CAPÍTOL 13

RIQUESA I DIVERSITAT

Índex Capítol

CAPÍTOL 13

RIQUESA I DIVERSITAT

En aquest capítol es fa un repàs a algunes de les mesures més utilitzades per mesurar la riquesa i diversitat d'un estil literari, que es fan servir per estudiar l'homogeneïtat en l'estil literari del *Tirant lo Blanc*. L'anàlisi de la riquesa i diversitat del vocabulari emprat en el *Tirant* es fa tant per blocs de 1000 paraules consecutives com per capítols de més de 200 paraules

13.1 Índexs de Diversitat

Definim N com el nombre total de paraules que apareixen en un text, i que en direm nombre d'ocurrències i V com el nombre de paraules diferents que formen el vocabulari observat en el text, i en direm formes. També definim V_d com el nombre de formes que apareixen d vegades en el text, de manera que tenim:

$$N = \sum_d d \cdot V_d ,$$

$$V = \sum_d V_d .$$

El quocient $\hat{p}_d = V_d/V$ és la proporció de paraules diferents que apareixen d vegades en el text, i en el límit, en augmentar el nombre de textos de llargada N, el promig d'aquestes proporcions el denotarem com p_d, tenint que $p_d = E(V_d/V)$.

Per cada paraula j, existeix un valor π_j que és la proporció de vegades que apareixeria la paraula j en un text del mateix autor de llargada infinita. Implícitament es fa la hipòtesi que tant la llista de paraules com els valors π_j són constants al llarg de l'obra de l'autor en estudi, suposició que pot ser discutible quan els textos de l'autor comprenen èpoques o gèneres molt diferents. Denotem per n_j el nombre de vegades que apareix la paraula j

en el text, i $\hat{\pi}_j = n_j/N$ la proporció de vegades que hi apareix. Si la hipòtesi d'estacionarietat és vàlida, en augmentar la llargada N, $\hat{\pi}_j$ tendeix a π_j. Per a un text donat, s'acompleix que:

$$N = \sum_{j=1}^{V_d} n_j ,$$

$$\sum_{j=1}^{V_d} \hat{\pi}_j = 1,$$

$$\sum_{d=1}^{\infty} p_d = 1.$$

La seqüència de valors π_j està íntimament relacionada amb la seqüència de valors p_d. Si s'empra aquesta darrera per caracteritzar la distribució del vocabulari de l'autor en comptes de π_j, és perquè el suport de la variable aleatòria n_j és tot el vocabulari, tant l'observat com el no observat, i per tant és desconegut, mentre que el suport de V_d són els números enters. A més, p_d depèn de N, mentre π_j no.

Els valors de p_d i de π_j s'estimen a partir de les seqüències $\hat{p}_d$ i $\hat{\pi}_j$ que observem en els seus textos. La majoria de les mesures de diversitat del vocabulari es defineixen a partir de p_d.

Donat un text de llargada N, com més gran és V més ric i divers és el llenguatge; mentre que per un text amb N i V donades, com més gran és p_d per d petites i més petit és p_d per d grans, més divers és el vocabulari. Qualsevol índex de diversitat ha de respectar aquests dos nivells d'ordenació.

Patil i Taillie (1982) en un article en el que es repassa la diversitat com a concepte i les formes de mesurar-la, defineixen els índexs de diversitat (Δ) com a promig de les "*rareses*", entenent com a raresa una funció de $R=R(\pi_j)$ que expressa un concepte oposat al de abundància:

$$\Delta = \sum_{j=1}^{V} \pi_j \cdot R(\pi_j) ,$$

i els criteris que han de acomplir són:

- Per a un text amb una sola paraula es té que $R(1)=0$.
- R és una funció decreixent en π_j definida en l'interval $[0,1]$: en augmentar el nombre de paraules emprades (*riquesa*) augmenta la diversitat del vocabulari.
- D'entre dos vocabularis amb el mateix nombre de paraules, serà més ric aquell que tingui una més gran *uniformitat* en les π_j. Donats dos textos amb les mateixes V i N, el llenguatge serà més ric en aquell en el que totes les paraules hi apareguin N/V vegades, mentre que el menys ric serà aquell en el que una paraula aparegui $N-V+1$ vegades i les altres $V-1$ paraules apareguin una sola vegada.

Els índex més utilitzats són:

1. V: nombre de paraules diferents en un text de llargada N. És el més senzill, però la seva distribució depèn de la llargada del text N. Mentre N pot créixer sense límit, V queda limitat pel nombre de paraules que composen el vocabulari de l'autor. Per aquest motiu, V serveix només per a comparar textos de la mateixa llargada,

tot i que la dependència entre V i N també es pot aprofitar per a obtenir caracteritzacions de l'estil. Com més gran és V, més ric és el llenguatge. La riquesa és una component de la diversitat del llenguatge.

L'ús de V com a mesura de diversitat és anàleg a emprar el rang com a mesura de dispersió d'una variable quantitativa (Pielou, 1982). Com el rang, V està subjecte a grans errors deguts al mostreig.

2. *V/N*: quocient forma/ocurrència (o quocient tipus/token). És un índex molt senzill, però la seva distribució segueix depenen de la llargada del text, N, ja que no és proporcional a N perquè creix més lentament. Quan estudia "*El Don de plàcides aigües*", Kjetsaa (1979) restringeix l'estudi per a $N=500$. Troba que la distribució del quocient de tipus per a 500 ocurrències en tots els texts analitzats segueix aproximadament una llei Normal. El mateix fan McKinnon i Webster (1971) quan estudien mostres de textos escrits per Kierkegaard, que consten d'entre 6548 i 15432 ocurrències.

   A la pràctica, si es fixa el nombre d'ocurrències, N, el quocient forma/ocurrència és exactament igual que V, i com més gran és V/N, més ric és el llenguatge.

3. V_1: *Hapax legomena* o nombre de paraules que apareixen una única vegada en un text. Donat un text de llargada N, com més gran és V_1 més divers és el llenguatge, però com V, V_1 només és útil per a comparar textos de la mateixa llargada.

4. V_2: *Hapax dislegomena* o número de paraules que apareixen dues vegades en un text. Valen les mateixes consideracions que per a V_1.

5. R: índex d'*Honoré* (Honoré(1979)), mesura la propensió d'un autor a escollir entre emprar una paraula ja usada prèviament o utilitzar-ne una de nova. Es calcula mitjançant l'expressió:

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}}.$$

   Com més gran és R, més ric és el llenguatge, perquè un nombre major de paraules són usades poc freqüentment. La seva distribució també depèn de la llargada del text, N, però Honoré comprova empíricament com, a partir d'una certa llargada del text ($N>1300$), el seu valor esperat es satura.

6. V_2/V: Relació entre els *Hapax Dislegomena* i el nombre de paraules diferents. Holmes i Forsyth (1995) l'usen en l'anàlisi dels *Federalist Papers*.

7. *D:* índex de *Simpson* (Simpson, 1949): és la probabilitat de que prenent a l'atzar dues paraules siguin iguals, i es calcula:

$$D = \frac{\sum_r d(d-1)V_d}{N(N-1)} = \frac{\sum_j n_j(n_j-1)}{N(N-1)}.$$

   Es pot comprovar que:

$$E(D) = \sum_{j=1}^{V} \pi_j^{\,2}.$$

   D és un indicador tant de la diversitat del vocabulari del text analitzat com un estimador no esbiaixat de la riquesa del vocabulari de l'autor, per qualsevol

grandària de *N*. Aquesta característica el fa especialment útil per estudiar textos de llargades diferents, com poden ser els capítols del *Tirant*. El que sí que depèn de *N* és la variança de *D*.

L'índex *D*, proposat anteriorment en un altre context per Gini (1912), és tal que com més divers és el llenguatge més petit és el seu valor, atès que serà més petita la probabilitat de repetició. Per obtenir un índex tal que com més divers és el llenguatge més gran sigui el seu valor, a vegades es fa servir *1-D*, que és la probabilitat que dues paraules, preses a l'atzar en el text, siguin diferents.

8. *K*: índex de *Yule* (Yule, 1944): és una variant de l'índex de Simpson. Es calcula com:

$$K=10000D(1-1/N)\ .$$

Bennet (1969) el fa servir per comparar la riquesa de vocabulari de "*Juli Cèsar*" amb la de "*As you like it*" i va detectar diferències entre actes, però no diferències entre obres. La elecció arbitrària de la constant ha estat molt criticada, i Tallentire (1972) argumenta que aquestes mesures del grau de repetició no són un aspecte prou inconscient de l'estil d'un autor.

9. *H: Entropia* (índex de Shannon), basat en la teoria de la informació i que mesura el grau de concentració de la probabilitat en unes poques paraules de totes les que composen el vocabulari. Es calcula:

$$H=\sum_{j=1}^{V}\pi_j\cdot\log(\pi_j),$$

on el sumatori és sobre totes les formes del vocabulari, tant les observades com les no observades. Es pot estimar emprant:

$$H=-\sum_{j=1}^{V}\hat{\pi}_j\cdot\log(\hat{\pi}_j)=-\sum_{j=1}^{V}\frac{n_j}{N}\log(\frac{n_j}{N}).$$

Per una *N* donada, com més ric és el vocabulari, més gran és *V* i més petits són els n_j i, per tant, més gran és l'entropia i per tant més divers és el text. La seva distribució depèn de la llargada del text, *N*: en augmentar *N* l'entropia del text tendeix a disminuir, i per tant no és fàcil utilitzar-lo per a comparar textos amb llargades molt diferents. Per aquest motiu, de vegades s'estandaritza entre 0 i 100 fent:

$$H^s=-100\frac{\sum_{j=1}^{V}\hat{\pi}_j\cdot\log(\hat{\pi}_j)}{\log N},$$

tot i que la distribució de H^s segueix depenent de *N*.

És fàcil comprovar que per a qualsevol *N*, la diversitat absoluta ($N=V$ i $n_j=1/N$) implica que $H^s=100$, mentre que la uniformitat absoluta ($V=1$) implica que $H^s=0$.

El problema d'aquest índex, tant *H* com H^s, és que el seu valor esperat no coincideix amb l'entropia teòrica del vocabulari de l'autor, és a dir, la que s'avaluaria en un text de llargada infinita i que valdria:

$$H_{Teòrica}=-\Sigma\pi_j log\pi_j,$$

on el sumatori és sobre totes les formes de vocabulari, tant les observades com les no observades.

10. *B*: índex de *Brillouin*. És l'equivalent a l'entropia per a vocabularis formats per un nombre finit de formes. Es calcula:

$$B = \frac{1}{N} \log \frac{N!}{\Pi_j n_j!} .$$

Si *N* i totes les n_j creixen tant com per que *logN!* tendeixi a *NlogN-Nloge*, llavors *B* tendeix a *H*, amb $\hat{\pi}_j = n_j/N$. Com en el cas de l'entropia, per a una *N* donada, com més gran és *B* més divers és el llenguatge.

11. *M*: índex de *McIntosh* (McIntosh, 1967), basat en consideracions geomètriques:

$$M = \frac{N - \sqrt{\sum_{j=1}^{V} n_j^2}}{N - \sqrt{N}} .$$

Per a una *N* donada, com més gran és *M*, més ric és el llenguatge.

12. *W*: índex de *Brunet* (Brunet, 1978), utilitzat per Holmes i Forsyth (1995) en l'anàlisi dels *Federalist Papers*, es defineix com:

$$W = N^{V^{-a}} ,$$

on *a* és una constant que pren un valor comprés en l'interval *[0,165 ; 0,172]*. En l'estudi del *Tirant* usarem el mateix valor que fan servir Holmes i Forsyth, *0,170*. Brunet assegura que la distribució de *W* depèn molt poc de la llargada del text i que és específic de cada autor. Com en el cas de l'índex de Simpson, com més petit és *W*, més ric és el llenguatge.

Per més detalls sobre l'ús de les mesures de diversitat es pot consultar Efron i Thisted (1979), Thisted i Efron (1987), Good (1953), Good i Toulmin (1956), Patil i Taillie (1982), Harris (1982), Manly (1994), Holmes (1985), Margalef (1958) i Sichel (1986a,b).

13.2 Estudi dels Índexs per blocs

Per cadascun dels 396 blocs de 1000 paraules consecutives s'han calculat tots els índexs llistats en l'apartat 13.1 excepte:

- el quocient forma/ocurrència (*V/N*), perquè per *N* constant és igual al número de paraules diferents, *V*, multiplicat per una constant,
- l'índex de Yule, *K*, atès que per *N* constant és igual a l'índex de Simpson multiplicat per una constant.

13.2.1 Anàlisi descriptiva univariant de les dades

L'objectiu d'aquesta secció descriure les distribucions dels índexs de diversitat. La figura 13.1 mostra els histogrames dels deu índexs de diversitat. Alguns gràfics mostren formes molt properes a les de la distribució normal: V, V_1, V_2 i V_1/V_2 i R, mentre que els altres gràfics són clarament no simètrics.

La manca de simetria d'alguns d'aquests índexs pot associar-se o be a que tenen una distribució asimètrica o be a que tenint-la simètrica, ens trobem davant una mescla de poblacions que podria indicar estils diferenciats.

13.2.2 Gràfics de Control univariants pels índexs de diversitat

S'obté una sola mesura per bloc per cadascun dels índexs de diversitat. Aquesta mesura és discreta per V, V_1 i V_2, i contínua per a totes les altres. Sigui y_i el valor de qualsevol dels índexs en el bloc *i-èssim*, si l'estil no canvia en tot el llibre, en ser la llargada del bloc $N=1000$ constant, per cada índex es té que:

$$E(y_i)=\mu,$$

$$Var(y_i)= \sigma.$$

Quan l'estil no és homogeni en tot el llibre, tant el valor esperat com la variança dels índexs depenen del bloc, i.

