PROJECTE FINAL DE CARRERA

MACHINE LEARNING APLICAT A PREDICCIONS DURANT LA NIT ELECTORAL

(MACHINE LEARNING APPLIED TO ELECTORAL NIGHT FORECASTING)

Estudis: Enginyeria de Telecomunicació

Autor: Sergio Fernández Bertolín

Codirectora: Margarita Cabrera Beán

Codirector: Pol Blasco Moreno

Any: 2016

Índex general

Col·laboracions

Departament de Teoria del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Teoria del Senyal
i Comunicacions

Scytl, companyia de modernització electoral

Scytl
Innovating Democracy

Així es presenta Scytl a la seva plana web:

"Scytl es líder mundial en soluciones seguras de voto electrónico, gestión y modernización electoral. Nuestras soluciones emplean protocolos criptográficos únicos que garantizan la máxima seguridad, transparencia y auditabilidad en cualquier tipo de elecciones. La innovadora tecnología de seguridad electoral de Scytl está protegida por patentes internacionales y permite a las organizaciones realizar electrónicamente cualquier tipo de proceso electoral de manera totalmente segura y auditable, posicionando a Scytl como líder internacional en el sector."

Agraïments

Aquest treball no hauria estat possible sense Dataprix i el meu germà, que van fer possible la meva introducció al món del Data Science i R.

Un reconeixement particular mereixen els meus tutors: la Marga i en Pol. Amb el seu guiatge tranquil i comprensiu, el bon seguiment realitzat i els seus coneixements m'han ajudat a seguir les fites per a assolir el cim.

I per sobre de tots, l'agraïment a la paciència infinita i recolzament incondicional dels meus pares.

Resum del Projecte

Aquest projecte explora les dades electorals dels comicis estatals celebrats a Espanya entre els anys 2000 i 2011. Amb elles s'elabora i testeja un algorisme de predicció de resultats electorals amb un error de predicció petit.

La predicció a desenvolupar s'emmarca en l'anomenat *Electoral Night Forecasting* (ENF), prediccions durant la nit electoral. En concret, s'implementen durant l'escrutini, iniciant-se amb el recompte de vots de les primeres meses electorals i finalitzant amb la publicació de resultats definitius. Els còmputs a realitzar, a diferència d'altres més clàssics i coneguts, no es basen directament en dades sociodemogràfiques. Es combinen tècniques de *Machine Learning*, com *clustering* i PCA; amb dades històriques d'eleccions passades i les que s'obtenen del recompte de vots durant les eleccions actuals.

Hi ha dues fases ben diferenciades per a desenvolupar la predicció: el *clustering* de les dades històriques, que serveix com a entrenament del sistema, i la mateixa predicció. A la fase de *clustering* es fan grups de meses electorals, assignant al mateix conjunt les que tenen resultats similars (aquesta agrupació dependrà en gran mesura de la funció que es fa servir per a calcular la semblança entre resultats).

Tal com van arribant les dades durant l'escrutini, es van fent prediccions del resultat final. Aquestes assumeixen que els vots de les taules que encara no han finalitzat el recompte són similars a les ja escrutades del seu mateix clúster. D'aquesta manera es refina la predicció del resultat final i es redueix considerablement l'error provocat per l'arribada de meses amb patró de vot similar als primers estadis de l'escrutini.

El model de predicció escollit per a realitzar aquesta aproximació es va testejar a les eleccions sud-africanes del 2004 amb bons resultats.

Amb anterioritat a la fase de predicció s'estudien les dades, per a trobar el tipus de *clustering* i el nombre de clústers òptim a emprar. Els resultats obtinguts no són prou indicatius per a triar una bona configuració per al model descrit. Es passa llavors a cercar els millors paràmetres i nombre de clústers tot comparant-ne la reducció dels seus respectius errors de predicció directament.

Per a qualsevol configuració utilitzada, la predicció millora els resultats de l'escrutini pur, mostrant la seva utilitat. S'estudia el comportament de la predicció per a diferents valors de l'escrutini. Amb valors baixos, es troben millors resultats si les agrupacions tenen un menor nombre de clústers. Per a valors superiors d'escrutini, és millor fer servir un nombre més gran de clústers.

Es proposa també un càlcul alternatiu de l'algorisme fent servir PCA (*Principal Component Analysis*) per a alleugerir el volum de dades implicat en el càlcul de clústers i així obtenir temps d'execució més reduïts, comparant si afecta al resultat final i als paràmetres òptims.

Amb el processat de les dades d'entrenament amb PCA, el comportament del sistema millora notablement per la majoria de casos estudiats. Amb PCA obtenim també resultats òptims (o quasi òptims en el pitjor dels casos) amb un nombre de clústers grans, independentment del percentatge d'escrutini computat.

Resumen del Proyecto

Este proyecto explora los datos electorales de los comicios estatales celebrados en España entre los años 2000 y 2011. Con ellos se elabora y testea un algoritmo de predicción de resultados electorales con un error de predicción pequeño.

La predicción a desarrollar se enmarca en el llamado Electoral Night Forecasting (ENF), predicciones durante la noche electoral. En concreto, se implementan durante el escrutinio, iniciándose con el recuento de votos de las primeras mesas electorales y finalizando con la publicación de resultados definitivos. Los cómputos a realizar, a diferencia de otros más clásicos y conocidos, no se basan directamente en datos sociodemográficos. Se combinan técnicas de *Machine Learning*, como *clustering* y *PCA*; con datos históricos de elecciones pasadas y las que se obtienen del recuento de votos durante las elecciones actuales.

Hay dos fases bien diferenciadas para desarrollar la predicción: el *clustering* de los datos históricos, que sirve como entrenamiento del sistema, y la misma predicción. En la fase de *clustering* se hacen grupos de mesas electorales, asignando al mismo conjunto las que tienen resultados similares (esta agrupación dependerá en gran medida de la función que se utiliza para calcular la semejanza entre resultados).

Tal y como van llegando los datos durante el escrutinio, se van haciendo predicciones del resultado final. Éstas asumen que los votos de las mesas que aún no han finalizado el recuento son similares a las ya escrutadas de su mismo clúster. De esta manera se refina la predicción del resultado final y se reduce considerablemente el error provocado por la llegada de mesas con patrón de voto similar en los primeros estadios del escrutinio.

El modelo de predicción escogido para realizar esta aproximación se implantó en las elecciones sudafricanas de 2004 con buenos resultados.

Con anterioridad a la fase de predicción se estudian los datos, para encontrar el tipo de *clustering* y el número de clústeres óptimo a emplear. Los resultados obtenidos no son suficientemente indicativos para elegir una buena configuración para el modelo descrito. Se pasa entonces a buscar los mejores parámetros y número de clústeres comparando directamente la reducción de sus respectivos errores de predicción.

Para cualquier configuración utilizada, la predicción mejora los resultados del escrutinio puro, mostrando su utilidad. Se estudia el comportamiento de la predicción para diferentes valores del escrutinio. Con valores bajos, se encuentran mejores resultados si las agrupaciones tienen un menor número de clústeres. Para valores superiores de escrutinio, es mejor usar un mayor número de clústeres.

Se propone también un cálculo alternativo del algoritmo utilizando *PCA* (*Principal Component Analysis*) para aligerar el volumen de datos implicado en el cálculo de clústeres y así obtener tiempos de ejecución más reducidos, comparando si afecta al resultado final y a los parámetros óptimos.

Con el procesado de los datos de entrenamiento con PCA, el comportamiento del sistema mejora notablemente para la mayoría de casos estudiados. Con PCA obtenemos también resultados óptimos (o casi óptimos en el peor de los casos) con un número de clústeres grandes, independientemente del porcentaje de escrutinio computado.

Abstract

The current project explores electoral data from Spanish national elections from 2000 to 2011. Using these, it is developed and tested an algorithm to predict election results with a small error.

The analysed prediction is an Electoral Night Forecasting procedure, performed during the election night. Specifically, the implementation starts with the counting of the first votes and finishes with the publication of the final results. Differently from other widespread calculations, the basis of the study is not sociodemographic data. Machine Learning techniques, like clustering or PCA, are used in conjunction with electoral data from past elections and data from the current incoming votes.

There are two distinguishable phases in the prediction: clustering of historical data to train the system and the prediction itself. Within the clustering phase, polling districts are grouped according to the similarity between their voting behaviour (clustering implementation depends strongly on the function chosen to calculate this similarity).

Predictions of the final results are given as voting data is arriving. It is assumed that not computed polling districts have similar results to those already computed of the same cluster. With the previous assumption, predictions are improved. In addition, the error due to the bias of the early received results is reduced.

The prediction model chosen to develop this approach was tested with great outcomes in the 2004 South African elections.

A data study is conducted prior to the prediction, in order to find the optimal clustering type and number of clusters. Results obtained are not indicative enough to choose a good setting of the model. Consequently, the quest of the best parameters is done by the very same computation of predictions.

For any tested configuration, the prediction improves the results thrown by pure counting of incoming votes. Forecasted results are studied according to the percentage of votes arrived. The study reveals that fewer clusters are a better option for lower votes arrived. On the contrary, a larger number of clusters is more accurate for a large number of incoming votes.

An alternative implementation of the algorithm using PCA (*Principal Component Analysis*) is proposed. By using it, it is expected to lessen data volume and consequently, computational cost and execution time.

By using PCA with training data, most of the forecasting results are notably enhanced. With this technique, a great behaviour is guaranteed with a large number of clusters, independently from the percentage of votes arrived.

1 Introducció

Les prediccions electorals han demostrat ser de gran importància per a les organitzacions polítiques i poden marcar les seves estratègies de comunicació i d'acció de manera important.

Com a exemple més conegut d'aquestes, a Espanya destaquen les enquestes d'intenció de vot del CIS (*Centro de Investigaciones Sociológicas*). Aquestes corresponen a les prediccions que es poden implementar abans de votar, i que majoritàriament responen a dades sociodemogràfiques. No obstant, també hi ha tècniques que s'apliquen durant la jornada electoral, com les enquestes a peu d'urna que es realitzen als votants a l'escola electoral després de dipositar el seu vot. El focus d'estudi d'aquest treball són les prediccions durant la nit electoral, concretament un cop ha començat el recompte de paperetes a les diferents meses.

El que es vol aconseguir amb els plantejaments presentats és una estimació precisa dels resultats finals del procés electoral sense necessitat de tenir un volum de vots comptabilitzats gaire alt.

En països en vies de desenvolupament, per exemple, si el recompte de vots es dilata en el temps, allargant-se diversos dies o setmanes en alguns casos, el valor afegit de la proposta és ben evident. Si a més, alguns d'aquests estats viuen conflictes o tenen situacions polítiques convulses, es poden evitar situacions de risc si es pot anticipar el resultat final amb un escrutini poc significatiu.

En el cas de democràcies amb més recursos, malgrat la seva limitada aplicació en el temps, encara es poden destacar nombrosos avantatges i problemes resolts. Per començar, l'obtenció d'informació fiable i anticipada que ajudi als partits a elaborar polítiques de comunicació que no els deixin en evidència amb girs inesperats. Es pot també anticipar l'elaboració d'estratègies polítiques, avançant-les a la mateixa nit electoral. En entorns amb alta volatilitat, com la borsa, anticipar-se en el temps pot ser clau. Mirant els successos recents, es pot parlar del Brexit, que moltes enquestes deien que no es produiria. Un inversor amb actius a Regne Unit hagués pogut vendre les seves accions abans de trobar-se situacions de bloqueig o que el pànic s'estengués entre els altres inversors devaluant la seva inversió.

1.1 Context del projecte

El projecte fa servir les dades públiques de les eleccions estatals espanyoles entre els anys 2000 i 2011, ja que són les eleccions més recents de les que s'han publicat fitxers amb totes les dades per mesa electoral. Aquestes s'han extret de l'apartat de descàrregues de la plana web de consulta de dades electorals del *Ministerio del Interior* del *Gobierno de España* [1].

Espanya s'organitza en 52 circumscripcions electorals, cadascuna corresponent a una província diferent. A cada circumscripció hi pot haver partits diferents, pel que s'ha de considerar cada una d'elles com un escenari independent. Per aquest motiu es treballa amb les dades de les meses electorals a cada província objecte d'estudi. En aquest projecte s'han estudiat amb major profunditat les dades de Girona i Barcelona, amb 800 i prop de 6000 meses electorals cadascuna, respectivament. S'han seleccionat aquestes per a representar regions de mida mitjana i gran.

1.2 Objectius

L'objectiu final és obtenir una predicció que millori substancialment els resultats obtinguts pel recompte de vots i permeti obtenir uns resultats acurats amb un percentatge de vots escrutats molt baix.

Els resultats depenen de com s'entrena el sistema amb les dades d'eleccions anteriors per fer la predicció. Per tant, s'ha d'avaluar i quantificar la tria adequada del tipus de *clustering* i dels seus paràmetres associats:

- Tipus d'agrupació de clústers: *fuzzy c-means* o *k-means*
- Nombre de clústers que minimitzi l'error
- Paràmetre m per *fuzzy c-means*

Es vol mesurar la precisió d'aquestes eleccions sense haver de realitzar la predicció, amb el còmput i interpretació d'uns índexs d'avaluació, diferents per *k-means* i *fuzzy c-means*. Amb aquests resultats es redueix el nombre de possibilitats i paràmetres a calcular.

S'obté una mesura de l'error comès per comprovar empíricament quina és la millor manera de predir els resultats, amb les estimacions dels millors paràmetres obtinguts a priori.

Comparar la configuració donada per l'estimació a priori amb els índexs d'avaluació i la òptima obtinguda amb la predicció és un altre objectiu.

Atesa la complexitat dels càlculs a implementar, es vol mesurar també la bondat d'una versió modificada del model original, reduint l'espai de dades utilitzat per a l'entrenament del model de predicció amb l'ús de PCA (*Principal Component Analysis*).

1.3 Estructura de la memòria

El projecte s'inicia explicant la part teòrica essencial per a entendre què s'hi fa. Aquesta introducció teòrica correspon al capítol 2. Dins d'aquest, es presenta primer *l'Electoral Night Forecasting* i alguns exemples d'implementació. Continua el capítol amb més bases conceptuals, centrant-se en definir el *clustering*: tipus diferents a emprar i els índexs proposats per a avaluar la seva correcta aplicació. El bloc finalitza analitzant PCA, necessari per a reduir el volum de dades amb les que es treballa, minimitzant la pèrdua d'informació essencial en el procés.

El capítol 3 defineix el model de predicció proposat, explicant l'algorisme utilitzat i particularitzant-lo al cas d'estudi.

El bloc posterior, corresponent al capítol 4, presenta els resultats obtinguts, que es divideixen en dues parts ben diferenciades: avaluació dels índexs per a validar el nombre de clústers i resultats de la predicció.

La memòria finalitza al capítol 5, detallant les conclusions que es desprenen de l'estudi global en base als resultats obtinguts al capítol 4.

2 Estat de l'art

S'enumeren a continuació les eines teòriques necessàries per entendre el treball realitzat, dividides en quatre blocs: *ENF*, *Clustering,* índexs de validació i *PCA*. *ENF* es refereix només al procés electoral: com comptabilitzar els vots o com fer una previsió de resultats. Els altres tres blocs detallen les tècniques de *Machine Learning* que s'aplicaran a *ENF*.

2.1 ENF

Durant la jornada electoral es realitzen diversos processos per anar computant i estimant el resultat final de les eleccions. El més conegut és el recompte pur, on es van donant els resultats en percentatge de vot i escons a mida que es van comprovant, com si els vots comptabilitzats fins al moment fossin els totals. Els mecanismes més comuns de predicció del resultat final són:

- **Enquestes a peu d'urna**: Un grup d'enquestadors pregunta als electors, en sortir del col·legi electoral, quin ha estat el seu vot. S'aconsegueix així una predicció dels resultats abans de començar l'escrutini i també es fa servir per a obtenir dades sociodemogràfiques.
  Considerant les enquestes realitzades a les eleccions del 26 de juny del 2016 a Espanya, aquestes proporcionen una gran quantitat de dades (les mostres obtingudes van ser més de 100.000), però la seva qualitat és dubtosa i el cost molt elevat (les prediccions es van desviar fins a 20 escons per a algun partit i el cost fou superior a 300.000 €, també amb dades de juliol del 2016).

- **Recompte ràpid**: Consisteix en crear una xarxa de voluntaris, desvinculats de cap organització política i repartits en diferents col·legis electorals, que duran a terme un mostreig aleatori d'actes electorals [2].
  Aquests resultats, obtinguts abans que l'escrutini final, s'envien a una unitat central que s'encarrega de computar l'estimació del resultat final en escons i els intervals de confiança del resultat. S'organitzen com a mecanisme de validació del resultat provisional dels comicis i les duen a terme organitzacions civils que vetllen per la transparència democràtica.
  La fiabilitat dels resultats depèn de l'aleatorietat en l'elecció de meses i del nombre de mostres, però és considerablement més gran que la de les enquestes a peu d'urna. Elimina el biaix produït als primers estadis del recompte directe, on arriben primer les dades de meses petites amb un patró de vot similar.

