UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

ANNO ACCADEMICO 2004/2005

TESI DI LAUREA

Analisi dell'offerta bibliotecaria degli Atenei italiani

Relatore : Ch. ma Prof.ssa Laura Ventura

Laureando : Paolo Girardi

Ringraziamenti

Ringrazio di cuore i miei genitori, mia sorella e mia nonna, per avermi permesso di raggiungere questo importante risultato.
Un ringraziamento particolare alla Prof. ssa Ventura che mi ha continuamente sostenuto durante la stesura della tesi.
Un doveroso grazie alla Dott. Catinella, e a tutto il personale del CAB, per avermi dato la possibilità di effettuare questa interessante opportunità di studio.
Ringrazio gli amici e le amiche dell'Università per le giornate trascorse insieme.
Una dedica speciale a tutti gli amici del mio paese.

INDICE DEI CONTENUTI

Cap. 4 L'analisi fattoriale e delle componenti principali

Cap. 5 Cluster analysis

INTRODUZIONE

La tesi si propone di effettuare un'analisi di classificazione dei dati provenienti da un censimento sulle biblioteche di Ateneo eseguita nel 2002. Tale rilevazione si è svolta all'interno di un progetto proposto dal Gruppo Interuniversitario per il Monitoraggio dei sistemi bibliotecari (GIM), con lo scopo di valutare l'offerta bibliotecaria degli Atenei italiani.

Questa rilevazione risulta di estremo interesse. I dati così ottenuti rappresentano infatti una vera e propria anagrafe delle biblioteche universitarie, con dati aggiornati e attendibili, sui quali si possono eseguire elaborazioni altamente significative al fine di evidenziare gli aspetti più importanti del panorama bibliotecario universitario italiano.

L'indagine, svolta dal GIM, inizialmente si è articolata contattando figure amministrative dei singoli Atenei, dai quali è stata ottenuta una lista di tutte le biblioteche afferenti ad ogni Università. Ogni unità - che per l'appunto è la singola biblioteca – è stata contattata via web con un questionario *on-line*, che veniva poi inserito, con l'uso di un apposito software, in un database relazionale. Infine, tutti i dati sono stati agglomerati in un unico *dataset*, in cui ogni singola unità era la singola biblioteca. In questo modo si è ottenuta una considerevole massa di dati omogenei relativi a 1345 biblioteche afferenti a 77 Atenei. Inoltre si è ottenuta un alta percentuale di risposte (86.7 %), considerando che ben 1164 biblioteche hanno compilato almeno in parte il questionario *on-line*. In questo caso l'uso del *pc,* di *internet* e la metodica delle osservazioni hanno aiutato in modo basilare la raccolta delle informazioni rispetto alle precedenti indagini, costituendo una solida base per future indagini.

La struttura di questa rilevazione è suddivisa in diverse aree di interesse che spaziano dalle attrezzature, alle spese e ai servizi. Tuttavia, lo scopo della raccolta di queste misure e informazioni è legato alla compilazione di diversi indicatori, adottati a livello nazionale, per confrontare le singole biblioteche e gli Atenei tra loro.

Sviluppo della tesi

Il lavoro presentato in questa tesi è stato suddiviso in capitoli per dare una chiara rappresentazione, anche logica, di come è stata effettuata l'elaborazione e l'analisi partendo dai dati grezzi forniti dal GIM.

Nel Capitolo 1, si è evidenziata la struttura del *dataset*: la maggior parte delle variabili raccolte sono di tipo numerico, mentre, anche se presenti, risultano in minoranza le variabili dicotomiche e categoriali.

Un primo controllo sulla qualità dei dati ha evidenziato la presenza di dati mancanti e la necessità di un lavoro di "pulitura" del *dataset*, in modo da inserire nell'analisi le unità e le variabili effettivamente significative per le analisi successive. In realtà, in questa fase, è emerso che il valore di mancate risposte di talune domande del questionario è abbastanza elevato e si è dovuto decidere se tenere queste variabili o eliminarle. Analogo discorso è stato effettuato eliminando quegli Atenei di cui si hanno poche biblioteche rispondenti al questionario *on-line,* sicché le informazioni relative al complessivo di Ateneo potevano risultare poco attendibili e inadeguate.

Nel Capitolo 2 si è risolto il problema dei dati parzialmente e totalmente mancanti. Accingendo dalla letteratura in materia si è potuto risolvere il tutto, in modo opportuno, tramite tecniche statistiche che mirano a predire valori non osservati, conservando il più possibile la natura e la distribuzione dei dati stessi. Inoltre, tale lavoro di imputazione è stato svolto usando diverse tecniche a seconda dell'origine delle variabili (numeriche, dicotomiche e categoriali).