En la figura 13.2 representem l'evolució temporal dels deu índexs de diversitat al llarg del *Tirant*. Es fan servir gràfics de control I per observacions individuals. Per les variables discretes fem servir l'aproximació Normal de la distribució de Poisson. S'han obtingut els límits de control a partir dels rangs mòbils de dues observacions consecutives, tal i com s'ha explicat en el capítol 5.

Alguns d'aquests índexs $(V_2$ i $V_2/V)$ marquen de forma feble un canvi de nivell en el bloc 333. Per a tots els altres índexs de diversitat es pot observarar com l'evolució temporal senyala tres grans etapes perfectament diferenciades. Hi ha una zona central, que inclou pràcticament tota la tercera i gran part de la quarta part del llibre, que és on el llenguatge és més ric i divers. A partir del bloc 334, que correspon al capítol 383, i fins al final del llibre la diversitat disminueix en mitjana i augmenta en variabilitat, el que indica que el llenguatge d'aquesta part final és més pobre que a la resta del llibre. La novetat respecte al que troben Ginebra i Cabos (1998) i al que hem vist en els capítols anteriors és que en els gràfics de control per als índexs de diversitat hi apareix una altra frontera d'estil, més feble que la ja esmentada, entre els blocs 97 i 98, que correspon als capítols 118-119, poc després de l'inici de la tercera part del llibre. Tot i això, en general, la diferència de valors dels índexs entre el començament del *Tirant* i la zona central és més petita que la que hi ha entre aquesta i la part final del llibre. El llenguatge del començament del llibre (potser amb l'excepció dels primers 40 capítols) és més pobre que el de la zona central, entre els capítols 119 i 383, però més ric que el llenguatge de la zona final.

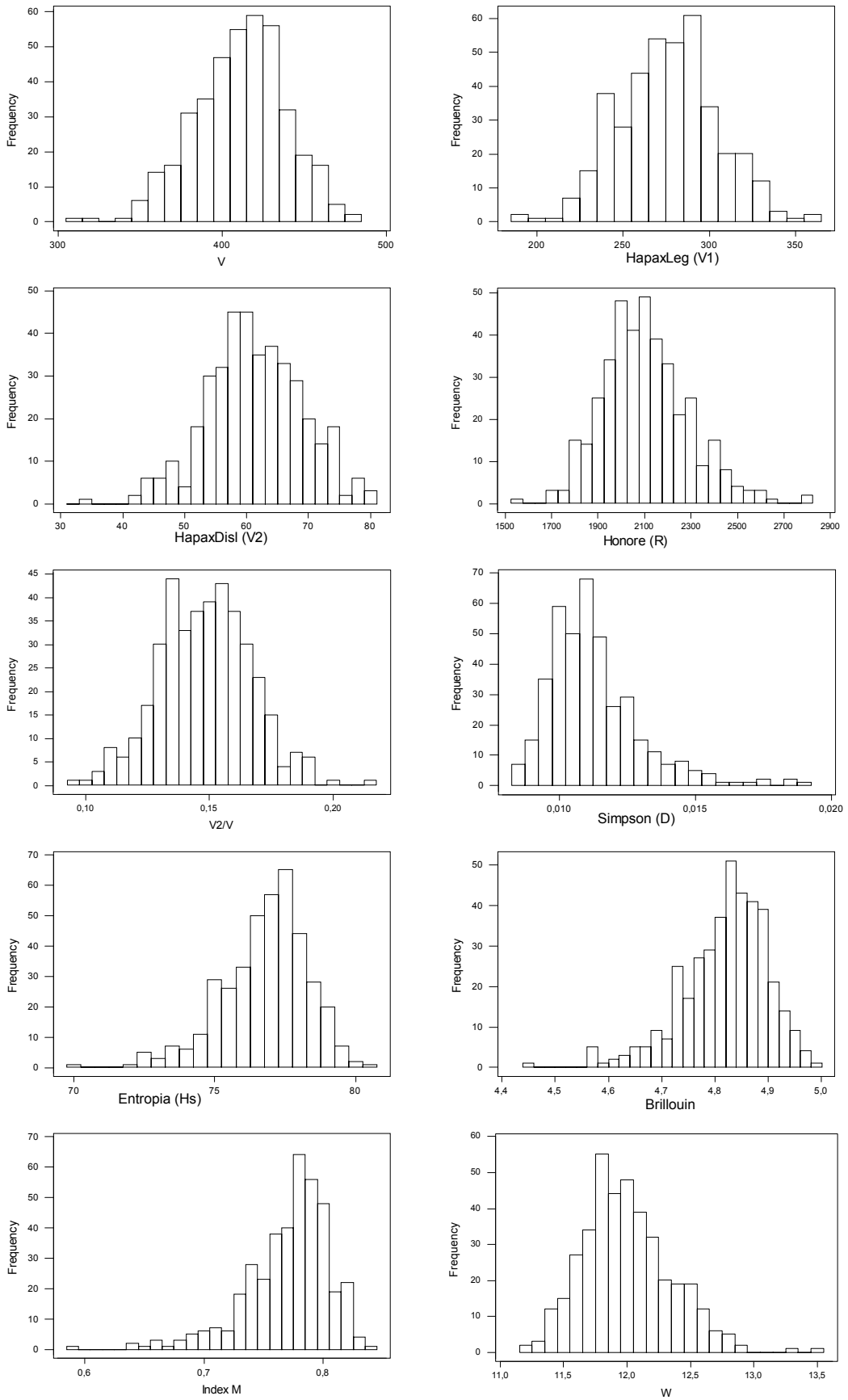

Figura 13.1: Histogrames dels índexs de diversitat pels blocs de 1000 paraules

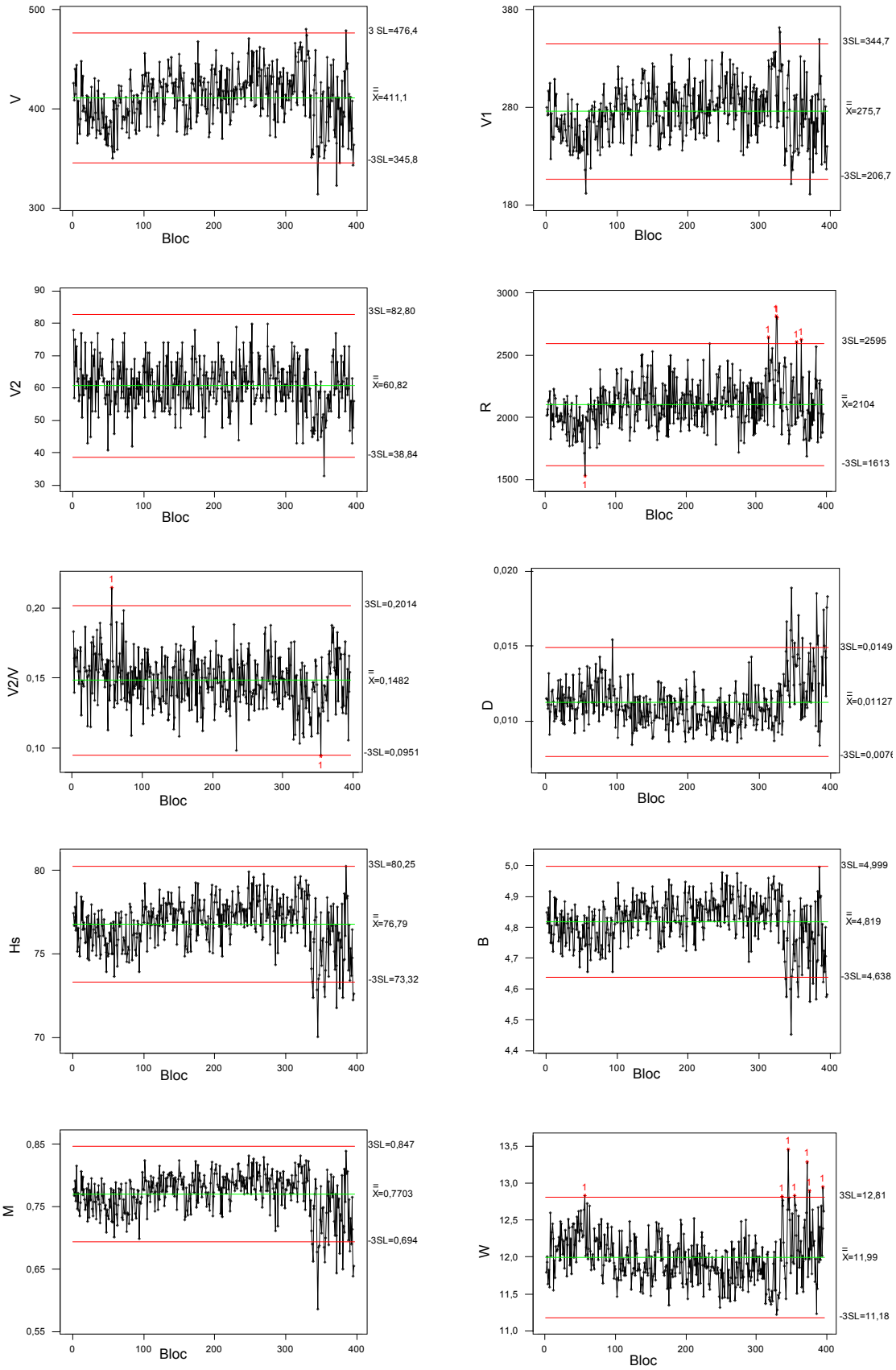

Figura 13.2: Gràfics d'observacions individuals per als índexs de diversitat per blocs de 1000 paraules consecutives en el *Tirant*.

La Figura 13.3 mostra els gràfics de control de sumes acumulades per als deu índexs de diversitat. En el capítol 5 hem descrit com aquests gràfics representen l'evolució temporal de:

$$S_m = \sum_{i=1}^{m} (y_i - \mu_0),$$

per $m=1,2,...,396$,on S_m rep el nom de suma acumulada fins al bloc m, y_i és el valor de l'índex pel capítol i-èssim i μ_0 és el valor nominal o objectiu. En no disposar d'un valor objectiu, s'ha triat com a μ_0 la mitjana de tots els 396 valors de l'índex de diversitat. Tres dels deu gràfics, les sumes acumulades de V_2, R i V_2/V, són poc coincidents amb els altres i poc coincidents també entre ells, com ja havíem observat pels gràfics d'observacions individuals, pel que hem de pensar que no estan mesurant correctament cap de les components de la diversitat de llenguatge. Els gràfics de sumes acumulades pels altres set índexs de diversitat mostren, de nou, tres zones diferenciades: una primera amb diversitat més petita que la mitjana i, per tant, la suma acumulada en aquesta zona és decreixent per aquells índexs que com més grans són més diversitat indiquen (V, V_1, H^s, B i M), i creixent per aquells índexs que com més petits són més diversitat indiquen (D i W). La segona zona, que abasta els blocs del 97 al 333, té una diversitat més gran que la mitjana, i per tant la tendència en els gràfics és contrària a la de la primera zona. La tercera zona, que va del bloc 334 a l'últim, torna a tenir una diversitat inferior a la mitjana, i, per tant, les sumes acumulades dels índexs tornen a comportar-se com en la primera zona.

El fet que S_m canviï de tendència en els blocs 97 i 333 ens fa pensar que en aquests punts hi pot haver canvis en la diversitat del llenguatge utilitzat.

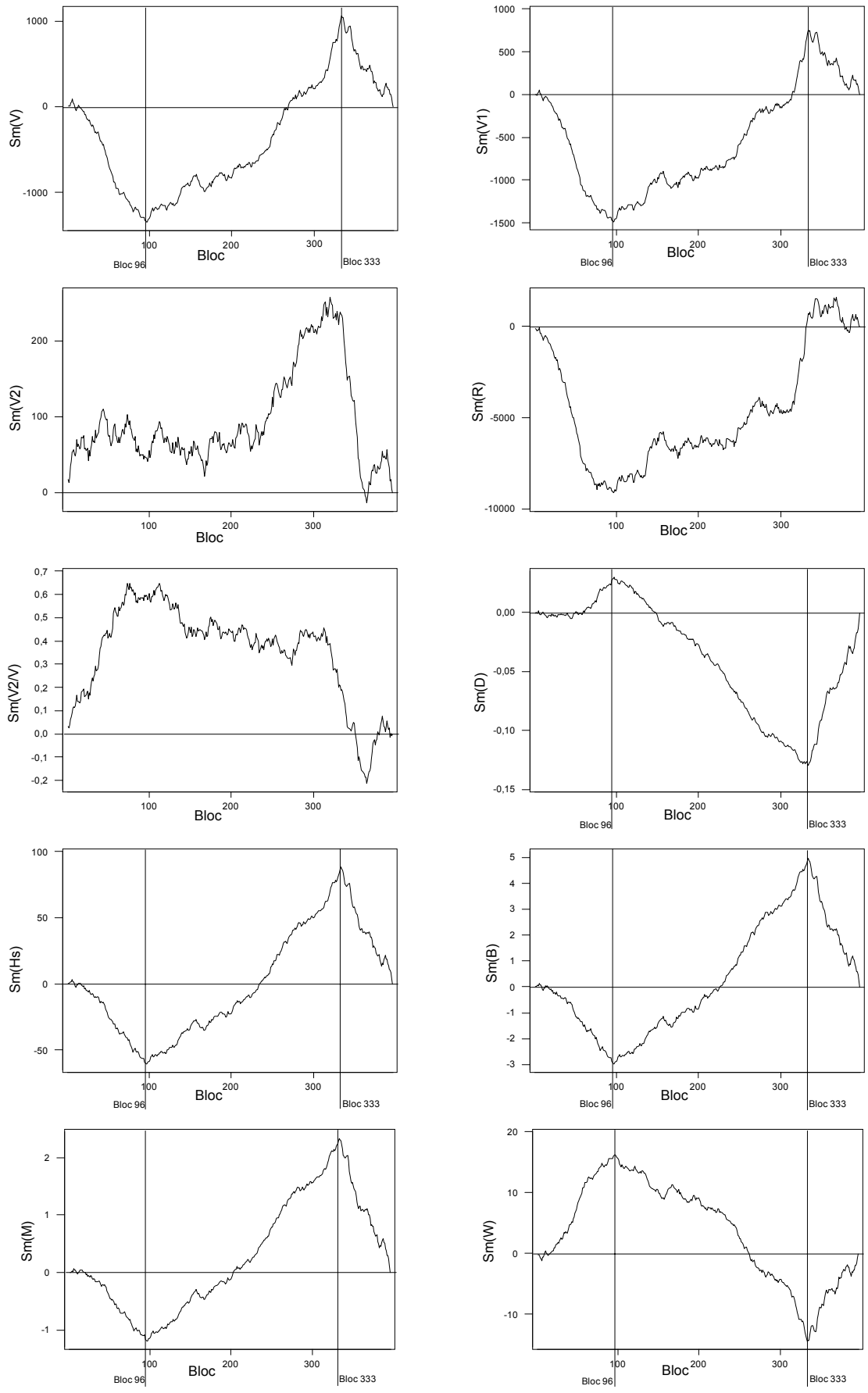

Figura 13.3: Gràfic CUSUM per als índexs de diversitat pels blocs de 1000 paraules.