- ***Electoral Night Forecasting* (ENF).** És un terme que serveix per anomenar les metodologies emprades per a predir resultats durant el recompte de vots, fent servir totes les dades que es van obtenint durant el propi escrutini.
  Es divideix en dues fases: entrenament i predicció.
  Per a entrenar l'algorisme de predicció es fan servir fonts diverses: dades d'eleccions anteriors, enquestes d'intenció de vot, a peu d'urna, o recompte ràpid, entre altres.
  La predicció es va actualitzant a mida que arriben dades de meses en les que el recompte ha finalitzat. Amb les dades auxiliars utilitzades per a l'entrenament s'intenta evitar el biaix que es produeix amb el recompte directe i el ràpid.
  A continuació es referencien diversos articles i estudis que implementen prediccions amb procediments variats:
  A València es va implementar un model Bayesià [3] per a predir el canvi de distribució de vots durant la nit electoral. Aquest estudi es recolzava en la identificació prèvia de patrons de vot segons àrea geogràfica.
  A les eleccions britàniques [4] també s'han desenvolupat mètodes per a predir el canvis en els patrons de vot durant el recompte dels mateixos.
  En una altra línia diferent s'alinea l'estudi amb dades d'una regió austríaca [5] a on les dades s'agrupen fent servir algorismes genètics per a optimitzar la predicció, en lloc d'estudis sociodemogràfics.
  L'estudi presentat aquí es basa en un model utilitzat per predir els resultats finals de les eleccions nacionals sud-africanes el 2004 [6]. En aquest, es fan servir les dades d'eleccions anteriors per a entrenar l'algorisme i agrupar les meses electorals segons similituds en el comportament de vot. Per predir els resultats de les meses que encara no tenen dades oficials, s'usen les dades del grup al que pertany. D'aquesta manera no es fa ús de cap dada sociodemogràfica. Un altre avantatge d'aquest algorisme respecte als basats en dades demogràfiques és que pot tenir un nombre de partits diferent per a les dades d'entrenament i els comicis a predir. Això és especialment interessant en l'escenari espanyol, en què s'ha passat d'un bipartidisme clar a un escenari multicolor amb partits molt rellevants que fa pocs anys no existien.

2.2 Clustering

El *clustering* és una de les tècniques més esteses de *Machine Learning*, concretament de l'aprenentatge sense supervisió. L'aprenentatge sense supervisió analitza un conjunt de característiques $X_1,X_2,...,X_p$ mesurades en n observacions. Aquestes tècniques permeten descobrir dades d'interès sobre un conjunt d'observacions [7].

El *clustering* és un tipus de mètode d'aprenentatge no supervisat que permet localitzar i agrupar dades en subgrups desconeguts fins al moment. A partir d'aquí es consideren dos algorismes de *clustering* diferents: *k-means*, el més clàssic i conegut, i *fuzzy c-means.*

2.2.1 K-means

K-means és un *clustering* que agrupa les observacions en funció de les característiques o variables, agrupant-les en un nombre determinat de grups o clústers sense solapament: cada observació ha de pertànyer exclusivament a un dels grups definits.

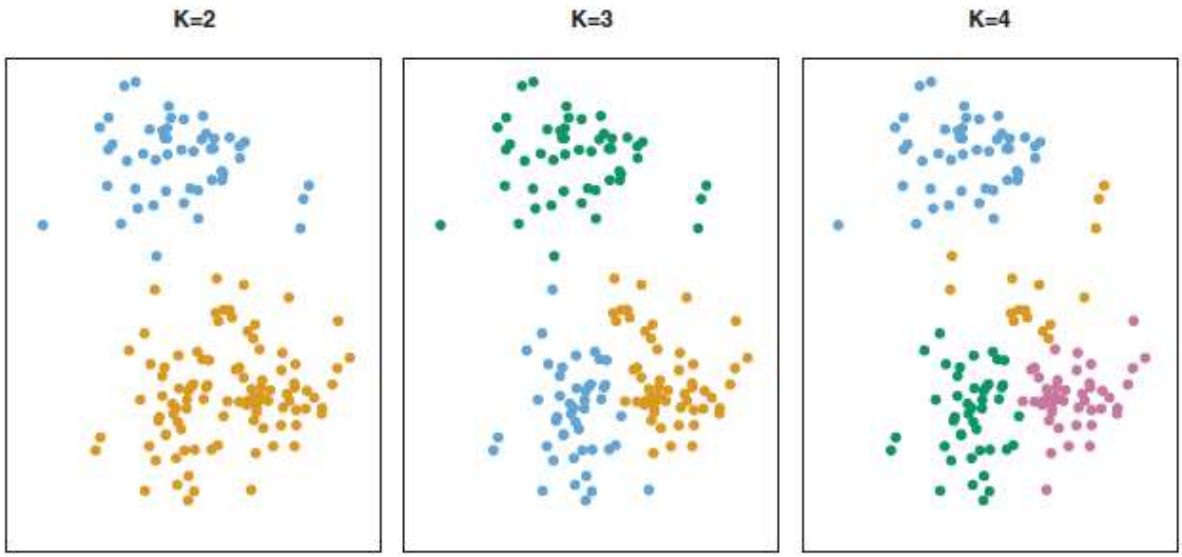

Figura 2.1: Conjunt de 150 observacions agrupat en 2, 3 i 4 clústers amb k-means.

A la figura 2.1 [7] hi ha 3 gràfiques amb clústers diferents per les mateixes 150 observacions. Les observacions que pertanyen al mateix grup s'han pintat amb un mateix color.

Per a assignar els elements de manera útil, la variació entre els pertanyents al mateix grup, o variació intra-cluster, ha de ser la mínima possible. S'ha d'establir una funció que mesuri les diferències entre aquests elements. Sigui C_k el conjunt d'observacions que pertanyen al clúster k. W(C_k) és la funció que mesura les diferències entre elements i K correspon al número de clústers del *k-means* implementat. Es minimitza l'equació (2.1).

$$\min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} \mathrm{W}(C_k) \right\} \tag{2.1}$$

Una implementació típica és utilitzar la distància euclídea entre els elements per a mesurar-ne la distància:

$$\min_{C_1,\dots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_K|} \sum_{i,i' \epsilon\, C_K} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\} \tag{2.2}$$

On $|C_K|$ és el nombre d'observacions del clúster k, p és el nombre de variables, x_{ij} i $x_{i'j}$ són els valors de les variables j per a diferents clústers.

Per garantir que es minimitza aquesta distància de manera eficient, *k-means* segueix el següent algorisme:

1. Assignació aleatòria de cadascuna de les observacions a un dels K clústers. Correspon a l'assignació inicial de clústers
2. Repetir fins que l'assignació de clústers deixi de variar:
    a. Calcular el centroide per a cada clúster. El centroide correspon a la mitjana de totes les observacions pertanyents al clúster en qüestió
    b. Assignar cada element al clúster que tingui el centroide més proper a ell. S'avalua amb la distància euclídea, equació (2.2)

Figura 2.2: Algorisme de càlcul del k-means.

Seguint les passes detallades a la figura 2.2 es redueix la distància a cada iteració, millorant l'agrupació fins que s'assoleix un punt òptim, on s'atura.

Cada assignació de valors inicials al pas 1 de la figura 2.2 garanteix que els resultats convergeixin, però no necessàriament dóna el millor *clustering* possible. Se selecciona un nombre determinat de punts inicials i s'hi implementa l'algorisme descrit a la figura 2.2 fins que l'assignació de clústers deixi de variar, seleccionant el valor mínim d'entre tots els calculats i els seus clústers associats.

2.2.2 Fuzzy C-means

Fuzzy c-means segueix uns principis molt semblants a *k-means.* Treballa també amb un nombre C definit de clústers i l'objectiu final és minimitzar una funció que mesura les diferències entre elements. Aquesta diferència es mesura, en molts casos, com la distància euclídea entre els elements. Les diferències conceptuals radiquen en què les observacions de l'espai de dades poden pertànyer a més d'un clúster alhora [8].

En concret el que fa aquest *clustering* és crear una matriu de pertinença (*membership*) de dimensions C x N, on N és el nombre d'observacions i C el nombre de clústers. Per cada columna d'aquesta matriu (una per observació) el *membership* dóna un coeficient de pertinença a cada clúster (una fila per clúster). Es denoten els memberships com u_{ik}, amb i indicant el clúster i k l'observació. Així, cada coeficient ha de complir les següents condicions:

$$\sum_{k=1}^{N} u_{ik} > 0 \quad per\ cada\ cluster\ i \tag{2.3}$$

I a més compleix que:

$$\sum_{i=1}^{C} u_{ik} = 1 \quad per\ a\ cada\ observació\ k \tag{2.4}$$

Com al *k-means*, els elements agrupats han de minimitzar una funció objectiu que quantifica la proximitat entre ells. Com que en aquest cas no és trivial computar una distància euclídea, es fa servir l'equació funcional d'errors quadràtics mitjans generalitzada:

$$J_m(U, v) = \sum_{k=1}^{N} \sum_{i=1}^{C} (u_{ik})^m \, (d_{ik})^2 \tag{2.5}$$

A on m (que ha de ser > 1) és l'exponent de pes de l'algorisme o "fuzzyficador" i d_{ik} són les distàncies entre cada element i el centre del clúster o centroide.

El valor de m ajuda a fer una interpretació ràpida de la distribució de *memberships* utilitzada.

Valor de m	Interpretació
Proper a 1	Maximitza la dispersió dels *memberships*, assignant nombres molt propers a 0 o 1. El comportament és anàleg al *k-means*
Infinit	Els *memberships* tendeixen a ser iguals entre sí, donant lloc a clústers idèntics

Taula 2.1: Interpretació del valor de m.

Es calculen els *memberships* amb l'equació (2.6)

$$u_{ik} = \frac{1/d_{ik}^{2/(m-1)}}{\sum_{i'=1}^{C} 1/d_{i'k}^{2/(m-1)}} \qquad (2.6)$$

I per a calcular els centroides v_i de cada clúster:

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} (u_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{N} (u_{ik})^m} \qquad (2.7)$$

Els elements $\mathbf{x}_k$ són els valors de les variables per a cada observació k.

Es fa servir un procediment de càlcul dels memberships i clústers molt semblant al del k-means.

1. Assignació d'un valor superior a 1 per al paràmetre m. Assignació aleatòria inicial de valors dels *memberships* u_{ik} que compleixen els requisits de les equacions (2.3) i (2.4).
2. Computar els centroides $\mathbf{v}_i$ de cada clúster fent servir l'equació (2.7)
3. Repetir fins que l'assignació de *memberships* deixi de variar:
    a. Calcular els nous *memberships* amb l'equació (2.6)
    b. Trobar els centroides $\mathbf{v}_i$ corresponents als *memberships* del pas anterior, de nou fent servir (2.7). Després comparar els nous *memberships* amb la matriu de l'iteració anterior per a seguir iterant en cas de variacions entre ells.

Figura 2.3: Algorisme de càlcul del fuzzy c-means.

Al capítol 3 es parla en profunditat de la particularització de les característiques del *fuzzy c-means* pel model estudiat.

Un cop estudiat el funcionament dels algorismes *k-means* i *fuzzy c-means*, s'ha de triar el nombre de clústers C o K més adient. Es pot fer per "força bruta", realitzant les prediccions per les que es dissenya el *clustering* per diferents valors de K, C i m i comparant-ne l'error de predicció.

Una altra manera de fer-ho, sense necessitat d'aplicar el *clustering* a cap sistema i estalviant càlculs, són els índexs de validació.

2.3 Índexs de validació

Els índexs de validació serveixen per a avaluar la bondat de les particions fetes per l'algorisme escollit. No s'ha d'oblidar que les tècniques de *clustering* són no supervisades i requereixen uns mecanismes de validació. Al context estudiat es poden entendre també com una mesura del número de clústers òptim a implementar per a que la predicció sigui el més acurada possible. A continuació es presenten breument els diferents índexs utilitzats.

2.3.1 Índexs per k-means

S'escullen dues tècniques diferents per a avaluar aquest tipus d'agrupació: la matriu de *Scatter* [9] i l'índex *Silhouette* [10].

- **Matriu de *Scatter***. S'obté la matriu de *Scatter*, que permet, estudiant la seva traça, maximitzar la distància entre clústers tot minimitzant la dispersió interna del clúster.
  Es fa servir notació vectorial per a explicar l'ús d'aquesta matriu. Considerant un conjunt de dades D format per n observacions o vectors de tipus $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Es divideix l'espai D en c clústers $D_1, D_2, \ldots, D_c$.
  El centroide o mitjana de cada clúster i es defineix:

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D} \mathbf{x} \qquad (2.8)$$

  I el vector mitjana total correspon a:

$$\mathbf{m} = \frac{1}{n} \sum_{x \in D} \mathbf{x} = \frac{1}{n} \sum_{i=i}^{c} n_i \, \mathbf{m}_i \qquad (2.9)$$

La matriu de *Scatter* pel clúster i, llavors, serà

$$\mathbf{S}_i = \sum_{x \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m})^T \quad (2.10)$$

I les matrius de *Scatter* intern del clúster (2.11), inter-clúster (2.12) i total (2.13) són, respectivament

$$\mathbf{S}_W = \sum_{i=i}^{c} \mathbf{S}_i \quad (2.11)$$

$$\mathbf{S}_B = \sum_{i=i}^{c} n_i \, (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \quad (2.12)$$

$$\mathbf{S}_T = \frac{1}{n} \sum_{i=i}^{c} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T = \mathbf{S}_W + \mathbf{S}_B \quad (2.13)$$

La matriu total de *Scatter* es compon de la suma de les equacions (2.11) i (2.12). Les expressions a optimitzar per a validar el bon funcionament dels clústers són:

$$\min Tr\,(\mathbf{S}_T^{-1}\mathbf{S}_W) \quad (2.14)$$

$$\max Tr\,(\mathbf{S}_W^{-1}\mathbf{S}_B) \quad (2.15)$$

Aquesta alternativa, que fa servir matrius, és molt més lleugera computacionalment parlant que els càlculs d'índexs tradicionals.

- **Coeficient *Silhouette***. De manera similar a l'anterior, aquesta mesura de validació també és un compromís entre càlculs de similaritat d'elements dins del mateix clúster i diferència amb elements de clústers externs.
  Per aquest cas cal obtenir dues mesures diferents per observació: a(i) i b(i), on i és l'observació en qüestió.

  a(i) correspon a la distància mitja de l'observació i amb els altres elements del seu clúster.

  b(i) correspon a la distància mitja de l'observació i amb tots els elements que no pertanyen al seu clúster.

Tenint aquestes mesures s'obté el coeficient *Silhouette* de la següent manera:

$$s_i = \frac{(b(i) - a(i))}{\max(a(i), b(i))} \tag{2.16}$$

Els valors dels coeficients varien entre -1 i 1, on els valors més alts del mateix indiquen el *clustering* més apropiat. El nombre de clústers òptim es dona pel valor K que maximitza l'expressió:

$$\max_K \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{2.17}$$

N correspon novament al nombre total d'observacions.

2.3.2 Índexs per fuzzy c-means

Existeix força documentació sobre els índexs de validació més adequats per a aquest tipus de *clustering*, però s'han seleccionat 3 dels més estesos per al seu còmput [11]: Fukuyama-Sugeno, Xie-Beni i partition coeficient normalitzat o nombre de Dunn normalitzat.

S'ha de trobar un valor òptim c de clústers i la m que permet obtenir la millor agrupació possible. A la taula-resum 2.2 es mostren les definicions dels índexs escollits.

Índex	*Equació*	*Optimització*
Coeficient partició normalitzat	$i_{CPN} = \frac{C\left(\frac{1}{n}\left(\sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^2\right)\right) - 1}{C - 1}$ (2.18)	*Maximitzar*
Fukuyama-Sugeno	$i_{FS} = \sum_{i=1}^{C}\sum_{k=1}^{n} u_{ik}^m(\|\mathbf{x}_k - \mathbf{v}_i\|^2 - \|\mathbf{v}_i - \mathbf{v}\|^2)$ (2.19)	*Minimitzar*
Xie-Beni	$i_{XB} = \frac{\sum_{i=1}^{C}\sum_{k=1}^{n} u_{ik}^m \|\mathbf{x}_k - \mathbf{v}_i\|^2}{n\left(\min\{\mathbf{v}_i - \mathbf{v}_j\}\right)}$ (2.20)	*Minimitzar*

Taula 2.2: Definició dels índexs de validació de fuzzy c-means.

Per a les definicions anteriors, n és el nombre d'observacions, C el nombre de clústers, u_{ik} els coeficients de *membership*, $\mathbf{x}_k$ és l'observació k, $\mathbf{v}_i$ i $\mathbf{v}_j$els centres dels clústers i, j respectivament i $\mathbf{v}$ la mitja de totes les dades de l'espai.

Els tres índexs seleccionats seran interessants perquè cap d'ells té un comportament monòton amb el nombre de clústers i ha de facilitar trobar el nombre de clústers que optimitzi l'índex corresponent. Això és així perquè aquests combinen en el seu còmput els memberships, les dades del conjunt i el nombre de clústers.

2.4 PCA

PCA (*Principal Component Analysis*) permet, en un entorn amb un ampli ventall de variables correlades, trobar un conjunt més petit de noves variables amb les què explicar la major part de la variabilitat de les dades globals [7].

L'objectiu és un nombre de dimensions més reduït, on cadascuna d'aquestes tingui una variabilitat el més gran possible. Les noves dimensions són una combinació lineal normalitzada de les variables originals.

El primer component principal per a un conjunt p de variables del tipus $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$ és un combinació lineal del tipus

$$\mathbf{Z}_1 = \emptyset_{11}\mathbf{X}_1 + \emptyset_{21}\mathbf{X}_2 + \cdots + \emptyset_{p1}\mathbf{X}_p \qquad (2.21)$$

On la combinació lineal tingui variança màxima.

