de mitjana freqüència o d'import moderat).
- Últim rebut (pagaments regulars).

- Trimestral (despeses de baixa freqüència o d'imports massa elevats per considerar-los bisetmanals o mensuals).
- Anual (despeses d'escassa freqüència o d'imports molt elevats).

D'aquesta manera, per tal de poder homogenitzar el període de temps i per facilitar-ne l'estudi, és necessari aplicar **Factors d'Elevació Temporal (FET)**, ja que la periodicitat de les despeses no és la mateixa per cadascuna d'elles. Els FET s'apliquen a les despeses amb la següent fórmula:

$$FET = \frac{T}{t}$$

Sent T la duració del període d'estudi (365 dies) i t la del període de referència, ambdues mesurades en dies. Els FET aplicats a cada periodicitat de despesa serien els següents:

Despeses	**Dies**	**FET**
Bisetmanals	14	26
Mensuals	30	12
Trimestrals	90	4
Anuals	360	1
Últim Rebut	...	Segons la periodicitat

Table 1: Taula de Factors d'elevació temporal en funció de la periodicitat de la despesa

El grup "Últim Rebut" representa un grup de despeses que es tenen amb una periodicitat variable i que es donen de forma regular en les llars, per exemple, el rebut de la llum, el rebut del gas, etc. Aquestes s'elevaran temporalment en funció de la periodicitat pròpia del mateix rebut.

Darrerament, per tal de fixar la despessa a un període anual s'aplica el FET de la següent manera:

$$Despesa\ Anual = \sum_{i=1}^{n_i} FET_i * Despeses_i$$

On i representa el nombre total de despeses.

Per altra banda, recordant que dins de cada secció censal es trien 10 llars mostrals per representar-la, també s'apliquen **Factors d'Elevació Espaial (FEE)** per tal d'inferir les dades mostrals a la població representada, és a dir, el FEE d'una llar mostral és el nombre de llars de la població que són representades per aquesta.

Els FEE s'utilitzen per trobar estimacions de la despesa total a nivell nacional. Aquests s'apliquen de la següent manera:

$$Despesa\ Anual\ Seccio\ Censal = FEE * Despesa\ Anual$$

Així doncs, per cada secció censal es trien 10 llars mostals i la suma de les seves despesa representa la despesa total de la secció censal.

3.1 Classificació del tipus de Despesa

D'altra banda, la despesa es classifica en funció del grup i subgrups al que pertanyen. És a dir, les despeses es classifiquen en uns grups principals que després es van subdividint un cert nombre de vegades fins a distribuir correctament la despesa en questió.

La codificació principal de les despeses és la coneguda com **COICOP** (*Classification of Individual Consumption According to Purpose*), estructurada en els següents dotze grans grups:

1. Aliments i begudes no alcohòliques.
2. Begudes alcohòliques, tabac i narcòtics.
3. Articles de vestir i calçat.
4. Llar, aigua, electricitat, gas i altres combustibles.
5. Mobiliari, equipament de la lar i despeses corrents de conservació de la llar.
6. Salut.
7. Transports.
8. Comunicacions.
9. Oci, espectacles i cultura.
10. Ensenyament.
11. Hotels, cafès i restaurants.
12. Altres béns i serveis.

Dins de cada gran grup de classificació existeixen un cert nombre de subgrups que podeu trobar dins de la pàgina web de l'INE[1].

[1] http://www.ine.es/daco/daco42/daco4213/anexoecpf06.pdf

3.2 Fitxers de Dades

Una vegada aclarit com es realitza el disseny de la mostra i com s'obtenen, s'eleven (temporal i espaialment) i es classifiquen les despeses cal parlar de les bases de dades resultants de l'EPF.

Concretament, s'obtenen tres bases de dades o fitxers anuals anomenats **Fitxer de la Llar**, **Fitxer dels Membres de la Llar** i **Fitxer de les Despeses**.

3.2.1 Fitxer de la Llar

Al Fitxer de la Llar s'inclouen característiques pròpies de la llar en sí, sense aprofundir molt en els seus membres més que en el sustentador principal. S'inclouen tants registres com llars mostrals hi ha a l'estudi, concretament hi ha 22.130 llars mostrals.

La informació del fitxer es pot resumir en vuit seccions diferenciables:

- Informació general: comunitat autònoma, grandària del municipi, densitat de població, etc.
- Característiques relatives a la llar: nombre de membres, grandària de la llar, nombre de fills, etc.
- Característiques relatives al sustentador principal: nombre d'hores que treballa, nivell d'estudis, sector d'activitat, etc.
- Característiques relatives a la vivenda principal: qüestions relatives a l'edifici, si la llar disposa d'aigua calenta, etc.
- Altres vivendes a disposició de la llar: mateixes qüestions que les relatives a la principal.
- Costos de consum de la llar: cost total anual segmentant en auto consum, auto subministrament, etc.
- Ingressos regulars mensuals de la llar: tipus de fonts dels ingressos.
- Nombre de menjars i sopars durant la bisetmana: es recullen el nombre de menjars i sopars per tots el membres de la llar que no són del servei domèstic, convidats, hostes, etc.

3.2.2 Fitxer dels Membres de la Llar

El Fitxer dels Membres de la Llar es centra en cadascun dels membres que habiten a la llar sol·licitant sexe, edat, nacionalitat, etc. Cada registre representa un membre de la llar concatenat-lo amb la llar a la qual pertany. Hi ha un total de 59.517 registres.
Les principals característiques d'aquesta base de dades són:

- Ingressos: per cada membre de la llar es recull si percep o no ingressos i, en cas afirmatiu, es demana l'import exacte o l'interval d'ingressos nets mensuals.
- Estudis: nivell d'estudis de cada membre de la llar que tingui una edat superior a setze anys.
- Nacionalitat: nacionalitat de cada membre així com la nacionalitat del pare i de la mare.
- Servei domèstic, hostes, convidats i altres membres que podrien ser considerats membres de la llar.
- Classificació dels membres: en funció de si són membres dependents, persones adultes, etc.

3.2.3 Fitxer de les Despeses

En el Fitxer de les Despeses estan recollides totes les despeses que ha tingut la llar al llarg de l'any. Cada registre representa un tipus de despesa diferenciada segons la codificació COICOP i alhora concatenada amb la llar a la qual pertany. Hi ha un total de 1.947.709 registres.
Les principals característiques que recull aquesta base de dades són:

- Despesa: la despesa de cada tipus de cost diferent elevada temporal i espaialment.
- Codi: codificació de la despesa.
- Quantitat: en cas de ser necessari, indiquen la quantitat de productes comprats en cada cas, també elevada pels dos factors.
- Naturalesa monetària: classifica la despesa segons el tipus de pagament.

4 Procesament de Dades

Per tal d'assolir el nostre objectiu i poder realitzar l'anàlisi adientment, és prioritari que les dades siguin procesades, és a dir, que a través de certs criteris de depuració i combinant informació que ens brinda l'EPF poguem assolir una Base de Dades pròpia amb la que treballar.

Aquest apartat tractarà sobre com s'ha gestionat i transformat la informació actual per tal d'assolir el nostre objectiu de manera eficient. Cal matitzar que moltes de les decisions preses a l'hora de realitzar depuracions o processaments s'han près de forma arbitrària, caldria certa supervisió segons necessitats contextuals.

4.1 Combinació de bases de dades

Cadascun dels tres fitxers de dades resultants de l'EPF conté variables que podrien ser interessant a l'hora de realitzar el nostre anàlisi, de tal manera que es triaran les que semblin que poden aportar més informació. A continuació, es tractarà d'explicar la informació que s'ha obtingut de cadascun dels fitxers.

Del Fitxer de la Llar es guardarà informació referent a la geolocalització de la llar, tant la referent a la comunitat autònoma com la referent a la regió a la qual pertany. També es guardarà informació referent a la mida del municipi, a la zona de residència (urbana de luxe, urbana inferior, rural, etc.), al nombre de membres que habiten a la llar i als ingressos, imputats i no imputats, que rep mensualment.

Del Fitxer de les Despeses es guardarà informació pròpia de les despeses, agrupant cadascuna d'elles dins dels grans grups de classificació COICOP (aliments, begudes alcohòliques, articles de vestir, etc..), és a dir, s'agruparà totes les despeses diferents (incloent despeses monetàries i no monetàries) que pertanyin al mateix subgrup de despesa quedant-nos així amb els 12 grans grups ja indicats.

Del Fitxer dels Membres de la Llar es guardarà informació relativa a l'edat i el sexe de cada membre de la llar per classificar-lo segons la categoria a la qual pertanyi.

La classificació dels membres s'ha dut a terme de la següent manera:

- Nounat: Menys de 2 anys.
- Infant: De 2 a 5 anys.
- Nen/a: De 5 a 12 anys.
- Adolescent Dona: De 12 a 16 anys i sent Dona.
- Adolescent Home: De 12 a 16 anys i sent Home.
- Jove Dona: De 16 a 25 anys i sent Dona.
- Jove Home: De 16 a 25 anys i sent Home.
- Adult Dona: De 25 a 65 anys i sent Dona.
- Adult Home: De 25 a 65 anys i sent Home.
- Sènior Dona: Més de 65 anys i sent Dona.
- Sènior Home: Més de 65 anys i sent Home.

Addicionalment, també es conservarà informació referent al nombre de membres que són dependents dins de la llar. Es considera membre dependent a tots aquells membres que tinguin menys de 16 anys o que tinguin entre 16 i 25 anys i siguin inactius econòmicament.

4.2 Elevacions espaials i temporals

Com s'ha comentat prèviament, les despeses de l'estudi estan elevades espaial i temporalment. Per tal de realitzar l'anàlisi de les llars de forma individual, s'eliminarà els factors d'elevació espaial, ja que aquests ponderen les llars mostrals de maneres diferents en funció del nombre de llars a les que aquesta representa, d'aquesta manera cada llar es representa a si mateixa i totes elles tindran el mateix pes.

L'elevació temporal sí és necessària de manera que totes les despeses seran homogeneitzades temporalment. D'aquesta manera, l'estudi analitza el consum de llars individuals en un període d'un any.

4.3 Despesa Total

Una decisió a pendre a l'hora de considerar les despeses una variable de l'estudi és escollir si es vol treballar amb despeses monetàries únicament o amb despeses monetàries i no monetàries alhora entenent ambdues com una sola despesa.

Primerament, cal diferenciar despesa monetària de despesa no monetària. La definició contextualitzada a l'EPF seria:

- Despesa monetària: acció per la que s'entreguen diners a canvi de béns i serveis, els diners entregats provenen de la llar.
- Despesa No monetària: Acció per la que s'entreguen diners a canvi de béns i serveis, els diners entregats provenen de membres o entitats alienes a la llar.