13.2.3 Anàlisi descriptiva multivariant

En aquesta secció, en primer lloc es veurà una anàlisi gràfica de les relacions entre els índexs i a continuació una anàlisi en components principals, per explotar aquestes relacions per reduir la dimensionalitat.

Relacions bivariants

Els diferents índexs de diversitat mesuren aspectes complementaris de la diversitat del llenguatge. En la figura 13.4 hi ha totes les relacions bivariants entre els índexs de diversitat pels 396 blocs de 1000 paraules.

De l'anàlisi de les relacions bivariants entre índexs podem concloure:

- V_2/V no està correlacionat amb cap altre índex tret dels *Hapax dislegomena* V_2 i, de forma molt més feble, amb l'índex d'*Honoré*, *R*.
- Els *Hapax Dislegomena* V_2 tampoc no estan correlacionats amb altres índexs.
- L'índex *R*, presenta correlacions febles amb gairebé tots els altres índexs de diversitat, tret dels *Hapax Legomena* (V_1).
- L'*Entropia,* H^s, l'índex de *Brillouin, B*, i l'índex *M* tenen una fortíssima relació lineal entre ells, amb correlació positiva i molt propera a 1 (en tots els casos superior a 0,98): els tres índexs mesuren la mateixa component de la diversitat de llenguatge. En l'anàlisi de les relacions bivariants amb els altres índexs es comentarà només les de l'Entropia.
- El nombre de paraules diferents, *V* estan fortament correlacionats amb V_1, atès que les dues magnituds no són independents. Les relacions que presenten aquestes dues variables amb la resta d'índexs són molt semblants, pel que s'analitzarà només *V*.
- L'índex *W* té una relació funcional amb *V* i, les relacions amb els altres índexs són molt semblants a les que té *V*: una forta relació lineal negativa amb H^s i una certa correlació positiva amb l'índex de *Simpson, D.*
- H^s i *D* tenen una molt forta correlació negativa. *D* té correlació, també negativa i més feble, amb *V*, mentre que *V* i H^s tenen una forta correlació lineal positiva. Les relacions entre *D,* H^s *i V* estan ampliades en la figura 13.5. S'observa com els blocs de la tercera zona, que inclou els capítols del 334 al final, en vermell al gràfic, tenen un comportament diferent als anteriors, amb valors dels índex de diversitat que denoten un llenguatge més pobre que l'emprat fins aquell punt.

Fig. 13.4 Relacions bivariants entre els índexs de diversitat. El color blau correspon als primers 96 blocs, el color verd als blocs 97-333 i els vermells als blocs del 334 al final.

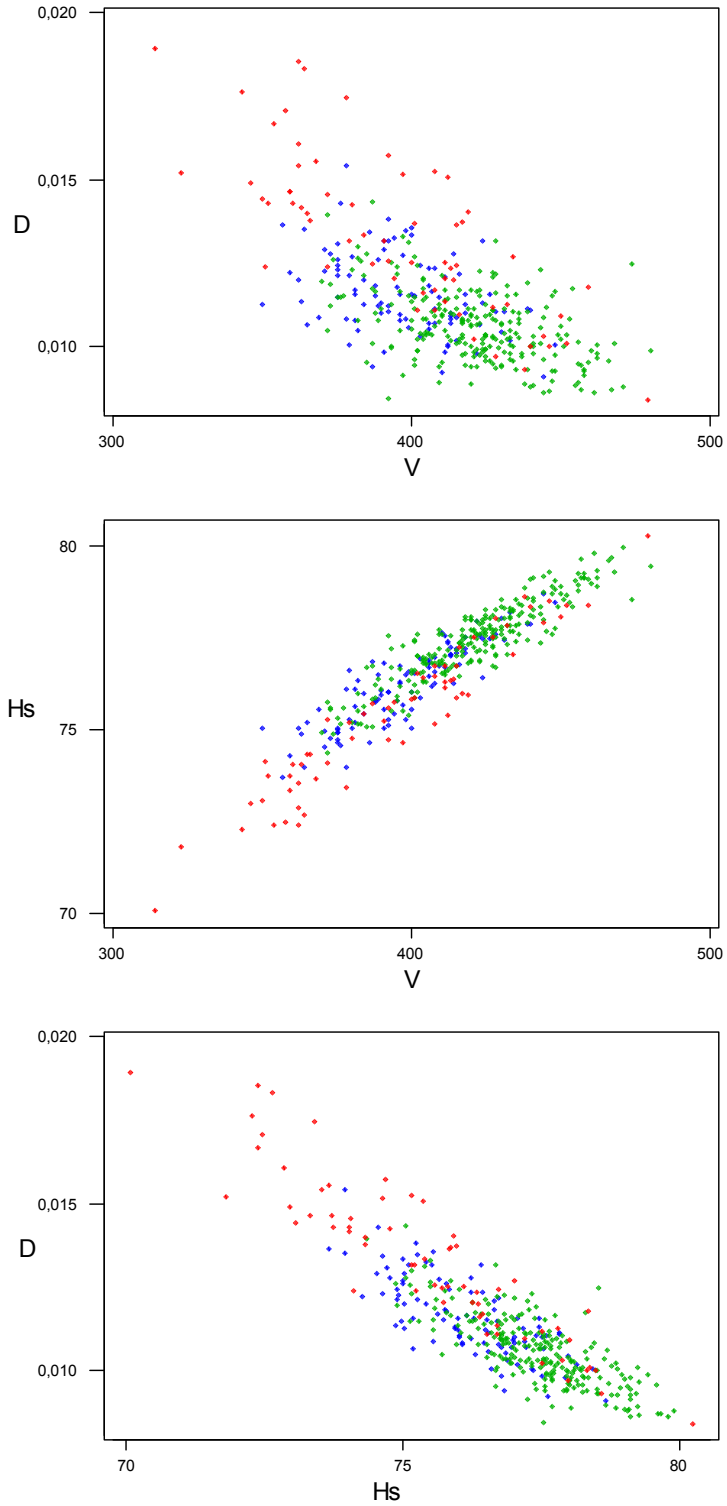

Figura 13.5. Relacions bivariants entre els índexs de diversitat *D, H^s i V* per als blocs de 1000 paraules. Com més grans són *V* i H^s i més petit *D*, més ric i divers és el llenguatge. Hem representat en blau els primers 96 blocs, que corresponen als primers 118 capítols, en verd els blocs 97-333 (capítols 119-383) i en vermell els posteriors.

13.2.3.1 Anàlisi en Components Principals

El fet que els índexs de diversitat estiguin molt correlacionats entre sí permet pensar que, tot i que mesuren quantitats diferents, comparteixen una part important de la informació. És doncs raonable plantejar una reducció en la dimensionalitat de les dades, de forma que es pugui retenir la informació relevant continguda en els índexs en unes poques components, i eliminar la que només aporta soroll. Això ho fem mitjançant l'anàlisi en components principals.

Per a l'anàlisi en components principals s'han fet servir els 10 índexs de diversitat que hem emprat en l'estudi dels blocs. Com que tenen ordres de magnitud molt diferents, per exemple V té mitjana 411 i desviació tipus $s=27,5$; H^s té mitjana $76,9$ i $s=1,5$, i D té mitjana $0,011$ i $s=0,0017$, s'estandaritzen les variables, de forma que totes tinguin mitjana 0 i desviació tipus 1, i evitar així que aquelles amb una variabilitat més elevada tinguin un pes superior a l'hora d'obtenir les components principals.

Els resultats de l'anàlisi mostren com la informació continguda en els índexs de diversitat es pot resumir en les dues primeres components principals. Els coeficients que prenen les variables en la primera component són molt semblants en valor absolut, tret dels Hapax Dislegomena, V_1, i de V_2/V que són molt més petits i R que també té un valor absolut més petit que els altres índexs, tot i que més elevat que els dos esmentats. Tots els coeficients, tret de D i W, tenen el mateix signe. En la figura 13.6 hi ha on tros del llistat de sortida del MINITAB per a l'anàlisi en components principals, on s'hi poden llegir els coeficients que pren cada índex en les components principals.

La primera component és la riquesa i diversitat del llenguatge. S'obté com a un promig dels índexs de diversitat estandaritzats, amb el ben entès que D i W disminueixen quan la diversitat augmenta i, per tant, els coeficients que prenen tenen el signe oposat a tots els altres índexs. Si representem la projecció dels blocs, que es troben en l'espai de dimensió 10 definit pels deu índexs de diversitat, sobre l'espai bidimensional definit per les dues primeres components principals (PCs), veure figura 13.8, observem una ordenació al llarg de la primera component principal en funció de la riquesa i diversitat, com més a l'esquerra es troben en el gràfic el llenguatge és més pobre, com més a la dreta, més ric. A més a més, en el gràfic de control de la figura 13.8 s'observa com la primera PC senyala, de la mateixa manera que ho feien els índexs individualment, una frontera en la diversitat d'estil: els blocs de la part central (del 97 al 333) en verd en el gràfic, són els que tenen un valor més alt de la primera PC i, per tant, són els que tenen una major diversitat; els blocs en blau, corresponents als primers 96 blocs, tenen valors inferiors als de la part central i, per tant, una menor diversitat, mentre que els que estan en vermell, que corresponen als blocs del 334 al final, en vermell són els que tenen un llenguatge més pobre de tots, tot i que n'hi ha alguns, com ja havíem detectat en l'estudi dels gràfics de control, que tenen valors semblants als de la zona central del llibre. Això permet identificar ens quins blocs de la tercera zona la diversitat de l'estil és semblant a la de les dues primeres zones del llibre. La segona PC no es bona per marcar diferències en la homogeneïtat de l'estil: les úniques variables que tenen pes en la segona component són V_2, V_2/V i R, que són els que, com ja s'havia vist anteriorment, comparteixen poca informació amb els altres índexs.

Principal Component Analysis
Eigenanalysis of the Correlation Matrix

Eigenvalue	6,7974	2,3253	0,7764	0,0820	0,0121	0,0045
Proportion	0,680	0,233	0,078	0,008	0,001	0,000
Cumulative	0,680	0,912	0,990	0,998	0,999	1,000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
V	0,373	-0,015	-0,242	0,242	-0,049	-0,345
Simpson	-0,306	-0,180	-0,587	0,500	-0,039	0,379
HapaxLeg	0,354	-0,179	-0,310	-0,085	-0,052	-0,402
HapaxDis	0,096	0,596	-0,374	-0,089	0,088	-0,303
Entropia	0,374	0,113	0,126	0,266	-0,398	0,148
Brilloui	0,366	0,137	0,217	0,232	-0,473	0,253
Index M	0,370	0,101	0,224	-0,128	0,533	0,358
Honore	0,266	-0,405	-0,365	-0,622	-0,224	0,306
V2/V	-0,103	0,611	-0,257	-0,274	-0,166	0,344
W	-0,374	0,006	0,221	-0,271	-0,491	-0,239

Figura 13.6: Coeficients que prenen els índexs de diversitat en les 6 primers components principals.

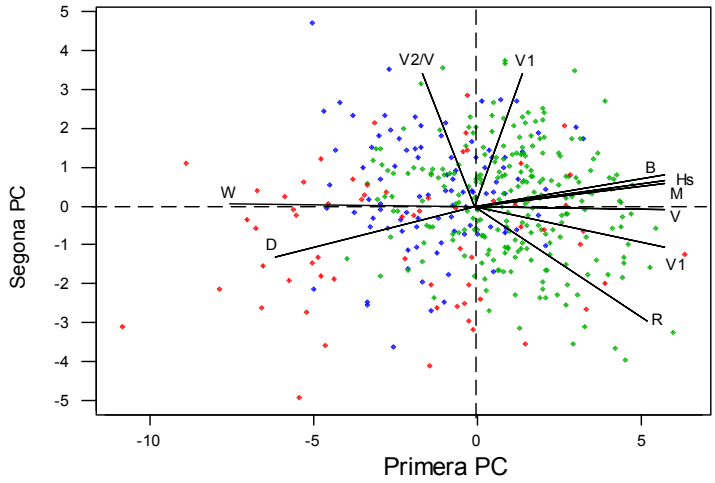

Fig. 13.7 Projecció dels blocs en l'espai definit per les dues primeres components principals (PC). El color blau correspon als primers 96 blocs, fins al capítol 118, el color verd als blocs 97-333 que comprenen els capítols 119 a 382 i els vermells als blocs del 334 al final, corresponents als capítols a partir del 383. Com més a la dreta es trobin els blocs en el gràfic, més gran és la diversitat de l'estil. S'ha superposat la projecció dels deu índexs en el pla format per les dues primeres components principals.