S'anomenen els elements $\emptyset_{11}, \emptyset_{21}, \dots, \emptyset_{p1}$ pesos (*loadings*) del primer component principal. Que l'expressió està normalitzada vol dir que els seus pesos compleixen el següent requisit:

$$\sum_{j=1}^{p} \emptyset_{j1}^2 = 1 \qquad (2.22)$$

Es considera un espai de p variables amb n observacions diferents, conformant un espai de dimensions n x p. Per a eliminar errors no desitjats en el còmput i centrar l'estudi en la variança, les variables han d'estar centrades, és a dir, tenir mitja zero. Les combinacions lineals de les noves variables centrades es poden expressar amb la forma següent

$$z_{i1} = \emptyset_{11}x_{i1} + \emptyset_{21}x_{i2} + \cdots + \emptyset_{p1}x_{ip} \qquad (2.23)$$

Un cop fet aquest procés, es cerquen les combinacions lineals de les variables centrades i amb els pesos normalitzats que maximitzin la variança per a trobar el primer component principal

$$\max_{\emptyset_{11},\ldots,\emptyset_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \emptyset_{j1} x_{ij} \right)^2 \right\} \qquad (2.24)$$

Sempre subjectes a la condició de normalització de l'equació (2.22). El conjunt de pesos (loadings) del primer component principal defineixen la direcció en l'espai de variables de màxima variabilitat.

El segon component principal és el vector incorrelat amb el primer component que té variança màxima. Es computen així successivament fins a completar el nombre total de variables.

S'obté així l'espai de dimensions més reduïdes possibles que més s'acosta a les n observacions. O interpretat d'una altra manera, s'aproxima el nou espai obtingut amb PCA de manera òptima a l'espai original de dades amb un nombre mínim de dimensions.

S'ha de considerar si les variables a estudiar tenen unitats similars entre elles o no, per decidir si és necessari escalar-les per a obtenir desviació estàndard 1 o no. En el cas que les variables estiguin expressades amb les mateixes unitats no cal escalar perquè les dimensions existents ja són comparables. A l'estudi actual es treballa per definició amb percentatge de vots, fent innecessari aquest pas.

Per veure com de representatiu és l'espai reduït, s'utilitza el terme de proporció de variança associada. Conceptualment, el que mostra, és la variança total de les dades que hi ha dintre del component estudiat. Es mostra l'expressió que serveix per a calcular el percentatge de variança associada al component principal m-èsim en un espai de p variables i n observacions

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \emptyset_{jm} x_{ij} \right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^{\ 2}} \qquad (2.25)$$

Per veure quina proporció de dades hi ha al nostre nou espai respecte del total, se sumen els percentatges de variança associada de tots els components principals que decidim incloure-hi. No hi ha cap estàndard o norma per a decidir el nombre òptim de components a fer servir.

Amb PCA s'aconsegueix reduir el cost computacional dels programes i eliminar soroll del sistema.

3 Model de predicció

Conegudes les eines matemàtiques i tècniques de Machine Learning necessàries per l'estudi, és hora de conèixer els elements concrets que constitueixen el model de predicció, extrets de la referència [6]. Per a agrupar les dades d'entrenament s'utilitza un algorisme de tipus *fuzzy c-means.*

Com que també es vol testejar com funciona amb un *clustering* de tipus *k-means*, però l'algorisme no està adaptat per a aquests, es provarà el mateix amb un valor del paràmetre m molt proper a 1. Aquesta configuració resulta en un *clustering* gairebé idèntic al *k-means*. És per això que en la part de resultats, quan parlem de "k-means" ens referim sempre a aquest *fuzzy c-means* amb un valor de m suficientment proper a 1 com per a obtenir un comportament molt similar al d'aquest *clustering*.

Amb l'organització electoral a Espanya, s'ha de considerar cada circumscripció o província com un problema de predicció diferent, ja que cada una d'elles té un nombre de meses electorals i partits diferents.

Com ja s'ha comentat en altres apartats, el càlcul del model de predicció es divideix en la fase d'entrenament (*clustering* amb dades d'eleccions anteriors) i la de predicció.

3.1 Entrenament del model

El format de les dades electorals respon a matrius de N files i P columnes, on N és el nombre de meses electorals de la província estudiada i P el nombre de partits a l'exercici electoral amb què entrenem. Per a cada mesa, es treballa amb el percentatge de vots obtinguts per cada partit sobre el total de la mesa. Així, els coeficients de la matriu de cada fila han de sumar el 100 % dels vots, complint la següent condició

$$\sum_{p=1}^{P} x_{np} = 100 \qquad per \quad n = 1, \dots, N \tag{3.1}$$

Sigui C_n el nombre de votants censats i el de vots vàlids V_n . Es defineix la participació A_n de cada mesa

$$A_n = {}^{V_n}/_{C_n} \quad per \quad n = 1, \dots, N \tag{3.2}$$

Es defineix la mesura de la distància entre elements del *clustering* com la distància euclídea entre resultats $x_{n_1 p}$ i $x_{n_2 p}$ de les meses n_1 i n_2

$$d_{n_1 n_2} = \sqrt{\sum_{p=1}^{P} \left(x_{n_1 p} - x_{n_2 p}\right)^2} \tag{3.3}$$

Tal com s'ha vist a l'apartat anterior, cal una funció objectiu per a minimitzar. A l'apartat 2 s'ha vist l'equació (2.5), que és una funció objectiu genèrica. Es particularitza incloent els vots vàlids V_n de cada mesa

$$J_m(u,n) = \sum_{n=1}^{N} V_n \sum_{c=1}^{C} (u_{cn})^m \, (d_{cn})^2 \quad amb \quad m > 1 \tag{3.4}$$

On d_{cn} correspon a la distància entre l'element x_{np} i el centre del clúster v_{cp}.

Els coeficients del *membership* u_{cn} han de complir la condició de l'equació (2.4), particularitzades pel sistema

$$\sum_{c=1}^{C} u_{cn} = 1 \quad per\ a\ cada\ mesa\ n = 1, \dots, N \tag{3.5}$$

El còmput dels *memberships* es fa tal com s'ha il·lustrat al capítol anterior, a l'equació (2.6), que aquí es particularitza

$$u_{cn} = \frac{1/d_{cn}^{2/(m-1)}}{\sum_{c'=1}^{C} 1/d_{c'n}^{2/(m-1)}} \ per\ a\ cada\ clúster\ c = 1, \dots, C\ i\ mesa\ n = 1, \dots, N \tag{3.6}$$

S'afegeixen els vots vàlids de cada mesa per a adaptar el càlcul de centroides realitzat amb l'equació (2.7), quedant

$$v_{cp} = \frac{\sum_{n=1}^{N} V_n (u_{cn})^m x_{np}}{\sum_{n=1}^{N} V_n (u_{cn})^m} \ per\ a\ cada\ clúster\ c = 1, \dots, C\ i\ partit\ p = 1, \dots, P \tag{3.7}$$

Amb aquests elements (*memberships* i centroides) es pot computar el *fuzzy c-means* seguint les passes detallades a l'algorisme de la figura (2.3). D'aquesta manera obtenim una agrupació de les dades d'entrenament, les de les eleccions anteriors a les d'estudi. Amb elles podrem obtenir una predicció millorada de les eleccions a estudiar, sense necessitat que el nombre de partits d'uns comicis i els següents coincideixen i evitant el biaix que suposa fer prediccions durant les etapes inicials de l'escrutini.

3.2 Predicció

El primer que es defineix a la predicció de resultats és un conjunt variable amb el temps, $\Omega(t)$, que guarda el conjunt de meses electorals de les que ja s'han rebut les dades del recompte de vots. Els resultats arribats en un moment determinat es denoten

$$y_{np} \; amb \; p = 1, \dots, P_{nou} \;\; i \; els \; districtes \; n \epsilon \Omega(t) \qquad (3.8)$$

On P_{nou} és el nombre de partits participants a les eleccions a predir, que com ja hem vist, pot ser diferent a P, el nombre de partits de les dades d'entrenament.

Es calcula també un nou conjunt de centres del clúster pel conjunt de resultats computats fins al moment

$$v_{cp}(t) = \frac{\sum_{n \epsilon \Omega(t)} V_n \, u_{cn} y_{np}}{\sum_{n \epsilon \Omega(t)} V_n \, u_{cn}} \; pels \; clústers \; c = 1, \dots, C \; i \; partit \; p = 1, \dots, P_{nou} \quad (3.9)$$

A l'equació (3.9) es multipliquen els vots vàlids amb els resultats arribats i el *membership*. El *k-means* pur no té matriu de *membership* perquè no té sentit per a aquest *clustering*. No és possible calcular aquesta equació amb el *clustering k-means*. És per això que per a aproximar-nos-hi, es fa servir una implementació amb m molt propera a 1.

S'obté la participació efectiva del clúster c

$$A_c(t) = \frac{\sum_{n \epsilon \Omega(t)} V_n \, u_{cn}}{\sum_{n \epsilon \Omega(t)} C_n \, u_{cn}} \; pels \; clústers \; c = 1, \dots, C \qquad (3.10)$$

S'han computat els valors dels nous centres de clústers i s'han guardat els valors arribats. S'han de predir els resultats de les meses que encara no han finalitzat el recompte. Per començar, la predicció del percentatge de vots per als partits d'una mesa no escrutada s'obté trobant el centroide del seu clúster

$$\hat{y}_{cp}(t) = \frac{\sum_{c=1}^{C} u_{cn} v_{cp}(t) A_c(t)}{\sum_{c=1}^{C} u_{cn} \, A_c(t)} \qquad (3.11)$$

$$per \; p = 1, \dots, P_{nou} \; i \; els \; districtes \; n \notin \Omega(t)$$

La predicció de la participació per les meses que encara no tenen dades s'expressa ponderant-la pel seu pes o *membership* a cada clúster

$$\hat{A}_c(t) = \sum_{c=1}^{C} u_{cn} \, A_c(t) \quad pels \; districtes \; n \notin \Omega(t) \qquad (3.12)$$

Amb tots aquests càlculs auxiliars, que tenen en compte les meses electorals que han arribat i fan servir els clústers de les dades d'entrenament per a predir els resultats que encara no han arribat, es prediuen els resultats per a un partit concret en un temps determinat

$$\hat{y}_p(t) = \frac{\sum_{n\epsilon\Omega(t)} V_n\, y_{np} + \sum_{n\notin\Omega(t)} C_n \hat{A}_c(t)\, \hat{y}_{cp}(t)}{\sum_{n\epsilon\Omega(t)} V_n + \sum_{n\notin\Omega(t)} C_n \hat{A}_c(t)} \qquad (3.13)$$

L'equació (3.13) combina les dades que ja s'han comptabilitzat amb els percentatges que predits per les meses sense acabar de recomptar. D'aquesta manera construeix la predicció que ha de millorar l'escrutini pur.

Els resultats d'aquesta predicció per partit satisfan la condició de que la suma de vots obtinguts per tots els partits és de 100

$$\sum_{p=1}^{P} \hat{y}_p(t) = 100 \qquad per \quad n = 1, \dots, N \qquad (3.14)$$

Per a fer la predicció, cada vegada que arriben noves dades escrutades, es calculen els nous centres i participacions de les dades arribades $v_{cp}(t)$ i $A_c(t)$. Amb aquestes, es prediuen el resultats per les meses que encara no tenen dades, amb $\hat{y}_{cp}(t)$ i $\hat{A}_c(t)$. Finalment l'equació (3.13) serveix per a calcular el resultat global final $\hat{y}_p(t)$.

4 Resultats

Per a obtenir els resultats mostrats s'ha fet servir el programari lliure R, per computació estadística i gràfica. En una primera fase, amb les dades brutes de diverses eleccions estatals al Congrés, es construeix un espai de dades que pugui llegir el software i amb ell, per a cada província a calcular, una matriu de dimensions N x P. N és el nombre de meses de la província a les eleccions computades i P el nombre de partits a les mateixes. La matriu **M** conté el percentatge de vots obtinguts v_{np} per a cada partit p a cada mesa n.

$$\mathbf{M} = \begin{bmatrix} v_{11} & \cdots & v_{1P} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NP} \end{bmatrix} \tag{4.1}$$

Seguint el model presentat al capítol 3, el pas previ a la predicció és crear els clústers més adequats per a les dades d'entrenament del sistema. L'assignació de clústers no és un procés purament determinista i cada vegada que es fa els resultats són diferents. Això duu a computar 100 realitzacions diferents de cada procés per a tots els índexs i prediccions. S'obté un resultat més fiable amb el càlcul de la mitjana de les esmentades realitzacions.

Una altra particularitat de les dades és que el vots es concentren en una porció bastant reduïda dels partits, amb partits petits que no s'haurien de comptabilitzar de la mateixa manera que els majoritaris, però tampoc s'haurien de menysprear. Una manera objectiva d'atenuar aquestes dades menys significatives és fent servir PCA. Així s'obté un nou espai de dades que concentra les dades més vitals en un número més reduït de dimensions. A la taula 4.1 es veuen les desviacions estàndards, proporcions de variança i proporcions de variança acumulada per al resultats de les eleccions de 2011 a Girona. Per a fer-les més fàcils de llegir, s'arrodoneixen a 1 decimal en el cas dels percentatges.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Desviació estàndard	14,664	5,788	4,343	2,428	1,217	0,950	0,847
Proporció de Variança	77,5%	12,1%	6,8%	2,1%	0,5%	0,3%	0,3%
Proporció acumulada	77,5%	89,5%	96,3%	98,4%	99,0%	99,3%	99,6%

	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Desviació estàndard	0,654	0,526	0,483	0,437	0,247	0,160	0,000
Proporció de Variança	0,2%	0,1%	0,1%	0,1%	0,0%	0,0%	0,0%
Proporció acumulada	99,7%	99,8%	99,9%	100,0%	100,0%	100,0%	100,0%

Taula 4.1: Desviació estàndard, proporció de variança i proporció acumulada del PCA de les dades de 2011 a Girona.

La taula 4.1 té 14 components principals. El nombre màxim de components principals ha de coincidir amb el nombre total de dimensions: els 14 partits de la província de Girona a 2011. Llegim a la taula que als primers 5 components del nou espai amb PCA, s'associa un 99% del total de la variança acumulada. Ampliant el rang a 10 components, el percentatge puja fins al 99,9%. És per això que en tots els apartats es calculen els resultats de la manera convencional i després simplificant-los amb un PCA per veure si aquesta reducció és capaç de millorar-los.

L'exposició de resultats comença a l'apartat 4.1, amb un estudi del tipus de *clustering* i nombre de clústers més adequat segons els índexs de validació presentats a l'apartat 2.3 del capítol 2. Després, a l'apartat 4.2, es mostren els resultats d'aplicar la predicció presentada al capítol 3 amb les diverses configuracions estudiades al 4.1. Finalment es comparen els resultats.

4.1 Avaluació dels índexs de validació

Per a determinar el nombre de clústers ideal per la predicció, se seleccionen els índexs de validació enumerats als apartats 2.3.1 i 2.3.2. S'han calculat 100 realitzacions per a cada cas. Inicialment es discuteix la matriu de *Scatter* i el *Silhouette*, tècniques amb les quals testejar el *k-means*, sobre les dades més recents de les què es disposa: les de les eleccions de 2011.

4.1.1 Avaluació per K-means

4.1.1.1 Matriu de Scatter

S'avalua primer la matriu de *Scatter*. Tenim dues magnituds diferents per comprovar, corresponents a les equacions (2.14) i (2.15), que transcrivim aquí

$$\min Tr(\mathbf{S}_T^{-1}\mathbf{S}_W) \tag{4.2}$$

$$\max Tr\,(\mathbf{S}_W^{-1}\mathbf{S}_B) \tag{4.3}$$

S'obtenen resultats per a les províncies de Girona i Barcelona, ja que Girona té un volum mitjà de dades i Barcelona un volum molt alt, considerant les dues bastant representatives de províncies tipus a Espanya.

Per ambdues regions, el valor òptim es dóna per a 2 clústers. Es representen gràficament els resultats de Girona a la figura 4.1. Aquesta mostra el valor a minimitzar en blau, resultant de l'equació (4.2). La corba vermella correspon a l'equació (4.3) a maximitzar. L'eix horitzontal indica el nombre de clústers pels que es calcula cada índex.

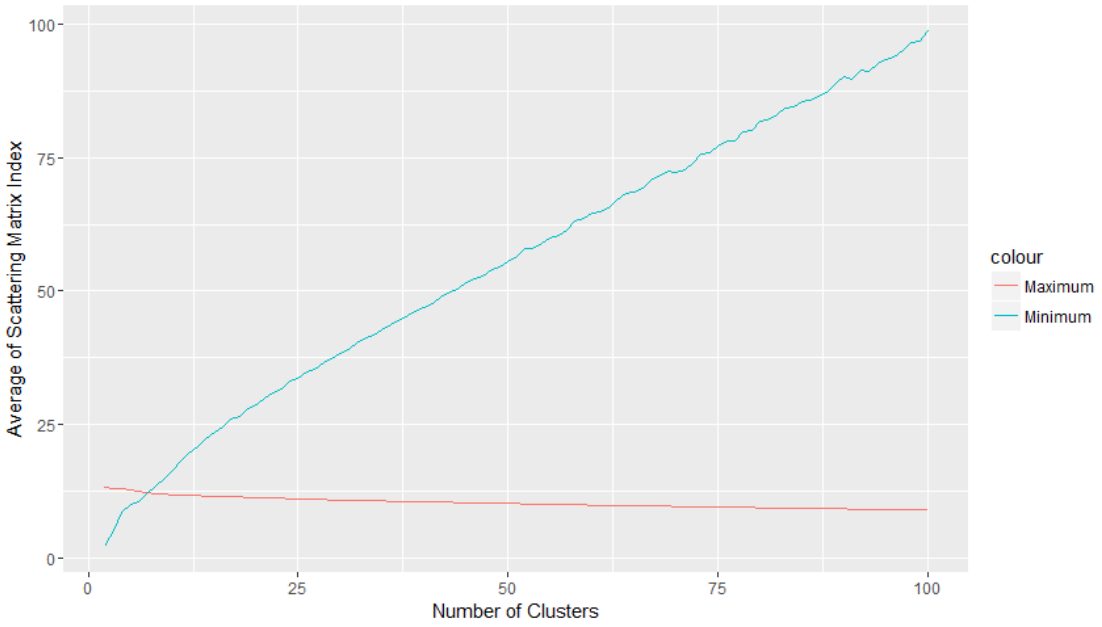

Figura 4.1: Mitjana per a 100 realitzacions dels dos índexs de la matriu de Scatter per a les dades de Girona de 2011.