D'aquesta manera, és possible que una llar tingui un nivell de despesa molt baix en productes alimentaris ja que l'empresa on treballen alguns dels seus membres els proporciona el menjar.

Per tal de reduir la variabilitat que implica no considerar la despesa no monetària i donat que les llars que tenen despesa no monetària són massa elevades com per eliminar-les de la base de dades s'ha près la decisió de treballar amb la **Despesa Total de la Llar**, de tal manera que:

$$Despesa\ total = Despesa\ monetria + Despesa\ no\ monetria$$

A l'hora de realitzar aquesta transformació de la despesa s'ha de tindre en compte que el valor no monetari afegit a aquesta s'ha d'afegir també al sou de la llar, és a dir, si una empresa paga cada dia el menjar a un membre d'una llar s'interpretarà que aquest membre té un sou una mica més elevat. És a dir:

$$Sou\ total\ anual = Sou\ anual + Despesa\ no\ monetria$$

4.4 Depuració de les dades

Una vegada s'ha combinat els diferents fitxers de dades, ajustats els factors d'elevació i transformat les despeses en despeses totals es procedirà a depurar les dades.

Es començarà depurant la base de dades en funció del nombre de membres de la llar, donat que hi ha llars que tenen molts membres i això podria generar massa desviació provocant així un mal ajust.
A continuació es mostra la freqüència de membres per llar:

N. de Membres	**Freqüència**	**Freqüència Acumulada**
1	3.951	3.951
2	7.049	11.000
3	5.081	16.081
4	4.614	20.695
5	1.068	21.763
6	246	22.009
7	73	22.082
8	21	22.103
9	12	22.115
10	8	22.123
11	3	22.126
12	2	22.128
13	1	22.129
16	1	22.130

Table 2: Taula de Freqüències del Nombre de Membres de la Llar

S'eliminarà de la base de dades totes les llars amb més de 6 membres ja que les seves freqüències són força baixes. El tamany de la mostra passa de 22.130 llars a 22.009.

Seguidament es consideraran els ingresos. Uns ingressos excessivament alts o excessivament baixos podrien generar biaixos crítics.

Dins de cada llar hi ha membres que aporten ingresos i hi ha membres que són econòmicament dependents. Es consideren membres econòmicament dependents a tots aquells menors de 16 anys o tots aquells membres entre 16 i 25 anys que no aporten ingresos.

Dit això, el criteri seguit a l'hora de depurar els ingresos es centra en els membres que aporten ingresos a la llar, donat que no coneixem la xifra que aporta cada membre definirem una nova variable que representa la mitjana d'aportació econòmica de cada membre que treballa:

$$Mitjana\ Ingresos\ Mensuals\ per\ Membre\ que\ treballa = \frac{1}{12} * \frac{Sou\ Total\ Anual}{Nombre\ Membres\ que\ treballen}$$

Segueix la següent distribució:

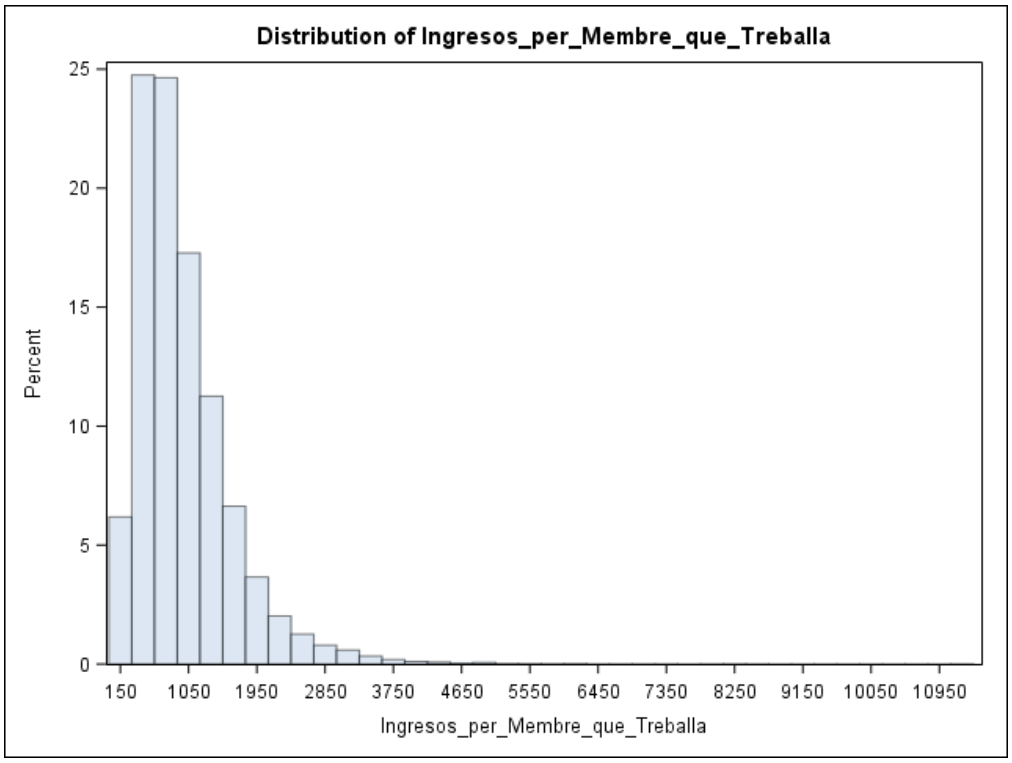

Figure 1: Histograma de la Mitjana dels Ingresos Mensuals per Membre que treballa

Depurarem totes les llars que tinguin una mitjana d'ingresos mensuals inferiors a 300 euros i totes les que tinguin una mitjana superior a 5.000 euros ja que no hi ha punts d'inflexió clars i així s'eliminarà els outliers més extrems. La distribució final és la següent:

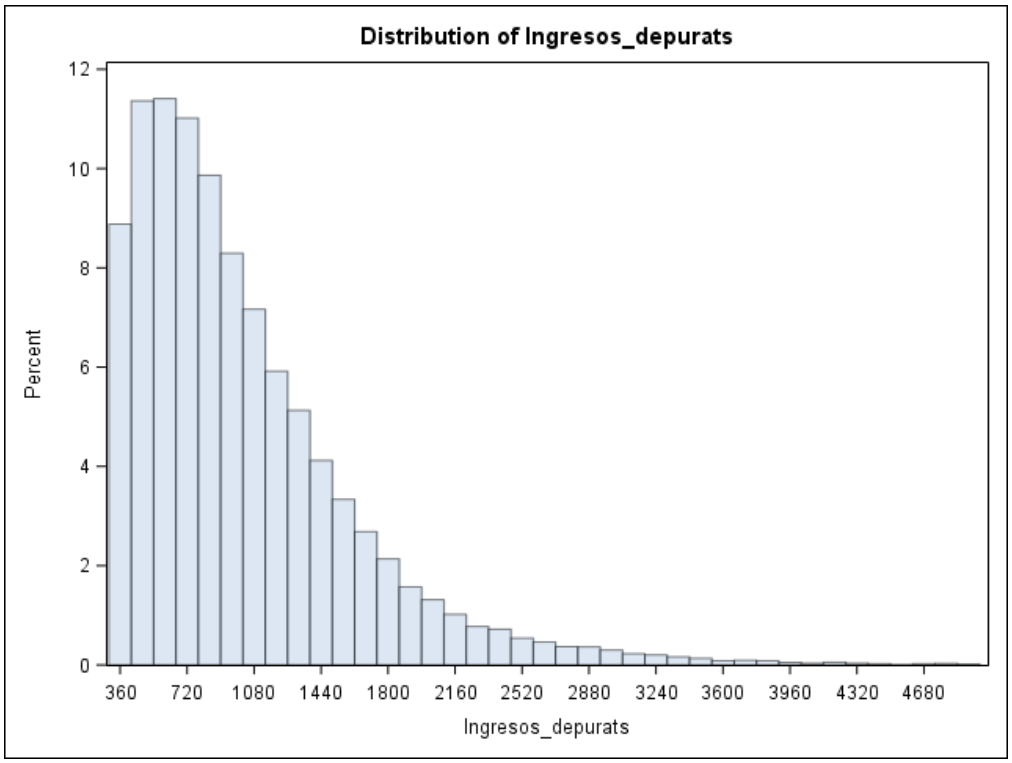

Figure 2: Histograma de la Mitjana dels Ingresos depurats Mensuals per Membre que treballa

De les 22.009 llars de la mostra es passa a 20.415.

Una altre criteri de depuració és observar com les llars gestionen la seva despesa, per exemple, si una llar empleа gran part del seus ingresos en un sol grup de despesa del criteri COICOP, serà depurada. Es depuraran les llars que gastin més del 80% dels seus ingressos en un sol grup de despesa.

Després de realitzar aquesta darrera depuració pasem de 20.415 llars a 20.376.

El darrer criteri de depuració empleat fa referència a les categories de despesa en les que una llar no ha gastat res, per exemple, és possible que una llar de dos persones adultes all llarg d'un any no gasti en educació.

D'aquesta manera i de forma arbitrària s'eliminarà totes les llars que tinguin 7 o més grups de COICOP sense despesa dels 12 possibles.

Finalment es passa de 20.376 llars a 20.154 llars amb les que es treballarà. En total s'han depurat 1.976 llars de la base de dades.

4.5 Base de dades de treball

En aquest apartat es parlarà sobre les variables que formen part de la base de dades. Totes han estat seleccionades seguint el criteri propi identificant les que poden resultar rellevants per poder donar una anàlisi més interesant.

Les variables seleccionades són:

- Grups de classificació COICOP: són 12 variables diferents que indiquen el cost que ha realitzat la llar dins de cada gran grup de despesa al llarg de l'any.
- Comunitat Autònoma: Comunitat Autònoma de residència de la llar. S'inclouen les illes Canàries i Balears i també Ceuta i Melilla.
- Regió: regió de residència de la llar (nord-est, sud, etc.). Es distingeixen també com a regions la Comunitat de Madrid i les illes Canàries.
- Mida del municipi: mida del municipi al qual pertany la llar. Els diferents grups estan classificats de la següent manera:
  - Municipi Molt gran: Més de 100.000 habitants
  - Municipi Gran: Entre 50.000 i 100.000 habitants
  - Municipi Mitjà: Entre 20.000 i 50.000 habitants
  - Municipi Petit: Entre 10.000 i 20.000 habitants
  - Municipi Molt Petit: Menys de 10.000 habitants
- Nombre de membres: nombre de membres que habiten a la llar.
- Zona de residència: tipus de residència a la qual pertany la llar. Cada residència està dintre d'una de les següents categories:
  - Urbana de Luxe.
  - Urbana Alta.
  - Urbana Mitjana.
  - Urbana Baixa.
  - Rural industrial.
  - Rural Pesquera.
  - Rural Agrària.
- Membres dependents: nombre de membres dependents que habiten a la llar.
- Tipus de membre: són 11 variables diferents que fan referència al nombre de nounats, infants, homes sèniors, etc. que habiten a la llar.
- Sou: ingressos que rep la llar anualment elevats temporalment, és a dir, multiplicats per dotze. No inclou pagues extres, ni pagues dobles, ni altres fonts d'ingressos que es puguin donar en un moment determinat com, per exemple, la venta d'un bé propi.