A la figura 13.7 s'ha superposat la projecció dels deu índexs de diversitat sobre l'espai format per les dues primeres components principals. S'hi observa l'estructura de correlacions, ja comentada a l'apartat 13.2.5: tots els índexs, tret de V_2 i de V_2/V, estan fortament correlacionats entre sí. *D* i *W* estan correlacionats negativament amb tots els altres i positivament entre sí, atès que els valors d'ambdós índexs augmenten en disminuir la diversitat de l'estil. V_2 i V_2/V, estan fortament correlacionats positivament entre sí i, encara que de forma feble i negativa, amb *R*. Aquest està moderadament correlacionat amb els altres índexs. A més s'observa com els blocs que es troben més a l'esquerra en el gràfic són els que tenen valors més alts de *D* i *W* i més baixos de *B*, H^s, *M*, *V* i V_1, denotant una diversitat més baixa.

Si l'estil és el mateix en tot el llibre, la distribució de la primera component principal, i per tant dels seus valors esperats i variances, haurien de ser la mateixa per tots els blocs.

L'evolució temporal de la primera component principal, representada en el gràfic de control per observacions individuals de la figura 13.8 obtinguda amb *l=7* índexs de diversitat mostra comportaments diferenciats en les tres zones separades pels blocs 96 i 333: major riquesa i diversitat d'estil ens els blocs centrals en verd en la figura, lleugerament menor en els inicials en blau, força més pobre en els finals, en vermell.

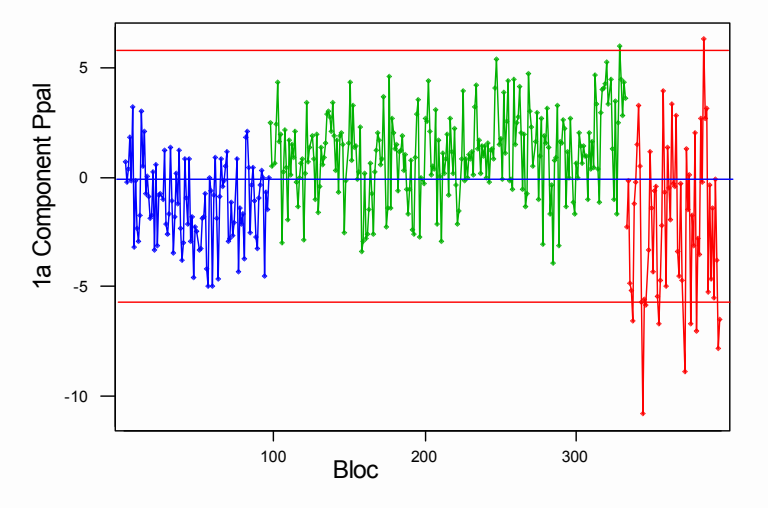

Figura 13.8: Gràfic de control d'observacions individuals per als valors de les projeccions dels blocs sobre la primera component principal obtinguda amb *l=7* índexs de diversitat. S'han representat en blau els primers 96 blocs, en verd els blocs del 97 al 332 i en vermell del 333 al final.

13.2.4 Gràfics de Control multivariants

Si per cada punt a representar en un gràfic de control es disposa d'una mostra de *m* unitats, i per cada unitat es prenen mesures en *l* variables possiblement correlacionades, com és el cas dels índexs de diversitat, en el capítol 5 s'ha descrit com el gràfic T^2 representa l'evolució temporal de l'estadístic:

$$T_i^2 = m(\bar{y}_i - \bar{\bar{y}})' S^{-1} (\bar{y}_i - \bar{\bar{y}}),$$

on $\bar{y}_i = (\bar{y}_{i1},...,\bar{y}_{il})$ és el vector de les mitjanes mostrals per les *l* variables en la mostra i-èssima, per *i=1,...,n*, i on $\bar{\bar{y}} = (\bar{\bar{y}}_1, \bar{\bar{y}}_2,..., \bar{\bar{y}}_l)'$ és el vector de mitjanes històriques per al període en estat de control. $S=(S_{jh})$ és la matriu de covariances de les *l* variables a controlar. Sota la hipòtesi de que les covariances són constants durant els *n* períodes, la covariança entre la variable *j* i la variable *h* es calcula com un promig de les estimacions en les *n* mostres:

$$S_{jh} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{m-1} \sum_{k=1}^{N} (y_{ijk} - \bar{y}_{jk})(y_{ihk} - \bar{y}_{hk}),$$

que en el cas que *j=h*, $S_{jj} = S_j^2$ és la variança mostral de la variable j-èssima. Observem com per al càlcul de la matriu de covariances cal que *m* sigui més gran que 1.

Com que per cada bloc es disposa d'una sola mesura per cadascun dels índexs de diversitat, *m=1*, i, per tant, no és possible calcular la matriu de covariances *S*. Com a aproximació, s'han considerat *mostres mòbils* formades per *m* blocs consecutius, amb

m=2,3,...,10. Així, la primera mostra està formada per els blocs *1, 2, ...m*, la segona pels blocs *2,3,...,m+1*, i la i-èssima *i, i+1, ...i+m*, per *i=1,2,..,n-m*.

Les estimacions que s'obtenen de *S* i, per tant, els gràfics T^2 per a diferents valors de *m* són molt semblants. La figura 13.9 mostra la relació en els valors de T_i^2 per a *m=2* i *m=10*, per a *l=7* índexs de diversitat: *V, V_1, H^s, B*, *M*, *D* i *W*.

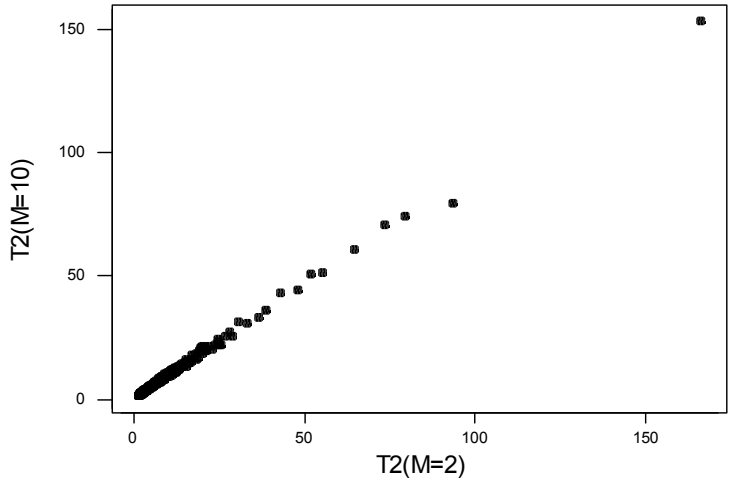

Figura 13.9: Relació entre els valors de T_i^2 per a *m=2* i *m=10*, per a *l=7* índexs de diversitat.

La figura 13.10 mostra el gràfic en seqüència temporal dels valors de l'estadístic T_i^2 per a la seqüència dels *l=7* índexs de diversitat i mostres mòbils de *m=5*. S'han representat en blau els primers 96 blocs, en verd els següents fins al 333 i en vermell des d'aquest punt i fins al final.

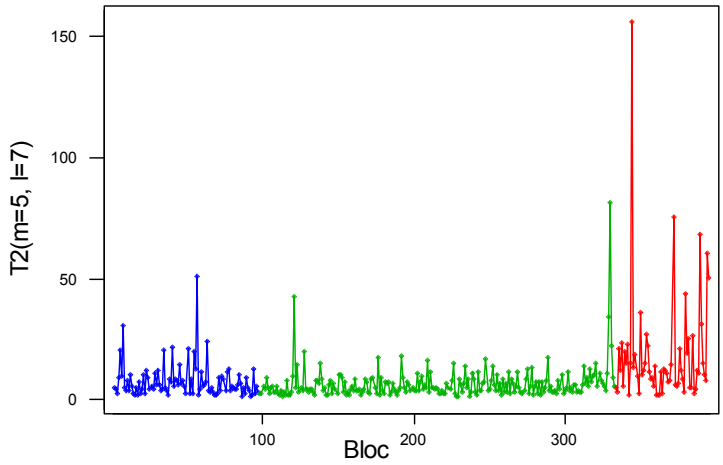

Figura 13.10: Gràfic de control T^2 per a *m=5* i *l=7* índexs de diversitat. S'han representat en blau els primers 96 blocs, en verd els blocs del 97 al 333 i en vermell del 334 al final.

S'observen tres comportaments diferenciats en les tres zones identificades amb colors diferents: els blocs de la zona final, en vermell, són els que tenen valors de T^2 més grans, el que denota un llenguatge més pobre, a més a més d'una major variabilitat, els blocs en la zona central, del 97 al 333, en verd en el gràfic, són els que prenen valors més baixos i, per tant, és on l'estil és més ric i divers, mentre que els 96 blocs inicials, en blau, prenen valors, en general, que es situen entre els altres dos.

13.2.5 Estimació dels punts de canvi

Per a l'estimació del punt de canvi es fa servir, en primer lloc el mètode proposat per a estimar un punt de canvi en seqüències de Normals, i s'aplica a les seqüències dels deu índexs de diversitat i a la primera component principal.

Com que en gràfics de control de set dels deu índexs i de la primera component principal observem la possible presència d'un segon punt de canvi, a continuació fem servir el mètode d'estimació de més d'un punt de canvi per a seqüències Normals i l'apliquem a aquests set índexs i a la primera component principal

13.2.5.1 Estimació d'un punt de canvi

Sota les hipòtesi de independència i Normalitat, els índexs de diversitat es distribuirien:

$$y_i \mid Ind_{1i} \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \sigma^2),$$

per *r=1,2,...,n-1*, on y_i és l'i-èssim element de la seqüència i $Ind_{1i}^{(r)}$ és la variable indicadora que pren valor *0* per les mostres *1,2,...,r* i valor *1* per les mostres *r+1,...,n*.

Per estimar *r*, ajustem el model lineal *n-1* vegades, per $1 \le r \le n-1$, i escollim el valor de *r* que millor ajusta les dades fent servir com a mesura de bondad d'ajust l'estadístic *F* de la taula ANOVA. Sigui F_r l'estadístic *F* obtingut per a *r*. Llavors *r* és estimat com:

$$\hat{r}_N = \max_{1 \le r \le n-1} F_r .$$

La figura 13.11 mostra els gràfics de F_r en funció de *r* per als 10 índexs de diversitat estudiats. Per a quatre d'aquests índexs (*D, H^S, B i M*) el punt de canvi es troba en el bloc 333. Els gràfics mostren un màxim perfectament definit. En altres dos casos, (*V* i *W*) el gràfic presenta dos màxims, un per *r=96* i l'altre per *r=333*. Per a la seqüència de *W* el màxim global es troba per *r=333* mentre que per *V* es troba en *r=96*, tot i que els valors de F_r per a ambdós punts no és molt diferent. Per a la seqüència de *Hapax Legomena*, V_1, és té un sol màxim per *r= 96*, mentre que per l'índex d'*Honoré*, *R*, i per a la relació V_2/V també només hi ha un màxim clarament definit, que s'avança fins als blocs *76* i *73* respectivament. Per a la seqüència de *Hapax Dislegomena*, V_2, és tenen dos màxims molt propers, per a *r=320* i *r=333*.

Fem servir el mateix mètode per a estimar un únic punt de canvi en la seqüència univariant formada pels valors que prenen els blocs un cop projectats sobre la primera component principal. Observem, figura 13.13 esquerra, com el màxim es troba en *r=333* i un màxim local per *r=96*, el que ens fa pensar en un possible doble punt de canvi.