Els resultats d'aquesta matriu són monòtonament ascendents o descendents amb el nombre de clústers. S'extreuen dues lectures diferents: el número mínim de clústers, 2, és l'òptim, o l'avaluació presentada és insuficient per concloure sobre el número adequat de clústers. Esperem a l'avaluació de prediccions per a comprovar quina de les dues opcions és la més apropiada.

Podem veure resultats anàlegs per a la província de Barcelona a la Figura A1.1, a l'apartat d'annexos 1.

Reduint l'espai amb PCA, el comportament és exactament el mateix que el mostrat a la figura 4.1. Les figures A1.2 i A1.3 de l'annex 1 ho il·lustren per a les dades de Barcelona i Girona de 2011 amb 10 i 5 components del PCA. En aquest cas no es comparen diferències entre els valors obtinguts amb PCA i sense perquè els resultats depenen de les realitzacions i no són comparables directament.

4.1.1.2 Coeficient Silhouette

La segona proposta per a avaluar les agrupacions realitzades amb el k-means és el coeficient *Silhouette*, mesurat amb les equacions (2.16) i (2.17). La configuració òptima correspon a la que obtingui un valor màxim.

Per aquest cas es comprova que la solució també es monòtona i n'hi ha prou amb 10 realitzacions per a calcular-ne la mitjana.

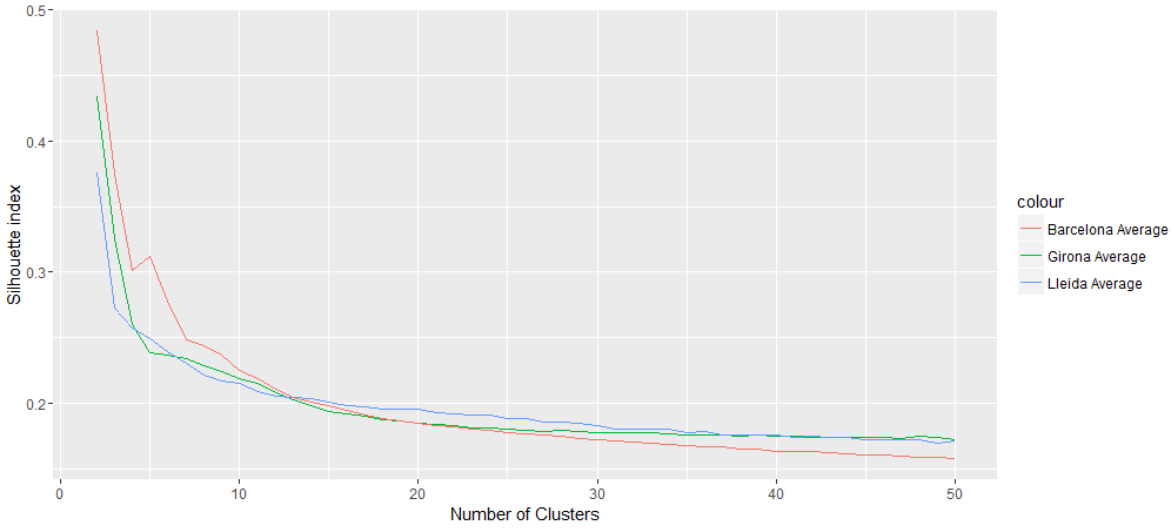

Figura 4.2: Mitjana per a 10 realitzacions del coeficient Silhouette per a les dades de Barcelona, Girona i Lleida de 2011.

A la Figura 4.2 es representa el coeficient *Silhouette* per a les dades de 2011 de les províncies de Barcelona, Girona i Lleida. Les tres tenen un volum de dades molt diferent: Barcelona prop de 6.000 meses, Girona més de 800 i Lleida unes 550. Es veu com aquest índex decreix clarament de manera molt similar per a les 3 i situa l'òptim novament a 2 clústers. Les conclusions són les mateixes que per l'estudi de la matriu de *Scatter*: sembla que s'ha de fer la predicció i comprovar la configuració més adequada en aquesta segona fase.

Es realitza el mateix estudi amb una reducció de les dades amb PCA, considerant el 99,9% de la variança acumulada. Amb aquest paràmetre, el nombre de columnes amb les que es treballa es redueix (es mostra a la taula 4.2).

	Barcelona	**Girona**	**Lleida**
Partits (dimensió original)	13	14	15
Dimensions amb PCA del 99,9 %	7	6	7

Taula 4.2: Partits per a les províncies de Barcelona, Girona i Lleida el 2011 i dimensions reduïdes amb PCA del 99,9%.

Com en el cas de la matriu de *Scatter*, el comportament tampoc varia fent servir PCA (s'inclou la gràfica corresponent a l'annex: Figura A1.4). Novament, no té cap sentit conceptual comparar els índexs amb els resultats anteriors i s'han de fer servir procediments més contundents, ajustant la predicció amb assaig i error.

4.1.2 Avaluació per Fuzzy c-means

Els índexs de l'apartat 2.3.2, que es defineixen amb les equacions (2.18), (2.19) i (2.20) no depenen només de les dades d'entrada i del nombre de clústers, a diferència dels de k-means. En aquest cas també s'estudia com varien els índexs per diferents valors del paràmetre m, que mesura el grau de "*fuzzyficació*".

Es realitza un estudi paral·lel als desenvolupats fins ara. Fixem una m=2 i s'observa com varien els paràmetres per als resultats de Girona de 2011. Igual que en casos anteriors, es calcula la mitjana de 100 realitzacions per a obtenir valors fiables.

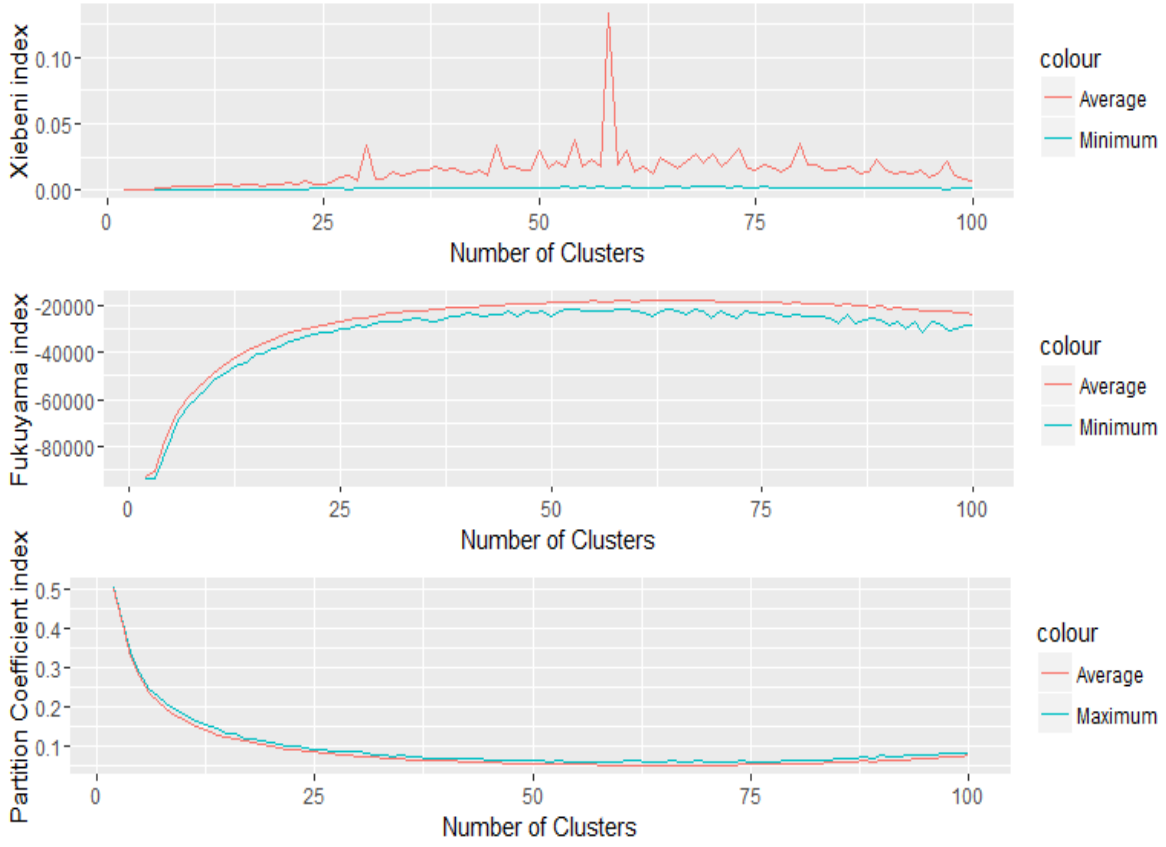

Figura 4.3: Mitjana i òptims per 100 realitzacions dels índexs de Xie-Beni, Fukuyama-Sugeno i del Partition coefficient normalitzat per a dades de Girona de 2011 amb m=2.

La figura 4.3 mostra tant la mitjana de les realitzacions com el valor òptim de cadascuna d'elles, ja sigui el màxim o el mínim obtingut. Els millors valors de Xie-Beni i Fukuyama-Sugeno corresponen als mínims i per l'altre indicador al màxim. Succeeix doncs el mateix que amb les validacions de *k-means*: el millor valor s'obté per a 2 clústers.

Es fa servir la interpretació que ha dominat fins ara: que l'avaluació realitzada per aquests índexs no és gaire indicativa. No es pot concloure res i s'ha de fer servir la "força bruta" amb la predicció. Com a la resta de casos analitzats prèviament, l'estudi sobre l'espai reduït PCA deriva en els mateixos resultats, que podem veure a la figura A1.5 de l'Annex 1.

Ens centrem ara en l'estudi de com influeix l'elecció del paràmetre m en l'obtenció d'aquests índexs. Els valors a estudiar seran m=1.2, 1.4, 1.6, 2 i 2.4. S'avaluen primer els índexs per a Barcelona, amb dades de 2011.

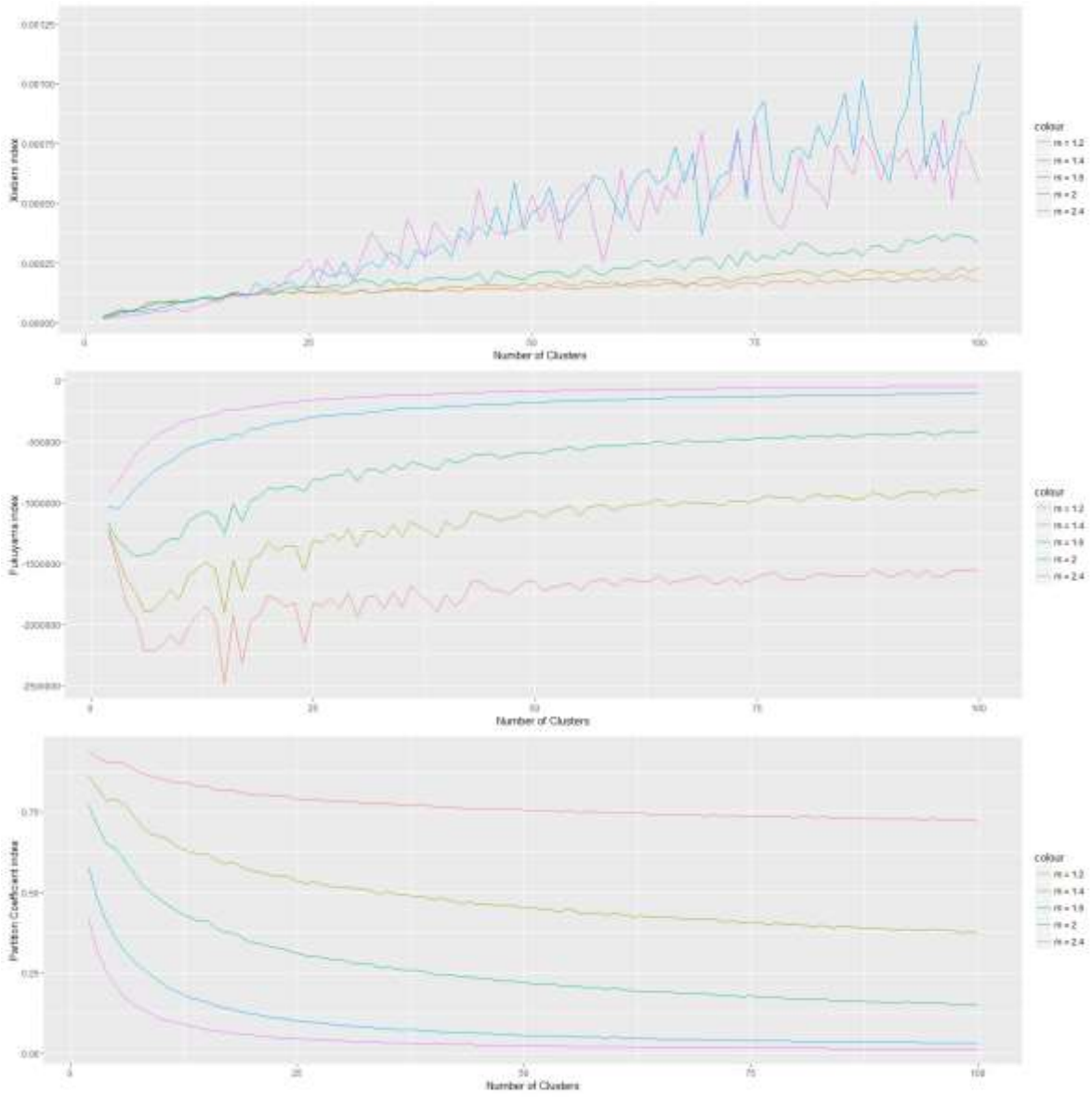

Figura 4.4: Mitjana i valors òptims per a 100 realitzacions dels índexs de Xie-Beni, Fukuyama-Sugeno i del Partition coefficient (de dalt a baix) normalitzat per a les dades de Barcelona de 2011 i segons el paràmetre m.

A la figura 4.4 els millors índexs de Xie-Beni i Fukuyama-Sugeno són els mínims, de manera que els resultats són millors com més petits són els valors de m.

Pel *Partition coefficient* normalitzat també els valors més grans de m són els que es comporten pitjor. Observem fins i tot que per l'índex de Fukuyama i valors de m inferiors a 2, obtenim òptims fent servir agrupacions d'entre 3 i 20 clústers, aproximadament. S'obté un primer resultat a comprovar diferent als anteriors, malgrat que la tendència general continua sent que no trobem gaires indicis que ens recomanin un nombre de clústers adequat abans de predir resultats.

Per als resultats de Girona de 2011 (figura 4.5) els valors de m petits milloren també el comportament del sistema.

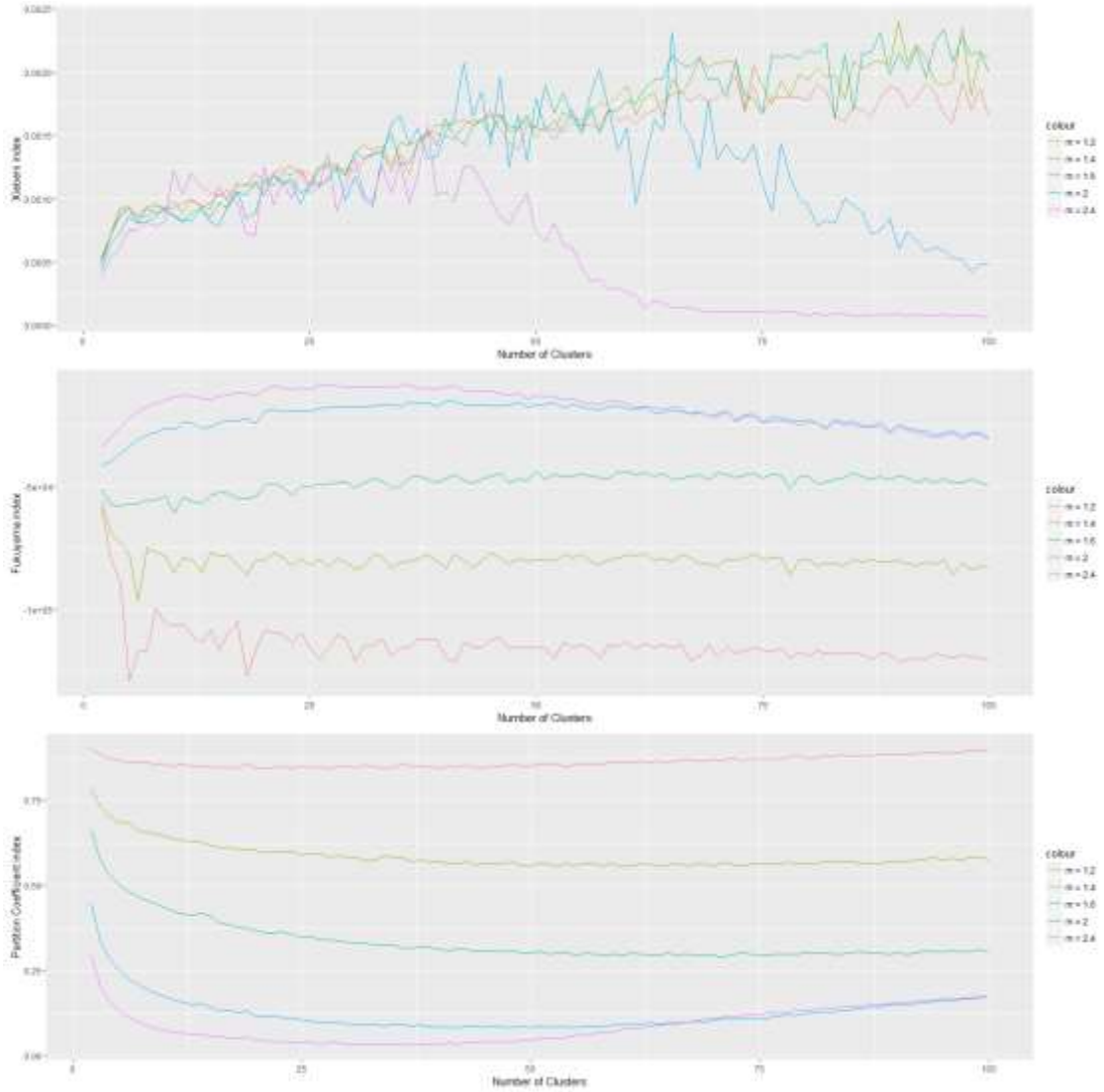

Figura 4.5: Mitjana i valors òptims per a 100 realitzacions dels índexs de Xie-Beni, Fukuyama-Sugeno i del Partition coefficient (de dalt a baix) normalitzat per a les dades de Girona de 2011 i segons el paràmetre m.