Con il *dataset* così completo, nel Capitolo 3 si sono potuti calcolare gli indicatori di valutazione per gli Atenei. Tali indicatori, presenti e usati a livello internazionale, hanno lo scopo di fornire misurazioni su specifiche aree di interesse in cui gli indici, standardizzati, possono essere utilizzati per condurre delle sommarie analisi descrittive. In questo modo, facendo leva sulle rappresentazioni grafiche si cerca di cogliere gli aspetti fondamentali dei dati, senz'altro utili per affrontare la successiva parte di analisi e per capire i fenomeni che si stanno studiando.

"There is no single statistical tool that is as powerful as a well chosen graph", è quanto affermano *Chambers et al. (1983)* sottolineando che un grafico ben costruito è in grado di fornire una grande quantità di informazioni, che consentono di mettere in evidenza aspetti, caratteristiche e legami esistenti tra i dati e non visibili apparentemente.

Infine, nell'ultima parte della tesi, sono state effettuate le analisi classiche per capire le associazioni esistenti tra unità in cui sono presenti un elevato numero di variabili. In tal caso si cerca di riassumere le peculiarità esistenti delle biblioteche tra i differenti gli Atenei attraverso l'uso di un ristretto numero di fattori. Inoltre, tramite l'analisi *cluster*, si cerca di raggruppare le unità in base misure di similarità. Ciò consente di individuare gruppi di biblioteche o di Atenei che hanno caratteristiche, tra loro, paragonabili.

Per effettuare le elaborazioni e le analisi è stato usato l'ambiente statistico R. L'uso di un software così versatile e la disponibilità di una vasta gamma di pacchetti statistici ha sicuramente agevolato il mio lavoro. Inoltre, bisogna considerare che il programma statistico R è un software " *open – source* "; ciò permette il libero utilizzo del programma e il suo continuo sviluppo.

CAPITOLO 1

IL QUESTIONARIO GIM

1.1 Struttura e caratteristiche del questionario

Il questionario elaborato dal GIM si prefigge di ricavare informazioni da ciascuna biblioteca per poter effettuare successivamente la compilazione degli indicatori. Per questo motivo esso risulta suddiviso in sette parti, ognuna delle quali ha un ramo di interesse specifico. Esso risulta così ramificato:

- informazioni generali;
- spazi e attrezzature;
- dotazione documentaria;
- personale;
- spese (impegno finanziario dell'anno 2002);
- orari e servizi;
- utenza potenziale non istituzionale.

La prima parte del questionario - *informazioni generali* - consiste nel localizzare in modo univoco la biblioteca nel questionario e definisce le peculiarità di base delle unità, come la tipologia e l'articolazione dei punti di servizio.
La parte centrale dell'indagine svolta dal GIM consiste nel ricavare informazioni sui caratteri descrittivi: i valori raccolti in questa sezione sono tutti di tipo numerico ed esprimono i caratteri oggettivi relativi alla biblioteca, e sono poi utilizzati nel calcolo degli indicatori.
Infine, nella parte finale, il questionario si rivolge ad estrarre informazioni relative alle attività e ai servizi offerti dalla biblioteca e alla tipologia di utenza coinvolta in tale servizio.

Da come si può capire, l'indagine svolta risulta nelle sue parti abbastanza ben formata e chiara in tutte le sue sessioni: difatti ogni singola domanda risulta semplice e ben specificata e la comprensione dei quesiti è facilitata da brevi spiegazioni che identificano lo scopo della richiesta e la metodologia di rilevazione.

1.2 Presentazione del *dataset*

Il *dataset*, contenente i dati grezzi, inizialmente si presentava composto da 1345 biblioteche, con 77 variabili. La parte dei dati che in questa fase ci interessa maggiormente è quella relativa alle informazioni generali. Infatti, in tale sessione troviamo indicazioni sul nome della biblioteca e sulla sua tipologia. Queste generalità sono le uniche, oltre al numero di questionario, di cui si è riusciti a non avere dati mancanti nel *dataset* e quindi possono risultare estremamente utili nelle successive analisi.
Da una analisi descrittiva del numero di biblioteche per Ateneo, si notano subito che vi sono alcuni Atenei che spiccano per la grande quantità di biblioteche: in questo caso si è scelto un modello di strutturazione decentralizzato. Esempi sono l'Università "*La Sapienza*"di Roma e l'Università degli Studi di Milano. Dall'altra parte sono ben 17 gli Atenei che hanno condiviso la scelta di avere una sola biblioteca comune: questo risulta molto interessante perché le università interessate non sono soltanto di media – piccola dimensione, ma anche di una certa levatura come l'Università degli Studi di Milano "*Bicocca"*.