5 Mètodes Estadístics

En aquesta secció s'emplearà eines estadístiques per tal d'afrontar l'objectiu marcat. Es començarà amb una anàlisi descriptiva de les dades i, posteriorment, es procedirà a realitzar una classificació de les tipologies de despesa empleant mètodes de clusterització.

5.1 Anàlisi descriptiva

L'objectiu d'aquest punt és descriure la base de dades i donar un primer cop d'ull per tal de realitzar una primera anàlisi exploratòria de les variables. Donat que la base de dades conté més de vint-i-cinc variables, solament es comentaran trets peculiars o que es considerin importants.

Inicialment, com es pot comprobar a la figura 2, la distribució que segueixen els ingresos és una distribució molt assimètrica on les dades estan centrades entre els 350 i els 1.000 euros aproximadament. A partir d'aquí, la cua va baixant gradualment.
Per tal d'escalar les dades, aplicarem una transformació logarítmica en base 10. La idea és aconseguir que els ingresos segueixin una distribució Normal aproximadament.

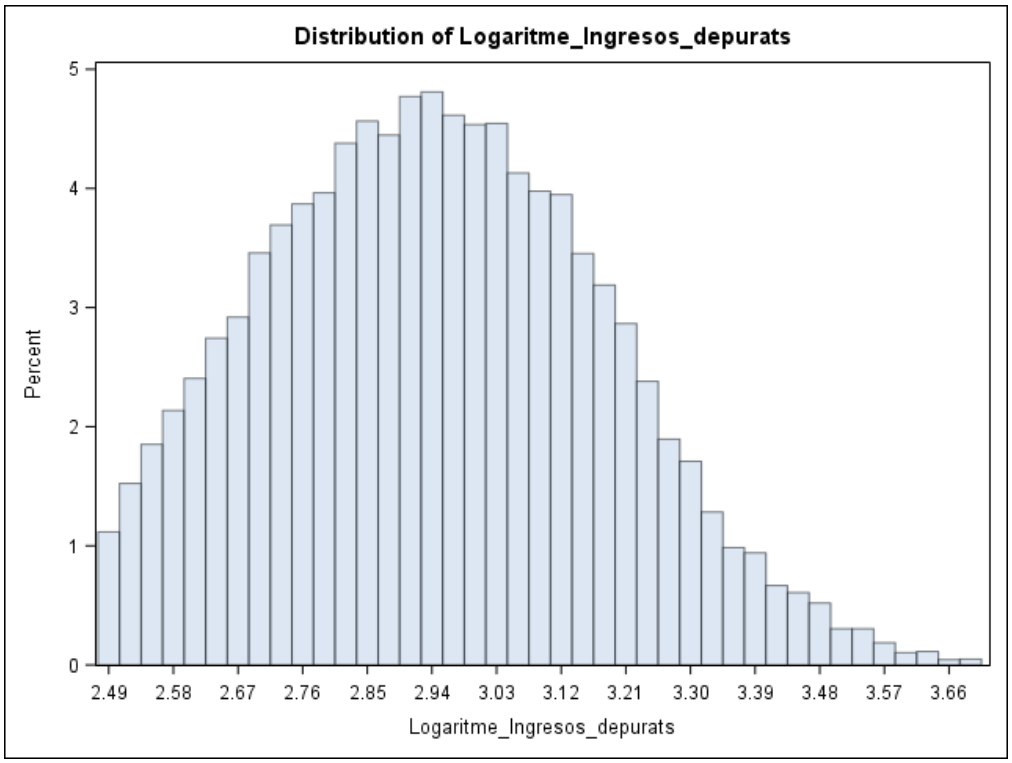

Figure 3: Histograma del logaritme de la Mitjana dels Ingresos depurats Mensuals per Membre que treballa

Aplicarem també la transformació logarítmica en base 10 a les despeses. Donat que per aquesta variable tenim algunes dades que prenen valor zero, se'ls hi sumarà una unitat per tal que el logaritme es pugui avaluar.
La majoria de les despeses té com a mínim un 10% de llars que no l'han consumit. Es parlarà d'aquesta inflació de zeros més endevant.

Categoria COICOP de despesa	Mínim	Percentil 10%	Mediana	Percentil 90%	Màxim
Aliments i begudes no alco...	0	3.15	3.59	3.89	4.69
Begudes alcohòliques, taba...	0	0	2.16	3.23	4.17
Articles de vestir i calça...	0	0	2.88	3.58	4.60
Llar, aigua, electricitat, ...	0	3.00	3.34	3.75	4.80
Mobiliari, equipament de l...	0	1.73	2.77	3.46	4.55
Salut	0	0	2.45	3.38	4.86
Transports	0	0	3.20	3.90	4.90
Comunicació	0	2.37	2.87	3.16	3.92
Oci, espectacles i cultura	0	0	2.93	3.63	4.71
Ensenyament	0	0	0	3.06	4.38
Hotels, cafès i restaurants	0	0	3.15	3.80	4.82
Altres béns i serveis	0	2.69	3.18	3.62	4.99

Table 3: Logarítme de la despesa en cada categoria COICOP

Seguidament farem una anàlisi univariant de la variable Comunitat Autònoma. El seu histograma és el següent:

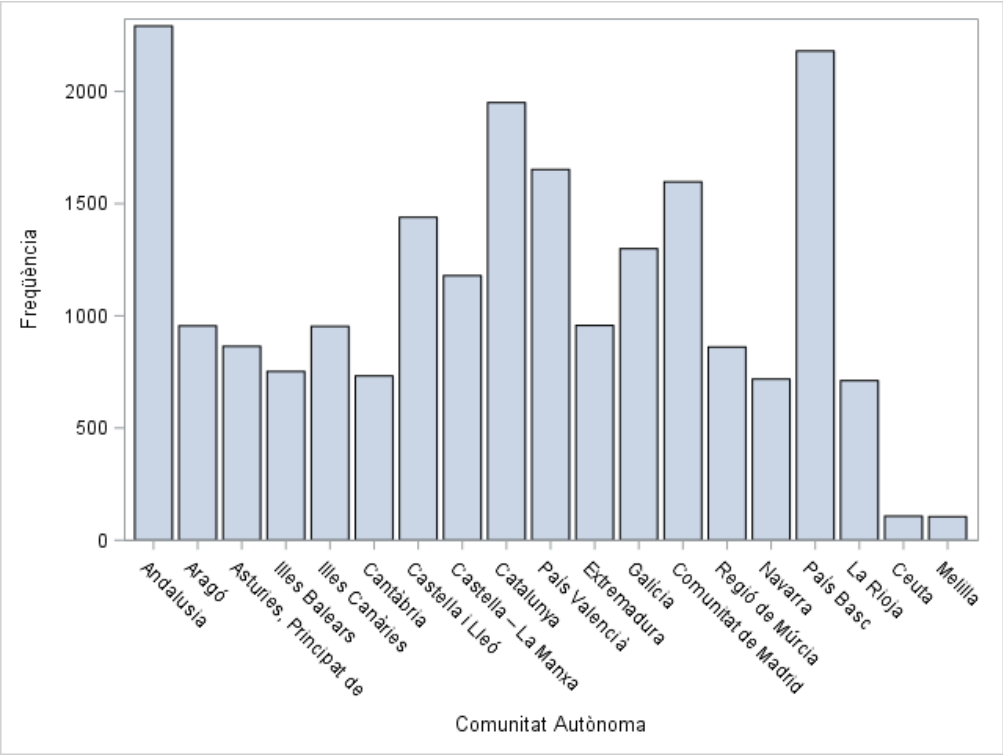

Figure 4: Histograma de Llars per Comunitat Autònoma

Les Comunitats Autònomes de les que tenim més mostres són Andalusia i País Basc, seguides per Catalunya, el País Valencià i la Comunitat de Madrid.

S'observa que tant de Ceuta com de Melilla la mostra és força baixa, això és degut a que hi ha molt poques seccions censals en aquestes comunitats, principalment, a causa del seu tamany. Amb una mostra d'aquestes dimensions es podria plantejar eliminar ambdues comunitats de la Base de Dades tot i que, com aquestes no afecten directament els objectius de l'estudi, es decideix deixar-les. No obstant, la seva interpretació podria ser dubtosa.
També s'observa una sobre representació de llars del País Basc, donat que en aquest cas la mida de la mostra s'ha duplicat a propòsit en col·laboració amb l'Institut d'Estadística d'aquesta comunitat autònoma.

Seguidament, l'histograma de la variable Zona de Residència:

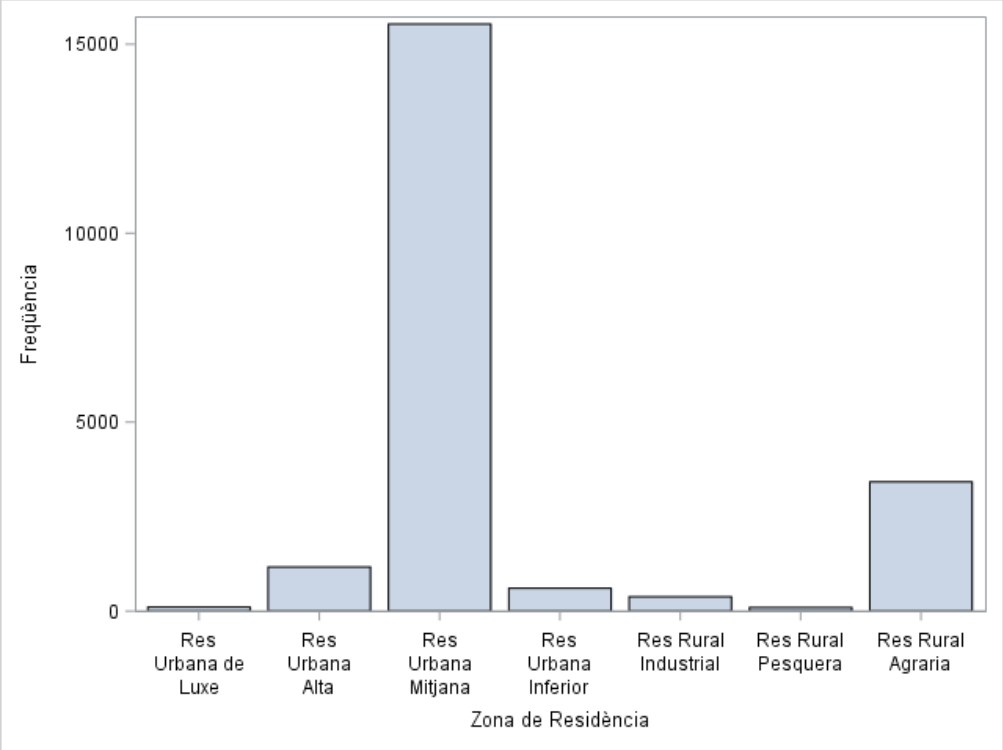

Figure 5: Histograma de Llars per Zona de Residència

Es veu clarament com la gran majoria de les llars, prop d'un 75%, estan ubicades en una Zona de Residència Urbana Mitjana mentre que Recidències Urbanes de Luxe o Rurals Pesqueres tenen 105 i 95 observacions respectivament.