Coincidint amb el que hem observat per Barcelona, per l'índex de Fukuyama també obtenim valor òptims per a clústers entre 3 i 20 sempre que el paràmetre m sigui inferior a 2. En el cas de l'índex de Xie-Beni es troben òptims també per un nombre de clústers gran, superior a 60. Aquests valors són exagerats tenint en compte que el nombre de meses a Girona se situa per sobre de 800, suposant gairebé un 10% d'aquestes. Considerem que aquest òptim tampoc serà un bon resultat a estudiar.

Per resumir els apartats 4.1.1 i 4.1.2 es presenta presentem aquesta taula amb el nombre de clústers òptims obtingut per província, tipus de *clustering* i índex, per a m=2.

Província	**Tipus de Clustering**	**Índex**	**Nº de clústers òptim**
Barcelona	*K-means*	*Silhouette*	2 clústers
		Matriu de *Scatter*	2 clústers
	Fuzzy c-means (m=2)	*Xie-Beni*	2 clústers
		Fukuyama-Sugeno	3 clústers
		Partition coefficient normalitzat	2 clústers
Girona	*K-means*	*Silhouette*	2 clústers
		Matriu de *Scatter*	2 clústers
	Fuzzy c-means (m=2)	*Xie-Beni*	2 clústers
		Fukuyama-Sugeno	98 clústers
		Partition coefficient normalitzat	2 clústers

Taula 4.3: Resum dels valor òptims obtinguts per cada tipus de clustering, índex i província.

Provar un model amb dos clústers dóna un model massa simple, que no aporta gaires millores predictives al tenir molt pocs grups per a millorar els resultats. Per contra, el resultat aïllat de 98 clústers com a òptim a Girona, resulta massa complex i dificultaria la obtenció de resultats.

4.2 Resultats de la predicció

Aquest apartat estudia i avalua el comportament final de l'algorisme de predicció descrit al capítol 3. Als apartats previs s'han considerat les particularitats de *k-means* i *fuzzy c-means* per a avaluar els *clusterings* abans d'aplicar-los. L'algorisme emprat, si no es modifica, no admet *clusterings* de tipus *k-means*, pel que es fa una aproximació a aquests amb els valors viables més petits de m. Després de realitzar diverses simulacions, es determina que un valor de m = 1,015 pot ser suficientment petit com per comportar-se gairebé com un *k-means* i tenir un funcionament acceptable. A partir d'ara denotarem aquest tipus d'aproximació com a "*k-means*".

També s'ha vist, al final de l'apartat 4.1, que els índexs d'avaluació per *fuzzy c-means* donen millors resultats per a valor petits de m. Així, per a comparar resultats de prediccions se seleccionen dos valors de m: 1,015 i 2. Considerem m = 1,015 com la m viable més petita. L'elecció de m = 2 és relativament aleatòria, cercant un valor gran, però no excessivament. Per a reduir el cost computacional i intentar agrupar les dades de manera més intel·ligent, també provarem a aplicar PCA amb un percentatge de variança acumulada del 99,9 % sobre les dades d'entrenament. Seleccionem, llavors, 4 combinacions possibles de paràmetres per testejar.

Mode	Dessignació	Paràmetre m	Clustering	PCA
1	KmeansPCA	1,015	"K-means"	Sí
2	Kmeans	1,015	"K-means"	No
3	FuzzyPCA	2	Fuzzy c-means	Sí
4	Fuzzy	2	Fuzzy c-means	No

Taula 4.4: Conjunt de paràmetres seleccionats per la predicció i designació dels mateixos.

Tots els càlculs de resultats estan destinats a millorar la predicció amb l'arribada progressiva del recompte de meses. Per avaluar el model, es compta en aquest cas amb les dades finals de totes les eleccions. Així, el que es fa és obtenir prèviament el resultat final de les eleccions i comparar-lo a cada moment amb les prediccions i amb els resultats de l'escrutini pur a mida que es va simulant l'arribada de dades. Es calcular el MSE (Mean Square Error) entre el resultat final i la predicció.

$$MSE = \frac{1}{P}\sum_{j=1}^{P}(r_j - p_j)^2 \qquad (4.4)$$

A l'equació (4.4), P és el nombre de partits de la província i any a predir, r_j és el resultat final (percentatge total de vots) del partit j i p_j és la predicció del partit j.

Anàlogament, el MSE (Mean Square Error) entre el resultat final i l'escrutini parcial és:

$$MSE = \frac{1}{P}\sum_{j=1}^{P}(r_j - e_j)^2 \qquad (4.5)$$

A l'equació (4.5), P és el nombre de partits de la província i any a predir, r_j és el resultat final (percentatge total de vots) del partit j i e_j és l'escrutini parcial del partit j.

Es poden pensar diverses maneres de comptabilitzar el percentatge de vots escrutats, però es fa servir la convenció de les eleccions a Espanya, què respon al nombre de meses que han enviat el seu recompte respecte del total de les eleccions.

Un factor que pot influir de manera determinant en l'obtenció de resultats és l'ordre amb què arriben les dades. Per a eliminar el biaix, es fan 100 realitzacions per a cada predicció, cadascuna amb un ordre aleatori i diferent d'arribada de dades. Totes les simulacions obtingudes responen al comportament de la figura 4.6, que mostra el MSE de les equacions (4.4) i (4.5) per a la predicció de la província de Girona el 2011. Es fan servir les dades de 2008 com a entrenament i 100 realitzacions amb ordres diferents d'arribada de dades. Per a cada ordre es computen 6 nombres de clústers diferents: 2, 5, 10, 15, 20 i 30.

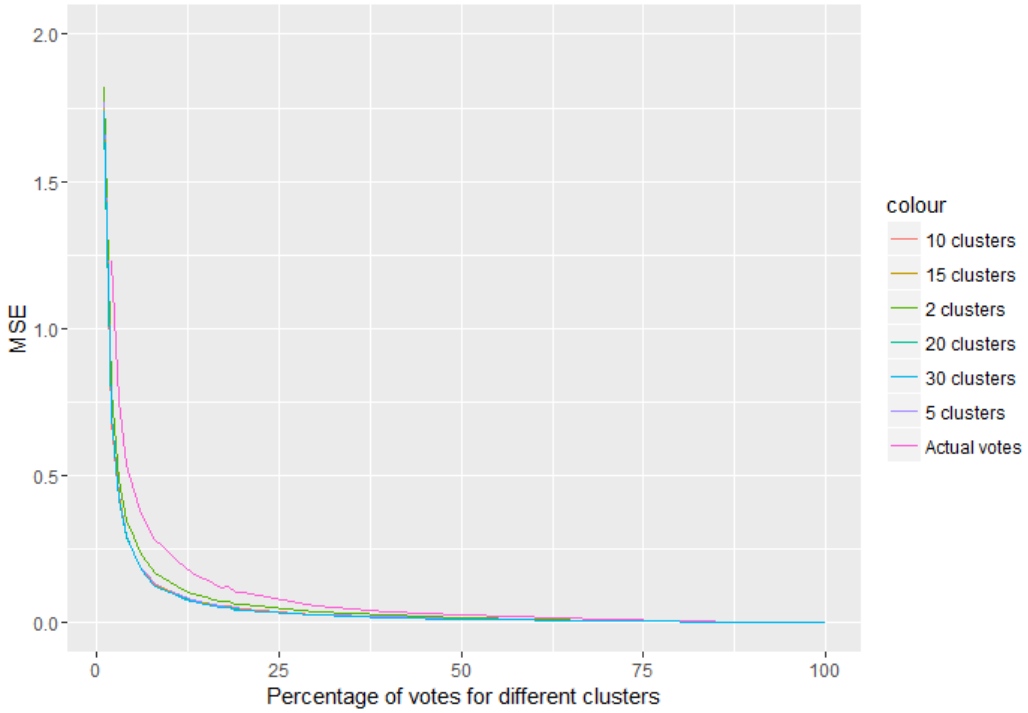

Figura 4.6: MSE dels resultats predits i de l'escrutini parcial respecte al resultat final de les eleccions a Girona al 2011 amb Fuzzy clustering i m=2 sense PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

A la figura 4.6 s'observa com es redueix el MSE a mida que augmenta l'escrutini. També queda molt clar que totes les prediccions fetes milloren l'escrutini pur (corba rosa, "Actual votes"). A continuació es detalla una anàlisi equivalent a la de la figura 4.6, però més profunda, per als resultats de Girona i de Barcelona. Es considera un nombre de clústers màxim inferior a $\sqrt{N}$, sent N el nombre de meses electorals [12]. Totes les gràfiques presenten dades dels primers estadis de l'escrutini, fins a un 5% o un 20%, tram a on és més present el biaix produït per la ràpida arribada de resultats de meses petites. S'inicia l'estudi amb Girona, per ser una província de mida mitjana, suficientment representativa però no gaire gran. Es fa servir com a banc de proves per a preparar l'estudi de Barcelona, passant de 800 meses electorals de la primera a les gairebé 6.000 de la segona.

4.2.1 Resultats a Girona

Després de diverses simulacions es determina utilitzar 2, 5, 10, 15, 20 i 30 clústers per a entrenar la predicció de Girona. A partir dels 30 clústers, l'algorisme perd efectivitat per a aquesta província, amb MSE major que pels altres. S'estudien 3 parells diferents de dades d'entrenament i de predicció:

- Entrenament amb dades de 2000 i predicció amb dades de 2004
- Entrenament amb dades de 2004 i predicció amb dades de 2008
- Entrenament amb dades de 2008 i predicció amb dades de 2011

Com que els resultats són prou similars entre ells independentment del parell de dades estudiades, només s'adjunten a aquesta part els resultats amb dades de 2008 com a entrenament i predicció amb dades de 2011. La resta s'ha adjuntat a l'Annex 2 (figures A2.1 a A2.8). Comencem per fer un cop d'ull al resultat per *clustering Fuzzy c-means* i m=2 entre l'1 i el 5% d'escrutini.

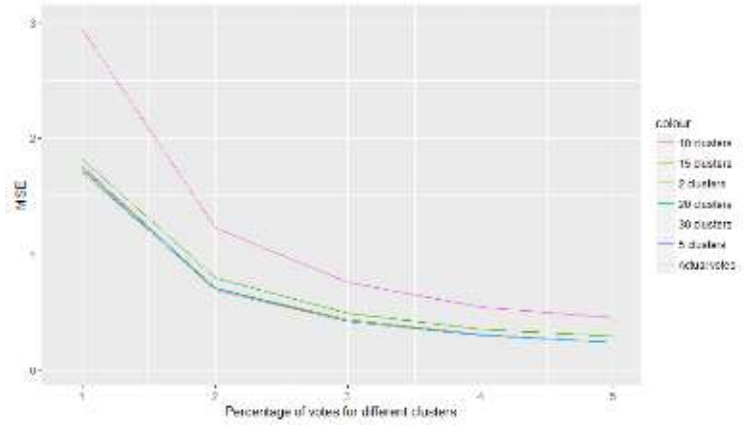

Figura 4.6: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2011 amb Fuzzy clustering i m=2 sense PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

La figura 4.6 deixa clar que qualsevol elecció en el nombre de clústers millorarà les dades de l'escrutini de manera important. No sembla vital escollir-ne el millor, tot i que sí es descarta l'entrenament amb 2 clústers, que inicialment es veu bastant pitjor.

Es mostra ara la mateixa gràfica, però amb PCA per a computar l'entrenament de les dades del 2008.

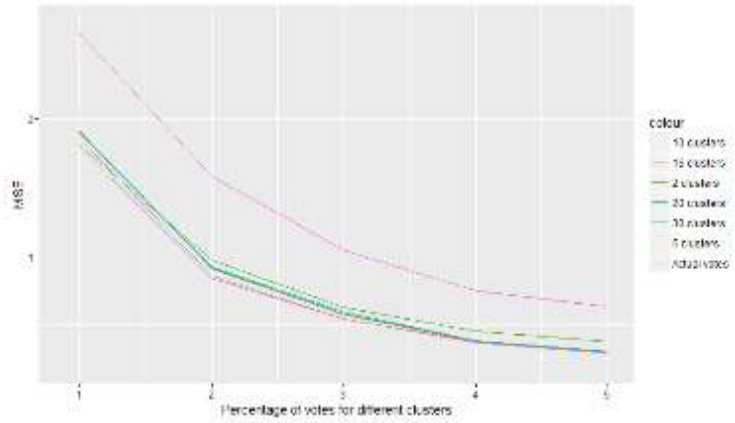

Figura 4.7: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2011 amb Fuzzy clustering i m=2 amb PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Amb PCA s'observa una tendència pràcticament idèntica, però l'agrupació amb 2 clústers ha millorat respecte al cas anterior per a escrutinis fins al 2%. Després s'analitzaran les diferències quantitativament.

Ara s'analitza el resultat per a un *clustering "k-means"*, amb una m = 1,015 que l'aproxima molt a *k-means*, sense PCA

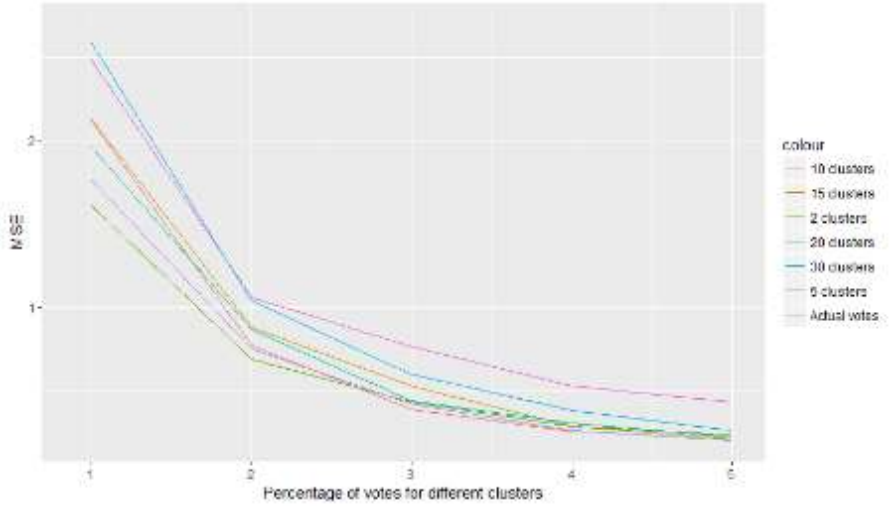

Figura 4.8: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2011 amb "k-means" clustering i m=1,015 sense PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

A la figura 4.8 es veu com a l'algorisme que s'aproxima a *k-means,* les prediccions empitjoren considerablement pels primers moments de l'escrutini.

Aplicant PCA (figura 4.9) sembla que es redueix el MSE, però s'haurà de fer una anàlisi més profunda per a determinar–ho.

