

Grau en Estadística

Títol: Anàlisi estadística de dades de producció de residus de les indústries de Catalunya

Autor: Helena Llorens Lluís

Director: Josep Anton Sànchez

Departament: Estadística i Investigació Operativa

Convocatòria: Gener 2024



AGRAÏMENTS

En primer lloc, vull agrair a en Josep Anton Sànchez pel seu suport i orientació al llarg d'aquest treball. També a l'Anna, la Verónica i la Noemí, de Símbiosy, per la seva confiança en el meu projecte i per haver-me proporcionat els coneixements essencials per a la implementació d'aquesta investigació.

RESUM

El creixement constant de la producció industrial té com a conseqüència l'augment de la quantitat de residus. Per poder mantenir un estil de vida semblant a l'actual, és imprescindible canviar el paradigma econòmic d'un model lineal a un de circular. Per aconseguir aquest objectiu, és essencial saber quins tipus de residus generen les empreses i com es poden convertir en recursos.

Aquest estudi té com a finalitat trobar una relació entre les activitats industrials i els tipus de residus que es generen. Aquesta caracterització es busca mitjançant l'anàlisi estadística multivariant, utilitzant tècniques com l'anàlisi de components principals i la clusterització jeràrquica. A més, es fa servir la modelització estadística amb un enfocament de dos models: un model lineal amb les tones de residus com a variable resposta, i un model lineal generalitzat amb la freqüència del residu com a variable resposta. Amb aquesta metodologia s'aconsegueix caracteritzar els tipus de residus generats per les activitats industrials.

Paraules clau: residus industrials, economia circular, anàlisi de components principals, *clustering* jeràrquic, model lineal, model lineal generalitzat.

ABSTRACT AND KEYWORDS

Title: Statistical analysis of industrial waste in Catalonia

Abstract: The constant growth of industrial production results in an increase in the quantity of waste. To maintain a lifestyle similar to the current one, it is essential to shift the economic paradigm from a linear model to a circular one. To achieve this goal, it is crucial to understand the types of waste generated by industries and how they can be transformed into resources.

This study aims to establish a relationship between industrial activities and the types of waste generated. This characterization is sought through multivariate statistical analysis, utilizing techniques such as principal component analysis and hierarchical clustering. Additionally, statistical modeling is employed with a dual-model approach: a linear model with waste tonnage as the response variable and a generalized linear model with waste frequency as the response variable. This methodology allows for the characterization of the types of waste generated by industrial activities.

Keywords: residual waste, circular economy, principal component analysis, hierarchical clustering, linear model, generalized linear model.

CLASSIFICACIÓ AMS

62H25 Factor analysis and principal components; correspondence analysis

62H30 Classification and discrimination; cluster analysis

62J12 Generalized linear models

62J05 Linear regression

ÍNDEX

1. INTRODUCCIÓ.....	4
1.1. Context	4
1.2. Justificació	6
1.3. Objectius.....	8
2. METODOLOGIA.....	9
2.1. Anàlisi de Components Principals (PCA).....	9
2.1.1. Càlcul de les components principals	9
2.1.2. Representació gràfica	10
2.2. Cluster jeràrquic	11
2.3. Modelització	13
2.3.1. Model lineal	13
2.3.2. Model lineal generalitzat.....	14
3. DESCRIPCIÓ DE LES DADES	15
3.1. Base de dades original	15
3.2. Base de dades resultant	16
4. RESULTATS.....	17
4.1. Descriptiva	17
4.1.1. Descriptiva univariant.....	17
Taula 4.2. Freqüències dels grups d'activitats industrials	19
4.1.2. Descriptiva bivariant.....	19
4.2. Anàlisi de Components Principals (PCA).....	21
4.2.1. PC1 vs PC2.....	23
4.2.2. PC3 vs PC4.....	25
4.2.3. PC5 vs PC6.....	27
4.3. Cluster jeràrquic	29
4.4. Modelització	34
4.4.1. Caracterització de les freqüències	34
4.4.2. Caracterització de les quantitats	39
5. CONCLUSIONS	43
6. BIBLIOGRAFIA	44
7. ANNEX.....	45

ÍNDIX DE FIGURES

Figura 1.1. Tones de residus industrials per any a Catalunya. Font: Agència de Residus de Catalunya..	4
Figura 1.2. Residus segons el destí. Font: Agència de Residus de Catalunya.....	5
Figura 1.3. Il·lustració de l'economia circular. Font: Símbiosy.....	6
Figura 1.4. Il·lustració de simbiosi industrial. Font: Símbiosy.....	7
Figura 3.1. Base de dades original.....	15
Figura 4.1. Freqüències dels grups de residus.....	17
Figura 4.2. Freqüències dels grups d'activitats industrials.....	18
Figura 4.3. <i>Heatmap</i> de residu per cada activitat industrial.....	19
Figura 4.4. <i>Heatmap</i> ampliat de residu per activitat industrial.....	20
Figura 4.5. <i>Scree plot</i> amb la variabilitat de cada component principal.....	21
Figura 4.6. Variables en els eixos de les components 1 i 2.....	23
Figura 4.7. <i>Biplot</i> de les components 1 i 2.....	24
Figura 4.8. Variables en els eixos de les components 3 i 4.....	25
Figura 4.9. <i>Biplot</i> de les components 3 i 4.....	26
Figura 4.10. Variables en els eixos de les components 5 i 6.....	27
Figura 4.11. <i>Biplot</i> de les components 5 i 6.....	28
Figura 4.12. Dendograma del <i>cluster</i> jeràrquic.....	29
Figura 4.13. <i>Clusters</i> sobre les components principals 1 i 2.....	30
Figura 4.14. Resultat del test de sobredispersió.....	34
Figura 4.15. Gràfics de validació del model.....	35
Figura 4.16. Coeficients del model per les activitats industrials.....	36
Figura 4.17. Coeficients del model pels residus.....	37
Figura 4.18. Gràfics de validació del model lineal.....	39
Figura 4.19. Coeficients del model lineal per les activitats industrials.....	40
Figura 4.20. Coeficients del model pels residus.....	41

ÍNDEX DE TAULES

Taula 4.1. Freqüències de residus.....	17
Taula 4.2. Freqüències dels grups d'activitats industrials.....	18
Taula 4.3. Taula de contingència de residus i activitats industrials.....	20
Taula 4.4. <i>Loadings</i> de les variables a les 6 primeres components principals.....	22
Taula 4.5. Residus del primer <i>cluster</i>	31
Taula 4.6. Residus del segon <i>cluster</i>	31
Taula 4.7. Residus del tercer <i>cluster</i>	31
Taula 4.8. Residus del quart <i>cluster</i>	32
Taula 4.9. Residus del cinquè <i>cluster</i>	32
Taula 4.10. Residus del sisè <i>cluster</i>	33
Taula 4.11. Residus del setè <i>cluster</i>	33

1. INTRODUCCIÓ

1.1. Context

La producció industrial experimenta un creixement constant, i cada any es generen més de 3.5 milions de tones de residus industrials a Catalunya. Una gestió eficient d'aquests residus és necessària per a minimitzar-ne l'impacte ambiental. En aquest context, la implementació de pràctiques sostenibles i estratègies en l'eficiència de gestió dels recursos és essencial per a preservar el seu esgotament al planeta.

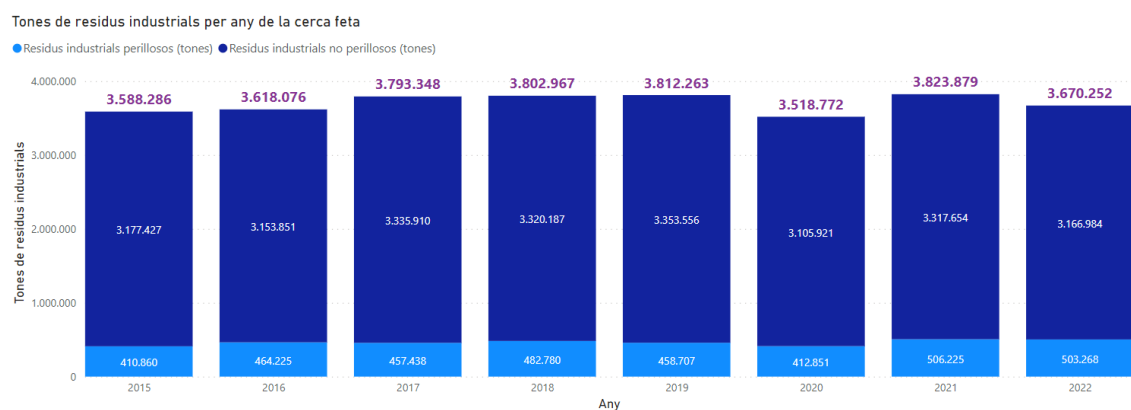


Figura 1.1. Tones de residus industrials per any a Catalunya. Font: Agència de Residus de Catalunya

Aproximadament, un 80% d'aquests residus van a processos de valorització, és a dir, s'aprofiten per fer nous productes. Aquesta estratègia busca aprofitar els materials que hi ha als residus per donar-ne una nova vida i poder reduir-ne la quantitat final, i així disminuir l'impacte que tenen els residus en el medi. Molts dels residus que van a valorització requereixen processos de selecció segons els materials per anar al mercat de productes secundaris, no tots poden aprofitar-se. Hi ha una part dels residus que no passen per cap procés de valorització i es destinen directament a dipòsit controlat (7%) o a incineració (1%).

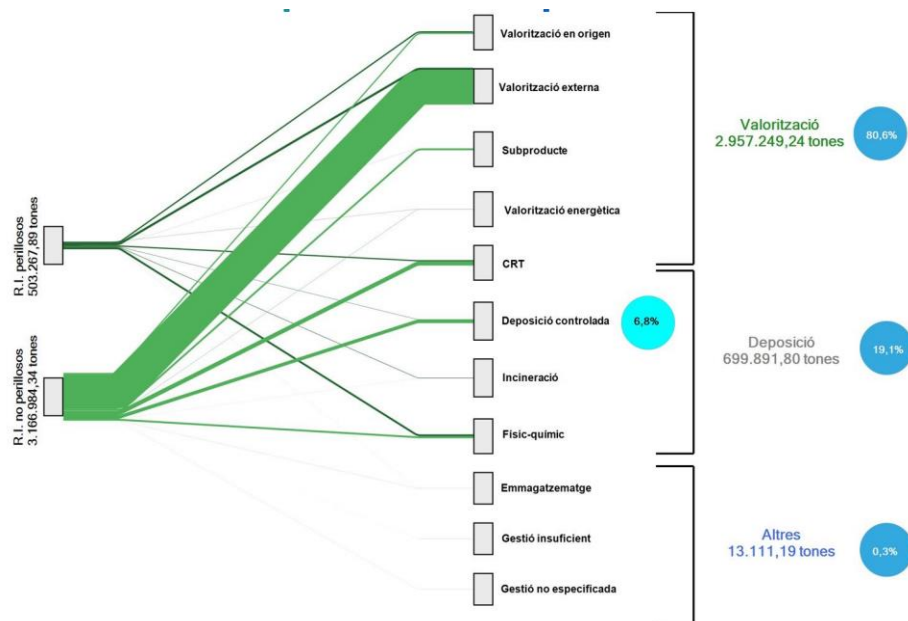


Figura 1.2. Residus segons el destí. Font: Agència de Residus de Catalunya

L'organisme que s'encarrega de recollir la informació sobre dades de residus és l'Agència de Residus de Catalunya, una entitat de dret públic que té competència sobre els residus que es generen a Catalunya i els que es gestionen al seu àmbit territorial. Les empreses registren les seves dades a la Declaració Anual de Residus Industrials (DARI), un document on es recullen les dades dels residus produïts per cada centre de producció en el període d'un any natural. Estan obligades a declarar al DARI les indústries productores de residus industrials (indústries extractives, indústries manufactureres, subministrament d'energia elèctrica, gas, vapor i aire condicionat, i subministrament d'aigua: activitats de sanejament, gestió de residus i descontaminació) i productores de residus perillosos.

Les funcions de l'Agència de Residus de Catalunya són promoure la minimització de residus i la seva perillositat, el foment de la recollida selectiva, la valorització dels residus, la disposició del rebuig i la recuperació d'espais i sòls degradats per descàrregues incontrolades de residus o per contaminants.

1.2. Justificació

És evident que, si aspirem a mantenir un estil de vida semblant a l'actual, la generació que fem de residus és insostenible. Per aquest motiu, s'està promocionant l'economia circular, un model econòmic de producció i consum en què els recursos i serveis estan dissenyats per mantenir sempre el seu valor mentre circulen per l'economia, eliminant d'aquesta manera el concepte de residu. Aquest enfocament implica també l'eliminació de substàncies tòxiques i la incorporació d'energies renovables, contribuint a una gestió més sostenible de tot el conjunt de recursos necessaris per a la producció industrial.

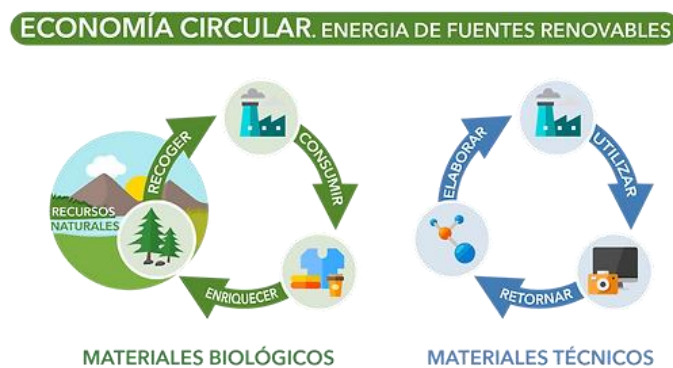


Figura 1.3. Il·lustració de l'economia circular. Font: Símbiosy

Símbiosy, una empresa de consultoria especialitzada en l'economia circular, ha estat la principal impulsora d'aquest estudi. Aquesta empresa es dedica a la implementació de projectes de simbiosi industrial, que són l'eina clau per aplicar els principis de l'economia circular al teixit productiu d'un territori. Per entendre el concepte de simbiosi industrial, podem imaginar el conjunt d'indústries com un ecosistema natural, un bosc per exemple, on no hi ha residus, l'energia flueix i, com més divers sigui l'ecosistema, millor es desenvolupa. En el context industrial, es tracta d'encoratjar les empreses a col·laborar per maximitzar l'ús dels recursos i fer els processos més eficients comptant amb les energies necessàries.

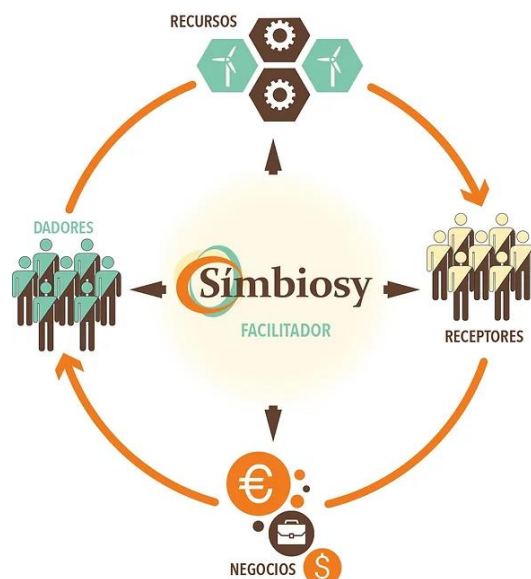


Figura 1.4. Il·lustració de simbiosi industrial. Font: Símbiosy

Símbiosy desenvolupa la seva funció en un ecosistema industrial explorant possibles simbiosis entre empreses. Aquesta responsabilitat implica la identificació d'oportunitats de col·laboració entre diverses entitats, amb l'objectiu de localitzar empreses amb recursos excedents que puguin ser donats i altres que puguin rebre aquests recursos de manera beneficiosa.