1.3 Regole di selezione per variabile e biblioteche

Nel *dataset* descritto è presente un consistente numero di dati mancanti. Delle 1345 biblioteche rispondenti ai requisiti indicati dal GIM, ben 195 biblioteche non hanno compilato il questionario *on-line*, e ben 32 di quest'ultime, pur essendosi collegate, non hanno fornito nessuna risposta. La selezione di unità e variabili si basa, almeno in questa prima fase, sui suggerimenti proposti dal GIM in materia e utilizzati nella precedente indagine svolta dallo stesso Gruppo Interbibliotecario per il Monitoraggio.

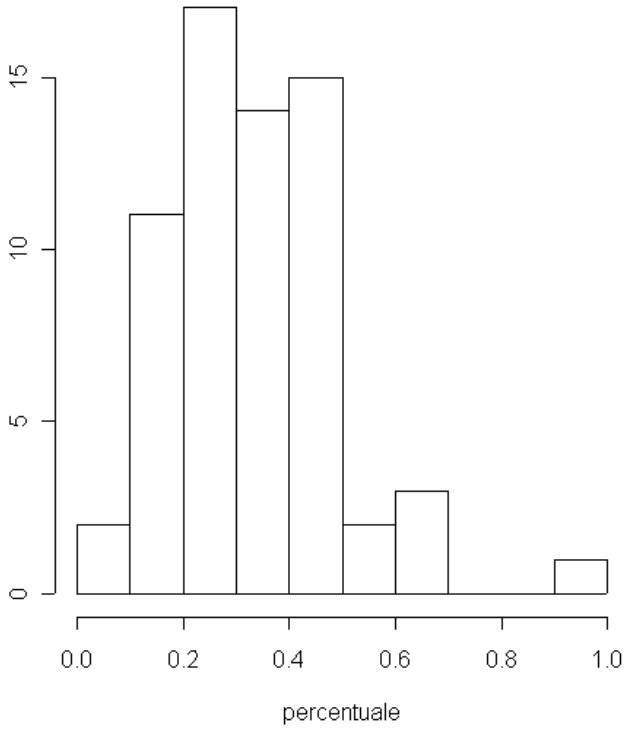
Si è reso necessario fissare un tetto di selezione per includere gli Atenei che hanno una significativa percentuale di biblioteche. Concordi alla precedente analisi svolta dal GIM, si è deciso di fissare questo valore di soglia al 30%: in questo modo, calcolando le percentuali di mancate risposte totali per Ateneo, si è proceduto all'eliminazione delle università con troppe biblioteche mancanti.
In totale sono state escluse dall'analisi 233 biblioteche distribuite su 9 Atenei, includendo anche l'Università della Valle d'Aosta, della quale non si ha alcuna informazione nel *dataset*.

Tabella 1.1 Atenei esclusi dall'analisi, relativa numerosità di biblioteche e percentuale di non risposte.

Nome dell'Ateneo	Numero di biblioteche	Percentuale mancante
Università degli Studi di Cassino	8	62.5 %
Università della Valle d'Aosta	0	100 %
Università degli Studi del Sannio	4	50 %
Università degli Studi di Napoli Federico II	92	33.7 %
Istituto Universitario Navale di Napoli	12	41.7 %
Università degli Studi di Salerno	29	34.5 %
Università degli Studi di Foggia	4	50 %
Università degli Studi di Messina	55	50.9 %
Università degli Studi di Catania	29	41.4 %

Dalla tabella precedente si nota quanto ha inciso la percentuale di dati mancanti sul totale. Con questa estromissione, la numerosità del *dataset* risulta di 1112 biblioteche.
La successiva analisi del *dataset* è volta ad analizzare le percentuale di mancate risposte relative a ciascuna variabile (quesito del questionario) per capire se vi è una numerosità campionaria tale da giustificare i risultati delle successive analisi basandosi sui dati aggregati. In questo caso, non essendo stato effettuato alcun studio precedente, sembra importante l'opportuna scelta di un valore soglia di inclusione.

Figura 1.2 Istogramma della distribuzione delle percentuali di mancate risposte di ciascuna variabile del dataset.