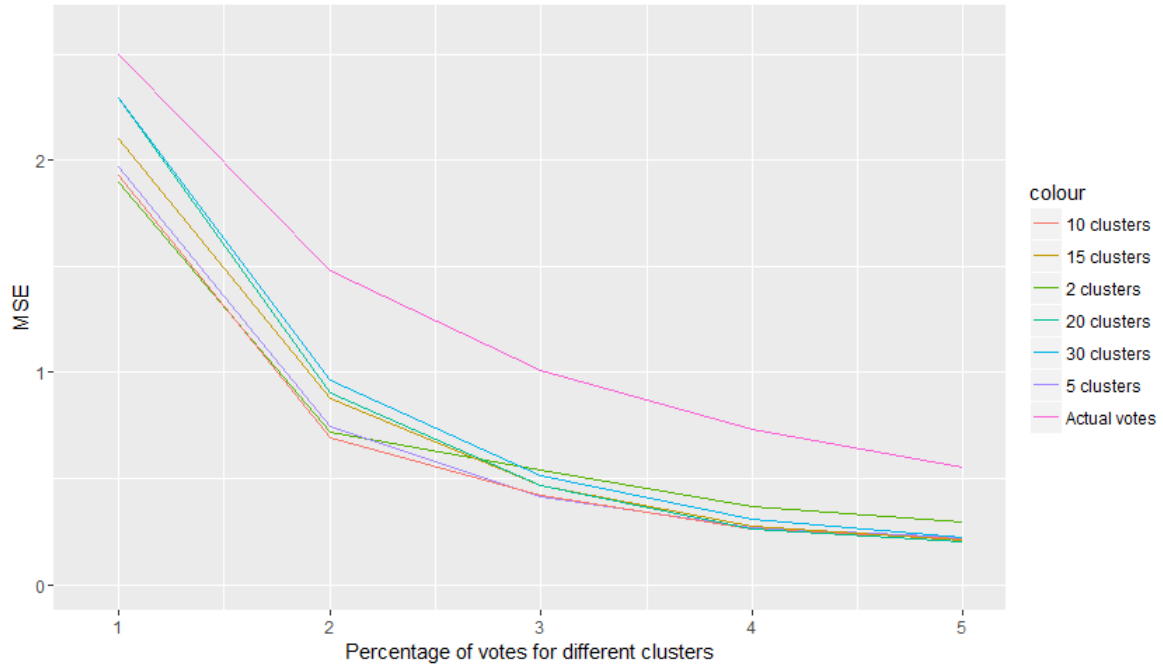

Figura 4.9: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2011 amb "k-means" clustering i m=1,015 amb PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

S'elabora una taula-resum a on comparar la bondat de cadascun dels modes de predicció especificats a la taula 4.4. Per a comparar-los cal trobar una mesura de l'error comès, ja que el MSE és diferent per a cada especificació degut al caràcter aleatori del càlcul.

Es mesura la millora com la diferència entre el MSE de l'escrutini (MSE_e) i el de la predicció (MSE_p).

$$M = \frac{MSE_e - MSE_p}{MSE_e} \qquad (4.6)$$

A la taula-resum s'inclou la mida de millora i el número de clústers que la maximitza per a cada percentatge estudiat de l'escrutini.

	Fuzzy c-means				"K-means"				
	No PCA		PCA 99,9 %		No PCA		PCA 99,9 %		
Percentatge	Nº de clústers	Millora	Nº de clústers	Millora	Nº de clústers	Millora	Nº de clústers	Millora	Millor configuració
1 %	20	42%	5	32%	2	35%	2	24%	Fuzzy
2 %	10	44%	5	47%	2	35%	10	53%	K-means PCA
3 %	10	45%	5	48%	10	49%	5	59%	K-means PCA
4 %	20	46%	5	50%	10	52%	20	64%	K-means PCA
5 %	20	48%	5	53%	10	55%	20	64%	K-means PCA
6 %	30	52%	20	55%	5	58%	20	64%	K-means PCA
7 %	30	53%	20	56%	15	58%	20	68%	K-means PCA
8 %	30	57%	15	54%	5	67%	15	70%	K-means PCA
9 %	30	57%	30	55%	10	69%	20	71%	K-means PCA
10 %	30	57%	15	56%	15	72%	15	72%	K-means
11 %	30	57%	30	58%	15	72%	20	74%	K-means PCA
12 %	30	58%	30	58%	30	71%	15	73%	K-means PCA
13 %	30	57%	30	59%	30	71%	15	76%	K-means PCA
14 %	30	58%	30	58%	30	71%	15	77%	K-means PCA
15 %	30	58%	30	59%	30	72%	20	77%	K-means PCA
16 %	30	58%	30	59%	30	73%	15	77%	K-means PCA
17 %	30	56%	30	59%	30	74%	15	77%	K-means PCA
18 %	30	58%	30	60%	30	74%	15	77%	K-means PCA
19 %	30	58%	30	60%	30	76%	15	77%	K-means PCA
20 %	30	60%	30	60%	30	76%	15	79%	K-means PCA
30 %	30	58%	30	61%	30	74%	15	81%	K-means PCA
40 %	30	56%	30	60%	30	73%	15	82%	K-means PCA
50 %	30	58%	30	60%	30	78%	20	79%	K-means PCA
70 %	30	54%	30	60%	30	82%	30	76%	K-means
90 %	30	58%	30	63%	20	81%	30	78%	K-means

Taula 4.5: Clústers òptims i comparativa de configuracions per a la predicció de dades de Girona pel 2011 amb entrenament del 2008 en funció de l'escrutini.

Observem que el número de clústers òptim és de 15, 20 o 30 en la major part dels casos. No obstant, per a escrutinis molt baixos pot ser 2, 5 o 10 clústers. En quant a la millor configuració per a la predicció, les implementacions amb PCA gairebé sempre tenen una mesura de la nostra millora amb l'equació (4.6) més gran: comparant els *clusterings*, el "k-means" computa millores més grans, excepte per un escrutini del 1, 2 o 3 %.

4.2.2 Resultats a Barcelona

Per a Barcelona es trien 20, 50, 75 i 100 clústers per a entrenar la predicció. A partir dels 100 clústers, l'algorisme perd efectivitat i el MSE augmenta. Després de veure amb Girona que les dades no ofereixen diferències significatives en funció de l'any de predicció, es decideix estudiar nomes la combinació de:

- Entrenament amb dades de 2008 i predicció amb dades de 2011

S'analitza el resultat per *clustering Fuzzy c-means* i m=2 entre l'1 i el 5% d'escrutini.

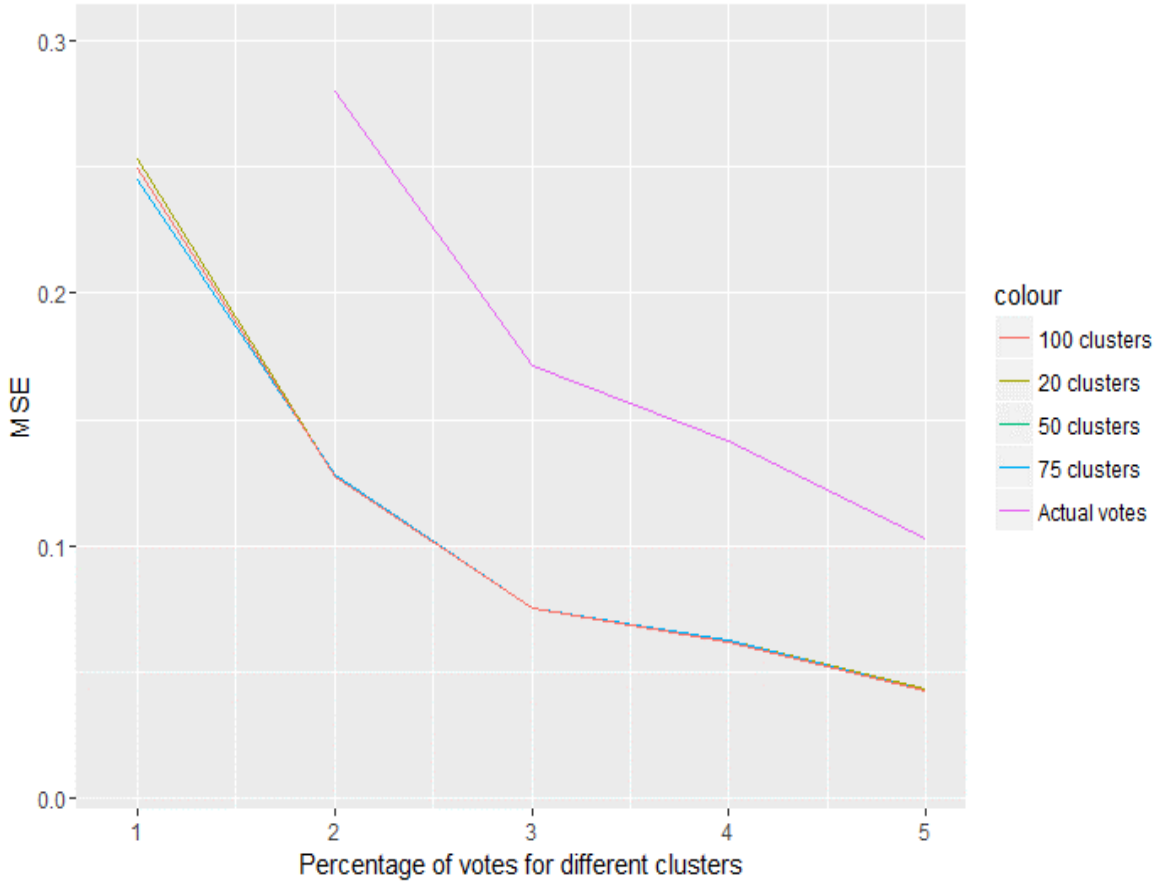

Figura 4.10: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Barcelona al 2011 amb Fuzzy clustering i m=2 sense PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 20, 50, 75 i 100 clústers.

La figura 4.10 no deixa gaire clar quina és la millor elecció en quant a nombre de clústers. Sembla més aviat que totes les possibilitats siguin bones, ja que totes milloren significativament les dades del còmput d'escrutini.

S'observa ara la mateixa gràfica, però fent servir PCA per a computar l'entrenament de les dades del 2008.

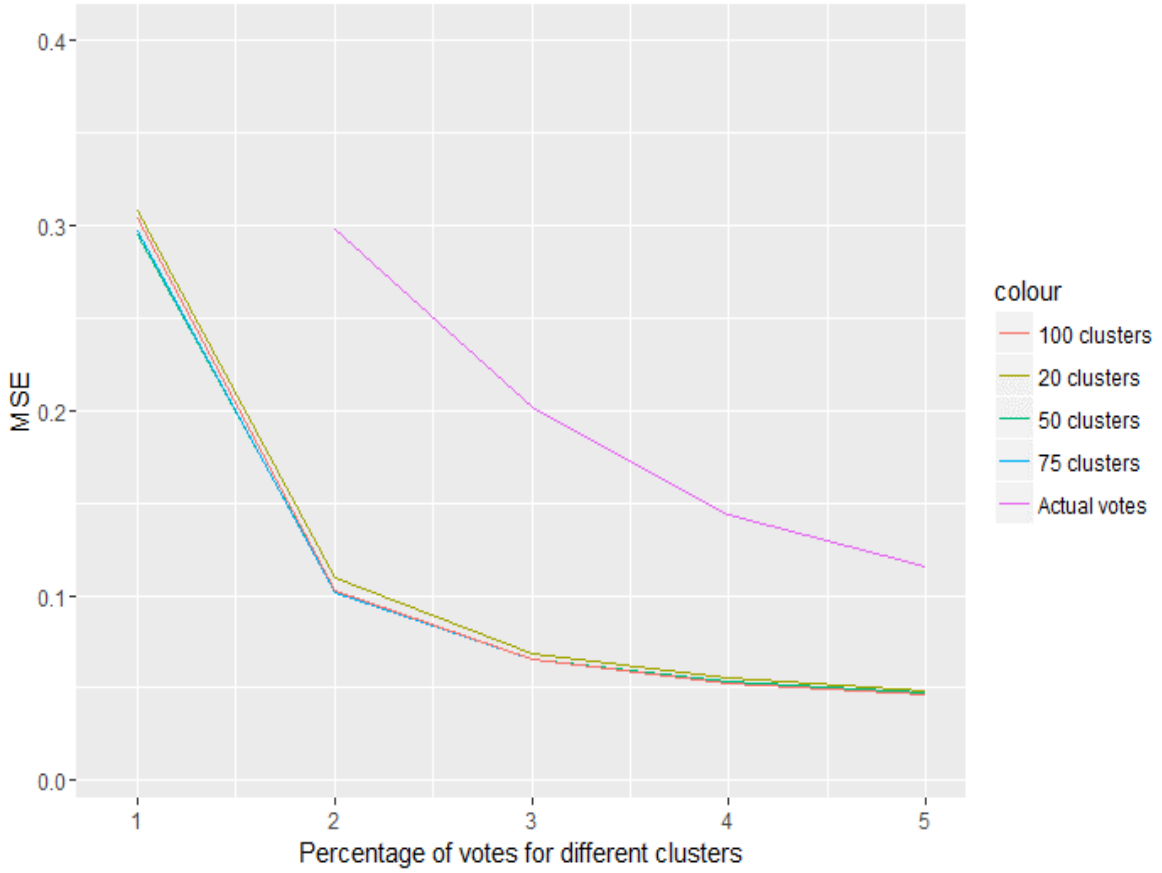

Figura 4.11: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Barcelona al 2011 amb Fuzzy clustering i m=2 amb PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 20, 50, 75 i 100 clústers.

Amb PCA es detecta una petita variació: les prediccions amb major nombre de clústers es destaquen molt lleugerament amb un MSE menor, però sense aportar un canvi gaire significatiu.

Ara s'analitza el resultat per a un *clustering "k-means"*, amb una m = 1,015 que l'aproxima molt a *k-means*, sense PCA.

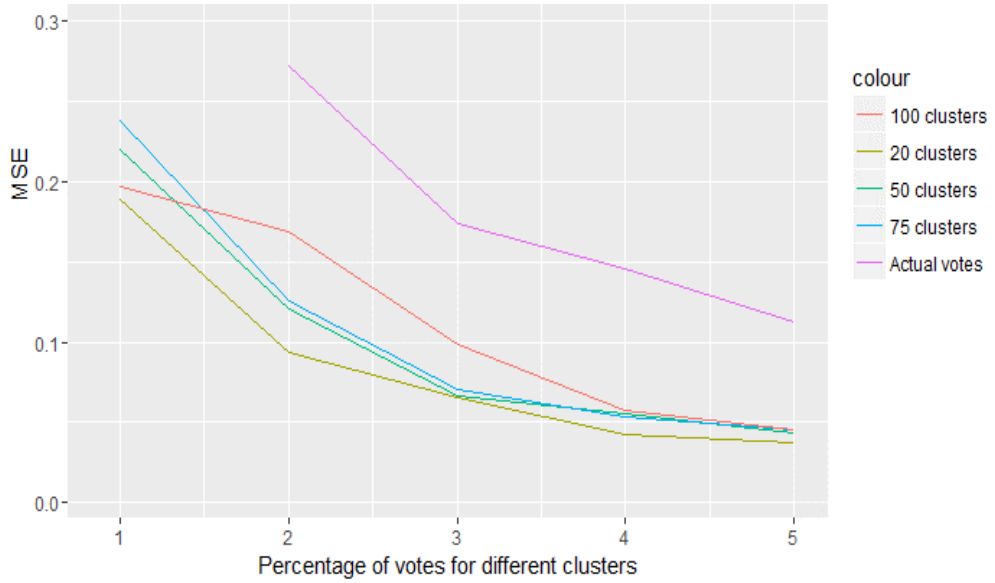

Figura 4.12: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Barcelona al 2011 amb "k-means" i m=1,015 sense PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 20, 50, 75 i 100 clústers.

A la figura 4.12 es visualitzen uns resultats més separats entre ells que en el cas del fuzzy c-means amb m = 2. S'obté, en la major part dels casos, millor resultat com menor és el nombre de clústers de l'entrenament.

Veiem el resultat aplicant PCA (figura 4.13).

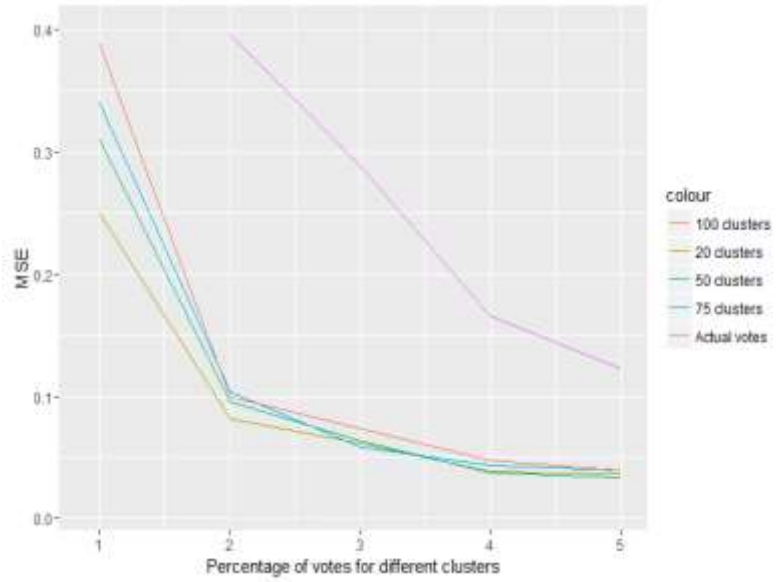

Figura 4.13: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Barcelona al 2011 amb "k-means" i m= 1,015 amb PCA. Les dades d'entrenament són de 2008 i la predicció es fa per 20, 50, 75 i 100 clústers.

El comportament observat pel "k-means" amb m = 1.015 mostra la mateixa tendència a resultats més bons per a nombre menor de clústers, però només dins un escrutini del 2 %. Després les diferències entre els valors obtinguts semblen poc significatives. Per a mesurar objectivament aquestes diferències i trobar la configuració més adequada, es calcula el nombre òptim de clústers i el valor de la millora amb l'equació (4.6). La taula 4.6 posa de manifest aquestes diferències.