Un altre cop tampoc es depuraran aquestes categories ja que no afecten als objectius.

5.2 Inflació de zeros

Com s'ha comentat prèviament, la majoria de categories COICOP de despesa tenen més d'un 10% d'inflació de zeros, això pot ser un greu problema a l'hora de treballar amb distàncies i poder construir tipologies de consum.

Una despesa de zero indica que una llar al llarg de l'any d'estudi no ha tingut cap cost d'aquella determinada despesa. Com ja s'ha comentat, aquesta situació es pot donar força sovint. Observem que es dona sobretot en el grup de despesa de l'ensenyament, tot i que també es dóna en grups com l'oci, la salut o els transports.

Aquesta situació es pot donar per diferents motius:

- La temporalitat de les despeses està fixada a un període màxim d'un any. Això implica que depeses de periodicitat superior a un any, per exemple la compra d'un automòbil, pateixin particularment de zeros inflats.
- La periodicitat de despeses de cost menor, com el tabac o l'alcohol, pel fet de ser despeses de cost menor se'ls atribueix una periodicitat menor a la que realment tenen. Això segurament es fa per reduir la probabilitat que la persona oblidi si ha comprat o no i reduir la variabilitat de l'obtenció de dades.
- Les despeses es donen amb una certa probabilitat, per exemple, és possible que una llar no consumeixi una despesa de periodicitat mensual durant el mes acotat a l'EPF, tot i que el següent mes sí que la consumeixi.
- Té sentit que certes llars no consumeixin mai segons quins tipus de despeses.

Totes aquestes situacions provoquen que moltes despeses tinguin un número excesivament elevat de zeros inflats, el qual dificulta en gran mesura l'objectiu d'aquest treball ja que a l'hora de realitzar clústers els zeros ponderaran molt i les conclusions tindran cert biaix.

Un cop entès el problema de la inflació i explicades les possibles causes d'aquesta, resta extreure una conclusió:

Sabent que totes les despeses es produeixen o no en funció de diferents factors: llars amb més membres dependents tenen una major probabilitat de consumir en educació, llars amb només dos persones adultes tenen una major probabilitat de viatjar, llars amb un major nombre d'ingresos gasten més en artícles de vestir, etc. Podem concluore que la periodicitat de la despesa moltes vegades ve fixada per la composició de la llar i pels ingresos de la mateixa, entre d'altres.

Això indica que podem estimar el cost esperat d'una llar en funció de les seves especificacions (número de membres, ingresos, comunitat autònoma, etc.) de manera que, donat que la gran majoria de zeros són en realitat dades censurades que en un futur consumiran la despesa en qüestió, aquests zeros poden ser tractats com a dades faltants i poden ser imputats per la seva estimació en funció d'algunes covariables.

5.3 Imputació de dades

Aquest apartat tractarà sobre el mètode empleat per a la imputació de les dades faltants, així com les variables empleades per a la imputació.

Es coneix com imputació la substitució de valors que no s'han obtingut (valors faltants o dades censurades) en un estudi per altres valors. Aquest pas és necessari per tal d'assegurar la màxima qualitat d'informació tot i que la interpretació de la variable despesa canviarà un cop imputem les dades. Es parlarà d'això més endevant.

Hi ha diverses tècniques d'imputació múltiple. Potser algunes de les més conegudes són: el mètode MLEM (*Maximum Likelihood Estimation with Missing Data*) que emplea a l'algoritme EM (*Expectation - Maximization*), el mètode d'imputació Monòtona o el MCMC (*Markov chain Monte Carlo*).

Es farà una breu introducció a cadascuna d'elles i darrerament s'exposarà el motiu pel qual s'ha escollit aplicar l'algoritme EM.

- Imputació de dades Monòtona: el procediment monòton consisteix en realitzar imputacions per dades faltants de manera seqüencial, és a dir, imputar dades triant variables amb pocs missings i aprofitar aquestes noves variables imputades per aconseguir imputar dades de variables amb un major nombre de missings i així trobar estimacions més robustes.
- Algoritme EM: l'algoritme EM és una tècnica que troba estimacions de màxima versemblança en models paramètrics per a dades incompletes. Aquest procediment aplica un procès iterant els següents pasos:
  - *Expectation* : Substitueix o imputa les dades per la seva esperança condicionada.
  - *Maximization*: Mitjançant estimacions de màxima versemblança es busca maximitzar el pas Expectation.

  D'aquesta manera, el procediment s'itera un cert nombre de vegades fins que les estimacions convergeixen i es considera que les imputacions són les màximes versemblants.
- MCMC: aquest procediment no només prediu, sinó que genera dades. És un mètode d'imputació múltiple i genera mostres de la base de dades imputada. D'aquesta manera permet tindre en compte la incertesa dels valors imputats.

El problema que es troba amb la Imputació de dades Monòtona és que aquesta no és tan potent com l'algorítme EM, que itera un procediment de màxima versemblança i generalment produeix resultats més robustos. D'altra banda el mètode MCMC, tot i ser un procediment molt complert, té una interpretació de les dades imputades més complexa que la de l'algorítme EM, que treballa amb esperances.

D'aquesta manera, tant per la seva simplicitat a l'hora d'interpretar els resultats com per la seva robustesa, s'aplicarà l'algoritme EM per tal d'imputar les despeses censurades de la base de dades.

Un altre punt important a comentar prèviament a aplicar l'algoritme EM és que, assumint que hi ha moltes despeses que tenen una periodicitat superior a un any (per exemple la compra d'un cotxe), s'ha d'ajustar aquesta periodicitat a un any, ja que volem que totes les despeses estiguin a la mateixa escala temporal i d'aquesta manera que totes tinguin el mateix pes.

Per exemple, si una llar es compra un cotxe que costa 10.000 euros cada 10 anys i ha coincidit que s'ha comprat el cotxe el mateix any de l'enquesta, aquesta llar tindrà com a despesa en vehicles 10.000 euros. Això provoca que, a l'hora d'imputar dades, sigui necessari un ajust doncs el valor a imputar no hauria de ser 10.000 euros sinó el promig que les llars gasten en compres de vehicle. Això implica que cal ajustar segons la periodicitat de la despesa. Aquesta periodicitat s'extreu a partir de justament el percentatge de llars sense despesa en el concepte pertinent.

Per tal d'ajustar la periodicitat de les dades a un any es trobarà l'esperança de que una despesa es produeixi, tornant amb l'exemple del cotxe, la compra d'un cotxe es produeix amb una periodicitat de 10 anys, d'aquesta manera dividint el preu del cotxe entre 10 obtindríem la despesa estimada del coche amb la periodicitat fixada a un any.

D'aquesta manera:

$$Despesa\ Anual = Despesa * periodicitat$$

On la periodicitat queda definida com l'esperança de compra d'un determinat producte en un període fixat d'un any. En el cas del coche, la periodicitat seria de 0.1. D'aquesta manera, la despesa esperada d'un cotxe és 1.000 euros a l'any, és a dir:

$$Periodicitat = \frac{Nombre\ Compres\ d'un\ producte}{Total\ producte}$$

On es considera compra d'un producte el fet d'haver gastat qualsevol quantitat d'ingresos en el mateix i el total del producte la suma del nombre de llars que compren i llars que no compren.

La periodicitat d'un producte varia en funció de les especificacions d'una llar (ingresos, número de membres, etc.). D'aquesta manera, s'ajusta la periodicitat de cada producte en funció de les especificacións de la llar i, un cop les despeses estiguin escalades a un període d'un any, podran imputar-se les despeses faltants.

Les despeses formen una mixtura entre una distribució binomial i una distribució normal. Mitjançant la binomial es decideix si la llar consumeix o no un determinat producte, ajustant així la seva periodicitat. En cas de que el consumeixi, la distribució normal determina la seva despesa. És a dir, s'ajustaran dotze models de mixtures on cadascun d'ells estimarà, en primer lloc, l'esperança de despesa d'una categoria COICOP en funció de covariables com els ingressos de la llar o el nombre de membres i, darrerament, estimant la periodicitat de la despesa, s'ajustarà aquesta a un període d'un any i s'estimarà el seu cost esperat.

A continuació s'aplicarà el procediment EM. Una manera de veure si les dades estan ben escalades és comprobant la mitjana abans i després de la imputació ja que, si s'ha fet bé, les mitjanes haurien de donar valors molt semblants.

Categoria COICOP de despesa	**Mitjana amb Missings**	**Mitjana amb dades Imputades**
Aliments i begudes no alco...	3.538	3.538
Begudes alcohòliques, taba...	1.753	1.753
Articles de vestir i calça...	2.398	2.402
Llar, aigua, electricitat, ...	3.348	3.348
Mobiliari, equipament de l...	2.635	2.634
Salut	1.977	1.980
Transports	2.774	2.800
Comunicació	2.741	2.743
Oci, espectacles i cultura	2.585	2.595
Ensenyament	0.834	0.804
Hotels, cafès i restaurants	2.711	2.724
Altres béns i serveis	3.153	3.153

Table 4: Taula de Mitjanes de les despeses amb Missings i de les despeses Imputades

Després d'imputar les dades no es pot seguir considerant la despesa com a despesa real de la llar. Com s'està treballant amb estimacions, la interpretació actual és estimació de la despesa.

Aquest canvi d'intepretació no té major rellevància tret que modifica l'objectiu d'aquest treball, que passa de "s'empraran eines d'anàlisi multivariant utilitzant anàlisi de clústers per tal d'assignar perfils de despesa a cada llar" a "**s'empraran eines d'anàlisi multivariant utilitzant anàlisi de clústers per tal d'assignar perfils de despesa estimada a cada llar**".