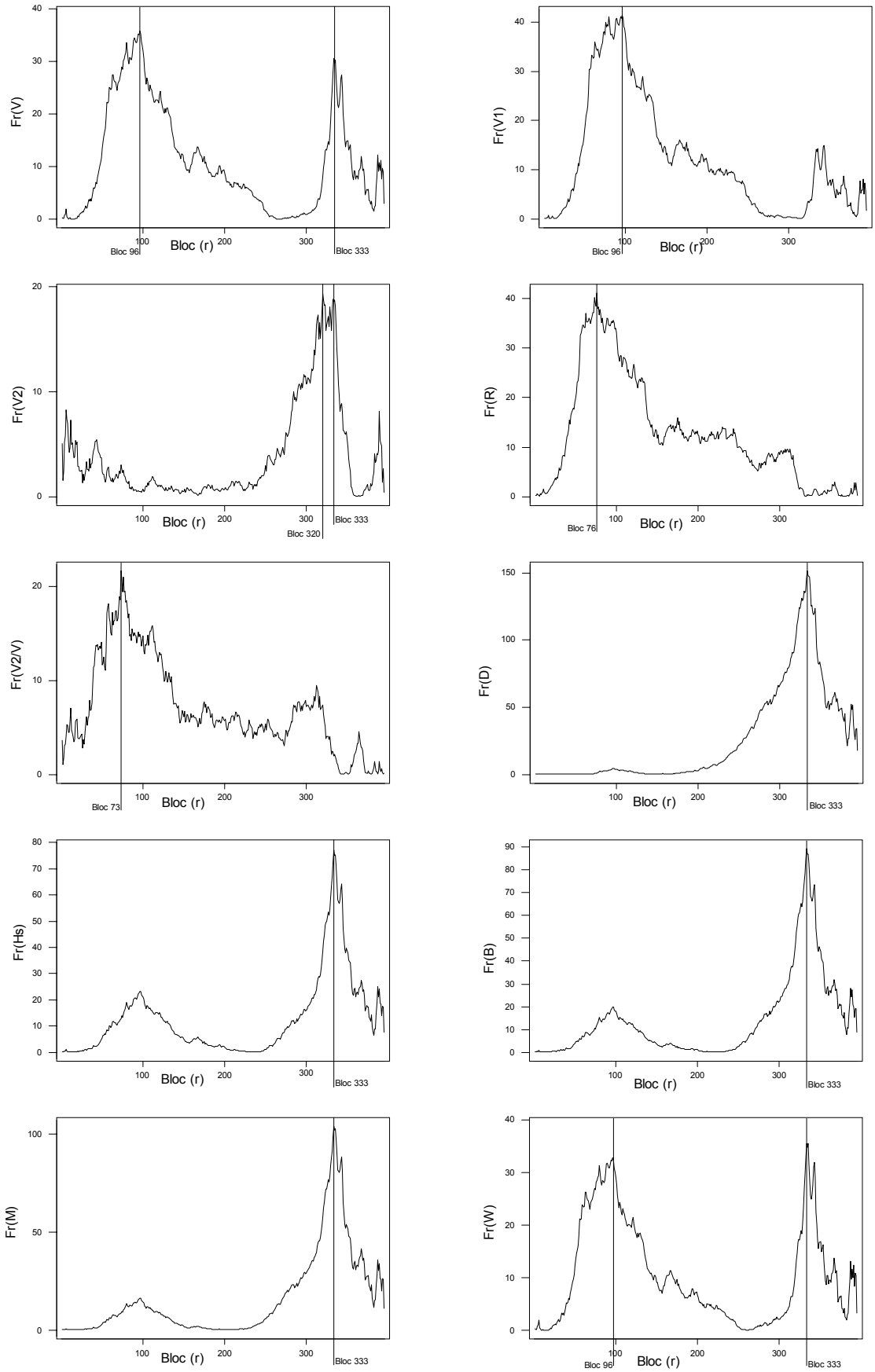

Figura 13.11: Gràfics de F_r en funció de r per a l'estimació d'un punt canvi en els índexs de diversitat.

13.2.5.2 Estimació de dos punts de canvi

La presència dels dos màxims, per *r=96* i per *r=333*, juntament amb el que havíem observat en els gràfics de control, fa pensar en la possible presència de dos punts de canvi en les seqüències dels índexs de diversitat. Per a comprovar-ho, fem servir el mètode d'estimació de més d'un punt de canvi, exposat en l'apartat 6.5, i que consisteix en ajustar $n\cdot(n-1)/2$ models de regressió lineal múltiple:

$$y_i \mid Ind^{(r)} \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + \beta_2^{(r)} Ind_{2i}^{(r)}, \sigma^2),$$

amb $\mathbf{r}=(r_1,r_2)$, $r_1=1,2,...,n-2$ i $r_2=r_1+1,..., n-1$, on y_i és l'i-èssim element de la seqüència i $Ind_{ki}^{(r)}$, per $k=1,2$, són les variables indicadores. Estimem la localització dels punts de canvi, $\hat{r} = (\hat{r}_1, \hat{r}_2)$, com aquella que millor ajusta les dades, fent servir com a mesura de bondad d'ajust F_r.

La figura 13.12 mostra els gràfics de contorn per al valor de F_r en funció de r_1 i r_2 per a les seqüències dels 7 índexs de diversitat que indiquen 2 punts de canvi: *V, Hapax Legomena (V_1), Simpson (D), Entropia (H^s), Brillouin (B), M i W*. Per a tots els 7 índexs $r_2=333$, mentre l'estimació del primer punt de canvi balla entre el bloc $r_1=96$ (per *V*, V_1 i *W*) i el $r_1=97$ (per *D*, H^s, *B* i *M*). Per als set índexs, tots dos punts de canvi són significatius i no hi ha ambigüitats en la determinació del màxim, $\hat{r} = (\hat{r}_1, \hat{r}_2)$.

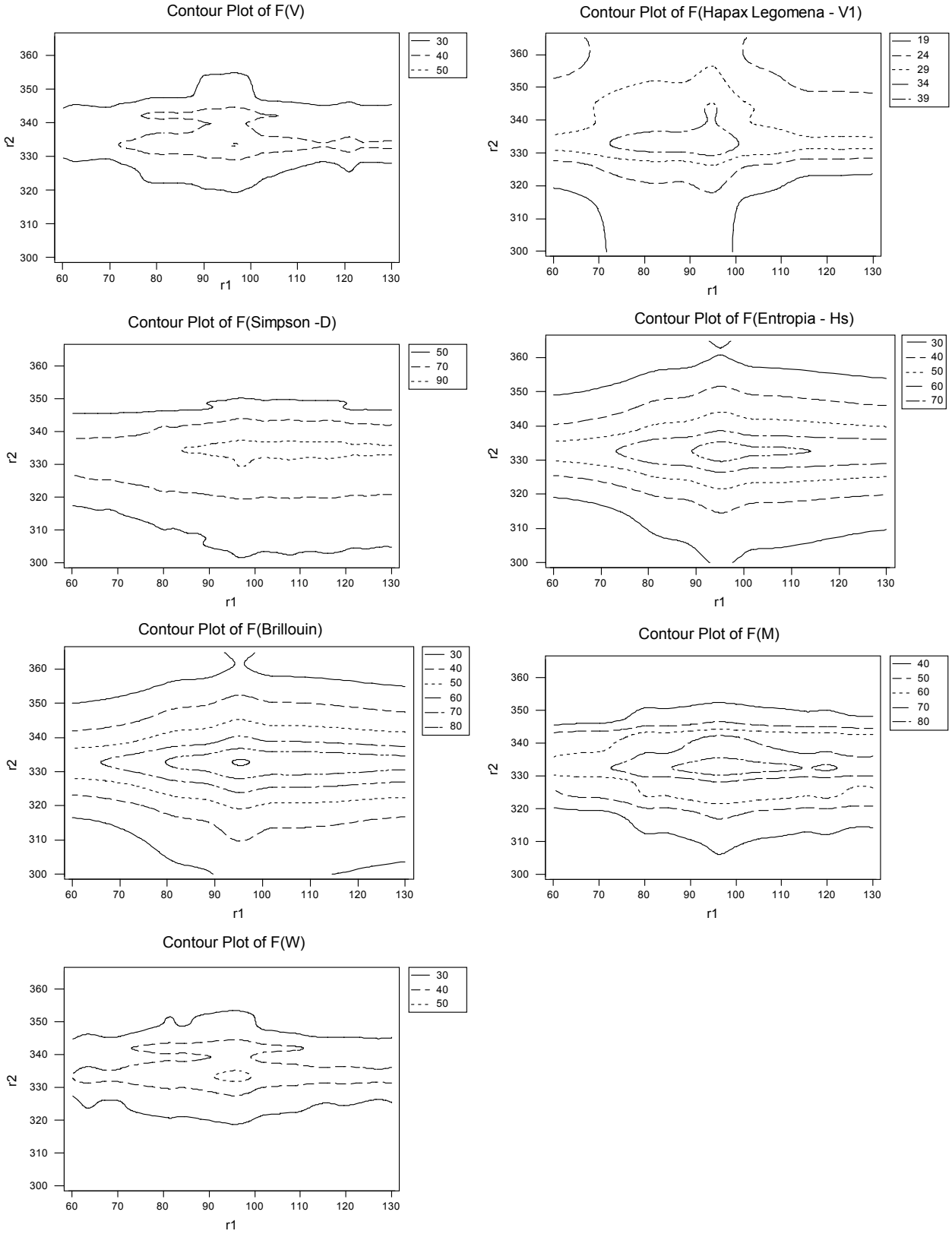

Figura 13.12: gràfics de contorn per al valor de F_r en funció de r_1 i r_2 per a V, V_1, *D*, *Hs, B, M* i *W*.

Si suposem dos punts de canvi en la seqüència de valors de la primera component principal, el millor model ajustat per dos punts de canvi, localitzats a $r_1=96$ i $r_2=333$, supera al millor model per un punt de canvi en els criteris de bondat de l'ajust (F_r, R^2_{Aj}). La figura 13.13 (dreta) mostra les corbes de nivell per al valor de l'estadístic F_r en funció dels valors de r_1 i r_2.

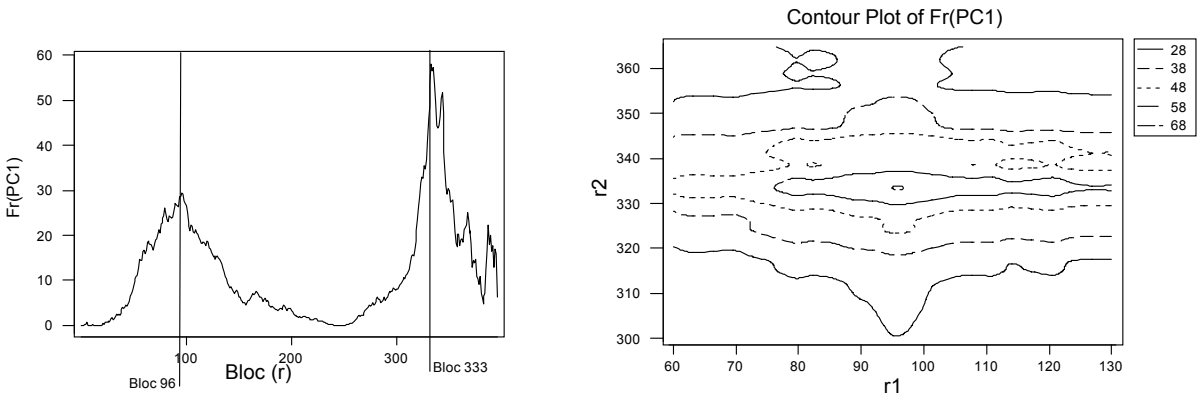

Figura 13.13: gràfics de la relació de F_r en funció de r per a l'estimació del punt de canvi per a la primera component principal. En el gràfic de l'esquerra es suposa un sol punt de canvi, mentre que el de la dreta representa les corbes de nivell de F_r quan es suposen dos punts de canvi a r_1 i r_2.

Per tot això, podem estar bastant segurs que la riquesa i diversitat d'estil presenta dos punts de canvi, el primer situat en el bloc 97 (o bé 96) i el segon en el bloc 333. Aquest segon punt de canvi és molt proper al detectat en l'anàlisi de la llargada de paraula, que oscil·lava entre el bloc 312 i el 327, i igual que l'estimat per les paraules eina.

13.2.5.3 Conclusions

En l'anàlisi de la riquesa i diversitat del vocabulari per blocs observem la presència de dos punts de canvi. El segon, que correspon al punt identificat en els capítols anteriors, es troba en el bloc 333, que correspon aproximadament al capítol 382. El primer canvi, més feble que el segon, l'estimem entre els blocs 96 i 97, que corresponen als capítols 118-119, poc després de l'inici de la tercera part del llibre. La diferència de valors mitjans dels índexs entre el començament del *Tirant* i la zona central és més petita que la que hi ha entre aquesta i la part final del llibre, però totes dues fronteres són clares. El llenguatge del començament del llibre (potser amb l'excepció dels primers 40 capítols) és més pobre que el de la zona central, entre els capítols 119 i 383, però més ric que el llenguatge de la zona final.

La taula 13.1 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de la riquesa i diversitat del vocabulari per blocs.

Distribució	Estadístic	Punt de canvi principal	Punt de canvi secundari
Normal	*Formes, V*	$\hat{r}_2$ =333	$\hat{r}_1$ =96
	Hapax legomena, V_1	$\hat{r}_2$ =333	$\hat{r}_1$ =96
	Hapax disleg., V_2	*r*=320	$\hat{r}_N$ =333
	Index d'HonoréR		$\hat{r}_N$ =76
	V_2/V		$\hat{r}_N$ =73
	Índex de Simpson, D	$\hat{r}_2$ =333	$\hat{r}_1$ =97
	Entropia, H^s	$\hat{r}_2$ =333	$\hat{r}_1$ =97
	Índex de Brillouin,B	$\hat{r}_2$ =333	$\hat{r}_1$ =97
	Índex M	$\hat{r}_2$ =333	$\hat{r}_1$ =97
	Índex de Brunet, W	$\hat{r}_2$ =333	$\hat{r}_1$ =96
	1ª Component Principal	*r=96*	$\hat{r}_N$ =333

Taula 13.1: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la riquesa i diversitat del vocabulari per blocs.

13.2.6 Anàlisi Cluster

En l'anàlisi gràfica dels índexs de diversitat s'ha pogut observar com hi ha blocs posteriors al segon punt de canvi, r_2, que tenen unes característiques, pel que fa a la diversitat del llenguatge, que els fan més semblants als anteriors a r_2 que als posteriors. Encara que d'una manera molt menys accentuada, també hi ha alguns blocs anteriors r_2 que poden semblar-se als posteriors. Aquesta situació s'ha observat en els gràfics de control tant pels índexs de diversitat presos individualment com per la primera component principal com per el gràfic T^2. Per aquest motiu, i tal i com havíem fet per a l'anàlisi de la llargada de paraula i de les paraules eina, hem decidit fer una anàlisi cluster, per veure si és possible recol·locar algun bloc en el grup en el que és més homogeni, fent servir les tècniques d'anàlisi cluster clàssica exposades en el capítol 7, a partir de les dades on les unitats són els 396 blocs de 1000 paraules i les columnes són els índexs de diversitat. De tots els índexs, farem servir els set que hem observat que expliquen la diversitat del llenguatge: *V, V_1, D, H^s, B, M* i *W*, que estandaritzem de manera que tinguin mitjana *0* i variança *1*, per tal que tots tinguin el mateix pes en l'anàlisi.