Per iniciar les millors tècniques en aprofitament de recursos cal conèixer el territori i identificar i quantificar els residus agroindustrials i urbans que potencialment es poden convertir en recursos a través de la simbiosi industrial, i identificar amb qui és la informació de base que tot projecte de simbiosi industrial hauria de disposar. Es tracta de *mapejar* les empreses que formen part de l'ecosistema i els fluxos de materials de les mateixes

Per trobar i quantificar aquests recursos, és essencial conèixer els residus que generen cadascuna de les empreses del territori a analitzar. En alguns casos, aquestes dades poden ser proporcionades directament per les empreses o extretes de declaracions oficials; no obstant això, a vegades no és possible. Per abordar aquesta problemàtica, una solució viable serà identificar els residus típics associats a cada tipus d'activitat. D'aquesta manera, mitjançant el coneixement del codi d'activitat d'una empresa, seria possible determinar quins tipus de residus produeix i, per tant, quins recursos podria utilitzar una altra empresa de manera eficient.

1.3. Objectius

L'objectiu principal d'aquest projecte és dur a terme una anàlisi per a poder determinar quins tipus de residus es generen per a cada sector d'activitat industrial.

Més en concret aquesta anàlisi pretén abordar temes com:

- Determinar quin tipus de residus es generen per a cada activitat industrial.
- Sabent l'activitat i la mida d'una empresa, trobar la quantitat i tipologia dels residus que genera.
- La possibilitat de trobar en quines activitats industrials es generen certs residus.
- Calcular les tones d'un cert residu es generen en una certa activitat industrial.

2. METODOLOGIA

2.1. Anàlisi de Components Principals (PCA)

L'anàlisi de components principals és una tècnica estadística multivariant que s'utilitza per a reduir la dimensionalitat d'una base de dades retenint, el màxim de variabilitat possible d'aquest conjunt de dades.

Aquesta metodologia implica una transformació lineal sobre les dades per trobar, de manera seqüencial, variables no correlacionades que seran les components principals. Aquestes noves variables constituïran un nou sistema de coordenades ortogonals que permetran que les dades es puguin representar en una dimensió inferior. Les components principals estan ordenades de manera que les primeres acumulen la major part de la variabilitat de les variables originals.

L'ús de l'anàlisi de components principals pot ser tan aplicat com a mètode de classificació no supervisat, sense informació a priori de les classes a què pertanyen les dades, com per interpretar les dades mitjançant les noves variables que obtingudes com a combinacions lineals de les variables originals.

2.1.1. Càlcul de les components principals

Abans de calcular les components principals ens hem d'assegurar que aquestes estiguin a la mateixa escala, ja que aquest mètode es basa en la variància entre variables, i variables amb unitats diferents podrien tenir un pes excessiu. Aquesta estandardització es fa de la següent manera:

$$Z = \frac{X - \mu}{\sigma}$$

Amb les dades estandarditzades, si considerem una sèrie de variables $(X_1, X_2, \dots, X_p)$ sobre un grup d'individus. Volem calcular un nou conjunt de variables $Z_1, Z_2, \dots, Z_p$, incorrelacionades entre si, com a combinació lineal de les variables originals, amb variàncies que vagin disminuint progressivament.

$$Z_j = \phi_{j1}X_1 + \phi_{j2}X_2 + \dots + \phi_{jp}X_p$$

Per garantir que les components tinguin una longitud unitària, aquesta combinació lineal ha d'estar normalitzada, i això implica que

$$\sum_{k=1}^p \phi_{jk}^2 = 1$$

La primera component es calcula buscant un ϕ_1 tal que Z_1 tingui la variància més gran possible i $\phi_1' \phi_1 = 1$. Un cop trobada la primera component, es calcula la segona component (Z_2) repetint el mateix procés, però afegint la condició que Z_1 i Z_2 han de ser perpendiculars. Aquest procés es compleix fins que totes les components són calculades.

Els pesos ϕ_j en les equacions de les components principals es coneixen com a *loadings*. Aquests indiquen la contribució de cada variable original a la construcció de cada component principal.

2.1.2. Representació gràfica

Per interpretar i visualitzar de manera efectiva les components principals, es fan servir dues eines: el gràfic de variables i el *biplot*. El primer gràfic es construeix a partir dels *loadings* calculats per a la formació de les components principals, i es representa en dues dimensions, corresponents a dues components. Es crea una coordenada utilitzant els *loadings* de cada variable en les dues components que es volen visualitzar. Aquesta representació gràfica proporciona una perspectiva clara sobre la contribució de cada variable a les components principals i en facilita la interpretació.

Pel que fa al *biplot*, aquest segueix una lògica similar, incorporant la representació de les variables. Aquesta inclusió es realitza mitjançant els *scores* de les variables, els quals es calculen multiplicant el vector d'observació de cada variable per la matriu de vectors propis (*loadings*). Així, cada individu té una coordenada que es construeix a partir del seu *score* a les dues components representades. Aquesta representació *biplot* proporciona una visió conjunta de les relacions entre les variables i les observacions.

2.2. Cluster jeràrquic

Quan ens enfrontem a un volum considerable de dades, una estratègia valuosa és la clusterització. Aquesta tècnica se centra en la categorització d'individus en diversos grups (*clusters*) de manera que els membres d'un mateix grup comparteixin característiques similars. Per aquest projecte, s'optarà per una metodologia jeràrquica i aglomerativa.

L'agrupació jeràrquica es basa en la idea que els individus estan més relacionats amb aquells pròxims a ells que amb els més distants. En la clusterització aglomerativa, inicialment cada objecte es considera un *cluster* individual. En cada pas, els *clusters* més semblants es fusionen per formar un nou *cluster*. Aquest procés es repeteix iterativament fins a arribar a un sol *cluster* que conté tots els individus.

Concretament, el mètode utilitzat per aquest projecte és el mètode de Ward. En aquest procediment, en cada etapa s'ajunten els dos *clusters* pels quals hi hagi un menor increment en el valor total en la suma de quadrats de les diferències, dins de cada *cluster*, de cada individu al centroide del *cluster*.

Abans d'explicar més detalladament el procés que se segueix, es volen donar unes notacions:

- x_{ij}^k és el valor de la j -èssima variable sobre l' i -èssim individu del k -èssim *cluster*
- n_k és el nombre d'individus al *cluster* k
- m^k és el centroide del *cluster* k , amb components m_j^k
- E_k és la distància euclídea al quadrat entre cada individu del *cluster* k al seu centroide.

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^n (x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k} \sum_{j=1}^n (x_{ij}^k)^2 - n_k \sum_{j=1}^n (m_j^k)^2$$

- E és la suma de les distàncies euclides per tots els clusters.

El procés comença amb m *clusters*, cadascun dels quals està format per un sol individu. Això implica que cada individu coincideix amb el centre del *cluster* i , per tant, en aquest primer pas $E_k = 0$ per cada *cluster* i , de la mateixa manera, $E = 0$. L'objectiu del mètode de Ward és trobar, en cada etapa, aquells dos *clusters* la unió dels quals suposi un increment mínim en la suma total d'errors (E).

Per visualitzar les agrupacions generades amb aquest algoritme, una eina amplament utilitzada és el dendrograma. Aquest diagrama d'arbre organitza les dades en subcategories que es desglossen successivament. Cada branca representa un individu, i l'altura de les branques indica la similitud entre dos individus: a major altura, menor similitud. A l'hora de

determinar el nombre de grups amb els quals es treballarà, es realitza un tall horitzontal al dendrograma. Aquesta decisió és subjectiva i pot basar-se en diversos mètodes.

2.3. Modelització

2.3.1. Model lineal

L'anàlisi de regressió és una tècnica estadística que investiga i modelitza la relació entre variables. En el cas de la regressió lineal múltiple, es considera un model lineal entre una variable aleatòria resposta Y i un grup de k variables no aleatòries $x_1, \dots, x_k$ explicatives, amb un terme aleatori ϵ . El model lineal es defineix com

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \dots, n$$

on β_0 és l'ordenada d'origen, el valor de y quan tots els predictors són zero; β_j és l'efecte que té l'increment d'una unitat de la variable predictora x_{ij} sobre la variable dependent y ; ϵ_i és l'error, la diferència entre el valor observat i l'estimat pel model.

En un model de regressió lineal, s'assumeix que els termes d'error ϵ_i segueixen una distribució normal amb una mitjana de zero i una variància constant ($\epsilon_i \sim N(0, \sigma^2)$). A més, es fa l'assumpció que la variable resposta Y també segueix una distribució normal.

És important validar el model per assegurar-ne l'adequació. Les assumpcions d'un model lineal són:

- Linealitat: la relació entre les variables predictores i la variable dependent ha de ser aproximadament normal.
- Homoscedasticitat: la variància dels errors ha de ser constant.
- Independència dels errors: l'error associat a una observació no pot estar influït pels errors d'altres observacions.
- Normalitat dels residus: els residus han de seguir una distribució normal.
- Absència de multicol·linealitat perfecta: les variables explicatives no han de tenir una relació de col·linealitat perfecta, ja que això dificultaria la identificació dels efectes individuals de cada variable.

2.3.2. Model lineal generalitzat

Els models lineals generalitzats (MLG) són una generalització de la regressió lineal ordinària, permetent que la variable resposta segueixi una distribució diferent de la normal. Aquests models tenen tres components:

- Família exponencial: la variable dependent Y pertany a la família exponencial, que inclou distribucions com la Normal, la Binomial, la Poisson i la Gamma. Això implica que la funció de densitat té la forma $\log(f(y; \theta, \phi)) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)$, on θ és el paràmetre canònic i ϕ és el paràmetre de dispersió.
- Predictor lineal (η): aquest vector representa una combinació lineal dels paràmetres desconeguts. El predictor lineal es pot expressar com $\eta = X\beta$.
- Funció d'enllaç: proporciona la relació entre el predictor lineal (η) i la mitjana de la funció de distribució (μ). La funció d'enllaç es pot notar com $\mu = g^{-1}(\eta)$.

Els MLG poden ser utilitzats per modelitzar una gran varietat de dades, incloses les contínues, binàries o de recomptes. En el context específic de recomptes, es fa servir la distribució de Poisson. El *link* canònic per aquesta distribució és $g(\mu) = \log(\mu)$, indicant que hi ha una relació log-linear entre la mitjana i el predictor lineal. La variància d'un model Poisson és idèntica a la mitjana, per tant, la dispersió es fixa a $\phi = 1$ i la funció de variància és $V(\mu) = \mu$.

No obstant això, pot passar que no es compleixi la premissa de $\phi = 1$, és a dir, es pot donar que hi hagi sobredispersió. Per comprovar-ho, es fa un test en què es fa un test d'hipòtesis on la hipòtesi nul·la assumeix equidispersió i la hipòtesi alternativa afirma que la variància té la forma $\sigma^2 = \mu + \alpha \cdot g(\mu)$. Aquest test d'hipòtesi es pot fer amb $g(\mu) = \mu$ o $g(\mu) = \mu^2$.

Si es fa el test amb $g(\mu) = \mu$, es comprova si la relació entre la variància i la mitjana és lineal. Si es té prou evidència per rebutjar la hipòtesi nul·la, es veu que $V[Y|X] = (1 + \alpha)E[X|Y]$. Aleshores, es pot fer servir el model de Quasi-Poisson, en què els coeficients s'estimen fent servir la distribució de Poisson, però els errors estàndards s'ajusten per tenir $\phi \neq 1$.

D'altra banda, si es fa el test amb $g(\mu) = \mu^2$, es comprova si la relació entre la variància i la mitjana és quadràtica. En aquest cas, ens trobaríem amb $V[Y|X] = E[Y|X] + \alpha E[Y|X]^2$. Aleshores, es podria ajustar una distribució Binomial negativa, on la variància depèn quadràticament de la mitjana.

3. DESCRIPCIÓ DE LES DADES

3.1. Base de dades original

La base de dades principal utilitzada per a aquest projecte és el DARI de l'any 2019. És important tenir en compte que, a causa de la COVID-19, es va registrar un nombre inferior de declaracions en comparació amb l'any anterior. En aquest context, l'Observatori de Residus va considerar oportú, a manera d'estimació de les declaracions que manquen, incorporar les dades corresponents a les declaracions de l'any 2018 dels centres productors que no van presentar declaració per a l'any 2019 (un total de 1856 empreses).

Per comprendre les variables amb les quals es treballarà, cal saber que tots els productors de residus tenen assignada una activitat industrial vinculada al CNAE (Classificació Nacional d'Activitat Econòmica). El CNAE és un codi de 4 xifres, de les quals cadascuna d'elles dona una informació addicional per detallar l'activitat industrial. D'altra banda, el CER (Classificació Europea de Residus) és un codi de 6 xifres que classifica els residus en funció del sector i subsector d'origen.

El DARI conté 23 variables, que serveixen per identificar tant l'empresa que està declarant com el residu declarat i la seva quantitat. Per a aquest projecte només es faran servir les variables que indiquen l'activitat industrial de l'empresa (CNAE), el tipus de residu declarat (CER) i la quantitat en tones que ha declarat l'empresa del residu corresponent. És important destacar que una empresa pot haver declarat el mateix residu en diverses ocasions; en aquests casos, es considerarà com una única declaració i les quantitats se sumaran.

any_decla	codi_prod	nom_prod	nif	adre_polig	adre_abrv	adre_carre	adre_num	adre_cp	adre_nom	nom_coma	cnae
2019	...	EXPLOTAD			CR	PLANA DEL	S/N	43815	AIGUAMÚR	Alt Camp	1089
2019	...	EXPLOTAD			CR	PLANA DEL	S/N	43815	AIGUAMÚR	Alt Camp	1089
2019	...	EXPLOTAD			CR	PLANA DEL	S/N	43815	AIGUAMÚR	Alt Camp	1089
2019	...	EXPLOTAD			CR	PLANA DEL	S/N	43815	AIGUAMÚR	Alt Camp	1089
2019	...	CELLERS VI			VP	MAS DE LA		43714	AIGUAMÚR	Alt Camp	1102
2019	...	CELLERS VI			VP	MAS DE LA		43714	AIGUAMÚR	Alt Camp	1102

descripcio	codi_residu	descripcio	descripcio	classe_resi	quantitat	tipus_gesti	forma_trac	descripcio	grup	grup_valor
Elaboració	150102	Envasos de	ENVASOS I	NE	0.00	GESTIÓ EXT	V12	Reciclatge	VALOR	RECICLATGE
Elaboració	150110	Envasos qu	BIDONS PL	ES	2.04	SUBPRODU	V51	Recuperaci	VALOR	RECICLATGE
Elaboració	191201	Paper i cart	RESIDU DE	NE	30.67	GESTIÓ EXT	V11	Reciclatge	VALOR	RECICLATGE
Elaboració	200301	Mescles de	Principalme	NE	9.25	GESTIÓ EXT	V99	Altres	VALOR	RECICLATGE
Elaboració	020701	Residus del	Brisa	NE	0.20	GESTIÓ EN	V81	Utilització	VALOR	AGRICOLA
Elaboració	020701	Residus del	Mares	NE	0.11	GESTIÓ EN	V81	Utilització	VALOR	AGRICOLA

Figura 3.1. Base de dades original

Hi ha un total de 629 codis CNAE i 843 codis CER i, per tant, per facilitar la presentació dels resultats, s'ha optat per agrupar-los. En el cas dels codis CNAE, s'ha decidit treballar amb els dos primers dígits del codi, creant així 88 grups. Pel CER s'han establert unes agrupacions en

funció del material declarat, en comptes de la seva procedència, generant un total de 26 grups. Aquestes agrupacions han estat proposades per l'empresa Símbiosy i estan fetes a partir del material declarat, sense tenir-ne en compte l'origen. No obstant això, no totes les activitats s'han declarat al DARI i, per tant, ens quedem amb 75 grups de CNAE.

Cal destacar que el DARI és una base de dades creada a partir de declaracions d'empreses i, per tant, poden existir errors. S'ha detectat que, en algunes ocasions, els CNAE declarats tenien menys de 4 dígit. Tot i això, com que per a aquest projecte es treballa només amb els dos primers dígit del codi, això no ha suposat un gran problema. Com que el nombre d'ocurrències d'aquests errors no era especialment elevat, s'han pogut examinar individualment els CNAE erronis amb les seves descripcions, per assegurar la coherència.