5.4 Tipologia de les despeses

Amb les despeses estimades sense dades faltants es pot iniciar l'anàlisi de clusters. La transformació logarítmica aplicada i el fet de que les dades estan a la mateixa escala temporal es pot considerar que és un millor ajust pel nostre objectiu que si estiguessin centrades i estandarditzades perquè, d'aquesta manera, els costos estan ponderats per la incidència econòmica estimada. És a dir, les diferents categories COICOP ponderen de forma diferent ja que hi ha algunes que representen una despesa estimada major i unes altres que representen una despesa estimada menor per les llars.

S'emplearan tres procediments diferents de clusterització començant pel procediment K-Means, després components principals amb el mateix procediment K-Means i darrerament ajustant un model de regressió i fent clúster sobre els residus.

5.4.1 K-Means

El mètode K-means és una eina força comuna dins del món dels clusters. Emplea un procediment iteratiu escollint un nombre concret de punts en funció del nombre de clusters desitjats i ubicant-los aleatòriament en un espai determinat. El procediment és simple: mitjançant distàncies euclidies compara cada observació amb els punts triats anteriorment. Cada punt representa un clúster i cada observació pertany al clúster que té més aprop. Un cop classificades totes les observacions, es calcula el centroide de cada clúster i es torna a iterar el procediment assignant cada observació al centroide que tingui més aprop, definint així nous clústers. Una vegada definits els nous clústers es tornarà a iterar el procediment un número determinat de vegades.

Tot i així, aquest procediment continua tenint un inconvenient molt gran en relació amb les nostres dades, ja que el que ens interesa és obtindre informació relativa als perfils de despesa de les llars. Empleant aquest mètode, les llars amb més membres tenen una despesa major que les llars amb menys membres, de la mateixa manera les llars amb més ingresos tenen una major despesa que les llars amb menys ingresos. D'aquesta manera, la classificació dels clústers els que ens està indicant realment són perfils de despesa en funció de les especificacions de la llar (número de membres, ingresos, etc.).

En resum, els clústers obtinguts de les variables originals es basen en magnituds i reflecteixen principalment la mida de la llar.

5.4.2 Components Principals

A continuació s'intentarà treure la correlació de les dades aplicant components principals. D'aquesta manera, es busca eliminar el biaix que provoca el nombre de membres i els ingresos de cada llar. Aquest procediment d'anàlisi multivariant emplea combinacions lineals de les variables per tal de definir unes noves variables incorrelades entre sí i així poder reduïr la dimensionalitat de la mostra. Cada component representa informació subjacent del sistema.

Una vegada s'ha aplicat el procediment, s'obtenen dotze components principals ordenades en funció de la variabilitat que explica cadascuna, sent així que la primera component explica un 30% de la variabilitat total, de la segona a la cinquena s'explica un altre 30% i el 40% de la variabilitat restant es troba de la sisena fins a la dotzena.

D'aquesta manera, la variància explicada per cada component es distribueix de la següent manera:

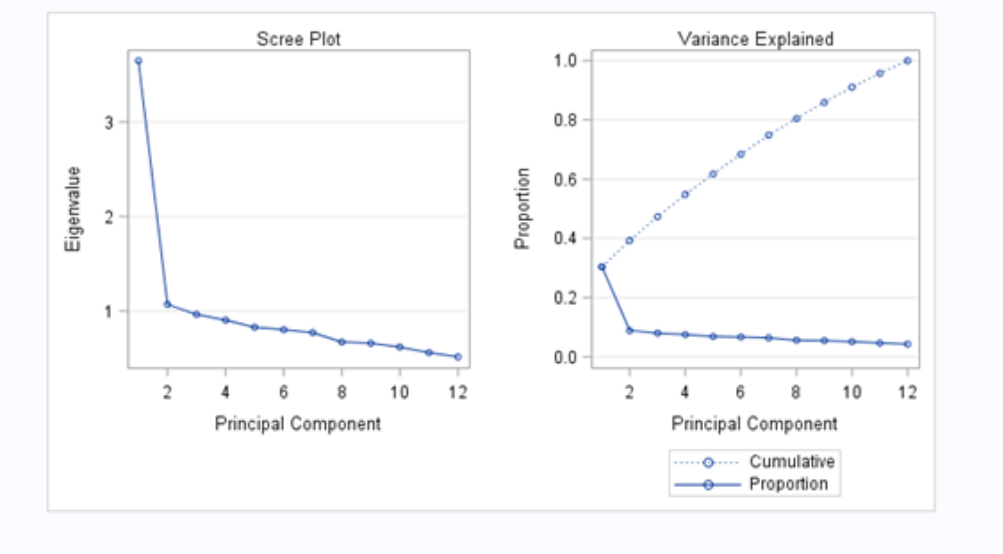

Figure 6: Variabilitat explicada per cada Component Principal

Les tres primeres components estan formades per:

Categora COICOP de despesa	Component 1	Component 2	Component 3
Aliments i begudes no alco...	0.2619	0.4813	0.2974
Begudes alcohòliques, taba...	0.1894	-.1129	0.8401
Articles de vestir i calça...	0.3098	-.0773	-.1283
Llar, aigua, electricitat, ...	0.1906	0.3221	-.2382
Mobiliari, equipament de l...	0.2775	0.4495	0.0822
Salut	0.2201	0.2835	-.1796
Transports	0.3381	-.3604	0.0147
Comunicació	0.2807	-.0179	-.1456
Oci, espectacles i cultura	0.3565	-.1906	-.0441
Ensenyament	0.2639	-.2167	-.2455
Hotels, cafès i restaurants	0.3542	-.3594	0.0373
Altres béns i serveis	0.3495	0.1432	-.0912

Table 5: Taula de Components Principals

A primera vista es pot observar com la primera component principal afecta a totes les categories de despesa de manera positiva. Això indica que augmentar en una unitat la primera component implicaria augmentar cadascuna de les despeses en el seu valor que ve representat per aquesta component. Com ja s'ha comentat, explica el 30% de la variabilitat total del sistema, així que és força probable que aquesta component expliqui els ingressos o el nombre de membres de la llar, ja que tots els seus coeficients prenen valors positius.

La segona component té una interpretació més complicada. Primerament, les despeses que augmenten amb aquesta component són els aliments, la llar (aigua, electricitat, etc.), el mobiliari, la salut i altres béns i serveis. Sembla aportar informació relativa a la estructura de la llar on l'objectiu principal és el menjar, la salut i el sostre.

Les despeses que augmenten amb la tercera component són els aliments, les begudes alcohòliques i el tabac, el mobiliari de la llar, transports i hotels, cafès i restaurants, cal remarcar que aquesta component té un índex altíssim en referència al tabac i les begudes alcohòliques. Addicionalment també té transports, hotels, cafès i restaurants, és probable que aquesta component intenti explicar una estructura de llar menys solida que la segona component.

Com s'acaba d'observar, la component que possiblement guarda informació referent al nombre de membres o als ingressos és la primera. Per tant, aquesta variable no es tindrà en compte a l'hora de realitzar l'anàlisis de clústers.

Per realitzar l'anàlisi s'utilitzaran les components compreses entre la dos i la sis, que otorguen un 38% de la variabilitat total de les despeses. Els resultats obtinguts són els següents:

Cluster Means					
Cluster	Prin2	Prin3	Prin4	Prin5	Prin6
1	-0.462349782	0.422489840	-0.056232910	-0.265440056	1.660588036
2	-2.387948305	-1.231430476	0.594589014	1.190638021	-0.097388839
3	0.165692545	-0.016983690	0.022074478	-0.001051654	-0.187877398
4	-4.658789352	1.305834123	-6.268856389	-1.030178828	-2.583220470
5	-7.043483788	0.735075304	-5.866399035	-0.769905193	1.319770505
6	-2.861386029	2.224760305	-6.539472227	-2.209583545	-2.088086459

Figure 7: Clusters amb PCA

El principal problema de l'anàlisi de components principals és la dificultat a l'hora d'interpretar els resultats. Per exemple, el primer clúster representa llars que prenen valors positius en la tercera component que, quan augmenta, ho fan despeses com el tabac i l'alcohol, els transports, hotels, etc. Mentre que la segona component pren valors negatius, penalitzant així una estructura de llar sòlida que pensa en la salut o en un bon equipament per la llar. Solament guiant-nos per aquestes components, el primer clúster sembla que fa referència a una estructura de llar algo caòtica.

Les freqüències de cada cluster són:

Cluster	**Freqüència**
1	2.232
2	679
3	18.285
4	22
5	8
6	76

Table 6: Taula de freqüències dels clústers de les components principals

Reduïnt el nombre de clusters les freqüències no milloren. El clúster número tres, l'únic cluster que pren un valor positiu per la segona component principal, que a l'hora representa una estructura familar sòlida, aborda la gran majoria de les llars de l'estudi.

Donada la seva recargolada interpretació, el fet de que les freqüències de cada clúster no són gaire bones i que no se sap si realment s'ha eliminat el pes que provoquen les característiques relatives a la llar, aplicarem un altre mètode de clusterització, aquesta vegada basat en models de regressió.

5.4.3 Clusteritzant Residus

La idea de modelitzar és eliminar el pes que poden tindre les característiques relatives de la llar sobre la despesa. Empleant un model de regressió lineal múltiple i escollint covariables que facin referència a aquestes categoríes (ingressos, membres dependents econòmicament, membres que treballen i edat del sustentador principal) per tal d'ajustar una esperança de la despesa basada en aquests criteris, dit d'una altra manera, s'ajustarà una regressió basada en característiques relatives a la llar, l'error que tingui aquesta regressió vindrà donat per característiques alienes a la llar.

D'aquesta manera s'elimina el pes que tenen els ingresos o el nombre de membres sobre la llar i les despeses ara estaran afectades per altres causes alienes a les que genera l'estructura de la llar.

Aquestes noves despeses seran residus de les despeses estimades amb les que s'ha treballat fins ara. Al ser residus estan centrats al zero i això provocarà que els clústers estiguin millor distribuïts.

Triant un altre cop sis clusers, els resultats obtinguts són els següents:

Categoria COICOP de despesa	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Aliments i begudes no alco...	0.000	0.020	-.073	-.230	0.040	0.018
Begudes alcohòliques, taba...	-.022	-.020	-1.203	0.618	-1.007	0.558
Articles de vestir i calça...	0.436	-1.432	0.291	-1.385	0.598	-.429
Llar, aigua, electricitat, ...	-.012	0.014	0.035	-.375	0.056	0.024
Mobiliari, equipament de l...	0.062	-.026	-.175	-.430	0.186	-.092
Salut	0.621	0.6425	-1.501	-1.134	0.076	-1.150
Transports	0.201	0.069	0.342	0.585	-1.592	-0.188
Comunicació	0.047	-.012	-.191	-1.475	-.021	0.066
Oci, espectacles i cultura	0.220	-.459	-1.126	-.475	-.296	0.116
Ensenyament	-.0 27	-.361	-.451	-.375	-.052	0.364
Hotels, cafès i restaurants	0.341	-1.307	-.696	0.562	-1.330	0.296
Altres béns i serveis	0.020	-0.025	0.014	-.218	-.021	-.014

Table 7: Taula de Clústers dels Residus

Les freqüències de cada clúster són:

Cluster	Freqüència
1	10.952
2	2.211
3	1.284
4	380
5	1.285
6	5.190

Table 8: Taula de freqüències dels clústers dels residus

Com es pot observar, les freqüències dels clústers estan millor distribuides.