L'anàlisi cluster clàssica basada en particions, tipus *k-means*, pressuposa que hem de saber quants clusters existeixen abans de realitzar l'anàlisi. Hem fet l'anàlisi en dos casos: suposant que hi ha dos grups i suposant que n'hi ha tres. Els resultats obtinguts en les dues anàlisi són compatibles: la divisió en clusters separa els blocs en funció de la seva diversitat, bàsicament en funció del valor de la primera component principal. La figura 13.14 mostra els gràfics amb els resultats les dues anàlisi: suposant dos (esquerra) i tres (dreta) clusters, on s'han representat en colors diferents els blocs que han estat assignats a clusters diferents. La diferència en les classificacions en dos i tres clusters està en que, en el segon cas, els blocs amb valors de la primera PC propers a *0* han estat

agrupats en el tercer cluster. Aquests blocs, en l'agrupació en dos clusters queden repartits entre les dues agrupacions, en funció de la primera component principal.

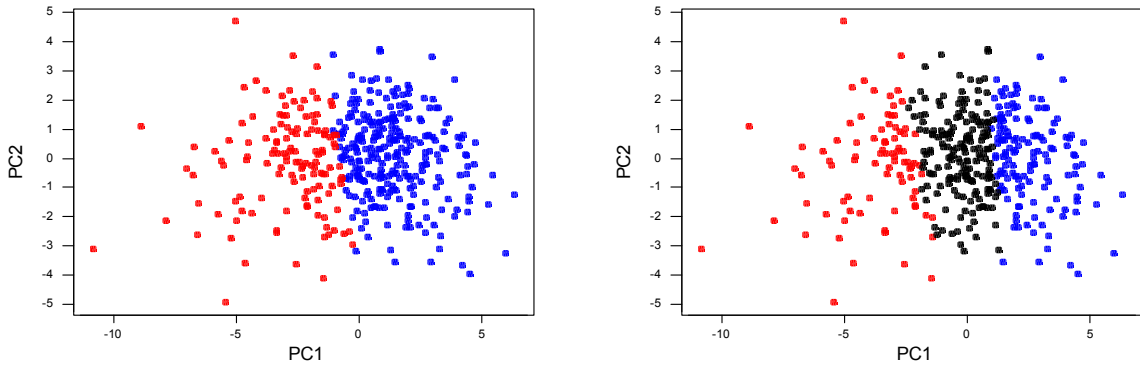

Figura 13.14: gràfics amb la separació dels blocs en 2 clusters (esquerra) i en 3 (dreta). S'han representat en colors diferents els blocs que han estat assignats a clusters diferents: en color blau els blocs amb una major riquesa de llenguatge i en vermell els de vocabulari més pobre. Els blocs en negre tenen una riquesa "mitjana".

En l'annex A13.1 hi ha els llistats en els que hi ha les assignacions dels blocs als clusters, tant per quan es suposen dos com tres clusters.

Per validar els resultats obtinguts quan es suposen dos clusters, es compara l'ajust del model emprat per a la determinació d'un punt de canvi:

$$y_i \mid Ind_{1i}^{(r=333)} \sim N(\mu = \beta_0^{(r=333)} + \beta_1^{(r)} Ind_{1i}^{(r=333)}, \sigma^2),$$

on $Ind_{1i}^{(r)}=0$ per $i=1,...333$ i $Ind_{1i}^{(r)}=1$ per $i=334,...,396$ amb el model:

$$y_i \mid Ind_{1i}^{(c)} \sim N(\mu = \beta_0^{(c)} + \beta_1^{(c)} Ind_{1i}^{(c)}, \sigma^2),$$

on $Ind_{1i}^{(c)}=0$ pels blocs assignats a un cluster i $Ind_{1i}^{(c)}=1$ pels blocs assignats a l'altre. Quan y_i és el valor de la primera PC per als blocs, l'ajust es veu molt millorat després de l'anàlisi cluster, passant el valor de F de la taula ANOVA de 57,8 pel model donat pel punt de canvi a 638 pel model donat pels clusters i el valor de la R^2 passa del 12,6% al 61,7%. Millores en l'ajust semblants es donen també si es prenen com a resposta cadascun dels 7 índexs de diversitat considerats en l'anàlisi cluster.

Quan es suposa que hi ha 2 punts de canvi i per tant tres clusters, comparem model de regressió lineal múltiple ajustat per a l'estimació dels punts de canvi:

$$y_i \mid Ind_i \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + \beta_2^{(r)} Ind_{2i}^{(r)}, \sigma^2) \qquad \text{per } r=(96,333),$$

on $Ind_{1i}^{(r)}=1$per $i=1,...96$ i és igual a *0* pels blocs *97, ... 396*, i $Ind_{2i}^{(r)}=0$ per $i=1,...,96$ i $i=334,...,396$, i és igual a *1* pels blocs *97, ..., 333*, amb el model:

$$y_i \sim N(\mu = \beta_0^{(c)} + \beta_1^{(c)} Ind_{1i}^{(c)} + \beta_2^{(c)} Ind_{2i}^{(c)}, \sigma^2)$$

on $Ind_{ki}^{(c)}=1$ pels blocs assignats al cluster k, per $k=1,2$, i $Ind_{ki}^{(c)}=0$ pels blocs assignats als altres dos clusters, per $i=1,...,396$. També quan suposem 2 punts de canvi i 3 clusters l'ajust millora molt després de realitzar l'anàlisi cluster, passant el valor de F de la taula ANOVA de 68,7 pel model ajustat en l'estimació dels punts de canvi a 734,3 pel model ajustat pels clusters, i el valor de la R^2 *ajustada* passa del 25,5% al 78,8%. Millores semblants en l'ajust es donen també, com ja s'observava pel cas d'un sol punt de canvi i dos clusters, si es prenen com a respostes cadascun dels 7 índexs de diversitat considerats en l'anàlisi cluster .

13.3 Estudi dels Índexs per capítols

Tal i com s'ha comentat a l'apartat 13.2, l'índex de Yule, K_i, i la relació forma/ocurrència, V_i/N_i, pel cas de llargada del text constant, $N_i=N$, són equivalents, respectivament, a l'índex de Simpson D_i i al nombre de paraules diferents V_i. Quan s'estudien els capítols, la llargada N_i és variable. Pel rang de llargades de capítol que estudiem, N_i més gran de *200*, K_i és pràcticament idèntic a D_i multiplicat per una constant:

$$K_i = 10000 D_i \left(1 - \frac{1}{N_i}\right)$$

pel que tampoc no l'analitzem. Dels índexs estudiats pels blocs de 1000 paraules, pels capítols no estudiarem aquells que s'han revelat poc útils per avaluar la riquesa i diversitat de l'estil literari: V_{2i}, V_{2i}/V_i i R_i.

13.3.1 Dependència dels índexs respecte a la llargada de capítol N_i

La distribució de tots els índexs de diversitat considerats depèn de N_i i, per tant, del capítol.

Per un autor amb diversitat de vocabulari donada, com més llarg és el text més gran és el valor esperat del nombre de paraules diferents, V_i, i de l'índex de *Brillouin*, B_i, i més petit és de l'Entropia, H^s_i, mentre que el valor esperat dels *Hapax Legomena*, V_{1i}, en el rang de llargades de capítols estudiat, de *0* a *6514*, augmenta en augmentar la llargada del text, però cal suposar que no creixeran indefinidament. La relació V_i/N_i decreix en créixer N_i, atès que per N_i petites gairebé totes les paraules són diferents, i la relació pren valors propers a *1*, mentre que en créixer la llargada de capítol, cada vegada trobem més paraules repetides.

El valor esperat de l'índex de Simpson, D_i, no depèn de N_i, i a la vista dels gràfics de la figura 13.15, el mateix passa amb l'índex M_i, però la seva variança sí que depèn de N_i.

La figura 13.15 mostra com varien els índexs de diversitat en funció de la llargada del capítol, N_i, després d'eliminar les paraules en cursiva. Com que dos dels capítols estan desdoblats i 19 estan escrits íntegrament en cursiva, el gràfic es basa en els 470 capítols amb N_i positiva. Els diagrames també estan estratificats amb tres colors per identificar els capítols de cadascuna de les zones limitades pels capítols 119 i 383 que havíem detectat en l'estudi dels blocs.

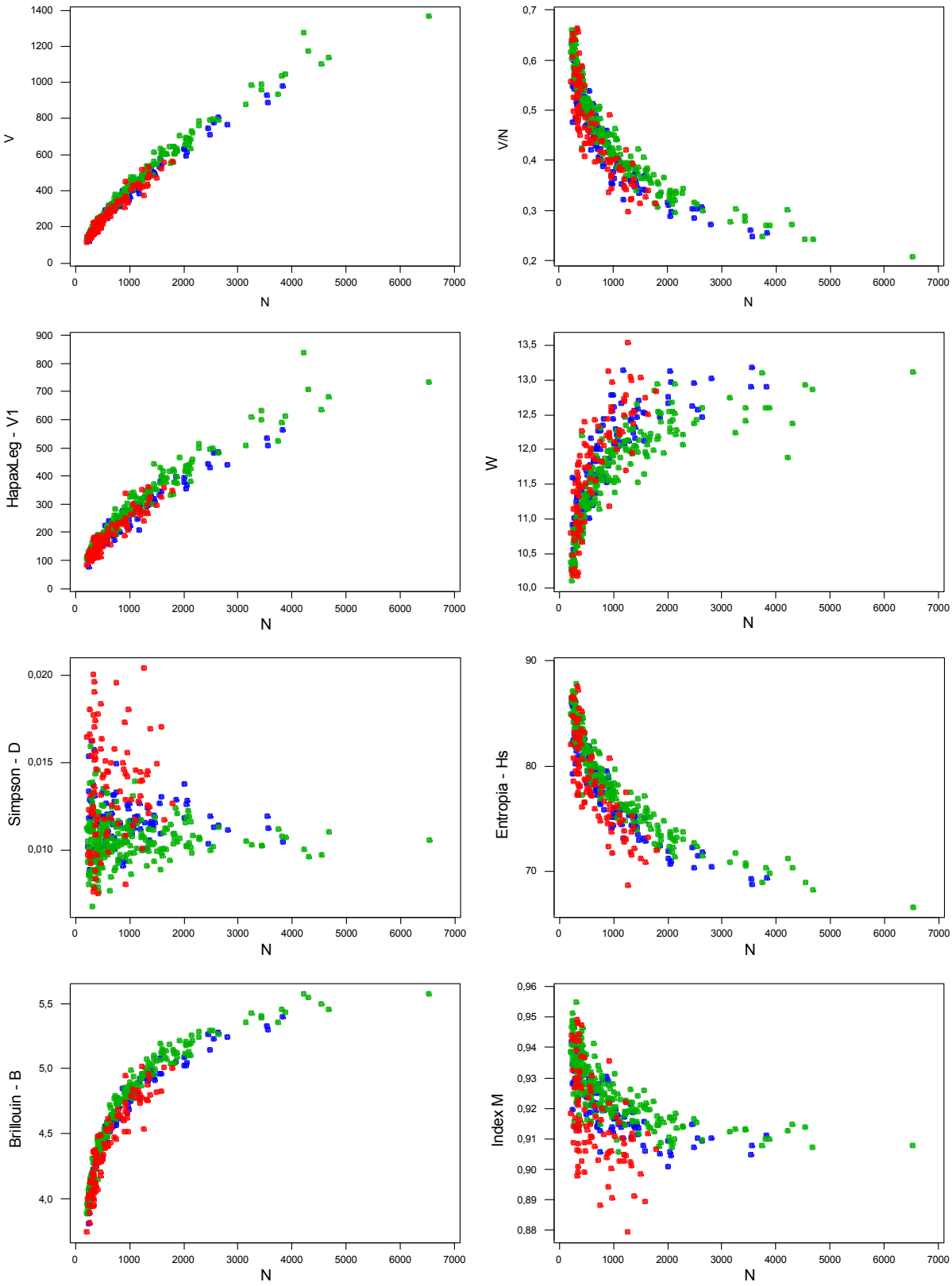

Figura 13.15: Relació dels índexs de diversitat amb la llargada del capítol després d'eliminar les paraules en cursiva. Els diagrames estan estratificats amb tres colors per identificar els capítols de cadascuna de les zones limitades pels capítols 119 i 383.

A la figura 13.15 s'observa empíricament com la variabilitat de D_i depèn N_i. Good (1982) dóna la fórmula de la variança de D_i en funció de la llargada del text N_i:

$$Var(D_i) = 2N_i^{-1}(N_i - 1)^{-2}\left[2N_i^2\left(\rho_{i3} - \rho_i^2\right) + N_i\left(\rho_i^2 + \rho_i - 6\rho_{i3}\right) + \left(4\rho_{i3} - \rho_i - 3\rho_i^2\right)\right]$$

on:

$$\rho_i = \sum_i \pi_{ji}^2,$$

$$\rho_{i3} = \sum_i \pi_{ji}^3,$$

i π_{ji} és la probabilitat d'aparició de la paraula j-èssima en el capítol i-èssim. Si en lloc dels valors desconeguts π_{ji} fem servir les seves estimacions,

$$\hat{\pi}_{ji} = \frac{n_{ji}}{N_i},$$

on n_{ji} és el nombre d'ocurrències de la paraula j-èssima en el capítol i-èssim, obtenim una estimació de $Var(D_i)$. En la figura 13.16 hem representat l'invers d'aquesta estimació de $Var(D_i)$ en funció de la llargada de capítol N_i. Observem com tenen una forta relació lineal.