3.2. Base de dades resultant

Per obtenir la base de dades final, s'han utilitzat tres fonts de dades: el DARI, una taula que inclou tots els codis CNAE amb les seves descripcions, i una altra taula que conté tots els CER amb les seves descripcions i l'agrupació a la qual pertanyen.

La base de dades resultant és una taula en què per cada combinació de CNAE i CER, s'indica la freqüència i la quantitat totals. Amb l'objectiu de facilitar el treball amb el PCA, es procedirà a transformar aquesta taula en una taula de contingència, on les columnes representen les agrupacions dels residus, les files els codis CNAE, i els valors continguts en la taula indicaran la freqüència de cada combinació. Aquesta reorganització millorarà la preparació del conjunt de dades per a l'anàlisi amb el PCA.

4. RESULTATS

4.1. Descriptiva

4.1.1. Descriptiva univariant

S'ha fet una anàlisi descriptiva de les dades amb l'objectiu d'obtenir una visió general de la distribució tant de les activitats industrials com dels residus a la base de dades del DARI.

Primer, analitzem la distribució de freqüències dels residus. S'ha declarat residus de 26 grups diferents. Entre els residus més recurrents, destaquen els pertanyents als grups de mesclats, paper i cartró, així com els metàl·lics. D'altra banda, observem una baixa incidència en la declaració dels residus com explosius, cendres i rentat de gasos, i pneumàtics.

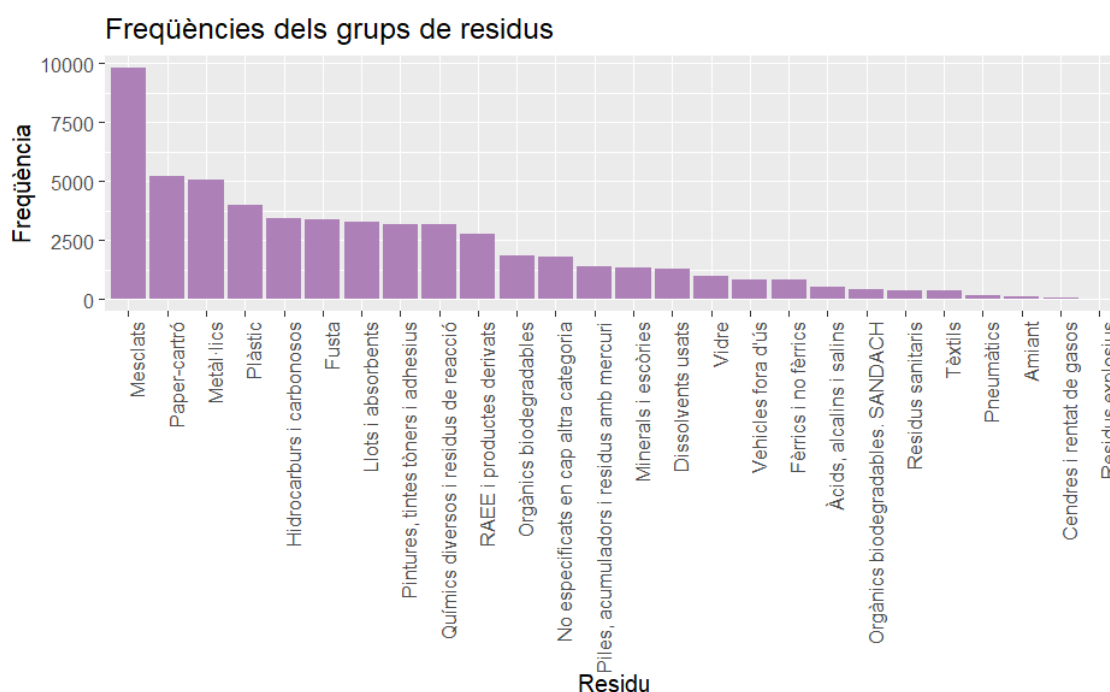


Figura 4.1. Freqüències dels grups de residus

També es pot veure aquest resultat numèricament a partir d'una taula amb les freqüències absolutes i relatives dels cinc residus més freqüents.

Residu	Freqüència absoluta	Freqüència relativa
Mesclats	9821	0.176
Paper-cartró	5241	0.094
Metàl·lics	5081	0.091

Plàstic	3996	0.072
Hidrocarburs i carbonosos	3445	0.062

Taula 4.1. Freqüències de residus

D'altra banda, el següent gràfic mostra la distribució de les freqüències de les activitats industrials declarades. Al DARI s'hi ha presentat declaracions de 75 activitats industrials diferents. Hi ha molts CNAE que s'han declarat en molt poques ocasions, destacant en aquest cas els codis 92 (Activitats de biblioteques, arxius, museus i altres activitats culturals), 63 (Serveis d'informació), 51 (Transport aeri) i 06 (Extracció de petroli brut i de gas natural), declarats només una vegada al DARI del 2019. En contrast, les activitats més declarades són la 25 (Fabricació de productes metàl·lics, excepte maquinària i equips), 10 (Indústries de productes alimentaris) i 20 (Indústries químiques).

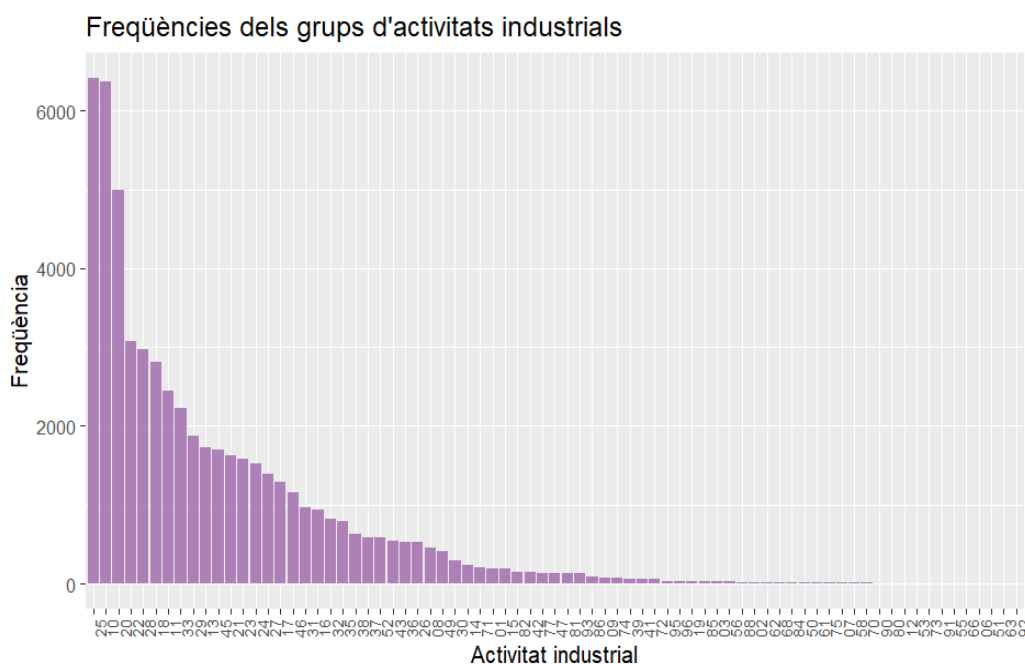


Figura 4.2. Freqüències dels grups d'activitats industrials

El resultat es pot veure numèricament amb una taula de freqüències de les cinc activitats industrials més importants.

Activitat industrial	Freqüència absoluta	Freqüència relativa
25	6415	0.115
10	6371	0.114
20	5002	0.090

22	3076	0.055
28	2975	0.053

Taula 4.2. Freqüències dels grups d'activitats industrials

4.1.2. Descriptiva bivariant

També és interessant veure les relacions entre les activitats industrials i els residus, així com les seves freqüències, a través d'un *heatmap*. Així, es veu de manera clara que els residus de tipus mesclats predominen en gran part de les activitats industrials.

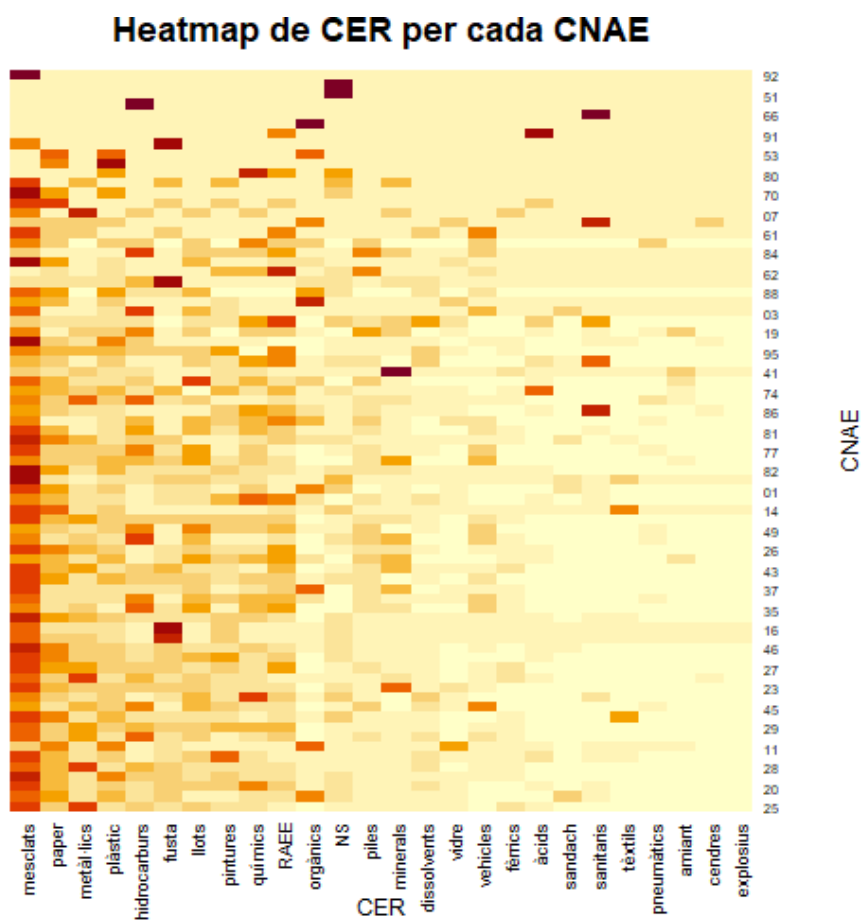


Figura 4.3. *Heatmap* de residu per cada activitat industrial

Davant la gran quantitat de variables, es recomana ampliar aquest *heatmap* per incloure tant les activitats com els residus més significatius. Així, es pot identificar de manera més detallada els residus importants associats a les activitats que es declaren amb més freqüència.

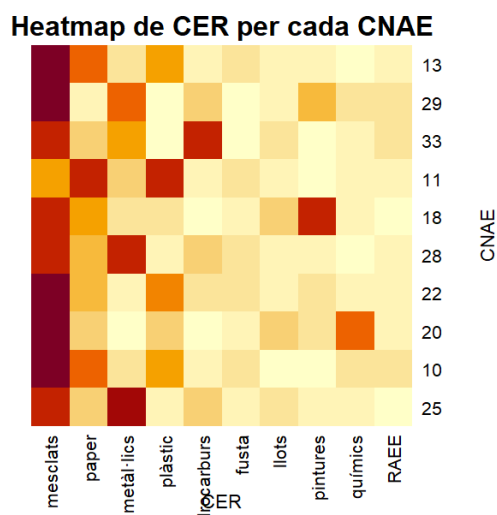


Figura 4.4. *Heatmap* ampliat de residu per activitat industrial

D'altra banda, també es pot explorar aquest resultat a partir d'una taula de contingència de les combinacions d'activitat industrial amb residus.

	mesclats	paper	metàl·lics	plàstic	hidrocarburs	fusta	llots	pintures	químics	RAEE
25	1285	482	1386	295	529	297	444	279	250	167
10	1055	760	334	563	227	366	174	189	307	329
20	911	401	236	385	181	300	386	331	659	305
22	693	327	179	437	222	219	166	194	157	147
28	546	284	555	134	260	177	149	167	108	142
18	619	340	171	173	53	90	230	581	88	79
11	218	396	123	396	34	82	39	31	52	60
33	327	163	234	90	328	90	144	73	107	128
29	253	128	197	110	155	112	122	158	138	142
13	385	265	92	187	68	94	63	80	46	61

Taula 4.3. Taula de contingència de residus i activitats industrials

4.2. Anàlisi de Components Principals (PCA)

Com s'ha explicat a la metodologia, l'anàlisi de components principals és una tècnica estadística que redueix la dimensionalitat d'una base de dades retenint-ne el màxim de variabilitat possible. És a dir, permet obtenir una combinació lineal de variables ortogonals entre elles, agrupant les variables originals que estan relacionades entre si.

És important destacar que no existeix un mètode per a trobar la quantitat òptima de components principals a utilitzar. En aquest treball, s'ha pres en consideració que estigui representat un 85% de la variabilitat, que és la quantitat de variabilitat que cada component explica.

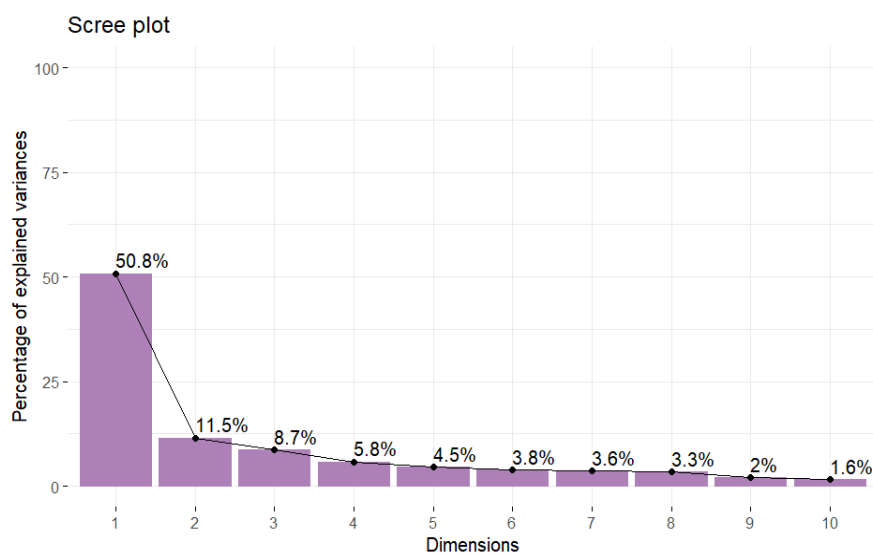


Figura 4.5. Scree plot amb la variabilitat de cada component principal

Segons la “rega del colze”, per aquesta anàlisi s'hauria de treballar amb 8 components principals, ja que és on es nota una decaiguda en el percentatge de variabilitat explicada. Malgrat això, per aquest estudi es faran servir només les 6 primeres components principals, ja que són suficients per explicar un 85% de la variabilitat de les dades. A continuació, es poden veure els valors propis associats a aquestes components.

[1] 3.63416134 1.72967933 1.50251159 1.22598115 1.08072749 0.99731068

Amb el nombre de components que s'ha d'utilitzar establert, es pot iniciar l'anàlisi.