No es poden interpretar directament els valors obtinguts a cada cluster, tot i que sí que es poden comparar entre ells. D'aquesta manera, es poden trobar perfils de despesa per cadascun d'ells. S'ha anomenat cada perfil de despesa de la següent manera:

- Cluster 1 - **Llars Càlides**: no tenen despeses molt més grans de l'esperat ni tampoc molt més petites. Aquest perfil de llar representa a més o menys la meitat de la població. Les seves prioritats de despesa són: artícles de vestir, salut, oci, transports i hotels, cafès i restaurants.

- Cluster 2 - **Llars Fredes**: aquest perfil de llar és el perfil que més consumeix en salut mentre que manté despeses molt baixes en artícles de vestir, hotels, cafès, restaurants i oci.

- Cluster 3 - **Estalviadores**: exceptuant artícles de vestir i transports, consumeixen molt menys de l'esperat en la gran majoria de despeses, sobretot en begudes alcohòliques o tabac, en salut i en oci.

- Cluster 4 - **Ocioses**: representen una minoria de 380 llars de la mostra. Aquest perfil de despesa es caracteritza per tindre el cost més elevat en begudes alcohòliques o tabac i en hotels, cafès i restaurants mentre que té els nivells més baixos de despesa en artícles de vestir i comunicació.

- Cluster 5 - **Còmodes**: no malgasten en hotels ni restaurants, tampoc en transports ni en begudes alcohòliques, i més o menys gasten l'esperat en educació.

- Cluster 6 - **Inquietes**: aquest perfil de despesa és el que més consumeix en ensenyament. També consumeix força més de l'esperat en begudes alcohòliques o tabac. Per contrapartida, té despeses molt baixes en salut i en artícles de vestir.

5.4.4 Clusterització dels perfils de despesa

Amb els perfils de despesa de les llars que s'han trobat es poden buscar associacions, és a dir, es poden intentar relacionar les diferents tipologies de despesa trobades amb altres variables d'interés com poden ser la geolocalització de la llar o el nivell d'estudis del sustentador econòmic principal.

Es pot relacionar la comunitat autònoma a la que resideix la llar amb la seva tipología de despesa?

El test de la chi-quadrat per comparar proporcions entre n grups independents dona un p-valor molt significatiu. D'aquesta manera, la comunitat autònoma afecta directa o indirectament als perfils de despesa de les llars.

Una forma de mirar quines comunitats autònomes es poden relacionar amb quins perfils de despesa és mirant els marginals fila i comparant proporcions entre grups. Una altra forma de relacionar les comunitats, juntament amb altres categories, és empleant arbres de classificació.

Els arbres de classificació, com bé indica la paraula, classifiquen variables entre sí trobant així similituds que serveixen per relacionar-les.

Desgraciadament, no s'ha trobat patrons que serveixin per relacionar apropiadament les tipologies de despesa esperada d'aquest estudi. S'ha començat buscant similituds entre les tipologies de despesa i la geolocalització de la llar (comunitat autònoma, tamany del municipi i zona de residència). La classificació ha agrupat la gran majoria de les llars en el primer clúster i la resta en el clúster número sis.

Seguidament, s'han intentat trobar relacions amb variables de caràcter més personal (nacionalitat i estudis del sustentador principal i sector on treballa). Els resultats obtinguts han sigut pràcticament els mateixos que els obtinguts a l'hora de relacionar la tipologia amb la geolocalització.

Finalment, s'ha buscat relació empleant geolocalització de la llar i característiques relatives al sustentador principal. Els resultats obtinguts tampoc han variat gaire.

L'arbre de classificació resultant és el següent:

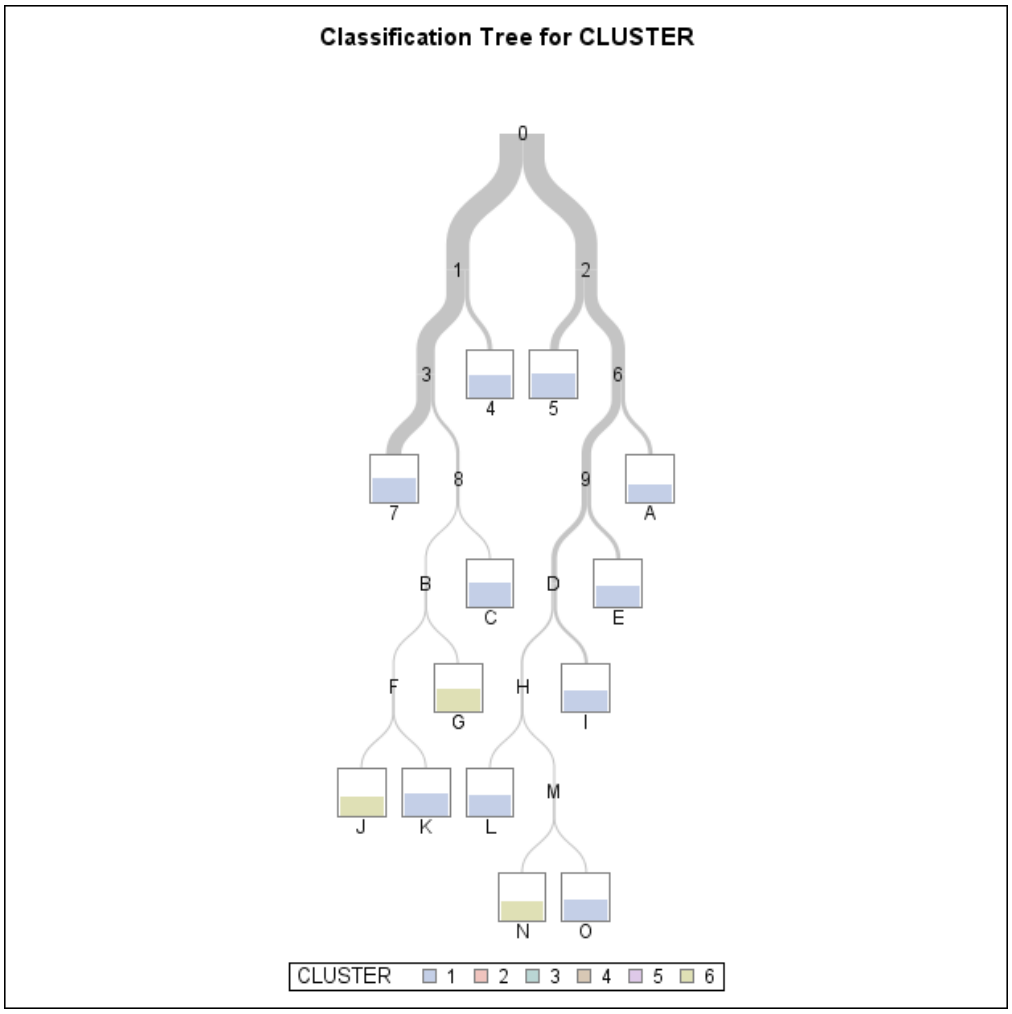

Figure 8: Arbre de classificació de les tipologíes de despesa

Com es pot observar, totes les llars s'han ubicat dins del primer i el sisè clúster.

6 Conclusions

El fet de que les tipologies de despesa siguin una característica explicada per molts factors que no es poden quantificar provoca que la seva classificació sigui força complicada. Addicionalment, el fet de que els clústers no estiguin equilibrats provoca un mal ajust si hi ha molta variabilitat no explicada.

Al marge de que la relació de la tipologia de despesa esperada amb factors com la geolocalització de la llar o característiques relatives als sustentador principal no dona els resultats desitjats, s'han aconseguit els objectius principals del treball.

S'ha donat un enfoc poc convencional i rarament empleat a l'EPF amb l'objectiu d'aportar noves eines estadístiques i mostrar així noves perspectives que poden ser empleades a l'hora de realitzar estudis sobre l'enquesta.

Els perfils de despesa esperada tenen característiques úniques i dispars entre elles, i s'han pogut realitzar sis grups:

- Llars Càlides
- Llars Fredes
- Estalviadores
- Ocioses
- Còmodes
- Inquietes

Finalment, al llarg de l'estudi s'han près algunes decisions de manera arbitrària i caldria una revisió d'una certa supervisió per tal d'asegurar la màxima qualitat d'informació.

7 Bibliografia

Enquesta de Pressupostos Familiars al web de l'INE:
http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176806&menu=resultados&secc=1254736195147&idp=1254735976608

Algoritme EM en SAS:
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htmstatug_mi_details02.htm

Imputació Monòtona en SAS:
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htmstatug_mi_details06.htm

MCMC en SAS:
http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htmstatug_mi_details21.htm

Finite Mixture Models en SAS:
http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htmstatug_fmm_syntax01.htm