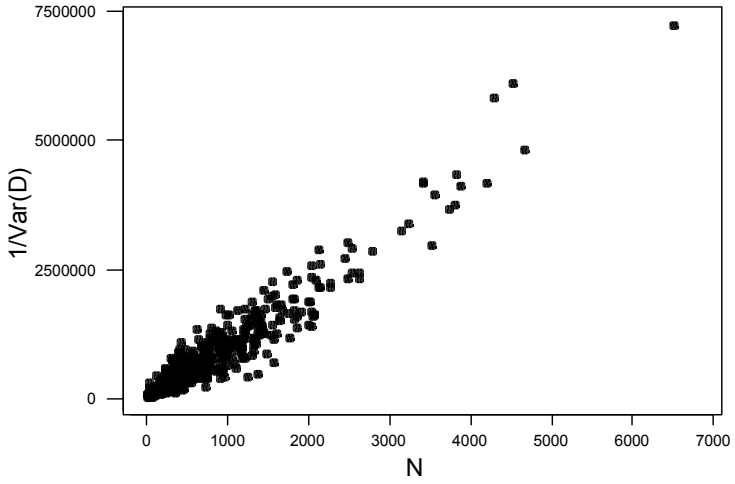

Figura 13.16: Relació entre l'invers de l'estimació de la *Var(D)* i la llargada de capítol. L'estimació de *Var(D)* s'ha fet prenent proporció d'aparicions de la paraula j-èssima en el capítol i-èssim en lloc de la probabilitat. S'observa una forta relació lineal.

Pel que fa a l'índex *M* no coneixem cap expressió de la seva variança, però a la vista del gràfic que el relaciona amb N_i podem pensar que aquesta es comporta com $Var(D_i)$.

13.3.2 Evolució temporal de D_i i M_i

La figura 13.17 mostra la representació gràfica de les seqüències temporals de D_i, esquerra, i M_i, dreta. Són els dos únics índexs que el seu valor esperat no depèn de N_i. En ella s'observa com a partir del capítol 383, D_i experimenta un canvi: la majoria dels valors es troben per sobre de la mitjana. També observem com alguns capítols posteriors al 383, en vermell al gràfic, prenen valors de D_i propers o inferiors a la mitjana. En els capítols del primer al 118, en blau en el gràfic, D_i es troba principalment per sobre de la mitjana encara que amb valors clarament inferiors als que pren en els capítols posteriors al 383, mentre que entre els capítols 119 i 382, en verd a la figura, es troben gairebé tots per sota del valor mig de D_i. Això indica que la part central del

Tirant, que abasta els capítols 119-382 és la més rica en llenguatge, que la última part és la més pobra o, al menys, la que té els capítols amb un llenguatge més pobre, mentre que els primers 118 capítols tenen una riquesa intermitja.

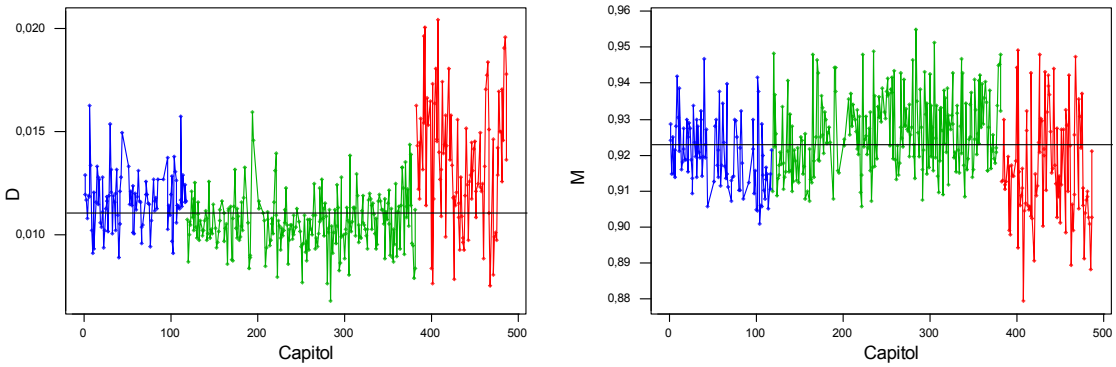

Figura 13.17: Evolucions al llarg del *Tirant* de D_i i M_i, per capítols de més de 200 paraules. En blau hi ha els capítols 1-118, en verd del 119 al 382 i en vermell del 383 al final.

L'anàlisi del gràfic temporal per a l'índex M de la figura 13.17, dreta, mostra exactament les mateixes característiques que el de D, amb la única diferència que major diversitat vol dir M més gran, mentre que implica D més petit. Per tant, els capítols des del 383 al final són els que tenen valors més petits de M_i, els capítols centrals, del 119 al 382 prenen valors bàsicament per sobre de la mitjana, mentre que els primers 118 capítols es troben en valors intermitjos.

13.3.3 Estimació del punt de canvi

S'han fet servir dues metodologies diferents per a estimar els punts de canvi. En primer lloc s'ha fet servir la mateixa metodologia emprada per als blocs per estimar els punts de canvi per a l'índex de Simpson, D_i, i s'aplica en el cas que es suposa un únic punt de canvi i quan se'n suposen dos. El mètode és el proposat per seqüències de Normals amb variança no constant, explicat al capítol 6.

A continuació, s'aprofita la dependència dels altres índex de N_i per estimar els punts de canvi per a les seqüències de V_i, V_{1i}, H^s_i i B_i mitjançant l'ajust de models lineals que expliquen la dependència d'aquests índexs en funció de N_i.

13.3.3.1 Estimació del(s) punt(s) de canvi per a D_i

Estimem la localització d'un sol punt de canvi per la seqüència de D_i ajustant *n-1* models lineal de regressió simple amb pesos, assumint independència entre observacions:

$$y_i \mid Ind_{1i}^{(r)} \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \sigma_i^2 = \frac{\sigma^2}{N_i}),$$

per *r=1,2,...,n-1*,on y_i és el valor de l'índex de Simpson pel capítol i-èssim i $Ind_{1i}^{(r)}$ és una variable indicadora que pren valor *0* per als capítols *1,2,...,r* i valor *1* per als capítols *r+1,...,n*. Estimem el punt de canvi com aquella *r* per la que el model ajusta millor les dades, pel criteri de màximitzar F_r.

Representem, en la figura 13.18, l'evolució de F_r en funció de *r*. El màxim del gràfic, i per tant la millor estimació de *r*, es té per *r=382*. Hi ha un altre màxim local a *r=118*, molt inferior en valor de F_r, però que coincideix amb el que hem trobat per blocs.

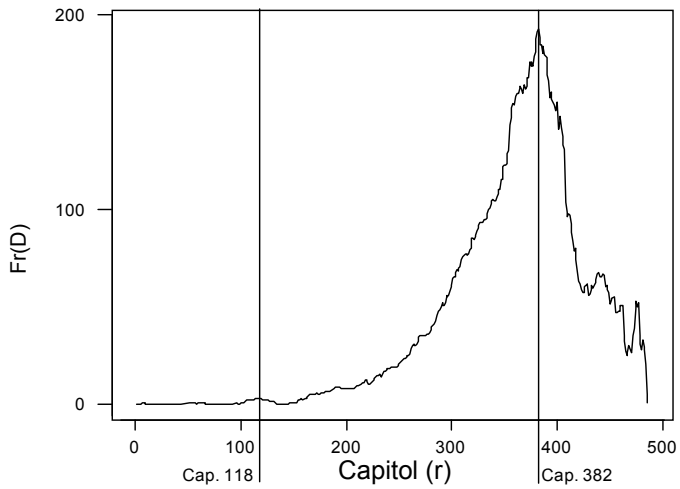

Figura 13.18: Gràfic de F_r en funció de r per l'índex de Simpson, D_i, per als capítols de més de 200 paraules. El màxim es troba a *r=382*, i es pot observar un màxim local a *r=118*.

La presència del màxim local a *r=118*, juntament amb el que havíem observat en els gràfics de control i de la informació que hem obtingut de l'anàlisi dels blocs, ens fa pensar en la possible presència de dos punts de canvi en la seqüència de dades. Per comprovar la possible existència dels dos punts de canvi emprem l'extensió del mètode de detecció a més d'un punt de canvi, consistent en ajustar els $n\cdot(n-1)/2$ models de regressió lineal múltiple amb ponderant les observacions:

$$y_i \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + \beta_2^{(r)} Ind_{2i}^{(r)}, \sigma_i^2 = \frac{\sigma^2}{N_i}),$$

amb $\mathbf{r}=(r_1,r_2)$, $r_1=1,2,...,n-2$ i $r_2=r_1+1,..., n-1$, on y_i és l'i-èssim element de la seqüència i $Ind_{ki}^{(r)}$ són les dues variables indicadores. El pes assignat a la observació i-èssima és $w_i=N_i$. Estimem els punts de canvi, $\hat{r} = (\hat{r}_1, \hat{r}_2)$, com aquells pels que el model ajusti millor les dades, essent el criteri de bondat de l'ajust el màxim de F_r.

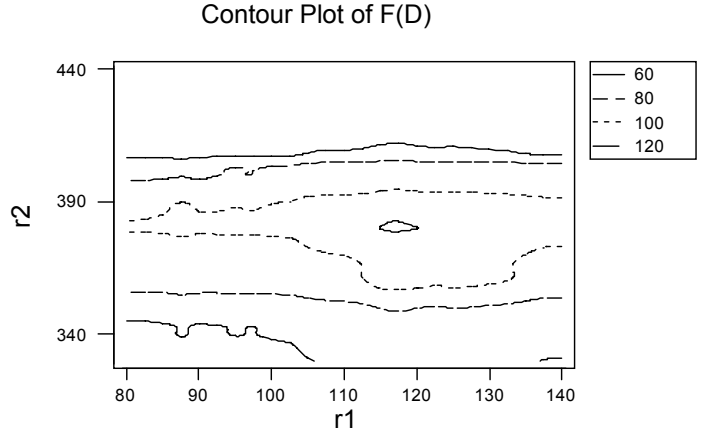

Figura 13.19: Gràfic de F_r en funció de $r=(r_1, r_2)$ per l'índex de Simpson, D. El màxim es troba per *r=(118,382)*.

13.3.3.2 Estimació dels punts de canvi per a V_i, V_{1i}, H^s_i i B_i

En la present secció s'estimen els dos punts de canvi en les seqüències dels índexs V, V_1, H^s i B, tenint en compte la dependència dels índexs respecte a la llargada del capítol.

En la figura 13.15 observem com els valors esperats de V_i i V_{1i} creixen en augmentar la llargada del capítol, fet força intuïtiu. $E(B_i)$ també creix, mentre que el valor esperat de l'Entropia, $E(H^s_i)$, disminueix en augmentar el nombre de paraules en el text. Observem en la figura 13.20 a més a més, com les transformacions logarítmiques de V_i, V_{1i}, H^s_i i B_i tenen una relació lineal força marcada amb el logaritme de N_i.

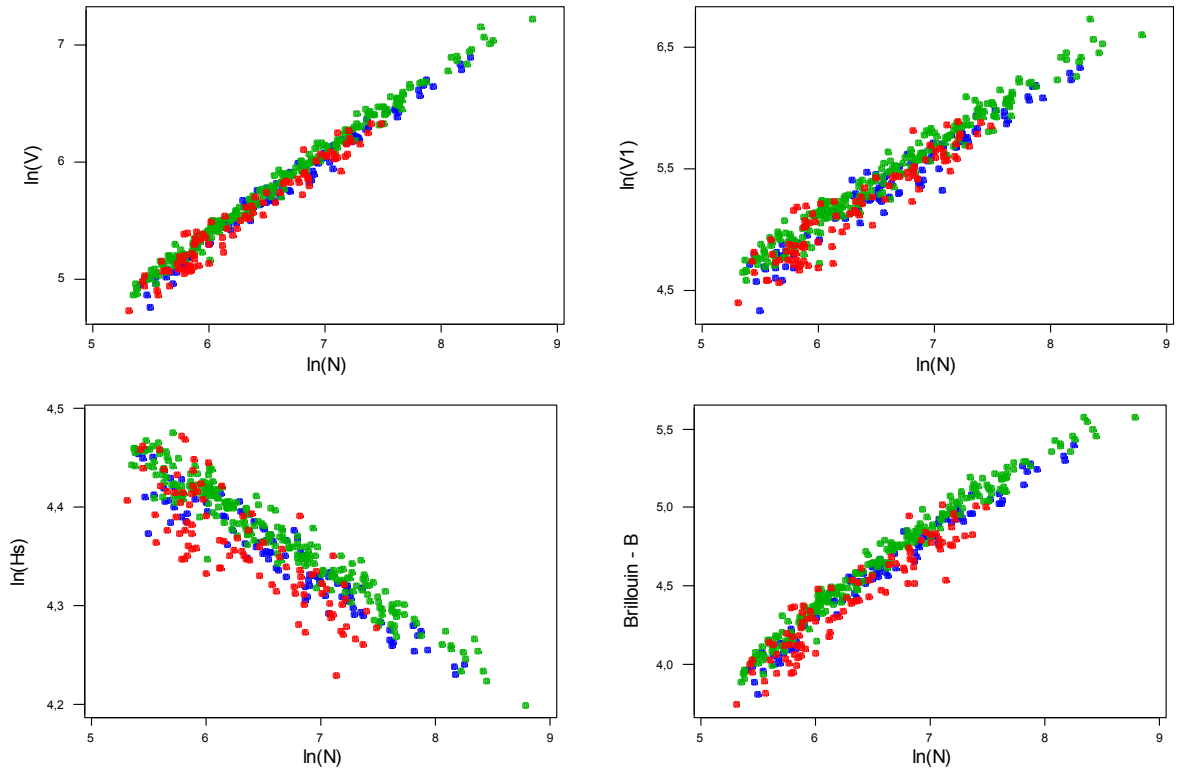

Figura 13.20: diagrames bivariants amb les relacions entre $ln(V_i)$, $ln(V_{1i})$, $ln(H^s_i)$ i B amb $ln(N_i)$. Els punts en blau corresponen als capítols 1-118, el punts en verd als capítols 119-383 i en vermell els capítols 384 i final.