A continuació, es mostren els *loadings* de les variables originals a les primeres 6 components principals.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
fusta	0.83500034	0.08384335	-0.134042261	0.040642829	0.0526506063	0.08941866
fèrrics	0.65742364	-0.56192749	-0.180363173	0.359733744	-0.0371431417	0.06021248
hidrocarburs	0.85407219	-0.38577850	0.165122737	0.193169488	0.0175753271	0.09539307
llots	0.93093942	-0.27545867	-0.056366159	-0.171545359	-0.0248246278	-0.03486021
mesclats	0.96567116	-0.06493568	-0.171289136	0.025085632	0.1142797647	0.01434865
metàl·lics	0.77849924	-0.45304900	-0.205714193	0.310727023	-0.0251070718	0.03146256
minerals	0.63834651	-0.09070090	-0.056457713	0.105544439	-0.3528545520	0.22382070
NS	0.94544288	0.19577921	-0.076489047	0.039077150	0.1337922257	0.12371952
orgànics	0.51750415	0.71033265	0.053458465	0.395353756	0.0079186141	-0.18343310
sandach	0.47478007	0.65468167	0.081360629	0.252955288	0.0420919438	0.07355307
paper	0.92506877	0.22185627	-0.120246088	0.127714422	0.1970609221	-0.10274362
piles	0.88094078	0.01109078	0.428698250	-0.072397239	-0.0610999203	0.07964745
pintures	0.77403397	-0.16596732	-0.193057976	-0.371100060	0.2291405584	-0.31947403
plàstic	0.87994754	0.31347962	-0.058163244	0.092800477	0.1333943171	-0.10601734
químics	0.86164914	0.12491706	0.001951367	-0.347685804	-0.1780383804	0.10513079
RAEE	0.93584477	0.17891080	0.091844308	-0.133758160	-0.0353326652	0.09035917
sanitaris	0.59914170	0.56962587	0.048432811	-0.225947889	-0.1722370777	0.11240163
tèxtils	0.08928586	0.06809904	-0.131965964	0.012342471	0.7629542178	0.47833338
vidre	0.44855435	0.30999445	-0.027278554	0.426693527	-0.0997011230	-0.41578913
dissolvents	0.88121436	-0.17318458	0.141614495	-0.294693996	0.0005539277	-0.07856142
vehícles	0.37827728	-0.23650795	0.852377746	0.059015672	0.0942407874	-0.07409602
amiant	0.59083884	0.23782814	0.057121873	-0.330565267	-0.3511357231	0.25713785
pneumàtics	0.22737239	-0.25646094	0.826583118	0.076921782	0.0891268148	-0.13679338
àcids	0.61214635	-0.31396649	-0.349558827	-0.339722752	0.1369046862	-0.41155422
cendres	0.50556517	-0.52120813	-0.272399360	0.337962094	-0.2511673259	0.17144353
explosius	0.07796554	-0.16915685	0.434843605	0.003620207	0.1244888905	0.08963384

Taula 4.4. *Loadings* de les variables a les 6 primeres components principals

Aquests *loadings* indiquen la contribució que té cada variable en cadascuna de les components. Per a interpretar el significat d'aquestes components, ho farem a partir d'aquesta matriu de correlacions i també mitjançant el gràfic que mostra les projeccions de les variables en els eixos de les noves dimensions.

4.2.1. PC1 vs PC2

En primer lloc, procedim a visualitzar les variables en relació amb els eixos de la primera i segona component principal (PC1 i PC2), les quals conjuntament acumulen un 62.3% de variabilitat.

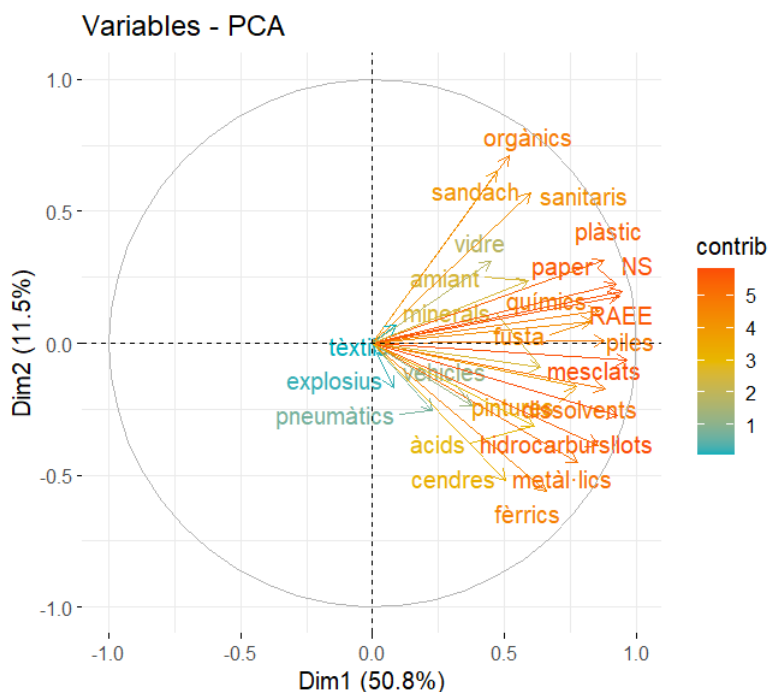


Figura 4.6. Variables en els eixos de les components 1 i 2

D'aquesta manera, es pot observar que la primera component presenta *loadings* positius per a totes les variables. Destaquen especialment les contribucions significatives dels residus de tipus mesclats, els no especificats en cap altra categoria, els d'aparells elèctrics i electrònics (RAEE), els llots i absorbents, així com el paper i cartró. Aquests són residus que es declaren en moltes ocasions, per tant, es podria interpretar aquesta component com a una component de mida, correlacionada positivament amb els grans productors de residus.

D'altra banda, els *loadings* més elevats de la segona dimensió corresponen als residus orgànics biodegradables i als SANDACH¹. Les variables amb pes negatiu que adquireixen una importància rellevant són els fèrrics i no fèrrics, les cendres i els metàl·lics. Aquesta dimensió sembla estar majoritàriament influenciada per la distinció entre els residus de tipus orgànic (amb pes positiu) i aquells de caràcter tècnic (amb pes negatiu).

Per aprofundir en aquesta anàlisi s'ha creat un *biplot*, que representa tant les variables com els individus en les dues primeres dimensions.

¹ Subproductes animals no destinats al consum humà

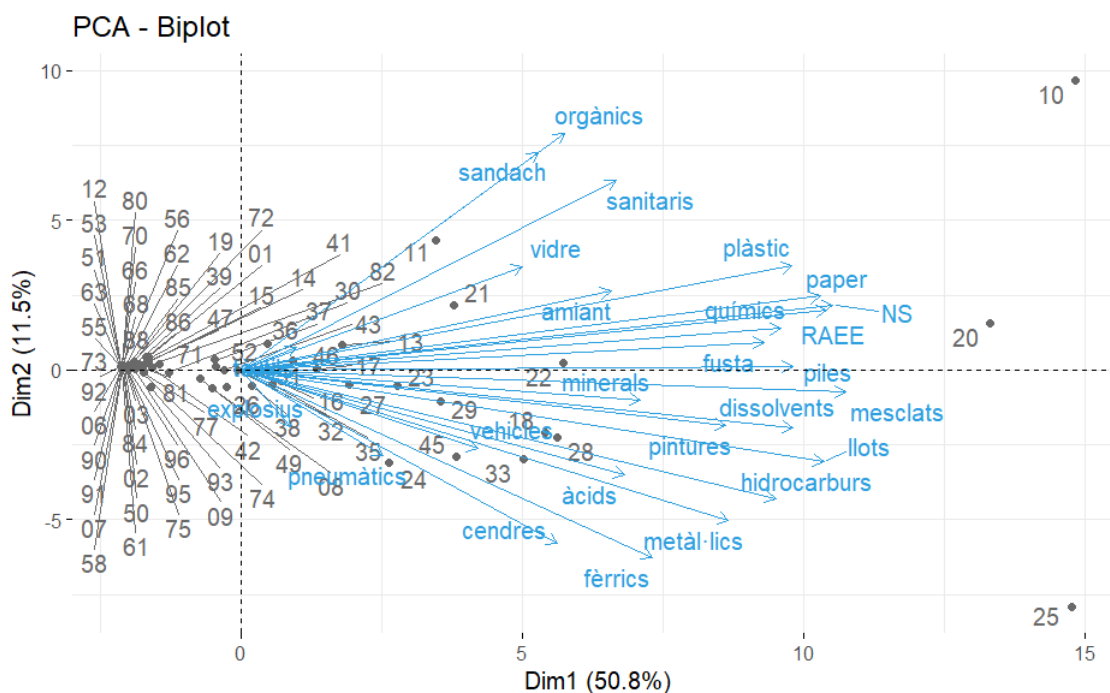


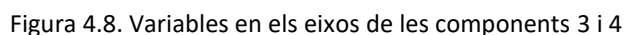
Figura 4.7. Biplot de les components 1 i 2

Amb aquest gràfic, podem veure que els codis d'activitat 10 (Indústries de productes alimentaris), 20 (Indústries químiques) i 25 (Fabricació de productes metàl·lics, excepte maquinària i equips) destaquen com a principals generadors de residus, ja que són els que tenen *scores* més grans per la primera component.

La indústria de productes alimentaris (CNAE 10) mostra una correlació positiva amb la segona component principal. Això suggereix que aquest sector tendeix a generar residus de tipus orgànic. En contrast, la fabricació de productes metàl·lics (CNAE 25) està negativament correlacionada amb la segona dimensió, indicant que aquesta indústria produeix principalment residus de naturalesa tècnica.

Pel que fa a la resta d'individus, la seva interpretació és més complexa, ja que no mostren valors destacats en aquestes dues components. Hi ha un grup de codis d'activitat amb valors negatius a la primera dimensió i propers a zero a la segona dimensió. Aquests podrien correspondre a aquells que no s'han declarat amb freqüència o bé que generen pocs residus.

A continuació, procedim a la representació gràfica de les variables en funció de la tercera i quarta dimensió. La proporció de variància acumulada per aquestes dues components és del 14.5%.



Pel que fa a la quarta component, les variables amb els pesos positius més destacats inclouen el vidre, els residus orgànics biodegradables i els fèrrics i no fèrrics. D'altra banda, les variables amb els pesos negatius més importants són les pintures, tintes, tòners i adhesius, els productes químics, l'amiant i els àcids, alcalins i salins. A partir d'aquesta informació, sembla que aquesta component podria estar negativament relacionada amb els residus perillosos i positivament amb els no perillosos.

25

4.2.3. PC5 vs PC6

Finalment, procedim a la visualització de les últimes components agafades, la cinquena i la sisena. Aquestes acumulen un 8.3% de la variabilitat total.

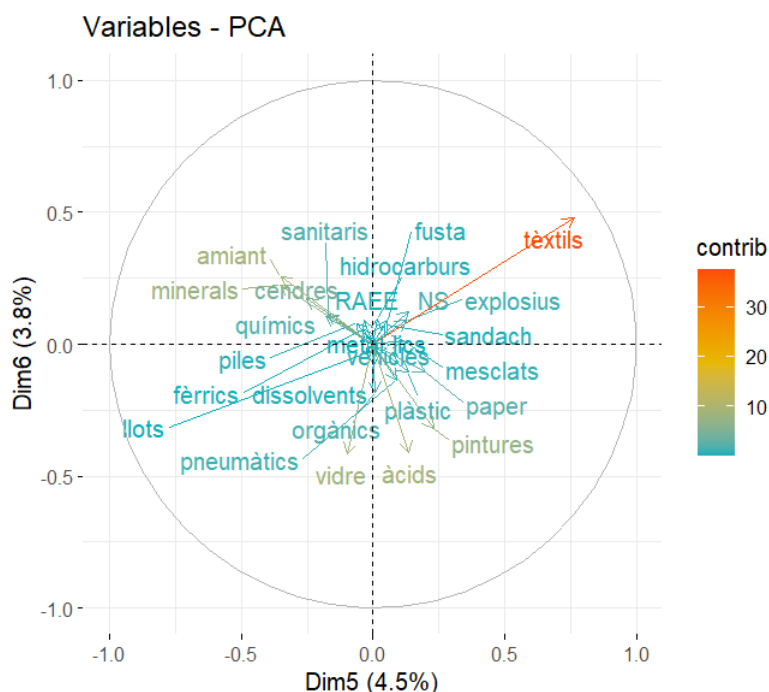


Figura 4.10. Variables en els eixos de les components 5 i 6

A partir d'aquest gràfic es veu evident que la cinquena component té una correlació positiva destacada amb els residus tèxtils.

Quant a la sisena component, la seva interpretació no és tan clara. Es pot destacar que té una correlació positiva amb els tèxtils, l'amiant i els minerals i escòries, i una correlació negativa amb el vidre, els àcids, alcalins i salins i les pintures, tintes, tòners i adhesius.

A continuació, es presenta la representació gràfica mitjançant el *biplot*:

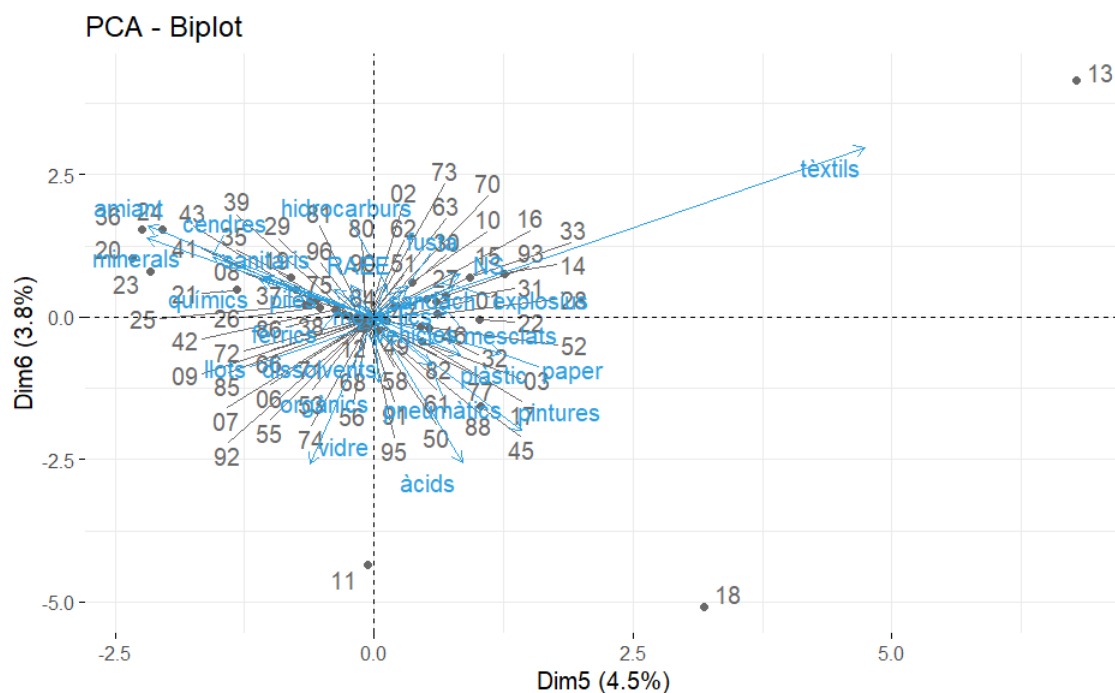


Figura 4.11. *Biplot* de les components 5 i 6

Es pot observar que les indústries tèxtils (CNAE 13) destaquen tant en la cinquena com sisena dimensió amb un *score* positiu. Aquesta indicació suggereix que els seus residus tindran predominantment una naturalesa tèxtil.

D'altra banda, les activitats 18 (Arts gràfiques i reproducció de suports enregistrats) i 11 (Fabricació de begudes) destaquen en la contribució a la sisena component amb correlació negativa. Això implica que entre els seus residus hi haurà materials com vidre, àcids, alcalins i salins i pintures, tintes, tòners i adhesius.

4.3. Cluster jeràrquic

Després d'haver dut a terme l'anàlisi de components principals, l'objectiu és identificar perfils potencials per a les activitats industrials. D'aquesta manera, podrem formar agrupacions de CNAE que presentin característiques similars, concretament en relació amb els residus.

Per aconseguir aquesta segmentació, utilitzarem un sistema de *clustering* jeràrquic, un mètode que proporciona una estructura organitzada i jerarquizada dels individus. A través d'aquesta tècnica, es realitzarà una identificació progressiva de subgrups, començant amb agrupacions més petites (o individuals) i evolucionant cap a agrupacions més àmplies, amb l'objectiu de reunir individus que comparteixin semblances rellevants en les característiques de producció de residus de les activitats industrials. Aquest *clustering* es farà amb el mètode de Ward i mètrica euclídea.