	Fuzzy c-means				"K-means"				
	No PCA		PCA 99,9 %		No PCA		PCA 99,9 %		
Percentatge	Nº de clústers	Millora	Nº de clústers	Millora	Nº de clústers	Millora	Nº de clústers	Millora	Millor configuració
1 %	75	57%	50	52%	20	70%	20	70%	K-means PCA
2 %	50	55%	75	66%	20	66%	20	79%	K-means PCA
3 %	50	56%	75	68%	20	62%	75	80%	K-means PCA
4 %	100	56%	100	63%	20	71%	50	77%	K-means PCA
5 %	100	59%	100	60%	20	66%	50	73%	K-means PCA
6 %	100	52%	100	58%	20	69%	20	73%	K-means PCA
7 %	100	47%	100	57%	20	65%	50	74%	K-means PCA
8 %	100	53%	100	56%	20	67%	20	75%	K-means PCA
9 %	100	53%	100	54%	20	65%	20	72%	K-means PCA
10 %	100	60%	100	54%	20	66%	50	76%	K-means PCA
11 %	100	60%	100	56%	20	70%	50	76%	K-means PCA
12 %	100	62%	75	61%	20	70%	50	79%	K-means PCA
13 %	100	62%	100	62%	100	72%	50	80%	K-means PCA
14 %	100	57%	100	63%	100	74%	50	80%	K-means PCA
15 %	100	57%	100	63%	100	75%	75	80%	K-means PCA
16 %	100	59%	100	62%	100	74%	50	79%	K-means PCA
17 %	100	59%	100	62%	100	74%	50	77%	K-means PCA
18 %	100	57%	100	62%	100	73%	100	78%	K-means PCA
19 %	100	52%	100	63%	50	73%	75	80%	K-means PCA
20 %	100	50%	100	63%	50	72%	50	80%	K-means PCA
30 %	100	63%	100	63%	100	75%	100	83%	K-means PCA
40 %	75	61%	100	69%	75	70%	100	80%	K-means PCA
50 %	50	63%	100	66%	100	66%	100	78%	K-means PCA
70 %	50	60%	100	64%	50	76%	50	74%	K-means
90 %	100	56%	75	65%	75	79%	75	77%	K-means

Taula 4.6: Clústers òptims i comparativa de configuracions per a la predicció de dades de Girona pel 2011 amb entrenament del 2008 en funció de l'escrutini.

El número de clústers òptim ara és de 75 o 100 en la major part dels casos. No obstant, per a escrutinis baixos i amb "k-means", pot ser 20 clústers. Com amb les dades de Girona, les implementacions amb PCA ofereixen una millora (equació (4.6)) més gran en quasi tots els casos. Per acabar, el "k-means" computa millores més grans que el fuzzy amb m = 2.

5 Conclusions

El principal objectiu era millorar les prediccions de resultats durant la nit electoral. Amb l'adaptació de l'algorisme implementat a les eleccions sud-africanes del 2004 [6] s'assoleix, reduint el biaix produït per l'arribada dels resultats de les meses electorals amb pocs electors abans que les que en concentren molts. Els resultats i gràfics de l'apartat 4 en són proves gràfiques i empíriques.

S'ha depurat l'aplicació per a aconseguir uns resultats encara millors escollint tipus de *clustering* i paràmetres de m adequats. El tipus de distribució de dades fan que totes les avaluacions a priori dels índexs d'avaluació resultin molt poc indicatives i gens fiables. Per contra, referint-se a l'estudi dels valors de m, els índexs sí revelen que els valors més petits de m es comporten millor, tal i com s'ha comprovat empíricament.

Els índexs mostren un nombre de clústers òptim pel nombre mínim d'aquests, revelant que les dades són prou similars entre elles i es poden simplificar. La utilització de la tècnica de PCA per a reduir les dimensions de les nostres dades ajuda en casos com aquest a reduir el soroll del sistema. Es podria considerar l'aplicació d'aquesta tècnica com el gran encert de l'estudi. Malgrat no estalviar gaire en temps d'execució, sí que es comprova experimentalment que aplicar PCA a les dades d'entrenament redueix l'error en la predicció. A les taules 4.5 i 4.6 es mostren dos exemples.

Amb l'escrutini entre un 1 i un 5%, els valors de m grans obtenen millors resultats. Per sobre d'aquest llindar, es comporten millor els *clusterings* amb valors de m més petits, assimilant-se a un *k-means*. Aquesta tendència canvia amb PCA: obtenim resultats òptims en la majoria d'ocasions pels valors de m petits, incloent les prediccions amb un escrutini inferior al 5%.

Finalment, destaquem que el nombre de clústers utilitzats és menys determinant si es fa servir PCA, sempre que ens movem dins d'un ordre de magnitud. Aquest ordre de magnitud, tal com es contempla en la referència [12] no ha de superar l'arrel quadrada del nombre de meses electorals o empitjora.

6 **Referències**

[1] Descàrrega de dades electorals del *Ministerio del Interior del Gobierno de España*: http://www.infoelectoral.interior.es/min/areaDescarga.html?method=inicio

[2] M. Estok, N. Nevitte, and G. Cowan, “The Quick Count and Election Observation” National Democratic Institute, 2002

[3] Bernardo and Giron. “Robust sequential prediction from non-random samples: the election-night forecasting case (with discussion)”. Bayesian Statistics 4, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 61{77. New York: Oxford University Press, 1992

[4] P. J. Brown, D. Firth, and C. D. Payne, “Forecasting on British election night 1997” J. R. Statist. Soc. A., 162, Part 2, pp. 211-226, 1999

[5] R. Hochreiter and C. Waldhauser, “Evolving Accuracy: A Genetic Algorithm to Improve Election Night Forecasts”, arXiv: 1401.4674v1, 19 Jan 2014

[6] J.M. Greben, C. Elphinstone, and J. Holloway, “A model for election night forecasting applied to the 2004 South African elections”, ORiON, 2006.

[7] G. James et al., “An Introduction to Statistical Learning: with Applications in R”, Springer Texts in Statistics, pp. 373-390, 2013

[8] JC Bezdek, R Ehrlich and W Full, “FCM: The Fuzzy c-means Algorithm Clustering”, Computers & Geosciences Vol. 10, No. 2-3, pp. 191-203, 1984

[9] PJ. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics 20 pp. 53-65, 1987

[10] R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification (2nd ed)”, John Wiley & Sons, 2000

[11] M. Ramze Rezaee, B.P.F. Lelieveldt and J.H.C. Reiber, “A new cluster validity index for the fuzzy c-mean”, Division of Image Processing, Department of Radiology, Leiden University Medical Center, 1997

[12] Nikhil R. Pal and James C. Bezdek, “On cluster validity for the Fuzzy c-Means Model”, IEEE Transactions on Fuzzy Systems, Vol.3, nº3, 1995

Annex 1: Índexs de validació

Índexs mitjans de la matriu de *Scatter* per número de clústers per a 100 realitzacions, per la província de Barcelona i amb dades de 2011.

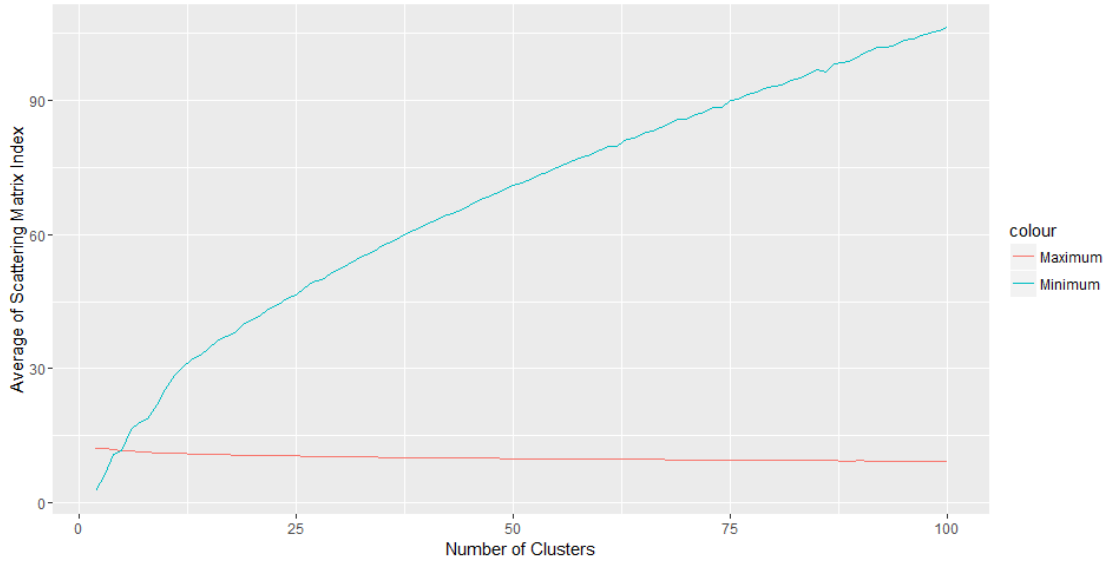

Figura A1.1: Mitjana per a 100 realitzacions dels dos índexs de la matriu de Scatter per a les dades de Barcelona del 2011.

Índexs mitjans de la matriu de Scatter per número de clústers per a 100 realitzacions, per la província de Girona, amb dades de 2011 i fent servir PCA amb el 99,9% de proporció acumulada, amb 10 components de 14.

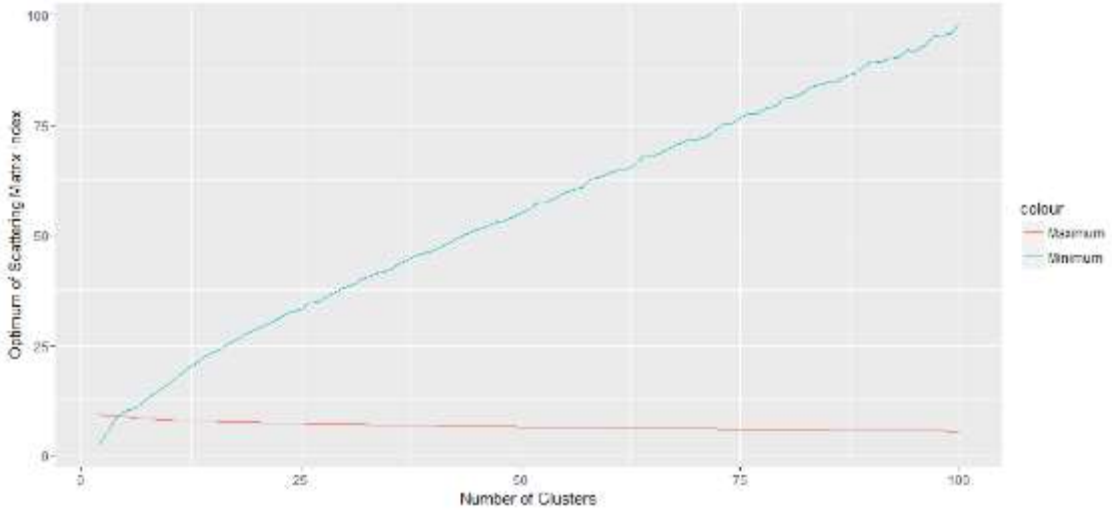

Figura A1.2: Mitjana per a 100 realitzacions dels dos índexs de la matriu de Scatter per a les dades de Girona del 2011, amb un PCA de 10 components.

Índexs mitjans de la matriu de Scatter per número de clústers per a 100 realitzacions, per la província de Girona, amb dades de 2011 i fent servir PCA amb el 99% de proporció acumulada, amb 5 components de 14.

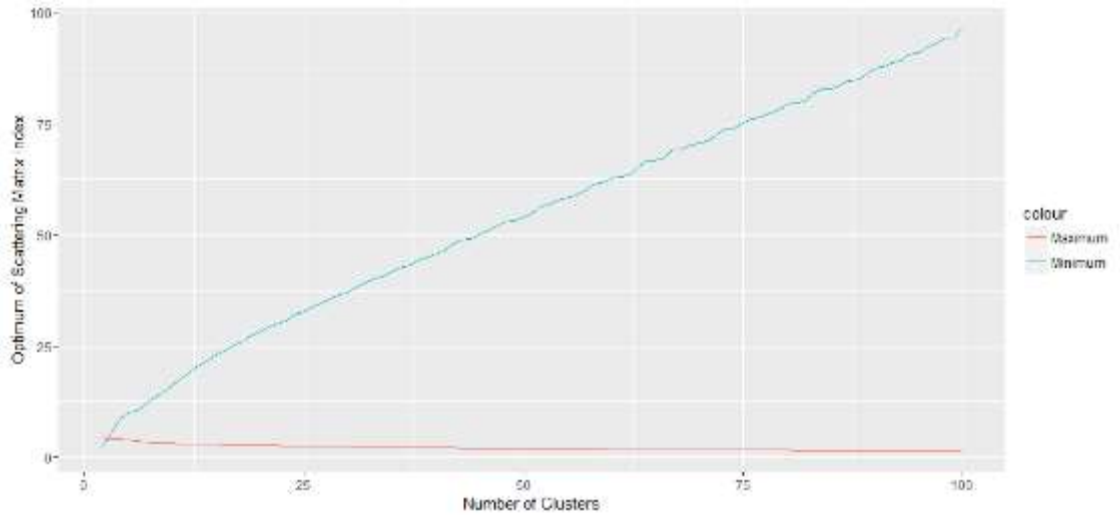

Figura A1.3: Mitjana per a 100 realitzacions dels dos índexs de la matriu de Scatter per a les dades de Girona del 2011, amb un PCA de 5 components.

Índexs mitjans del coeficient de Silhouette per número de clústers per a 10 realitzacions, per les províncies de Barcelona, Girona i Lleida, amb dades de 2011. Fem servir PCA amb el 99,9% de proporció acumulada, amb 7, 6 i 7 components de 12, 14 i 15, respectivament.

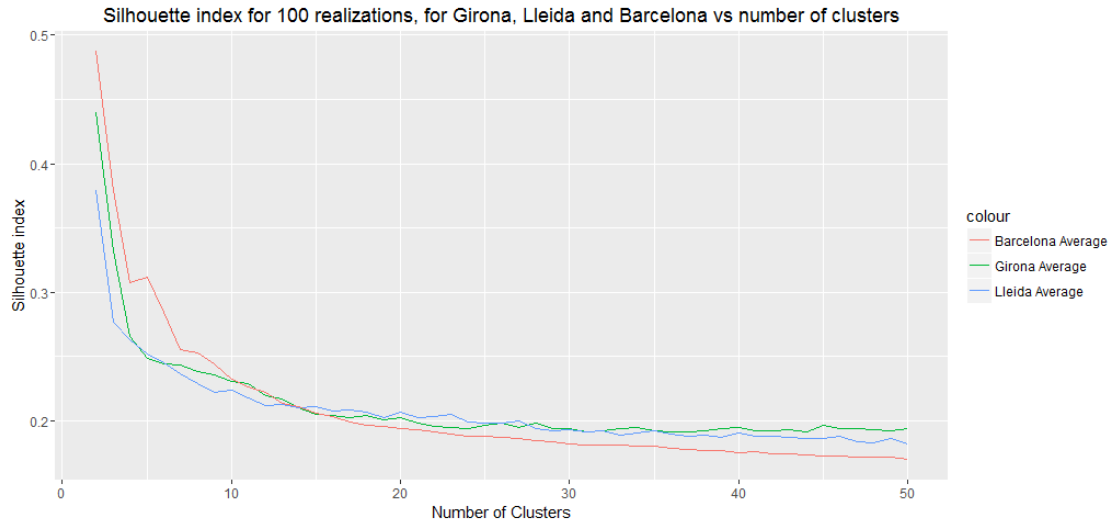

Figura A1.4: Mitjana per a 10 realitzacions del coeficient Silhouette per a les dades de Barcelona, Girona i Lleida del 2011 i PCA del 99,9%.

Índexs mitjans dels índexs de Xie-Beni, Fukuyama-Sugeno i Partition coefficient normalitzat per número de clústers per a 100 realitzacions. Es computa per la província de Girona amb dades de 2011. Fem servir PCA amb 6 components de 14, concentrant el 99,9% de variança acumulada.