Per cadascuna de les quatre variables ($ln(V_i)$, $ln(V_{1i})$, $ln(H^s_i)$ i B_i) ajustem els $n\cdot(n-1)/2$ models:

$$y_i \sim N\left(\beta_0 + \beta_1 Ind_{1i} + \beta_2 Ind_{2i} + \beta_3 \ln(N_i) + \beta_{13} Ind_{1i}\cdot\ln(N_i) + \beta_{23} Ind_{2i}\cdot\ln(N_i), \frac{\sigma^2}{N_i}\right), \qquad (13.1)$$

amb $\boldsymbol{r}=(r_1,r_2)$, $r_1=1,2,...,n-2$ i $r_2=r_1+1,..., n-1$.

La resposta y_i és un dels quatre índexs, i $Ind_{ki}^{(r)}$, per $k=1,2$, són les variables indicadores que prenen valor *0* per $i=1,2,...,r_k$ i *1* altrament. Estimem els dos punts de canvi com $\boldsymbol{r}=(r_1,r_2)$ tal que F_r, el valor de l'estadístic F de la taula ANOVA és màxim. Per a les quatre variables s'obté que $r=(118, 382)$.

Per tal de validar els resultats, contrastem la hipòtesi nul·la:

$$H_0 : \beta_1=\beta_2=\beta_{13}=\beta_{23}=0 ,$$

en front a l'alternativa de que el model correcte és *(13.1)*, per a les quatre variables rebutgem H_0 amb un nivell de significació inferior al 0,1%. Les proves per grups de coeficients amb hipòtesis nul·les:

$$H_0' : \beta_1=\beta_{13}=0 \qquad \text{i}$$

$$H_0'' : \beta_2=\beta_{23}=0 ,$$

també són rebutjades amb nivells de significació inferiors al 1% per a les quatre variables. Els resultats d'aquestes tres proves mostren com els models, i per tant les relacions entre els índexs de diversitat i la llargada del capítol són diferents per les tres zones en que hem dividit el llibre en funció de la seva diversitat. Aquest fet és un indicador de que podem diferenciar diversitats en l'estil en texts de llargades diferents a partir del modelat de la dependència d'aquests índexs ($ln(V_i)$, $ln(V_{1i})$, $ln(H^s_i)$ i B_i) en funció de logaritme de la llargada.

13.3.3.3 Conclusions

En l'anàlisi de la riquesa i diversitat del vocabulari per capítols de més de 200 paraules tornem a observar la presència de dos punts de canvi ja detectada en l'estudi dels blocs. El primer canvi, més feble que el segon, l'estimem a r_1=118, poc després de l'inici de la tercera part del llibre. El segon, que correspon al punt identificat en els capítols anteriors, l'estimem a r_2=382. El llenguatge del començament del llibre és més pobre que el de la zona central, entre els capítols 119 i 383, però més ric que el llenguatge de la zona final.

La taula 13.2 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de la riquesa i diversitat del vocabulari per capítols de més de 200 paraules.

Distribució	Estadístic	Punt de canvi principal	Pt.de canvi secundari
Normal	*Formes, V_i*	$\hat{r}_2$ =382	$\hat{r}_1$ =118
	Hapax legomena, V_{1i}	$\hat{r}_2$ =382	$\hat{r}_1$ =118
	Índex de Simpson, D_i	$\hat{r}_2$ =382	$\hat{r}_1$ =118
	Entropia, H^s_i	$\hat{r}_2$ =382	$\hat{r}_1$ =118
	Índex de Brillouin,B_i	$\hat{r}_2$ =382	$\hat{r}_1$ =118

Taula 13.2: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la riquesa i diversitat del vocabulari per capítols de més de 200 paraules.

13.3.4 Anàlisi Cluster

De forma anàloga al que hem fet en els capítols anteriors, fem anàlisi cluster, per veure si és possible recol·locar algun capítol en el grup en el que és més homogeni. Fem servir les tècniques clàssiques no jeràrquiques exposades en el capítol 7. De tots els índexs, només fem servir els dos que el seu valor esperat no depèn de la llargada de capítol: *D* i *M*, que estandaritzem de manera que tinguin mitjana *0* i variança *1* i el mateix pes en l'anàlisi.

L'anàlisi cluster clàssica basada en particions, mitjançant l'algorisme *k-means*, pressuposa que hem de saber quants clusters existeixen abans de realitzar l'anàlisi. Hem fet l'anàlisi en dos casos: suposant que hi ha dos grups i suposant que n'hi ha tres.

A l'hora de calcular els centres dels clusters, les tècniques clàssiques atorguen el mateix pes a totes les observacions. En l'estudi del *Tirant*, això implica donar el mateix pes a un capítol de 250 paraules que a un de 2500 paraules. Farem servir una adaptació de la tècnica clàssica que consisteix en ponderar les observacions de forma que tinguin un pes proporcional a la llargada del capítol. A la pràctica, el software utilitzat, MINITAB, no té possibilitat de fer una anàlisi cluster ponderant les observacions, pel que hem generat un nou full de dades en el que cada capítol hi surt un nombre de vegades proporcional a la seva llargada, i hem realitzat l'anàlisi mitjançant l'algorisme *k-means* clàssic amb aquestes dades. També hem fet l'anàlisi cluster sense ponderar els capítols. En tots dos casos, l'assignació inicial de capítols a clusters és la que hem obtingut de l'estimació del punt de canvi: quan suposem tres clusters un grup està format pels primers 118 capítols, un segon pels capítols 119-382, i el tercer pels capítols 383-final. Quan suposem només dos clusters, l'assignació inicial és: 1-382 i 383-final.

Els resultats obtinguts tant quan es suposen dos com tres clusters són molt semblants pel mètode que pondera les dades i pel que no les pondera. Veurem amb més detall el resultat obtingut mitjançant la ponderació dels capítols.

La divisió en clusters que s'obté separa els blocs en funció del valor de l'índex de Simpson, *D*. La figura 13.21 mostra els gràfics amb els resultats les anàlisi: suposant dos (esquerra) i tres (dreta) clusters, on s'han representat en colors diferents els blocs que han estat assignats a clusters diferents. La diferència en les classificacions en dos i tres clusters està en que, en el segon cas, els capítols amb valors de l'índex de Simpson intermig han estat agrupats en el tercer cluster. Aquests blocs, en l'agrupació en dos clusters queden repartits entre les dues agrupacions, en funció del valor de l'índex de Simpson.

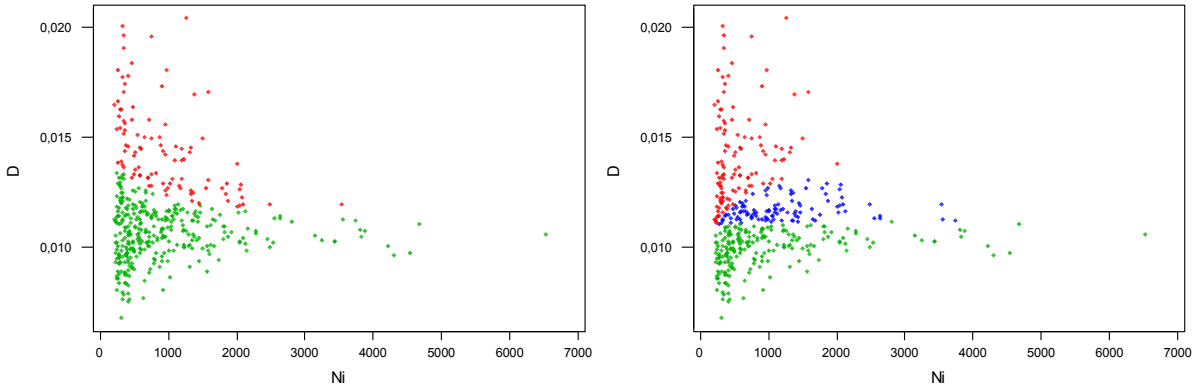

Figura 13.21: gràfics amb la separació dels blocs en dos clusters (esquerra) i en tres clusters (dreta). S'han representat en colors diferents els capítols que han estat assignats a clusters diferents: en color verd els capítols amb una major riquesa de llenguatge, en vermell els de vocabulari més pobre i blau els que tenen una riquesa intermitja.

La figura 13.22 mostra l'evolució de l'índex de Simpson al llarg del Tirant, en la que s'han representat en colors diferents els capítols assignats a clusters diferents. Es pot observar com, quan suposem que hi ha dos clusters, hi ha una bona part dels capítols posteriors al 382 que han sigut assignats al cluster que conté majoritàriament els capítols del principi, mentre que pocs capítols inicials han sigut assignats a l'altre cluster.

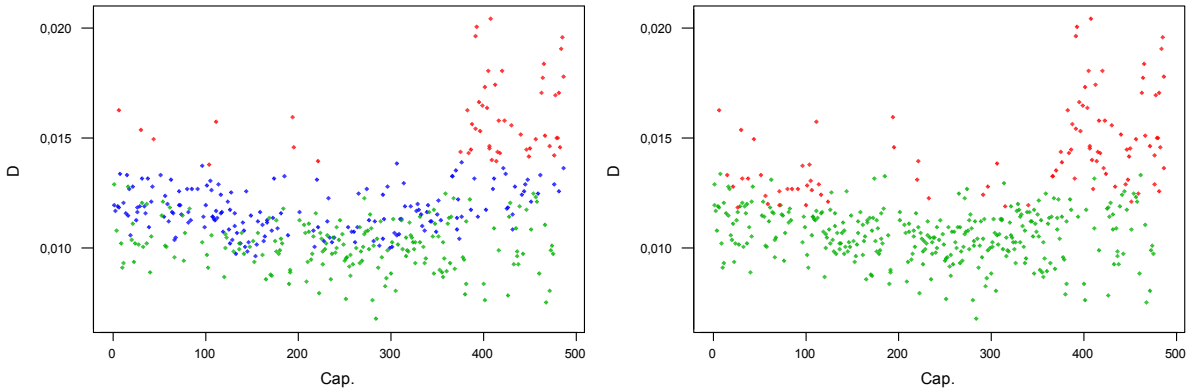

Figura 13.22: gràfics amb l'evolució temporal de l'índex de Simpson, *D*, al llarg del *Tirant* quan es suposen dos clusters (esquerra) i tres clusters (dreta). S'han representat en colors diferents els capítols que han estat assignats a clusters diferents: en color verd els capítols amb una major riquesa de llenguatge, en vermell els de vocabulari més pobre i blau els que tenen una riquesa intermitja.

En l'annex A13.2 hi ha els llistats en els que hi ha les assignacions dels capítols als clusters, tant per quan es suposen dos com tres clusters.

Per validar els resultats obtinguts quan es suposen dos clusters, es compara l'ajust del model emprat per a la determinació d'un punt de canvi:

$$D_i \mid Ind_{1i}^{(r=382)} \sim N(\mu = \beta_0^{(r=382)} + \beta_1^{(r)} Ind_{1i}^{(r=382)}, \sigma_i^2 = \frac{\sigma^2}{N_i}),$$

on $Ind_{1i}^{(r)}=0$ per $i=1,...382$ i $Ind_{1i}^{(r)}=1$ per $i=383,...,n$ amb el model:

$$D_i \mid Ind_{1i}^{(c)} \sim N(\mu = \beta_0^{(c)} + \beta_1^{(c)} Ind_{1i}^{(c)}, \sigma_i^2 = \frac{\sigma^2}{N_i}),$$

on $Ind_{1i}^{(c)}=0$ pels capítols assignats a un cluster i $Ind_{1i}^{(c)}=1$ pels capítols assignats a l'altre. Tots dos models s'han ajustat ponderant les observacions.

Quan es suposa que hi ha 2 punts de canvi i, per tant, tres clusters, comparem model de regressió lineal múltiple ajustat amb pesos per a l'estimació dels punts de canvi:

$$D_i \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)} + \beta_2^{(r)} Ind_{2i}^{(r)}, \sigma_i^2 = \frac{\sigma^2}{N_i}) \qquad \text{per } \boldsymbol{r}=(118,382),$$

on $Ind_{1i}^{(r)}=1$per $i=1,...118$ i és igual a *0* pels capítols 119, ... *487*, i $Ind_{2i}^{(r)}=0$ per $i=1,...,118$ i $i=383,...,487$, i és igual a *1* pels capítols *119, ..., 382*, amb el model:

$$D_i \sim N(\mu = \beta_0^{(c)} + \beta_1^{(c)} Ind_{1i}^{(c)} + \beta_2^{(c)} Ind_{2i}^{(c)}, \sigma_i^2 = \frac{\sigma^2}{N_i})$$

on $Ind_{ki}^{(c)}=1$ pels blocs assignats al cluster *k*, per $k=1,2$, i $Ind_{ki}^{(c)}=0$ pels blocs assignats als altres dos clusters, per $i=1,...,487$. Tant quan suposem 2 punts de canvi i 3 clusters com quan suposem 1 punt de canvi i 3 clusters l'ajust millora molt després de realitzar l'anàlisi cluster.