El dendrograma resultant d'aquesta clusterització serà una representació visual de les relacions jeràrquiques entre els diferents subgrups identificats. Aquesta eina gràfica proporcionarà una visió clara i intuïtiva de com les diferents activitats industrials es relacionen entre si en termes de producció de residus.

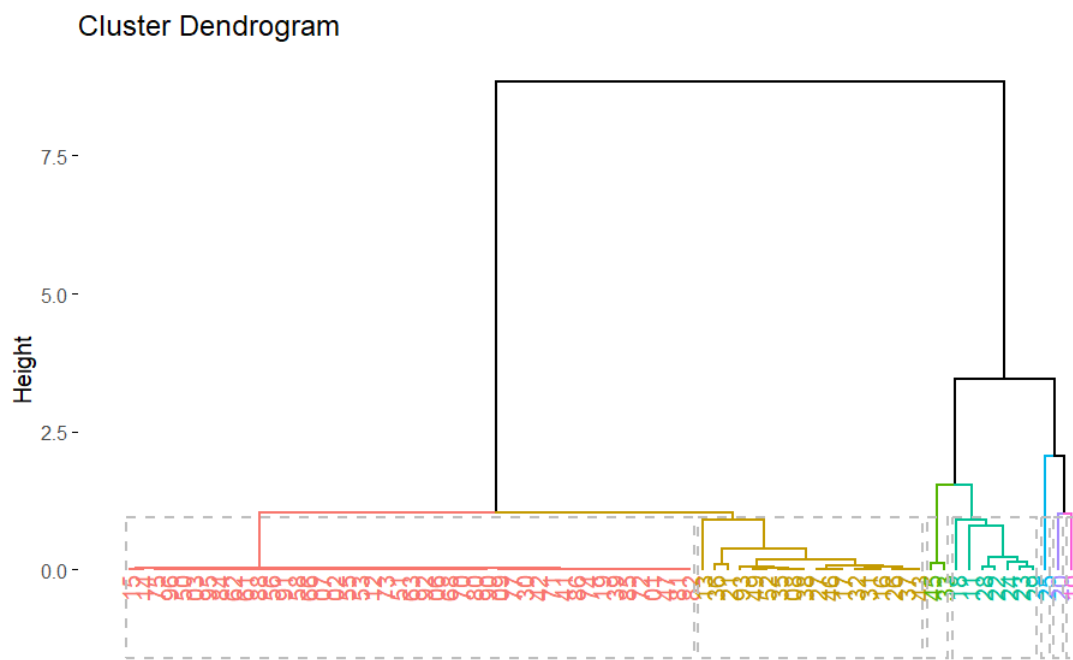


Figura 4.12. Dendrograma del *cluster* jeràrquic

Després de revisar el dendrograma, s'ha seleccionat el nombre de *clusters* com a 7. Aquesta elecció permet identificar característiques conjunes en cada grup sense generar massa grups on només hi hagi un sol individu, proporcionant una segmentació significativa i útil.

El següent gràfic mostra la distribució dels individus en els eixos de la primera i segona component principal, dividits pels *clusters* creats. Aquests eixos són rellevants, ja que acumulen gran part de la variabilitat, i són, per tant, els més influents en la creació dels *clusters*.

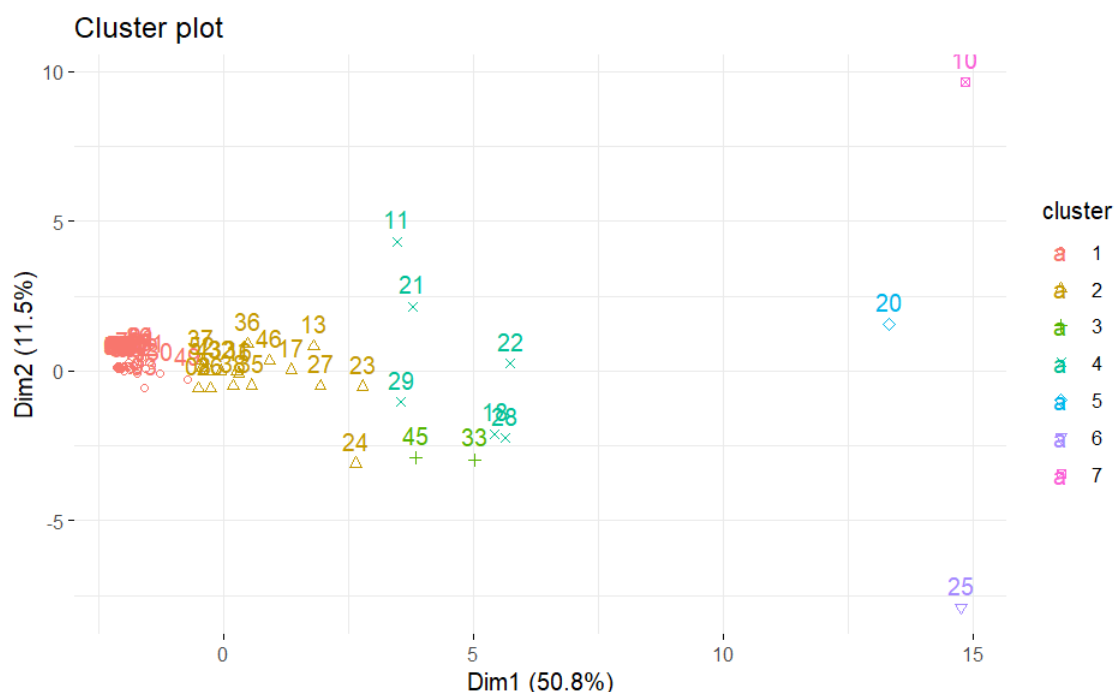


Figura 4.13. *Clusters* sobre les components principals 1 i 2

Per tal d'analitzar els *clusters* creats, s'ha d'explorar la descripció de cada grup en funció de les variables. Així, podem tenir una comprensió més completa de les característiques que defineixen cada *cluster* i es facilita la interpretació dels resultats.

Pel que fa al primer grup, és el més ampli i es caracteritza per la baixa producció de residus. Les mitjanes de les variables en aquest grup són significativament inferiors a les mitjanes totals.

A continuació es veu una taula amb els residus destacats pel primer grup amb el seu corresponent estadístic del *vtest*, la mitjana i desviació estàndard del residu al grup, la mitjana i desviació estàndard general i el *p*-valor associat.

\$`1`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
orgànics	-2.293027	2.23404255	24.9466667	4.6136374	110.393586	2.184643e-02
pneumàtics	-2.526123	0.48936170	1.9866667	1.3818326	6.606044	1.153290e-02
sanitaris	-2.733145	1.10638298	4.8266667	3.0682592	15.170474	6.273271e-03
vidre	-2.965590	0.89361702	12.9866667	1.5606060	45.447550	3.021028e-03
cendres	-3.014068	0.06382979	0.8266667	0.2444495	2.820749	2.577698e-03
àcids	-3.409679	0.70212766	7.2133333	1.7493451	21.283041	6.503935e-04
fèrrics	-3.614925	0.68085106	11.0000000	1.1129104	31.814881	3.004340e-04
vehícles	-3.628934	1.91489362	11.1733333	4.6532066	28.434427	2.845937e-04
metàl·lics	-3.977258	3.48936170	67.7466667	5.9496910	180.063033	6.971463e-05
amiant	-4.089406	0.29787234	1.8133333	0.8229396	4.130192	4.324788e-05
químics	-4.512009	4.06382979	42.0533333	6.3628050	93.838143	6.421638e-06
minerals	-4.647003	2.10638298	17.6800000	4.0012447	37.351005	3.367920e-06
pintures	-4.902155	3.06382979	42.4800000	5.0419696	89.613445	9.479108e-07
plàstic	-5.086742	4.19148936	53.2800000	5.7856694	107.553839	6.642668e-07
dissolvents	-5.130675	1.23404255	17.0400000	2.2141976	34.334605	2.887045e-07
NS	-5.157701	2.42553191	24.2933333	4.4993838	47.253578	2.499999e-07
hidrocarburs	-5.180105	4.72340426	45.9333333	8.4820264	88.664361	2.217612e-07
paper	-5.353614	5.85106383	69.8800000	9.2021707	133.295357	8.621478e-08
llots	-5.453779	4.34042553	43.5066667	8.2675387	80.038720	4.931048e-08
mesclats	-5.461170	11.40425532	130.9466667	15.9355879	243.961958	4.730076e-08
fusta	-5.606248	2.70212766	44.8933333	4.2018987	83.875554	2.067596e-08
RAEE	-5.799276	4.06382979	36.9066667	6.9325144	63.117916	6.660179e-09
piles	-5.868367	2.57446809	18.7733333	4.8891747	30.764730	4.401086e-09

Taula 4.5. Residus del primer *cluster*

Amb relació al segon *cluster*, també amb molts individus, destaca per la producció de residus del tipus minerals i fusta.

\$`2`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
minerals	2.944393	41.29412	17.68000	56.29728	37.35101	0.00323589
fusta	1.998301	80.88235	44.89333	87.46285	83.87555	0.04568402

Taula 4.6. Residus del segon *cluster*

El tercer grup es caracteritza per la producció de residus del tipus vehicles fora d'ús i de pneumàtics, és a dir, residus derivats de la indústria de l'automoció. A aquest *cluster* hi pertanyen empreses relacionades amb la reparació i instal·lació de maquinària i equips (CNAE 33), i la venda i reparació de vehicles de motor i motocicletes (CNAE 45).

\$`3`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
vehícles	7.577753	162.5	11.17333333	42.5	28.4344267	3.515923e-14
pneumàtics	6.684615	33.0	1.98666667	20.0	6.6060444	2.315328e-11
piles	4.430464	114.5	18.77333333	1.5	30.7647302	9.403036e-06
explosius	4.183323	0.5	0.02666667	0.5	0.1611073	2.872793e-05
hidrocarburs	3.558154	267.5	45.93333333	60.5	88.6643609	3.734701e-04
dissolvents	3.378172	98.5	17.04000000	21.5	34.3346045	7.296937e-04

Taula 4.7. Residus del tercer *cluster*

Quant al quart *cluster*, destaquen els residus de pintures, plàstic i dissolvents usats. Hi pertanyen empreses associades als grups d'activitat 11 (Fabricació de begudes), 18 (Arts gràfiques i reproducció de suports enregistrats), 21 (Fabricació de productes farmacèutics), 22 (Fabricació de productes de cautxú i matèries plàstiques), 28 (Fabricació de maquinària i equips ncaa), 29 (Fabricació de vehicles de motor, remolcs i semiremolcs).

\$`4`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
pintures	4.482507	200.83333	42.480000	179.11022	89.61345	7.377123e-06
plàstic	4.050064	225.00000	53.280000	137.85258	107.55384	5.120359e-05
dissolvents	3.912757	70.00000	17.040000	28.98275	34.33460	9.124838e-05
paper	3.652982	261.83333	69.880000	111.25109	133.29536	2.592120e-04
vidre	3.545042	76.50000	12.986667	131.35797	45.44755	3.925507e-04
àcids	3.450896	36.16667	7.213333	52.85646	21.28304	5.587294e-04
RAEE	3.118449	114.50000	36.906667	33.67863	63.11792	1.818054e-03
llots	3.026494	139.00000	43.506667	57.03800	80.03872	2.474075e-03
mesclats	3.012474	420.66667	130.946667	203.84689	243.96196	2.591276e-03
NS	2.605746	72.83333	24.293333	22.79924	47.25358	9.167452e-03
químics	2.539621	136.00000	42.053333	69.94998	93.83814	1.109727e-02
fusta	2.498312	127.50000	44.893333	52.19435	83.87555	1.247864e-02
hidrocarburs	2.362232	128.50000	45.933333	89.45157	88.66436	1.816526e-02
sanitaris	2.202741	18.00000	4.826667	29.69287	15.17047	2.761300e-02
piles	2.148762	44.83333	18.773333	18.09619	30.76473	3.165330e-02
metàl·lics	2.100299	216.83333	67.746667	156.48260	180.06303	3.570252e-02

Taula 4.8. Residus del quart *cluster*

Pel que fa al cinquè grup, està format únicament per un individu, les Indústries químiques (CNAE 20). En aquest *cluster*, destaca sobretot la presència de residus químics, seguit també per amiant i dissolvents usats. La desviació estàndard dins del grup és 0, ja que el grup està format per un sol individu.

\$`5`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
químics	6.574583	659	42.053333	0	93.838143	4.878969e-11
amiant	5.129705	23	1.813333	0	4.130192	2.901972e-07
dissolvents	4.658857	177	17.040000	0	34.334605	3.179705e-06
llots	4.279096	386	43.506667	0	80.038720	1.876542e-05
RAEE	4.247500	305	36.906667	0	63.117916	2.161694e-05
piles	3.582891	129	18.773333	0	30.764730	3.398128e-04
sanitaris	3.439137	57	4.826667	0	15.170474	5.835727e-04
pintures	3.219606	331	42.480000	0	89.613445	1.283668e-03
mesclats	3.197438	911	130.946667	0	243.961958	1.386540e-03
NS	3.146993	173	24.293333	0	47.253578	1.649591e-03
plàstic	3.084223	385	53.280000	0	107.553839	2.040847e-03
fusta	3.041490	300	44.893333	0	83.875554	2.354103e-03
minerals	2.846510	124	17.680000	0	37.351005	4.420138e-03
paper	2.484108	401	69.880000	0	133.295357	1.298766e-02
àcids	2.433236	59	7.213333	0	21.283041	1.496453e-02

Taula 4.9. Residus del cinquè *cluster*

El sisè grup també està format per només un individu, el codi 25 (Fabricació de productes metàl·lics, excepte maquinària i equips). Pel que fa als residus d'aquest grup, en ressalten els fèrrics i no fèrrics i els metàl·lics.

\$`6`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
fèrrics	7.512208	250	11.000000	0	31.814881	5.813839e-14
metàl·lics	7.321066	1386	67.7466667	0	180.063033	2.460090e-13
cendres	6.088219	18	0.8266667	0	2.820749	1.141740e-09
hidrocarburs	5.448262	529	45.9333333	0	88.664361	5.086440e-08
llots	5.003745	444	43.5066667	0	80.038720	5.622716e-07
mesclats	4.730464	1285	130.9466667	0	243.961958	2.240069e-06
àcids	3.842809	89	7.2133333	0	21.283041	1.216340e-04
NS	3.422104	186	24.2933333	0	47.253578	6.213850e-04
paper	3.091781	482	69.8800000	0	133.295357	1.989598e-03
fusta	3.005723	297	44.8933333	0	83.875554	2.649504e-03
dissolvents	2.940474	118	17.0400000	0	34.334605	3.277109e-03
minerals	2.819737	123	17.6800000	0	37.351005	4.806307e-03
pintures	2.639336	279	42.4800000	0	89.613445	8.306858e-03
plàstic	2.247433	295	53.2800000	0	107.553839	2.461240e-02
químics	2.216014	250	42.0533333	0	93.838143	2.669052e-02
piles	2.087672	83	18.7733333	0	30.764730	3.682743e-02
RAEE	2.061116	167	36.9066667	0	63.117916	3.929199e-02

Taula 4.10. Residus del sisè *cluster*

L'últim *cluster* conté un únic individu: les Indústries de productes alimentaris (CNAE 10). Entre els residus produïts per aquesta indústria, els més rellevants són els SANDACH i els orgànics biodegradables.