A Annexes

A.1 Codi

```
> # options ps=2000 ls=120 nodate nonumber;
> # %let r=C:\Users\Joan\Desktop\Encuesta Presupuestos Familiares;
> # libname presup "&r/Bases de dades";
> # *TotalTip2015 & Miembros2015;
> #
> # *******;
> # data membres;
> # set presup.miembros2015;
> # keep numero sexo edad ninodep;
> # run;
> #
> # data membres;
> # set membres;
> # NDEP =input(ninodep, 1.);
> # drop ninodep;
> # run;
> #
> # data membres;
> # set membres;
> # if edad <= 2 then nounat=1;
> # else nounat=0;
> # if edad >2 and edad <=5 then Infant=1;
> # else Infant=0;
> # if edad >5 and edad <=12 then Nen=1;
> # else Nen=0;
> # if edad >12 and edad <=16 and sexo="1"  then AdolescentH=1;
> # else AdolescentH=0;
> # if edad >12 and edad <=16 and sexo="6"  then AdolescentD=1;
> # else AdolescentD=0;
> # if edad >16 and edad  <=25 and sexo="1" then JoveH=1;
> # else JoveH=0;
> # if edad >16 and edad <=25 and sexo="6" then JoveD=1;
> # else JoveD=0;
> # if edad >25 and edad <=65 and sexo="1" then AdultH=1;
> # else AdultH=0;
> # if edad >25 and edad <=65 and sexo="6" then AdultD=1;
> # else AdultD=0;
> # if edad >65 and sexo="1" then VellH=1;
> # else VellH=0;
> # if edad >65 and sexo="6" then VellD=1;
> # else VellD=0;
> # if ndep not =1 then ndep=0;
> # else ndep=1;
> # run;
> #
> # proc sql;
> # create table membres2 as
> # select numero, sum(ndep) as ninodep,  sum(nounat) as nounat, sum(Infant)
```

```
> #  as Infant, sum(Nen) as Nen, sum(AdolescentH) as AdolescentH,
> #                 sum(AdolescentD) as AdolescentD, sum(JoveH) as JoveH, sum(JoveD) as
> #  JoveD, sum(AdultH) as AdultH
> #                 , sum(AdultD) as AdultD, sum(VellH) as VellH, sum(VellD) as VellD
> # from membres
> # group by numero;
> # run;
> #
> # data membres2;
> # set membres2;
> # nmiemb= sum(nounat + Infant + Nen + AdolescentH + AdolescentD + JoveH +
> #  JoveD + AdultH + AdultD + VellH  +VellD);
> # run;
> #
> # data membres3;
> # set membres2;
> # if                 nmiemb=1 and vellH=1 then TipLlar=1;
> # else if nmiemb=1 and vellD=1 then TipLlar=2;
> # else if nmiemb=1 and (JoveH=1 or AdultH=1) then TipLlar=3;
> # else if nmiemb=1 and (JoveD=1 or AdultD=1) then TipLlar=4;
> # else if nmiemb=2 and ninodep=0 and vellH=0 and vellD=0 then TipLlar=5;
> # else if nmiemb=2 and ninodep=0 and (vellH=>1 or vellD=>1) then TipLlar=6;
> # else if nmiemb=2 and ninodep=1 then TipLlar=7;
> # else if nmiemb=2 then TipLlar=999;
> # else if nmiemb=3 and ninodep=0 and vellH=0 and vellD=0 then TipLlar=10;
> # else if nmiemb=3 and ninodep=1 and vellH=0 and vellD=0 then TipLlar=11;
> # else if nmiemb=3 and ninodep=2 then TipLlar=12;
> # else if nmiemb=3 and ninodep=0 and (vellH=>1 or vellD=>1) then TipLlar=13;
> # else if nmiemb=3 and ninodep=1 and (vellH=>1 or vellD=>1) then TipLlar=14;
> # else if nmiemb=3 then TipLlar=16;
> # else if nmiemb=4 and ninodep=0 and vellH=0 and vellD=0 then TipLlar=17;
> # else if nmiemb=4 and ninodep=1 and vellH=0 and vellD=0 then TipLlar=18;
> # else if nmiemb=4 and ninodep>=2 and vellH=0 and vellD=0 then TipLlar=19;
> # else if nmiemb=4 and ninodep=0 and (vellH=>1 or vellD=>1) then TipLlar=21;
> # else if nmiemb=4 and ninodep=1 and (vellH=>1 or vellD=>1) then TipLlar=22;
> # else if nmiemb=4 and ninodep=3 and vellH=0 and vellD=0 then TipLlar=20;
> # else if nmiemb=4 and ninodep>=2 and (vellH=>1 or vellD=>1) then TipLlar=23;
> # else if nmiemb=4 and ninodep=3 and (vellH=>1 or vellD=>1) then TipLlar=24;
> # else if nmiemb=4 then TipLlar=25;
> # else if nmiemb=>5 and ninodep=0 and vellH=0 and vellD=0 then TipLlar=26;
> # else if nmiemb=>5 and ninodep=1 and vellH=0 and vellD=0 then TipLlar=27;
> # else if nmiemb=>5 and ninodep=2 and vellH=0 and vellD=0 then TipLlar=28;
> # else if nmiemb=>5 and ninodep=>3 then TipLlar=29;
> # else if nmiemb=>5 and ninodep=0 and (vellH=>1 or vellD=>1) then TipLlar=31;
> # else if nmiemb=>5 and ninodep=1 and (vellH=>1 or vellD=>1) then TipLlar=32;
> # else if nmiemb=>5 and ninodep=2 and (vellH=>1 or vellD=>1) then TipLlar=33;
> # else if nmiemb=>5 then TipLlar=36;
> # run;
> #
> # ************************;
> #
> # proc freq data=membres3;
```

```
> # *TITLE "Freqüències inicials";
> # tables tipllar;
> # run;
> #
> # data TotTip;
> # set presup.Totaltip2015;
> # run;
> #
> # data Llar;
> # merge TotTip membres3;
> # by numero;
> # sou=(GASTOT/FACTOR);
> # run;
> #
> #
> # data provan;
> # set Llar;
> # /*where numero='00001' or numero='00002' or numero='00003' or numero='00004'
> #  or numero='00005';*/
> #
> # array g _0: _1: ;
> # do i=1 to dim(g);
> # g{i} = log10(g{i}+1);
> # end;
> #
> # li = log10(sou+1);
> #
> # keep numero _01 _02 _03 _04 _05 _06 _07 _08 _09 _10 _11 _12 li gasto2 sou
> #  ccaa NUTS1 tamamu nmiemb tipllar ninodep tamano zonares edadsp
> #  nounat Infant  Nen  AdolescentH AdolescentD  JoveH  JoveD AdultH
> #  AdultD VellH VellD ESTUDIOSSP ACTESTB NACIONASP;
> # run;
> #
> # *La mediana de cost en menjar en espanya es de 3800???;
> # proc univariate data=provan ;
> # var _:  li ;
> # histogram / endpoints = 0 to 6 by 0.1;
> # run;
> #
> # *Definim una nova variable "Treb" que indica les persones de +25 anys
> #  o que aporten ingresos a la llar;
> # data prova2;
> # set provan;
> # Treb= NMIEMB - ninodep;
> # IngTreb= (sou/Treb)/12;
> # if IngTreb>300 and IngTreb<5000 and NMIEMB<6 and treb>=1 ;
> # run;
> #
> # proc means data=prova2 nolabels min p5 p10 median p90 p95 max;
> # var _0: _1: li sou;
> # run;
> #
```

```
> # ************************************************************;
> # ************************************************************;
> # ************************************************************;
> # ************************************************************;
> #
> # data validacio;
> # set prova2;
> # ojo=1;
> # if _01 > li + log10(0.8) then ojo=ojo*2;
> # if _02 > li + log10(0.8) then ojo=ojo*3;
> # if _03 > li + log10(0.8) then ojo=ojo*5;
> # if _04 > li + log10(0.8) then ojo=ojo*7;
> # if _05 > li + log10(0.8) then ojo=ojo*11;
> # if _06 > li + log10(0.8) then ojo=ojo*13;
> # if _07 > li + log10(0.8) then ojo=ojo*17;
> # if _08 > li + log10(0.8) then ojo=ojo*19;
> # if _09 > li + log10(0.8) then ojo=ojo*23;
> # if _10 > li + log10(0.8) then ojo=ojo*31;
> # if _11 > li + log10(0.8) then ojo=ojo*37;
> # if _12 > li + log10(0.8) then ojo=ojo*41;
> # if ojo >1 then delete;
> # run;
> #
> # proc freq data=validacio;
> # tables ojo;
> # run;
> #
> #
> # data validacio2;
> # set validacio;
> # ojo2=0;
> # if _01 <= log10(1) then ojo2=ojo2+1;
> # if _02 <= log10(1) then ojo2=ojo2+1;
> # if _03 <= log10(1) then ojo2=ojo2+1;
> # if _04 <= log10(1) then ojo2=ojo2+1;
> # if _05 <= log10(1) then ojo2=ojo2+1;
> # if _06 <= log10(1) then ojo2=ojo2+1;
> # if _07 <= log10(1) then ojo2=ojo2+1;
> # if _08 <= log10(1) then ojo2=ojo2+1;
> # if _09 <= log10(1) then ojo2=ojo2+1;
> # if _10 <= log10(1) then ojo2=ojo2+1;
> # if _11 <= log10(1) then ojo2=ojo2+1;
> # if _12 <= log10(1) then ojo2=ojo2+1;
> # if ojo2 >= 8 then delete;
> # run;
> #
> # proc freq data=validacio2;
> # tables ojo2;
> # run;
> #
> # proc freq data=validacio2;
> # tables tipllar;
```

```
> # run;
> #
> #
> # ********************************************;
> # *
> # * Anàlisi Descriptiva
> # *
> # ********************************************;
> #
> # proc format;
> # value $ comun
> # '01' ='Andalusia'
> # '02' ='Aragó'
> # '03' ='Asturies, Principat de'
> # '04' ='Illes Balears'
> # '05' ='Illes Canàries'
> # '06' ='Cantàbria'
> # '07' ='Castella i Lleó'
> # '08' ='Castella - La Manxa'
> # '09' ='Catalunya'
> # '10' ='País Valencià'
> # '11' ='Extremadura'
> # '12' ='Galícia'
> # '13' ='Comunitat de Madrid'
> # '14' ='Regió de Múrcia'
> # '15' ='Navarra'
> # '16' ='País Basc'
> # '17' ='La Rioja'
> # '18' ='Ceuta'
> # '19' ='Melilla';
> #
> # value $ dens
> # '1' = 'Municipi Molt Gran'
> # '2' = 'Municipi Gran'
> # '3' = 'Municipi Mitjà'
> # '4' = 'Municipi Petit'
> # '5' = 'Municipi Molt Petit';
> #
> # value $ res
> # '1' = 'Res Urbana Luxe'
> # '2' = 'Res Urbana Alta'
> # '3' = 'Res Urbana Mitja'
> # '4' = 'Res Urbana Inferior'
> # '5' = 'Res Rural Industrial'
> # '6' = 'Res Rural Pesquera'
> # '7' = 'Res Rural Agraria';
> #
> # value $ regio
> # '1' = 'Nord-oest'
> # '2' = 'Nord-est'
> # '3' = 'Comunitat de Madrid'
> # '4' = 'Central'
```

```
> # '5' = 'Est'
> # '6' = 'Sud'
> # '7' = 'Illes Canàries';
> # run;
> #
> # data validacio2;
> # set validacio2;
> # format CCAA comun. TAMAMU dens. ZONARES res. NUTS1 regio.;
> # drop tamano;
> # run;
> #
> #
> #
> # PROC SGPLOT DATA=validacio2 NOAUTOLEGEND;
> # VBAR CCAA;
> # YAXIS LABEL="Freqüència";
> # XAXIS LABEL="Comunitat Autònoma";
> # RUN;
> #
> # PROC SGPLOT DATA=validacio2 NOAUTOLEGEND;
> # VBAR CCAA/GROUP=ZONARES STAT=MEAN;
> # YAXIS LABEL="freqüència";
> # RUN;
> #
> #
> #
> # proc means data=validacio2 median nolabels noobs;
> # class CCAA;
> # var _0: _1:;
> # run;
> #
> #
> # proc means data=validacio2 nolabels min p5 p10 p20 median mean p90 p95 max;
> # var _0: _1: li;
> # run;
> #
> #
> # data validacio3;
> # set validacio2;
> # CTOTAL= _01+_02+_03+_04+_05+_06+_07+_08+_09+_10+_11+_12;
> # m01 = _01/CTOTAL ;
> # m02 = _02/CTOTAL ;
> # m03 = _03/CTOTAL ;
> # m04 = _04/CTOTAL ;
> # m05 = _05/CTOTAL ;
> # m06 = _06/CTOTAL ;
> # m07 = _07/CTOTAL ;
> # m08 = _08/CTOTAL ;
> # m09 = _09/CTOTAL ;
> # m10 = _10/CTOTAL ;
> # m11 = _11/CTOTAL ;
> # m12 = _12/CTOTAL ;
```

```
> # run;
> #
> #
> #
> # ********************************************;
> # * Model FMM *;
> # * Models per explicar la magnitud de la despesa i l'esdeveniment de la despesa;
> # ********************************************;
> #
> # data validacio3;
> # set validacio2;
> # zonares2 = zonares+0;
> # run;
> #
> # %macro fmm(var);
> # proc fmm data=validacio3 gconv=1e-5;
> # class ;
> # model &var = li treb ninodep zonares2 edadsp /dist=Normal;
> # MODEL &var = / DIST=constant(0);
> # PROBMODEL li treb ninodep zonares2 edadsp;
> # OUTPUT out=validacio3 predicted(component)=p MIXPROBS(component)=m;
> # run;
> # data validacio3;
> # set validacio3;
> # z&var = &var * (1 - m_2);
> # drop p_1 p_2 m_1 m_2;
> # run;
> # %mend;
> #
> # %fmm(_01);
> # %fmm(_02);
> # %fmm(_03);
> # %fmm(_04);
> # %fmm(_05);
> # %fmm(_06);
> # %fmm(_07);
> # %fmm(_08);
> # %fmm(_09);
> # %fmm(_10);
> # %fmm(_11);
> # %fmm(_12);
> #
> #
> # ********************************************;
> # * IMPUTACIO de 0s *;
> # ********************************************;
> # data validacio4;
> # set validacio3;
> # if _01 <= log10(1) then _01= .;
> # if _02 <= log10(1) then _02= .;
> # if _03 <= log10(1) then _03= .;
> # if _04 <= log10(1) then _04= .;
```

```
> # if _05 <= log10(1) then _05= .;
> # if _06 <= log10(1) then _06= .;
> # if _07 <= log10(1) then _07= .;
> # if _08 <= log10(1) then _08= .;
> # if _09 <= log10(1) then _09= .;
> # if _10 <= log10(1) then _10= .;
> # if _11 <= log10(1) then _11= .;
> # if _12 <= log10(1) then _12= .;
> # run;
> #
> # proc print data=validacio4 (obs=10);
> # var _: ;
> # run;
> #
> # proc mi seed=123 data=validacio4 nimpute=0 ;
> # em out=validacio5 outem=c INITIAL=CC;
> # var _: li treb ninodep zonares2 edadsp  ;
> # run;
> #
> # proc print data=validacio4 (obs=10);
> # var _: ;
> # run;
> # proc print data=validacio5 (obs=10);
> # var _: ;
> # run;
> #
> # proc means data=validacio3 nolabels;
> # var _: ;
> # proc means data=validacio5 nolabels;
> # var _: ;
> # run;
> #
> #
> # ********************************************;
> # * IMPUTACIO de 0s *;
> # *Amb els valors de despesa ponderats;
> # ********************************************;
> #
> # data validacio4;
> # set validacio3;
> # if z_01 <= log10(1) then z_01= .;
> # if z_02 <= log10(1) then z_02= .;
> # if z_03 <= log10(1) then z_03= .;
> # if z_04 <= log10(1) then z_04= .;
> # if z_05 <= log10(1) then z_05= .;
> # if z_06 <= log10(1) then z_06= .;
> # if z_07 <= log10(1) then z_07= .;
> # if z_08 <= log10(1) then z_08= .;
> # if z_09 <= log10(1) then z_09= .;
> # if z_10 <= log10(1) then z_10= .;
> # if z_11 <= log10(1) then z_11= .;
> # if z_12 <= log10(1) then z_12= .;
```

```
> # run;
> #
> # proc print data=validacio4 (obs=10);
> # var z_: ;
> # run;
> #
> # proc mi seed=123 data=validacio4 nimpute=0 ;
> # em out=validacio5 outem=c INITIAL=CC;
> # var z_: li treb ninodep zonares2 edadsp  ;
> # run;
> #
> # proc print data=validacio4 (obs=10);
> # var z_: ;
> # run;
> # proc print data=validacio5 (obs=10);
> # var z_: ;
> # run;
> #
> # proc means data=validacio3 nolabels;
> # var _: ;
> # proc means data=validacio5 nolabels;
> # var z_: ;
> # run;
> #
> #
> # ***********************************************;
> # ***********************************************;
> # ***********************************************;
> #
> #
> # *1) DISTANCIES BRUTES (Reproduiran l'estructura de la llar...!!! ;
> #
> # proc fastclus data=validacio5 out=clust1 maxclusters=8;
> # var _01 _02 _03 _04 _05 _06 _07 _08 _09 _10 _11 _12;
> # run;
> #
> # proc means data=clust1;
> # class cluster;
> # var li nmiemb ninodep TAMAMU;
> # run;
> #
> #
> #
> # *2) PCA ;
> # proc princomp data=validacio5 out=clust2;
> # var _0: _1: ;
> # run;
> # proc fastclus data=clust2 out=clust2a maxclusters=5;
> # var prin2 - prin6;
> # run;
> #
> #
```

```
> #
> # *3) Models ;
> #
> # data validacio6;
> # set validacio5;
> #
> # %macro resid(var);
> # proc glm data=validacio6;
> # model &var = li treb ninodep edadsp ;
> # OUTPUT out=validacio6 residual=res ;
> # run;
> # data validacio6;
> # set validacio6;
> # y&var = res;
> # drop res;
> # run;
> # %mend;
> #
> # %resid(_01);
> # %resid(_02);
> # %resid(_03);
> # %resid(_04);
> # %resid(_05);
> # %resid(_06);
> # %resid(_07);
> # %resid(_08);
> # %resid(_09);
> # %resid(_10);
> # %resid(_11);
> # %resid(_12);
> #
> #
> # proc fastclus data=validacio6 out=clust2 maxclusters=6;
> # var y_: ;
> # run;
> #
> #
> #
> # proc freq data=clust1;
> # tables ccaa * cluster / chisq;
> # run;
> #
> #
> # proc freq data=clust2 ;
> # tables ccaa * cluster / out=table1 outpct;
> # run;
> # proc means data=table1 noobs mean;
> # class CCAA CLUSTER;
> # var PCT_ROW ;
> # run;
> #
> #
```

```
> #
> # proc hpsplit data=clust2 cvmodelfit seed=123;
> # class  ccaa cluster  NUTS1 TAMAMU ZONARES/ upcase;
> # model cluster = ccaa NUTS1 TAMAMU ZONARES ;
> # prune costcomplexity(leaves=10);
> # run;
> #
> # proc hpsplit data=clust2 cvmodelfit seed=123;
> # class  cluster  NACIONASP ESTUDIOSSP ACTESTB/ upcase;
> # model cluster = NACIONASP ESTUDIOSSP ACTESTB ;
> # prune costcomplexity(leaves=10);
> # run;
> #
> #
> # proc hpsplit data=clust2 cvmodelfit seed=123;
> # class  cluster  NACIONASP ESTUDIOSSP ACTESTB ccaa NUTS1 TAMAMU ZONARES/ upcase;
> # model cluster = ESTUDIOSSP NACIONASP ACTESTB ccaa NUTS1 TAMAMU ZONARES;
> # prune costcomplexity;
> # run;
```