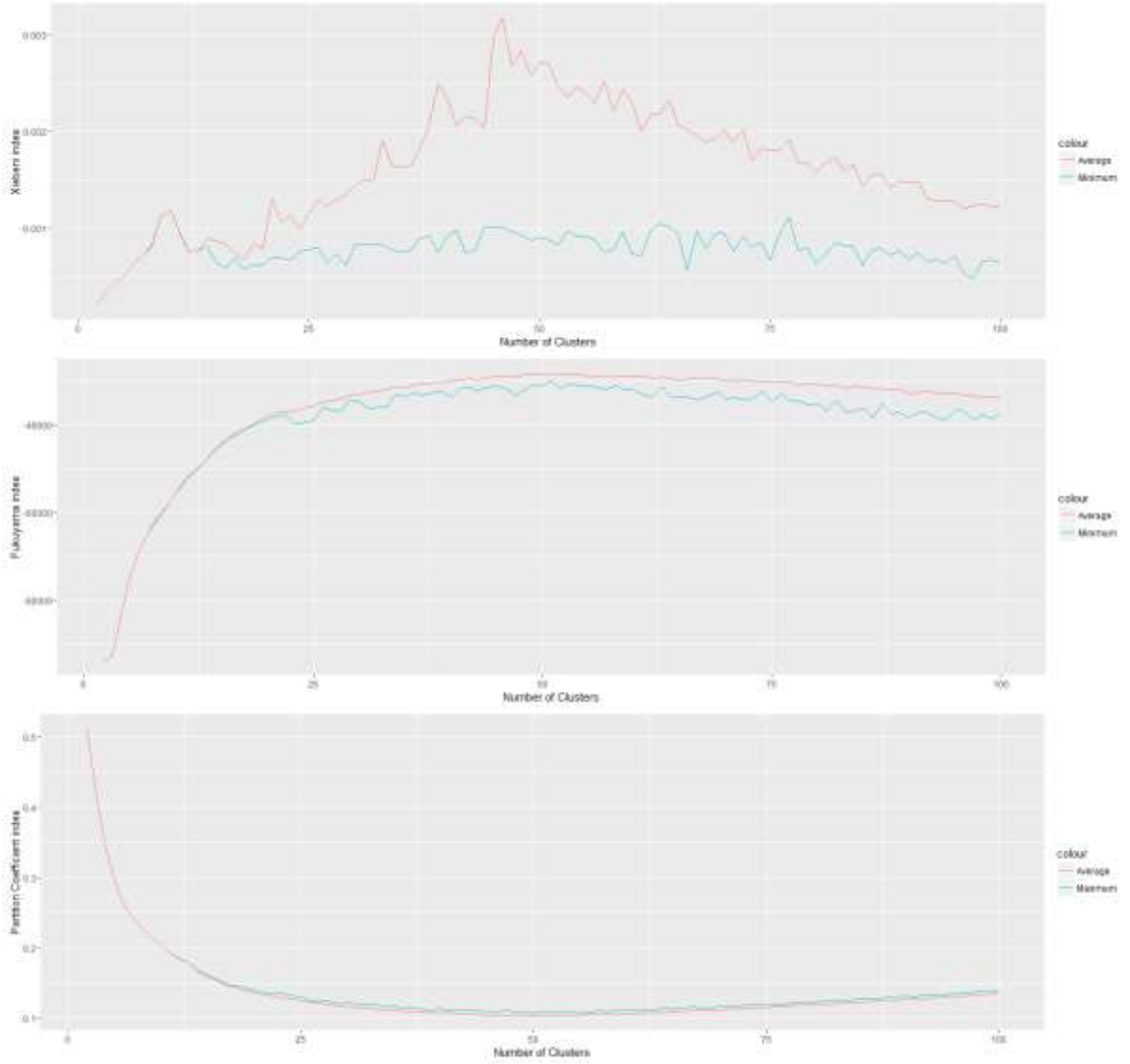

Figura A1.5: Mitjana i valors òptims per a 100 realitzacions dels índexs de Xie-Beni, Fukuyama-Sugeno i del Partition coefficient normalitzat per a les dades de Girona de 2011, amb PCA de 6 components i m=2.

Annex 2: Resultats de la predicció

Clustering Fuzzy c-means i m=2 entre l'1 i el 5% d'escrutini sense PCA. Entrenament amb dades de 2004 i predicció per 2008.

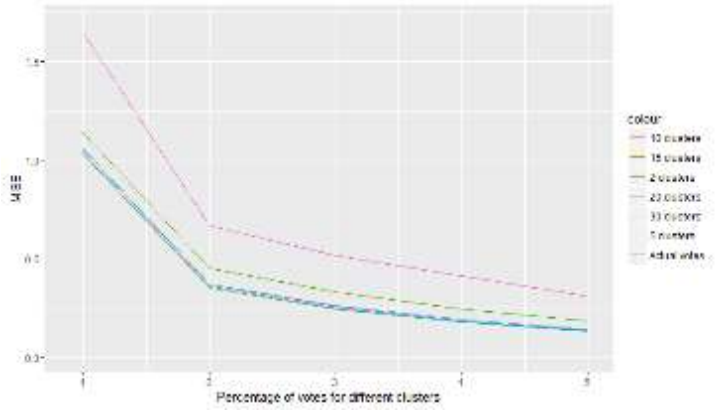

Figura A2.1: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2008 amb Fuzzy clustering i m=2 sense PCA. Les dades d'entrenament són de 2004 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering Fuzzy c-means i m=2 entre l'1 i el 5% d'escrutini sense PCA. Entrenament amb dades de 2000 i predicció per 2004.

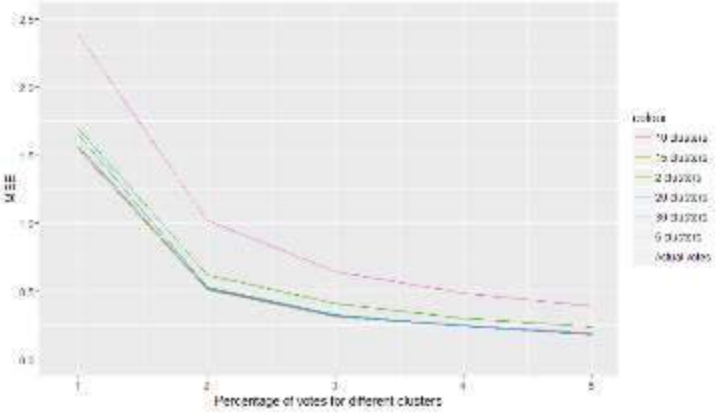

Figura A2.2: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2004 amb Fuzzy clustering i m=2 sense PCA. Les dades d'entrenament són de 2000 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering Fuzzy c-means i m=2 entre l'1 i el 5% d'escrutini amb PCA. Entrenament amb dades de 2004 i predicció per 2008.

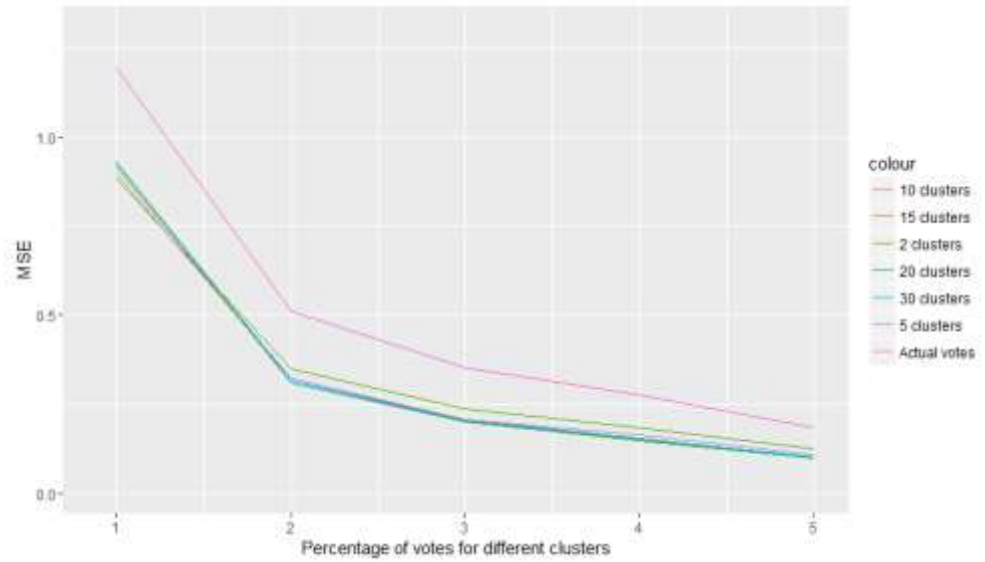

Figura A2.3: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2008 amb Fuzzy clustering i m=2 amb PCA. Les dades d'entrenament són de 2004 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering Fuzzy c-means i m=2 entre l'1 i el 5% d'escrutini amb PCA. Entrenament amb dades de 2000 i predicció per 2004.

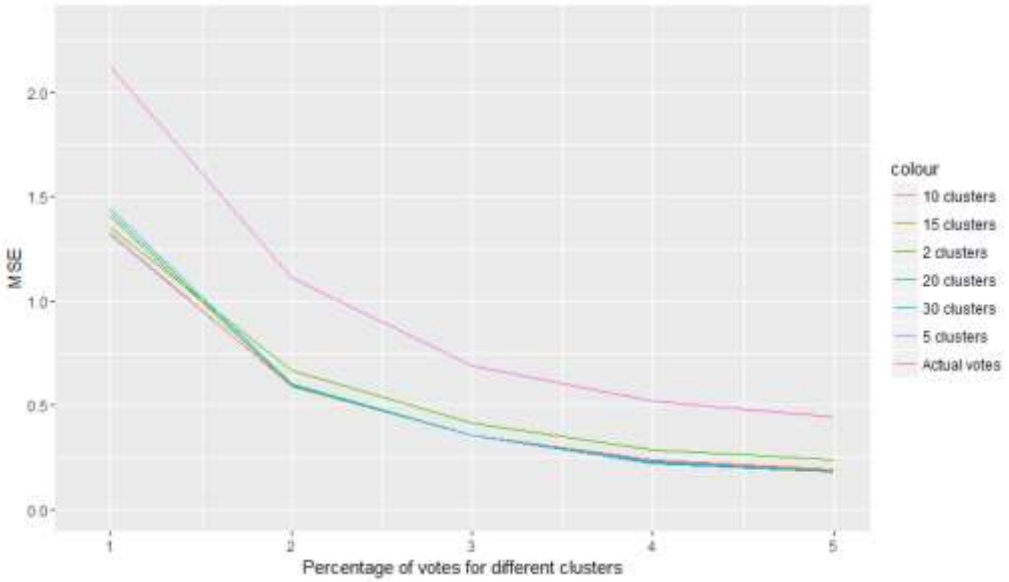

Figura A2.4: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2004 amb Fuzzy clustering i m=2 amb PCA. Les dades d'entrenament són de 2000 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering "k-means" i m=1,015 entre l'1 i el 5% d'escrutini sense PCA. Entrenament amb dades de 2004 i predicció per 2008.

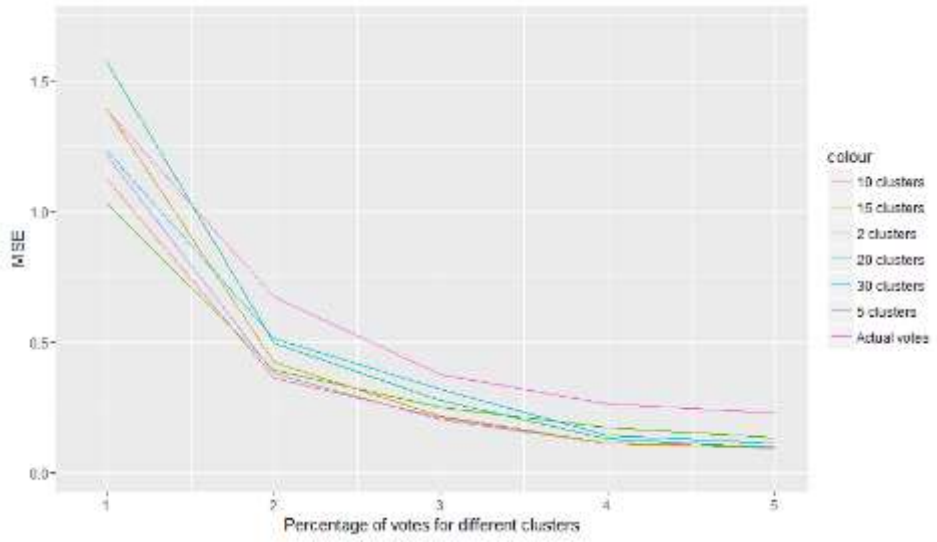

Figura A2.5: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2008 amb "k-means" clustering i m=1,015 sense PCA. Les dades d'entrenament són de 2004 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering "k-means" i m=1,015 entre l'1 i el 5% d'escrutini sense PCA. Entrenament amb dades de 2000 i predicció per 2004.

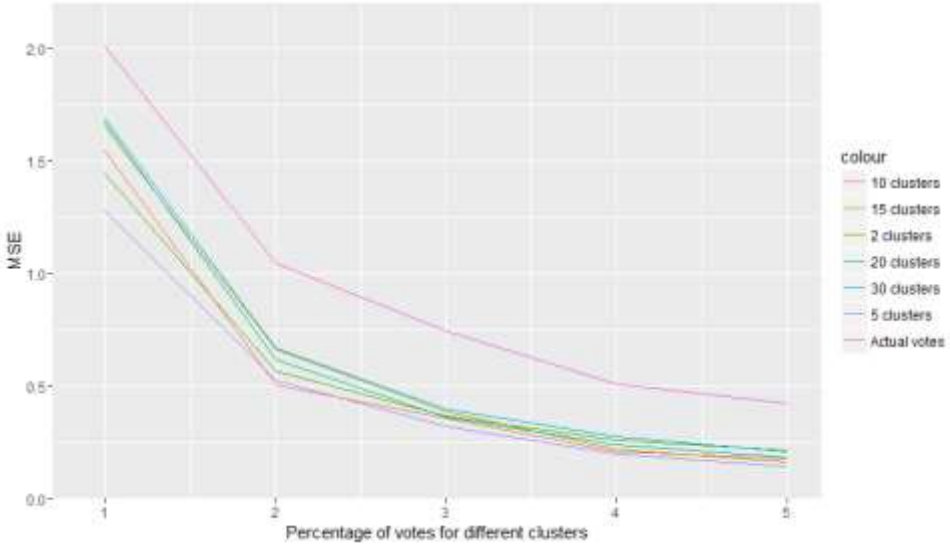

Figura A2.6: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2004 amb "k-means" clustering i m=1,015 sense PCA. Les dades d'entrenament són de 2000 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering "k-means" i m=1,015 entre l'1 i el 5% d'escrutini amb PCA. Entrenament amb dades de 2004 i predicció per 2008.

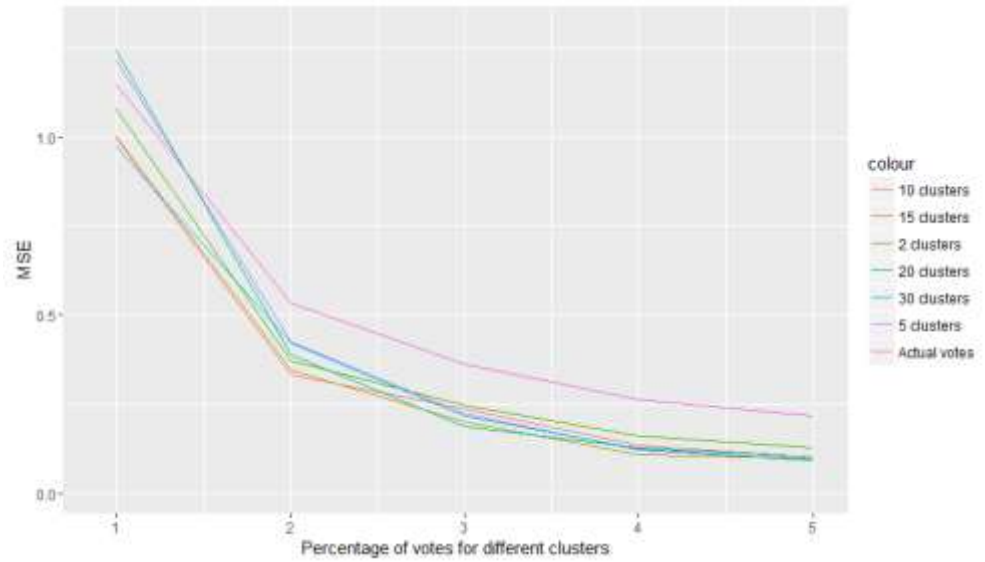

Figura A2.7: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2008 amb "k-means" clustering i m=1,015 sense PCA. Les dades d'entrenament són de 2004 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.

Clustering "k-means" i m=1,015 entre l'1 i el 5% d'escrutini amb PCA. Entrenament amb dades de 2000 i predicció per 2004.

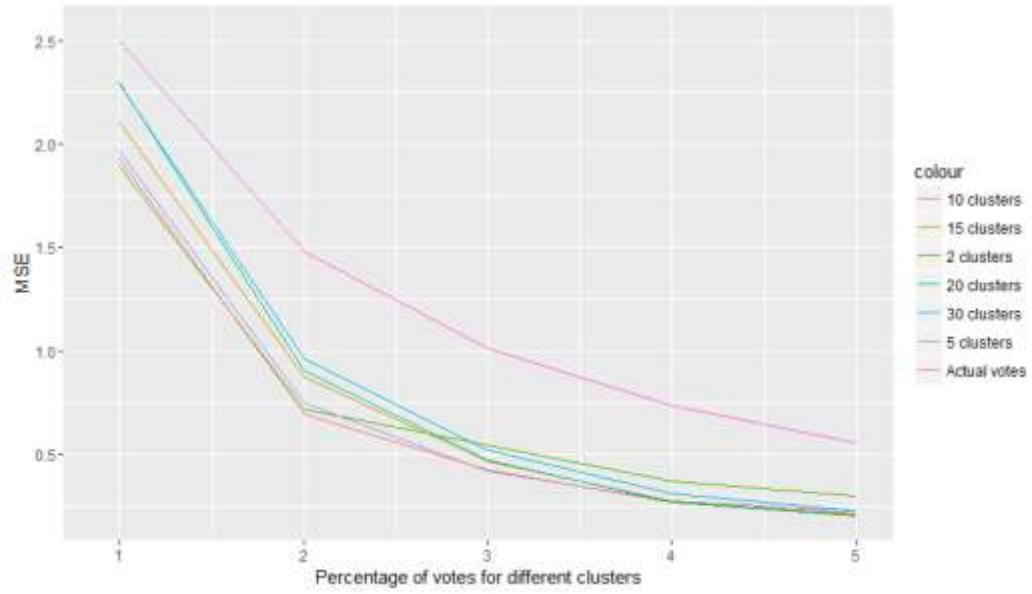

Figura A2.8: MSE dels resultats predits i de l'escrutini entre l'1 i el 5 % respecte al resultat final de les eleccions a Girona al 2004 amb "k-means" clustering i m=1,015 amb PCA. Les dades d'entrenament són de 2000 i la predicció es fa per 2, 5, 10, 15, 20 i 30 clústers.