\$`7`	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
sandach	8.589800	365	5.533333	0	41.848085	8.712110e-18
orgànics	7.301632	831	24.946667	0	110.393586	2.842981e-13
sanitaris	5.482580	88	4.826667	0	15.170474	4.191682e-08
NS	5.453696	282	24.293333	0	47.253578	4.933341e-08
paper	5.177375	760	69.880000	0	133.295357	2.250301e-07
plàstic	4.739208	563	53.280000	0	107.553839	2.145553e-06
RAEE	4.627740	329	36.906667	0	63.117916	3.696769e-06
fusta	3.828370	366	44.893333	0	83.875554	1.289947e-04
mesclats	3.787694	1055	130.946667	0	243.961958	1.520517e-04
piles	3.777919	135	18.773333	0	30.764730	1.581441e-04
químics	2.823443	307	42.053333	0	93.838143	4.751085e-03
amiant	2.466390	12	1.813333	0	4.130192	1.364825e-02
hidrocarburs	2.042158	227	45.933333	0	88.664361	4.113582e-02

Taula 4.11. Residus del setè *cluster*

4.4. Modelització

4.4.1. Caracterització de les freqüències

Una altra manera de caracteritzar les freqüències dels diferents tipus residus per a cada grup d'activitat industrial és a partir de la modelització estadística. Es planteja ajustar un model lineal generalitzat on la resposta és la freqüència, i les variables explicatives són el codi d'activitat i el codi de residu. Aquest model es considerarà additiu, és a dir, sense tenir en compte interaccions.

$$freq \sim cnae + cer$$

Se sap que aquest model segueix una distribució de $Poisson(\mu_{ij})$. No obstant això, és necessari validar aquest model mitjançant un test de sobredispersió. Amb una transformació amb $trafo = 2$, s'obté el següent resultat:

```
Overdispersion test
data:  mod0
z = 5.6792, p-value = 6.765e-09
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0.256245
```

Figura 4.14. Resultat del test de sobredispersió

Aquest resultat mostra un p-valor significatiu indicant que hi ha sobredispersió en el model i que els errors estàndard d'algunes variables podrien estar subestimats. Per abordar aquesta sobredispersió, es proposa ajustar el model amb una distribució Binomial Negativa, $NB(\mu_{ij}, \theta)$. El paràmetre θ es pot estimar mitjançant una funció auxiliar de R, i posteriorment, es pot ajustar el model amb aquesta nova distribució.

Amb el model ja ajustat, se sap que

$$\log(\mu_{ij}) = \beta_0 + \beta_{1i}\alpha_i + \beta_{2j}\gamma_j$$

On β_0 representa la contribució de l'activitat industrial i tipus de residu de referència (codi d'activitat 01, Agricultura, ramaderia, caça i activitats dels serveis que s'hi relacionen, i residus de tipus àcida, alcalins i salins), β_{1i} la contribució de l'activitat industrial i i β_{2j} la contribució del tipus de residu j . A més, α_i prendrà valors 0 o 1 en funció de si es té en compte l'activitat industrial i , i γ_j prendrà valors 0 o 1 en funció de si es considera el tipus de residu j . Per tant, la freqüència resultant es pot trobar mitjançant

$$\mu_{ij} = e^{\beta_0 + \beta_{1i}\alpha_i + \beta_{2j}\gamma_j}$$

Amb aquest model, es pot observar que els efectes de les activitats industrials i dels tipus de residus són independents. D'aquesta manera, es pot fer una anàlisi de freqüències tant de forma individual com en combinació.

Amb el model ja format, es procedeix a fer-ne una validació a partir dels gràfics que es presenten a continuació. Amb els dos gràfics de l'esquerra es veu que es pot acceptar la premissa de l'homoscedasticitat, és a dir, la variància dels residus és constant. D'altra banda, també es pot afirmar que els residus segueixen una distribució normal, tot i que hi ha certa asimetria a l'extrem superior. Finalment, amb l'últim gràfic es pot veure que hi ha algunes observacions influents que tenen una distància de Cook elevada. Per tant, hi ha alguns *outliers* en aquest model, però es compleixen les premisses.

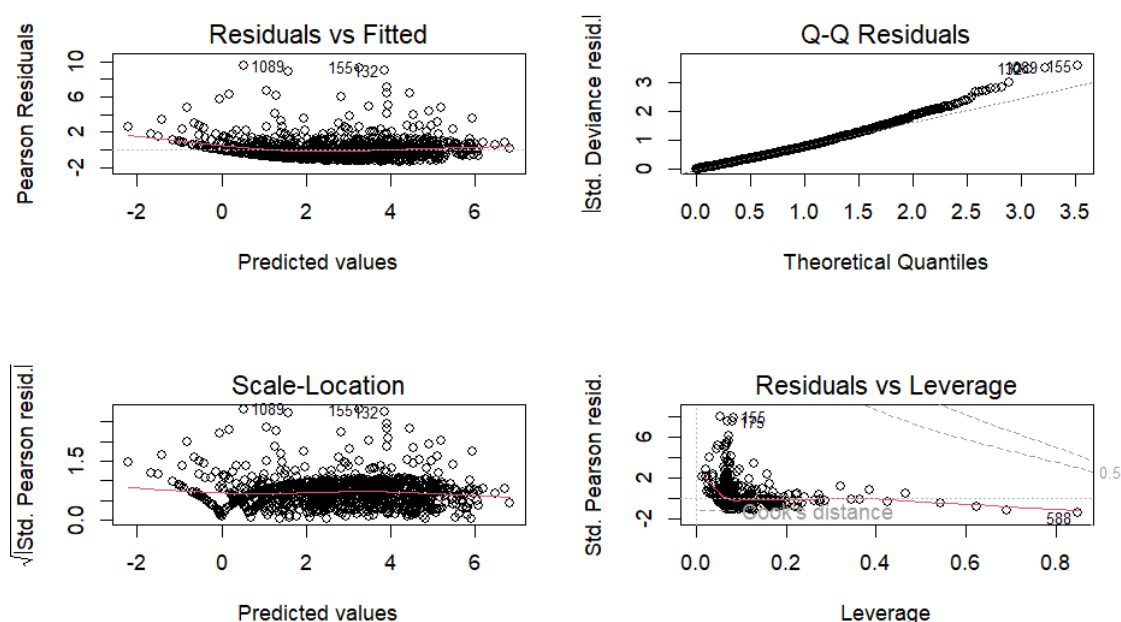


Figura 4.15. Gràfics de validació del model

Com s'ha dit, els coeficients d'aquest model representen contribucions. Per mostrar això de manera visual, es proposa un gràfic de barres dels coeficients dels CNAE per destacar aquells que tenen una importància més significativa.

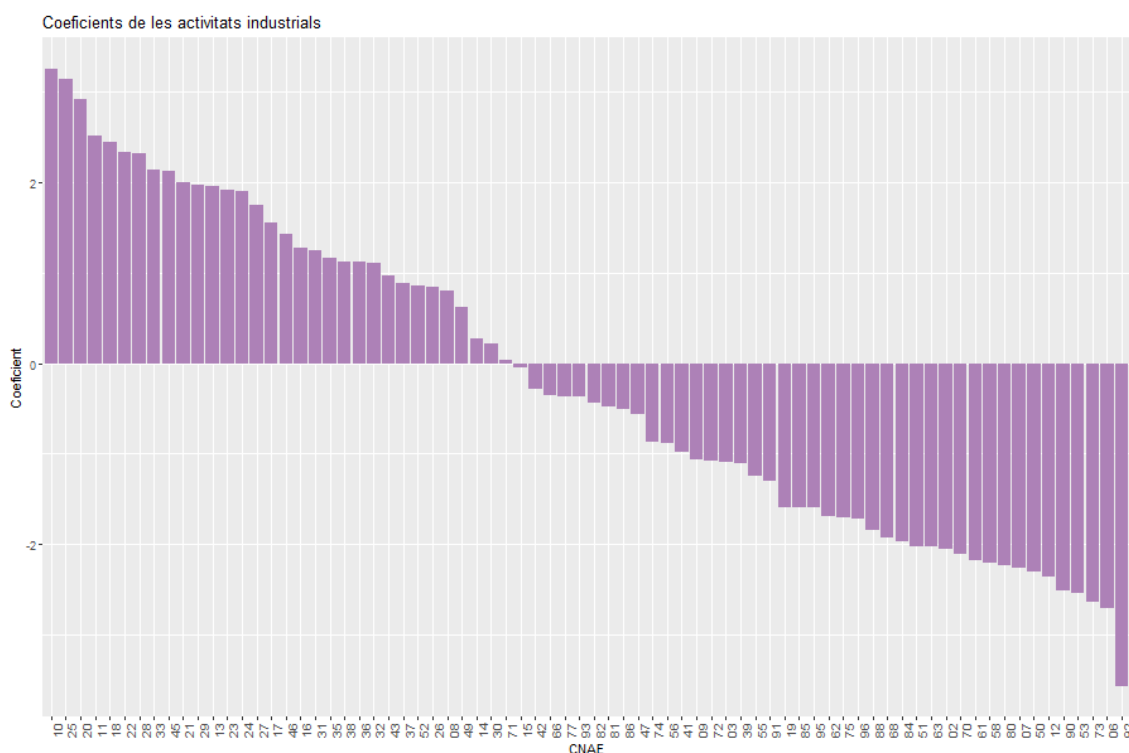


Figura 4.16. Coeficients del model per les activitats industrials

Aquest gràfic es pot interpretar de tal manera que les activitats amb coeficients positius contribueixen a augmentar la freqüència respecte a la categoria de referència, mentre que els coeficients negatius indiquen una disminució en la freqüència respectiva.

Com es veu, l'activitat industrial amb una contribució més destacada i, per tant, on hi haurà més freqüència de residus és la 10 (Indústries de productes alimentaris), seguida de la 25 (Fabricació de productes metàl·lics, excepte maquinària i equips) i la 20 (Indústries químiques). Destaquen com a menors generadors de residus al DARI els codis d'activitat 92 (Activitats relacionades amb els jocs d'atzar i les apostes), 06 (Extracció de petroli brut i de gas natural) i 73 (Publicitat i estudis de mercat).

D'altra banda, també es pot veure de manera gràfica la contribució de cada tipus de residu al model.

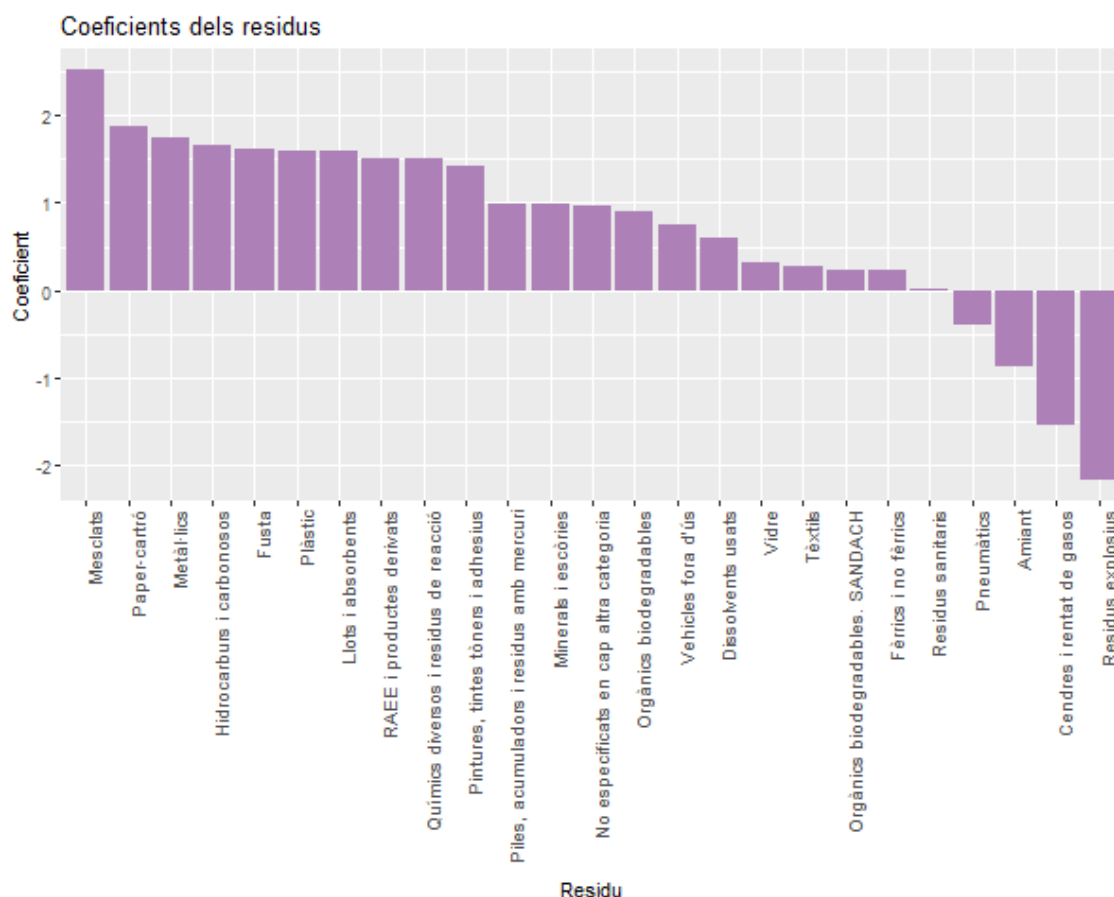


Figura 4.17. Coeficients del model pels residus

D'aquesta manera, es fa evident que els residus més freqüents són els mesclats, seguits del paper i cartró i dels metàl·lics. En contraposició, aquells residus que es generen amb menor freqüència són els explosius i les cendres i rentat de gasos.

Es proposa dos exemples per calcular les freqüències de les combinacions d'activitat industrial i tipus de residu:

- Pel codi d'activitat 10 i els residus mesclats (la combinació amb contribucions més altes).

$\mu = e^{1.042423+3.253450+2.520022} = 912$, on 1.042423 és el paràmetre de l'intercept, 3.253450 el coeficient corresponent al codi 10 i 2.520022 el coeficient corresponent als residus mesclats.

- Pel codi d'activitat 92 i els residus explosius (la combinació amb contribucions més baixes).

$\mu = e^{1.042423-3.562445-2.173754} = 0$, on 1.042423 és el paràmetre de l'intercept, -3.562445 el coeficient corresponent al codi d'activitat 92 i -2.173754 la β corresponent als residus explosius.

Els coeficients obtinguts per a les activitats industrials en aquest model haurien de mostrar similitud amb els *scores* obtinguts per la primera component principal en termes d'ordre. Cal recordar que la primera component proporcionava informació sobre la magnitud, és a dir, les activitats més generadores de residus tenien *scores* més elevats.

A continuació, es presenta una taula que ordena les activitats industrials segons els *scores* obtinguts per la primera component principal i, alhora, segons els coeficients obtinguts en el model.

Posició	PC1	Model
1	10	10
2	25	25
3	20	20
4	22	11
5	28	18
6	18	22
7	33	28
8	45	33
9	21	45
10	29	21

Taula 4.12. Comparació del model amb el PCA

Com es pot observar a la taula, les tres primeres posicions coincideixen en ambdues tècniques utilitzades. Amb l'excepció de les activitats industrials 29 i 11, que només apareixen en el rànquing del PCA i del model, respectivament, les altres estan presents en ambdós, tot i que amb ordres diferents.

4.4.2. Caracterització de les quantitats

Fins ara, ens hem centrat a caracteritzar els residus en funció de les seves freqüències, no obstant això, un altre aspecte a tenir en compte són les quantitats (en tones) generades d'aquests. Les quantitats no són per empresa, sinó que han estat agregades per caracteritzar la presència global d'un residu en una activitat industrial.

Per tant, es proposa un model lineal additiu de la quantitat en funció de l'activitat industrial i el tipus de residu, on $Q_{ij} \sim N(\mu_{ij}, \sigma^2)$. L'equació del model seria la següent:

$$\log(Q_{ij} + 1) = \beta_0 + \beta_{1i}\alpha_i + \beta_{2j}\gamma_j + \epsilon_{ij}$$

S'opta per calcular el logaritme de la quantitat, ja que hi ha asimetria en els valors. Aquesta transformació permet obtenir una escala uniforme per a comparacions i assegurar que els residus segueixin una distribució normal. Pel que fa als paràmetres, β_0 és la contribució de la categoria de referència (la mateixa que al model anterior), β_{1i} i α_i fan referència a la contribució i l'aparició, respectivament, de l'activitat industrial i , mentre que β_{2j} i γ_j fan referència als residus. Finalment, ϵ_{ij} representa l'error, que ha de seguir una distribució $N(0, \sigma^2)$.

Abans d'aplicar aquest model, s'ha de validar que compleix les premisses que s'assumeixen pels models lineals. És important que els residus segueixin una distribució normal, s'hi vegi homoscedasticitat i no hi hagi observacions massa influents.