A.2 Metadata

Titol: Tipologia de despesa a les llars Espanyoles

Autor: Joan Ferrando Ravella
Tutor: Llorenç Badiella

Resums:
L'Enquesta de Pressupostos Familiars (EPF) és un estudi molt rellevant que es du a terme anualment i a escala nacional. Alguns dels seus objectius són trobar estimadors de la despesa agregada o estimar el consum en quantitats físiques de diferents productes. Per altra banda, donada la magnitud de l'estudi, la informació que aquest estudi pot aportar als diferents àmbits de coneixement és molt elevada, motiu pel qual ha sigut analitzat i explotat per sectors molt diversos que generalment han empleat metodologies típiques del propi camp de coneixement.
Són molts els estudis realitzats sobre l'EPF i resten poques alternatives d'anàlisi que aportin conclusions innovadores i que no caiguin en redundància. L'objectiu d'aquest treball és aportar informació nova a un estudi de molt renom i notablement treballat. Per fer-ho, s'utilitzaran metodologies poc convencionals que escapen a les clàssicament empleades; concretament, s'empraran eines d'anàlisi multivariant utilitzant anàlisi de clústers per tal d'assignar perfils de despesa a cada llar.

La Encuesta de Presupuestos Familiares (EPF) es un estudio muy relevante que se lleva a cabo anualmente y a escala nacional. Algunos de sus objetivos son encontrar estimadores del gasto agregado o estimar el consumo en cantidades físicas de diferentes productos. Por otra parte, dada la magnitud del estudio, la información que este estudio puede aportar a los diferentes ámbitos de conocimiento es muy elevada, por lo que ha sido analizado y explotado por sectores muy diversos que generalmente han empleado metodologías típicas del propio campo de conocimiento.
Son muchos los estudios realizados sobre la EPF y quedan pocas alternativas de análisis que aporten conclusiones innovadoras y que no caigan en redundancia. El objetivo de este trabajo es aportar información nueva a un estudio de mucho renombre y notablemente trabajado. Para ello, se utilizarán metodologías poco convencionales que escapan a las clásicamente empleadas; concretamente, se usarán herramientas de análisis multivariante utilizando análisis de clusters para asignar perfiles de gasto en cada hogar.

The Family Budget Survey (EPF) is a very important study that is carried out annually and nationwide. Some of its objectives are to find estimators of the aggregate expenditure or to estimate the consumption in physical quantities of different products. On the other hand, given the magnitude of the study, the information that this study can contribute to the different fields of knowledge is very high, which has been analyzed and exploited by very diverse sectors that have generally used typical methodologies of the field of knowledge. There are many studies on EPF and there are few alternatives to analysis that provide innovative conclusions and do not fall into redundancy. The objective of this study is to provide new information to a very renowned and remarkably worked study. To do this, we will use unconventional methodologies that escape the classically used ones; Specifically, multivariate analysis tools will be used making cluster analysis to assign cost profiles to each home.

Paraules clau:

- Pressupostos Familiars
- Anàlisi Multivariant
- Clúster
- Perfils