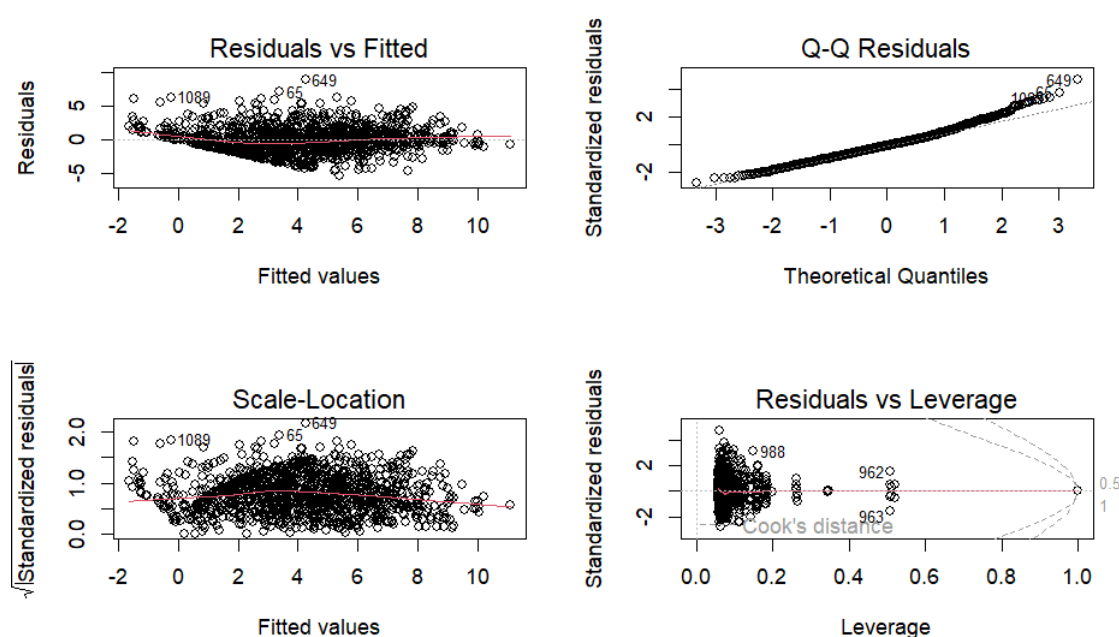


Figura 4.18. Gràfics de validació del model lineal

En el primer gràfic, s'hi representen els residus en funció dels valors predits. S'hi ha d'identificar un patró aleatori i uniforme per satisfer l'assumpció de l'homoscedasticitat. Així, es pot apreciar que els residus estan distribuïts de manera aleatòria. L'única tendència destacable es deu al fet que la resposta ha de ser estrictament positiva, la qual cosa pot generar un patró específic en la distribució dels residus. D'altra banda, el gràfic següent serveix per a verificar la normalitat dels residus. Com es pot apreciar, aquests s'ajusten adequadament a la línia de normalitat, tot i que es poden identificar cues pesants a l'extrem superior, indicant certa asimetria. Finalment, amb l'últim gràfic es pot veure que hi ha alguna observació influent.

Per la validació del model, també s'ha de tenir en compte la deviància. Aquest model té una deviància de 1181 amb 1063 graus de llibertat. Per fer el test corresponent per avaluar si és necessari incloure la interacció al model, calculem el punt crític $\chi^2_{1063} = 1139 > 1181$ i, per tant, s'hauria d'afegir la interacció. Malgrat això, es pren la decisió de no incloure-la, ja que podria fer que el model esdevingués massa complex, i els valors de la deviància i el punt crític no estan excessivament allunyats.

Amb el model ja validat, es pot fer un gràfic dels coeficients per veure les contribucions en la quantitat de residus tant per tipus de residu com per activitat industrial. Primer, es mostra el gràfic en funció dels codis CNAE.

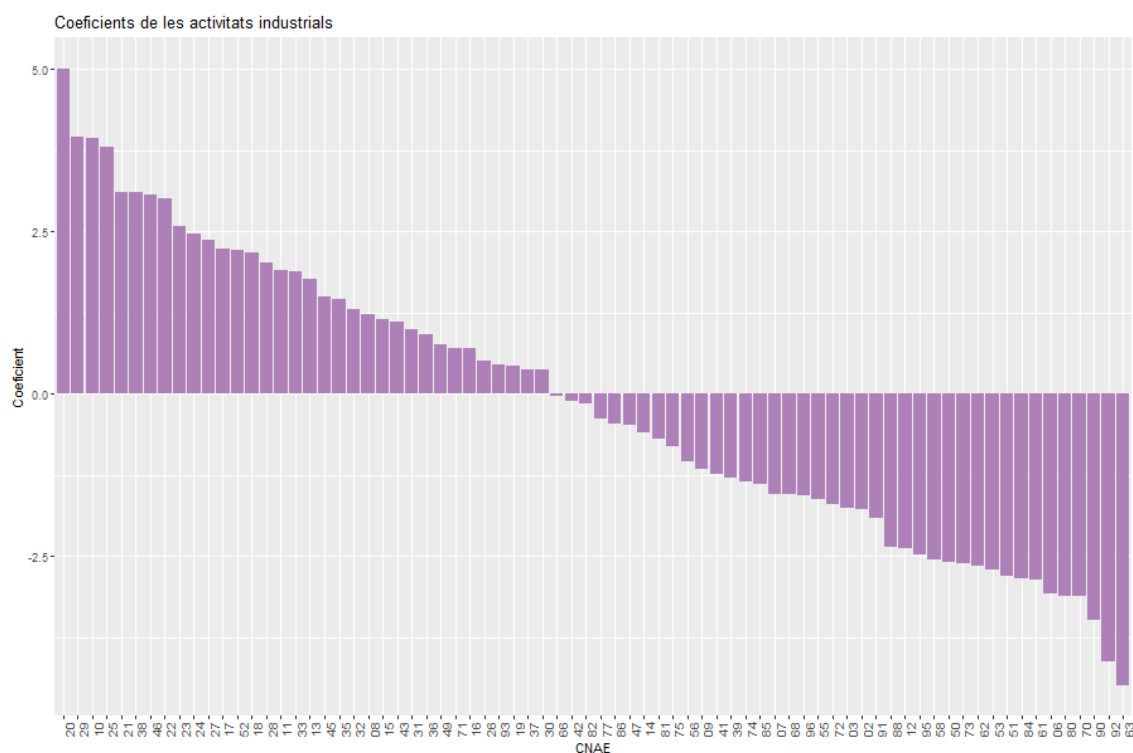


Figura 4.19. Coeficients del model lineal per les activitats industrials

D'aquesta manera, es pot veure que les activitats industrials que contribueixen de manera més significativa, en termes de tones de residus, són els grups 20 (Indústries químiques), 29 (Fabricació de vehicles de motor, remolcs i semiremolcs) i 10 (Indústries de productes alimentaris). D'altra banda, on es genera menys quantitat de residus és a la indústria dels Serveis d'informació (codi 63) i a les Activitats relacionades amb els jocs d'atzar i les apostes (codi 92).

Pel que fa als residus, les contribucions es veuen en el següent gràfic.

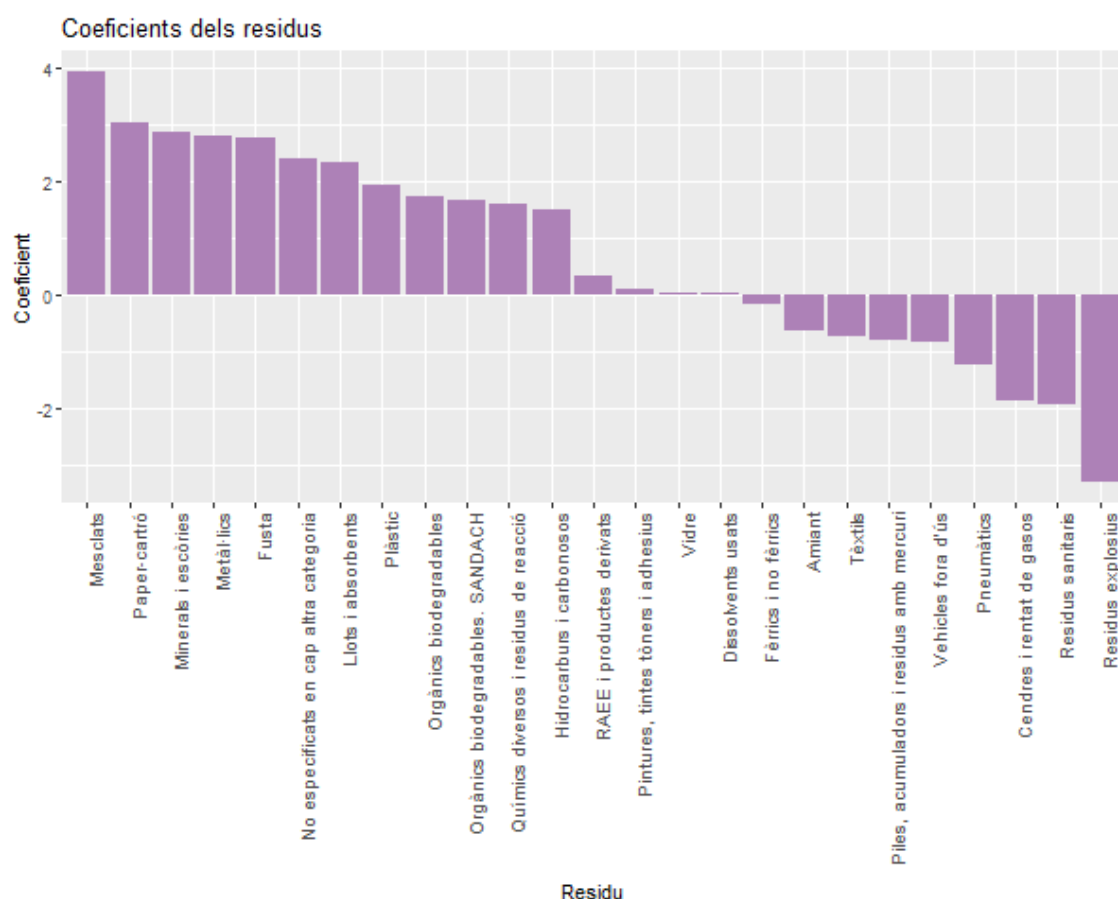


Figura 4.20. Coeficients del model pels residus

Així, es pot observar que els residus que es generen en més quantitat són els mesclats, el paper i cartró i els minerals i escòries. En contrast, aquells amb una quantitat més reduïda són els residus explosius i els sanitaris.

Els resultats presentats en aquest model són en general semblants als del model anterior. No obstant això, destaca la importància de considerar tant les freqüències com les quantitats dels residus, ja que es poden identificar algunes diferències rellevants.

Es proposa dos exemples per calcular les tones de residu que es generen en les combinacions d'activitat industrial i tipus de residu:

- Pel codi d'activitat 20 i els residus mesclats (la combinació amb contribucions més altes).

$Q + 1 = e^{2.16293+5.00331+3.93151} = 66022.44$ tones, on 2.16293 és el paràmetre de l'*intercept*, 5.00331 el coeficient corresponent al codi d'activitat 20 i 3.93151 la β corresponent als residus mesclats.

- Pel codi d'activitat 63 i els residus explosius (la combinació amb contribucions més baixes).

$Q + 1 = e^{2.16293-4.47771-3.30226} = 0.004$, on 2.16293 és el paràmetre de l'*intercept*, -4.47771 el coeficient corresponent al codi d'activitat 63 i -3.30226 la β corresponent als residus explosius.

5. CONCLUSIONS

En aquest estudi s'ha buscat caracteritzar les activitats industrials juntament amb els residus que s'hi generen de manera bivariant. Aquesta caracterització s'ha fet a partir de dues tècniques estadístiques: l'anàlisi multivariant i la modelització.

Per una banda, amb l'anàlisi de components principals, s'ha treballat amb 6 components principals. La primera component ha identificat les activitats industrials que destaquen com a principals generadores de residus, mentre que les altres components han permès caracteritzar les relacions entre les activitats industrials i els tipus de residus generats.

Mitjançant el *cluster* jeràrquic, s'han identificat 7 grups. Un d'aquests grups agrupa moltes activitats industrials que generen pocs residus, mentre que tres grups consisteixen en una única activitat industrial, proporcionant caracteritzacions més específiques.

D'altra banda, s'ha fet un model additiu de comptatges per analitzar la freqüència d'activitats industrials i residus. Els coeficients d'aquest model, tant per als residus com per a les activitats industrials, han permès caracteritzar la presència d'empreses segons certes característiques.

Finalment, s'ha implementat un model lineal per caracteritzar la quantitat total de residu generada en funció de la tipologia del residu i de l'activitat industrial.

Pel que fa a les conclusions extretes a partir dels resultats obtinguts en aquest estudi, s'ha vist que les activitats industrials més generadores de residus són les indústries de productes alimentaris (CNAE 10), la fabricació de productes metàl·lics, excepte maquinària i equips (CNAE 25) i les indústries químiques (CNAE 20). També s'ha trobat els residus més importants per aquestes indústries. Per la primera, en destaquen els residus orgànics biodegradables i els SANDACH. Pel segon codi d'activitat, en sobresalten els residus de tipus fèrrics i no fèrrics i els metàl·lics. Finalment, pel codi CNAE 20 destaquen els residus químics sobretot.

Per acabar, vull subratllar que aquest estudi m'ha permès aplicar les tècniques apreses durant el grau, tant en l'anàlisi multivariant com en la modelització, així com els coneixements suficients d'SQL per manipular la base de dades i R per aplicar les tècniques esmentades.

6. BIBLIOGRAFIA

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. CRC press.

Aluja Banet, T. (1999). *El análisis de componentes principales, una aproximación al Data Mining*. Barcelona: Ediciones Universidad de Barcelona.

Richard, A. J. (2007). *Applied multivariate statistical analysis*.

Pontaque, F. C. (2005). *Modelos lineales* (Vol. 3). Edicions Universitat Barcelona.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC press.

Nielsen, F., & Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195-211.

Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

7. ANNEX

Codificació de les activitats industrials

Codi	Descripció
01	Agricultura, ramaderia, caça i activitats dels serveis que s'hi relacionen
02	Silvicultura i explotació forestal
03	Pesca i aquicultura
05	Extracció d'antracita, hulla i lignit
06	Extracció de petroli brut i de gas natural
07	Extracció de minerals metàl·lics
08	Extracció de minerals no metàl·lics ni energètics
09	Activitats de suport a les indústries extractives
10	Indústries de productes alimentaris
11	Fabricació de begudes
12	Indústries del tabac
13	Indústries tèxtils
14	Confecció de peces de vestir
15	Indústria del cuir i del calçat
16	Indústria de la fusta i del suro, excepte mobles; cistelleria i esparteria
17	Indústries del paper
18	Arts gràfiques i reproducció de suports enregistrats
19	Coqueries i refinació del petroli
20	Indústries químiques
21	Fabricació de productes farmacèutics
22	Fabricació de productes de cautxú i matèries plàstiques

23	Fabricació d'altres productes minerals no metàl·lics
24	Metal·lúrgia; fabricació de productes bàsics de ferro, acer i ferroaliatges
25	Fabricació de productes metàl·lics, excepte maquinària i equips
26	Fabricació de productes informàtics, electrònics i òptics
27	Fabricació de materials i equips elèctrics
28	Fabricació de maquinària i equips ncaa
29	Fabricació de vehicles de motor, remolcs i semiremolcs
30	Fabricació d'altres materials de transport
31	Fabricació de mobles
32	Indústries manufactureres diverses
33	Reparació i instal·lació de maquinària i equips
35	Subministrament d'energia elèctrica, gas, vapor i aire condicionat
36	Captació, potabilització i distribució d'aigua
37	Recollida i tractament d'aigües residuals
38	Activitats de recollida, tractament i eliminació de residus; activitats de valorització
39	Activitats de descontaminació i altres serveis de gestió de residus
41	Construcció d'immobles
42	Construcció d'obres d'enginyeria civil
43	Activitats especialitzades de la construcció
45	Venda i reparació de vehicles de motor i motocicletes
46	Comerç a l'engròs i intermediaris del comerç, excepte vehicles de motor i motocicletes
47	Comerç al detall, excepte el comerç de vehicles de motor i motocicletes

49	Transport terrestre; transport per canonades
50	Transport marítim i per vies de navegació interiors
51	Transport aeri
52	Emmagatzematge i activitats afins al transport
53	Activitats postals i de correus
55	Serveis d'allotjament
56	Serveis de menjar i begudes
58	Edició
59	Activitats de cinematografia, de vídeo i de programes de televisió; activitats d'enregistrament de so i edició musical
60	Activitats d'emissió i programació de ràdio i televisió
61	Telecomunicacions
62	Serveis de tecnologies de la informació
63	Serveis d'informació
64	Mediació financera, excepte assegurances i fons de pensions
65	Assegurances, reassegurances i fons de pensions, excepte la Seguretat Social obligatòria
66	Activitats auxiliars de la mediació financera i d'assegurances
68	Activitats immobiliàries
69	Activitats jurídiques i de comptabilitat
70	Activitats de les seus centrals; activitats de consultoria de gestió empresarial
71	Serveis tècnics d'arquitectura i enginyeria; assajos i anàlisis tècnics
72	Recerca i desenvolupament
73	Publicitat i estudis de mercat

74	Altres activitats professionals, científiques i tècniques
75	Activitats veterinàries
77	Activitats de lloguer
78	Activitats relacionades amb l'ocupació
79	Activitats de les agències de viatges, operadors turístics i altres serveis de reserves i activitats que s'hi relacionen
80	Activitats de seguretat i investigació
81	Serveis a edificis i activitats de jardineria
82	Activitats administratives d'oficina i altres activitats auxiliars a les empreses
84	Administració pública, Defensa i Seguretat Social obligatòria
85	Educació
86	Activitats sanitàries
87	Activitats de serveis socials amb allotjament
88	Activitats de serveis socials sense allotjament
90	Activitats de creació, artístiques i d'espectacles
91	Activitats de biblioteques, arxius, museus i altres activitats culturals
92	Activitats relacionades amb els jocs d'atzar i les apostes
93	Activitats esportives, recreatives i d'entreteniment
94	Activitats associatives
95	Reparació d'ordinadors, d'efectes personals i efectes domèstics
96	Altres activitats de serveis personals
97	Activitats de les llars que donen ocupació a personal domèstic
98	Activitats de les llars que produeixen béns i serveis per a ús propi

99	Organismes extraterritorials
----	------------------------------

Llista de les agrupacions de residus

	Agrupació
1	Minerals i escòries
2	Llots i absorbents
3	No especificats en cap altra categoria
4	Orgànics biodegradables. SANDACH
5	Orgànics biodegradables
6	Plàstic
7	Químics diversos i residus de reacció
8	Metàl·lics
9	Dissolvents usats
10	Fusta
11	Mesclats
12	Àcids, alcalins i salins
13	Tèxtils
14	Pintures, tintes, tòners i adhesius
15	Piles, acumuladors i residus amb mercuri
16	Amiant
17	Residus Sanitaris
18	RAEE i productes derivats
19	Cendres i rentat de gasos
20	Vidre

21	Fèrrics i no fèrrics
22	Paper-cartró
23	Pneumàtics
24	Vehicles fora d'ús
25	Residus explosius
26	Hidrocarburs i carbonosos

Codi

```
library(readxl)
library(dplyr)
library(sqldf)
library(ggplot2)

##### BASE DE DADES #####

#bd ARC
arc <- read_excel("B_dari_19_tota_catalunya_dades-ok.xlsx", skip =
2)
colnames(arc)[13:16] <- c("descripcio_ccae", "cer", "descripcio_cer",
"descripcio_declarant")
arc$cnae <- as.character(arc$cnae)

# master cer
master_cer <- read_excel("Siner_PoC_BigQuery v2.0.xlsx", sheet =
"Master_CER")
master_cer <- master_cer[,1:6]
colnames(master_cer)[3:6] <- c("descripcio_ESP", "descripcio_CAT",
"agrupacio_ESP", "grup_cer")
master_cer$grup_cer[master_cer$grup_cer=="Residuos sanitarios"] <-
"Residus sanitaris"
master_cer$grup_cer[master_cer$grup_cer=="Varis"] <- "No
especificats en cap altra categoria"
cer <- sqldf("SELECT DISTINCT cer, descripcio_cer, grup_cer as
grup_cer, arc.classe_residu
FROM arc LEFT JOIN master_cer ON
arc.cer=master_cer.codigo")

# arreglar cnae
cnae_mal <- arc[which(nchar(arc$cnae)!=4),]
ccae <- read.csv("ccae-2009-ca-classes.csv", sep = ";", skip = 3)
ccae$Codi <- as.character(ccae$Codi)
ccae[which(nchar(ccae$Codi)!=4), "Codi"] <-
paste0("0", ccae[which(nchar(ccae$Codi)!=4), "Codi"])
```

```

tt <- sqldf("SELECT DISTINCT Codi, cnae, descripcio_ccae
FROM cnae_mal LEFT JOIN ccae ON
cnae_mal.descripcio_ccae=ccae.Descripció")
cc <- sqldf("SELECT DISTINCT arc.cnae, Codi
FROM tt LEFT JOIN arc ON arc.cnae=tt.cnae
WHERE Codi!='NA'")
for (i in 1:nrow(arc)) {
  if (nchar(arc$cnae[i]) != 4 && arc$cnae[i] %in% cc$cnae) {
    arc$cnae[i] <- cc$Codi[cc$cnae == arc$cnae[i]]
  }
}
cnae <- data.frame(substr(sqldf("SELECT DISTINCT cnae FROM
arc")$cnae, start = 1, stop = 2), sqldf("SELECT DISTINCT cnae FROM
arc")$cnae)
colnames(cnae) <- c("grup_cnae", "cnae")

# NA
cer[which(is.na(cer$grup_cer)),]
cer[which(is.na(cer$grup_cer)),]$grup_cer <- "Piles, acumuladors i
residus amb mercuri"

# taula amb nif, cnae, cer, quantitat
bd <- sqldf("SELECT nif, cnae, cer, sum(quantitat_anual) as
quantitat
FROM arc
GROUP BY nif, cnae, cer")
bd$grup_cnae <- substr(bd$cnae, start = 1, stop = 2)
bd[,6] <- sqldf("select grup_cer from bd left join master_cer on
bd.cer=master_cer.codigo")

##### TAULES #####

# Freqs cer i cnae
cnae_freq <- sqldf("SELECT grup_cnae, COUNT(grup_cnae) as freq
FROM bd
WHERE grup_cnae IS NOT NULL
GROUP BY grup_cnae")
cer_freq <- sqldf("SELECT grup_cer, COUNT(grup_cer) as freq
FROM bd
WHERE grup_cer IS NOT NULL
GROUP BY grup_cer")

# freqs combinació cer-cnae
cnae_cer <- na.omit(sqldf("SELECT grup_cnae, grup_cer,
COUNT(grup_cer) as freq_cer, sum(quantitat) as quantitat
FROM bd
GROUP BY grup_cnae, grup_cer"))

# taula de contingència
tcon <- matrix(0, nrow = length(unique(cnae_cer$grup_cnae)),
ncol = length(unique(cnae_cer$grup_cer)))

```

```

rownames(tcon) <- unique(cnae_cer$grup_cnae)
colnames(tcon) <- unique(cnae_cer$grup_cer)

for (i in 1:nrow(tcon)) {
  for (j in 1:ncol(tcon)) {
    subset_data <- cnae_cer[cnae_cer$grup_cnae == rownames(tcon)[i]
&
                                cnae_cer$grup_cer ==
colnames(tcon)[j], "freq_cer"]
    if (length(subset_data) == 1) {
      tcon[i, j] <- subset_data
    } else {
      tcon[i, j] <- 0
    }
  }
}

tcon <- as.data.frame(tcon)
colnames(tcon) <- c("fusta", "fèrrics", "hidrocarburs", "llots",
"mesclats", "metàl·lics", "minerals", "NS", "orgànics", "sandach",
"paper", "piles", "pintures", "plàstic", "químics", "RAEE",
"sanitaris", "tèxtils", "vidre", "dissolvents", "vehICLES",
"amiant", "pneumàtics", "àcids", "cendres", "explosius")

##### DESCRIPTIVA #####

#barplot cer
ggplot(data=cer_freq, aes(x=reorder(grup_cer, -freq), y=freq)) +
  geom_bar(stat = "identity", fill = "#AD81B7") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1))
+
  labs(title = "Freqüències dels grups de residus", x = "Residu",
y="Freqüència")

# barplot cnae
ggplot(data=cnae_freq, aes(x=reorder(grup_cnae, -freq), y=freq)) +
  geom_bar(stat = "identity", fill = "#AD81B7") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1,
size = 8)) +
  labs(title = "Freqüències dels grups d'activitats industrials", x
= "Activitat industrial", y="Freqüència")

# freqs relatives
cnae_freq$freq_rel <- round(cnae_freq$freq/sum(cnae_freq$freq), 3)
cer_freq$freq_rel <- round(cer_freq$freq/sum(cer_freq$freq), 3)

#heatmap
index_files_ordenades <- order(rowMeans(tcon_matrix), decreasing =
TRUE)
index_columnes_ordenades <- order(colMeans(tcon_matrix), decreasing

```

```

= TRUE)
tcon_matrix_ord <- tcon_matrix[index_files_ordenades,
index_columnes_ordenades]
heatmap(tcon_matrix_ord,
        Colv = NA, # Turn off column clustering
        Rowv = NA, # Turn off row clustering
        scale = "row", # Scale rows -> identifying trends or
outliers across different CER values
        #scale="none",
        main = "Heatmap de CER per cada CNAE",
        ylab = "CNAE", xlab = "CER")

#heatmap ampliati
heatmap(tcon_matrix_ord[1:15, 1:10],
        Colv = NA, # Turn off column clustering
        Rowv = NA, # Turn off row clustering
        scale = "row", # Scale rows -> identifying trends or
outliers across different CER values
        #scale="none",
        main = "Heatmap de CER per cada CNAE",
        ylab = "CNAE", xlab = "CER")

##### PCA #####

library(factoextra)
library(FactoMineR)
res.pca <- PCA(tcon, ncp = 6, graph=FALSE, scale.unit = T)
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 100), barcolor =
"#AD81B7", barfill = "#AD81B7")

# loadings
loadings <- data.frame(res.pca$var$coord)
loadings
names(sort(res.pca$ind$coord[,1], decreasing = T))

# PC1 vs PC2
fviz_pca_var(res.pca,
              col.var = "contrib", # Color by contributions to the PC
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE # Avoid text overlapping
)

fviz_pca_biplot(res.pca, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969" # Individuals color
)

# PC3 vs PC4
fviz_pca_var(res.pca,
              axes = c(3,4),
              col.var = "contrib", # Color by contributions to the PC

```

```

        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
        repel = TRUE      # Avoid text overlapping
    )

    fviz_pca_biplot(res.pca, repel = TRUE,
        axes = c(3,4),
        col.var = "#2E9FDF", # Variables color
        col.ind = "#696969" # Individuals color
    )

# PC5 vs PC6
fviz_pca_var(res.pca,
    axes = c(5,6),
    col.var = "contrib", # Color by contributions to the PC
    gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
    repel = TRUE      # Avoid text overlapping
)

fviz_pca_biplot(res.pca, repel = TRUE,
    axes = c(5,6),
    col.var = "#2E9FDF", # Variables color
    col.ind = "#696969" # Individuals color
)

# Scores
res.pca$ind$coord

##### CLUSTER #####

res.hc <- HCPC(res.pca, nb.clust = 7, graph = F)
res.hc$desc.var

fviz_dend(res.hc, rect = TRUE)
fviz_cluster(res.hc, ellipse = F, ggtheme = theme_minimal(),
show.clust.cent = FALSE)
fviz_cluster(res.hc, ellipse = F, ggtheme = theme_minimal(), axes =
c(3,4), show.clust.cent = FALSE)
fviz_cluster(res.hc, ellipse = F, ggtheme = theme_minimal(), axes =
c(5,6), show.clust.cent = FALSE)

##### GLM #####

mod0 <- glm(freq_cer~grup_cnae+grup_cer, cnae_cer, family=poisson)
library(AER)
dispersiontest(mod0, trafo=2) #té sobredispersió -> NB
library(MASS)
mod <- glm.nb(freq_cer~grup_cnae+grup_cer, cnae_cer)
mod2 <- glm(freq_cer~grup_cnae+grup_cer, cnae_cer,
family=neg.bin(theta=mod$theta))
par(mfrow=c(2,2))
plot(mod2)

```

```

par(mfrow=c(1,1))

# BARPLOT DE CNAE
coef_cnae <- data.frame(variable = names(mod2$coefficients)[2:75],
                        coefficient = mod2$coefficients[2:75])
coef_cnae <- coef_cnae[order(-coef_cnae$coefficient), ]
coef_cnae$variable <- sub("grup_cnae", "", coef_cnae$variable)

png("coef_cnae_glm.png", width = 900, height = 600)
ggplot(coef_cnae, aes(x = reorder(variable, coefficient, decreasing =
T), y = coefficient)) + #també exp(coef)
  geom_bar(stat = "identity", fill = "#AD81B7") +
  labs(title = "Coeficients de les activitats industrials",
       x = "CNAE",
       y = "Coeficient") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1,
size = 10))

# BARPLOT DE CER
coef_cer <- data.frame(variable =
names(mod2$coefficients)[76:length(mod2$coefficients)],
                      coefficient =
mod2$coefficients[76:length(mod2$coefficients)])
coef_cer <- coef_cer[order(-coef_cer$coefficient), ]
coef_cer$variable <- sub("grup_cer", "", coef_cer$variable)

png("coef_cer_glm.png", width = 600)
ggplot(coef_cer, aes(x = reorder(variable, coefficient, decreasing =
T), y = coefficient)) + #també exp(coef)
  geom_bar(stat = "identity", fill = "#AD81B7") +
  labs(title = "Coeficients dels residus",
       x = "Residu",
       y = "Coeficient") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1,
size = 10))
dev.off()

##### LM #####

mod_q <- lm(log(quantitat+1)~grup_cnae+grup_cer, cnae_cer) #log per
mateixa escala
summary(mod_q)
par(mfrow=c(2,2))
plot(mod_q)
Anova(mod_q)

# BARPLOT DE CNAE
coef_cnae <- data.frame(variable = names(mod_q$coefficients)[2:75],
                        coefficient = mod_q$coefficients[2:75])
coef_cnae <- coef_cnae[order(-coef_cnae$coefficient), ]

```

```

coef_cnae$variable <- sub("grup_cnae", "", coef_cnae$variable)

png("coef_cnae_lm.png", width = 900, height = 600)
ggplot(coef_cnae, aes(x = reorder(variable, coefficient, decreasing =
T), y = coefficient)) +
  geom_bar(stat = "identity", fill = "#AD81B7") +
  labs(title = "Coeficients de les activitats industrials",
       x = "CNAE",
       y = "Coeficient") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1,
size = 10))
dev.off()

# BARPLOT DE CER
coef_cer <- data.frame(variable =
names(mod_q$coefficients)[76:length(mod_q$coefficients)],
                      coefficient =
mod_q$coefficients[76:length(mod_q$coefficients)])
coef_cer <- coef_cer[order(-coef_cer$coefficient), ]
coef_cer$variable <- sub("grup_cer", "", coef_cer$variable)

png("coef_cer_lm.png", width = 600)
ggplot(coef_cer, aes(x = reorder(variable, coefficient, decreasing =
T), y = coefficient)) +
  geom_bar(stat = "identity", fill = "#AD81B7") +
  labs(title = "Coeficients dels residus",
       x = "Residu",
       y = "Coeficient") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust=1,
size = 10))
dev.off()

```