Dall'istogramma risulta evidente che la maggior parte delle variabili ha una percentuale di mancate risposte minore del 50 %. Tale dato complessivo risulta abbastanza soddisfacente, se si considera il fatto che le variabili con alte percentuali di mancate risposte sono state rilevate nell'ultima parte del *dataset*, quella che riguardava i servizi offerti. Si è deciso di escludere dal *dataset* le variabili con una elevata numerosità di mancate risposte e in cui non è possibile un lavoro di imputazione deterministica dei valori mancati. Infatti, in taluni casi, era possibile ricostruire il valore mancate con l'uso della risposta di una domanda collegata. Ad esempio, si pensi al numero di prestiti interbibliotecari e alla domanda collegata che qualifica se la biblioteca effettua prestiti interbibliotecari. In questo caso, se si conosce che una biblioteca non effettua prestiti interbibliotecari, si può dedurre che il numero di prestiti di questo tipo sarà nullo.

CAPITOLO 2

IMPUTAZIONE DEI DATI MANCANTI

2.1 Il problema dei *missing data*

La letteratura propone vari metodi per trattare dati incompleti, e l'applicazione dell'uno o dell'altro metodo dipende dal meccanismo generatore dei dati mancanti e dal tipo di analisi che si vuole condurre sul campione completo. La tipologia di mancate risposte si distingue in :

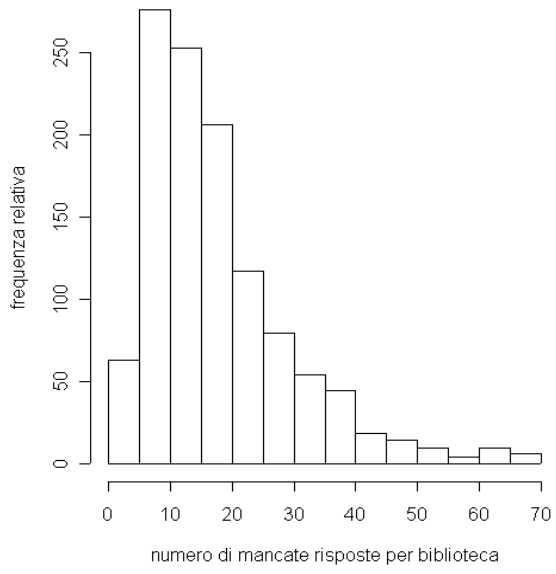
- mancata copertura;
- mancate risposte totali;
- mancate risposte parziali.

Per *mancata copertura* si intende l'esclusione dalla lista di campionamento di alcune unità appartenenti alla popolazione obiettivo. Queste unità hanno probabilità nulla di entrare nel campione e riflette la cattiva qualità delle liste di campionamento. Nel caso discusso in questa tesi, la probabilità di mancate risposte dovute all'assenza di copertura è molto bassa. Infatti, la lista di campionamento data da ogni singolo Ateneo al GIM risulta aggiornata.
Il secondo tipo di incompletezza dei dati deriva dalla *non risposta totale*. La mancata risposta totale può avvenire per rifiuto a collaborare da parte degli individui oggetto dell'indagine, o perché non si riesce, dopo ripetuti tentativi, a contattare l'individuo da intervistare o perché l'intervistatore è impossibilitato per motivi tecnici a comunicare i suoi risultati. Nel *dataset* originario la mancate risposte totali sono il 14.5%, per un totale di 195 biblioteche. Di queste, 163 unità non sembrano siano venute a conoscenza del questionario *on-line,* mentre 32 biblioteche, pur avendolo aperto, non hanno compilato in esso nessun campo.

Tuttavia il valore di biblioteche non rispondenti non è elevato, e ciò è dovuto soprattutto al buon lavoro effettuato dal GIM, dalla alta disponibilità riscontrata dalle biblioteche e dall'uso di strumenti informatici. Di queste biblioteche si hanno comunque le informazioni relative alla lista di campionamento quali il nome, l'Ateneo di cui fa parte e, dopo alcune ricerche, si è potuti risalire all'informazione relativa alla tipologia di biblioteca. Questo dato risulterà estremamente importante per l'imputazione e per le analisi successive.

L'ultimo tipo di incompletezza dei dati è dato dalla *mancata risposta parziale*. Con tale classificazione si indica la mancata risposta ad uno o più quesiti del questionario. Le cause di una non risposta parziale sono, in genere, dovute al rifiuto o all'incapacità di rispondere da parte dell'intervistato, dimenticanza da parte dell'intervistatore di porre una domanda o di registrarne la risposta, o errori alla trascrizione nei supporti informatici. Rispetto agli altri tipi di incompletezza questo può ritenersi il meno grave, in quanto si dispone di una serie di informazioni di contorno sull'individuo in questione. Nel *dataset* originario le biblioteche che hanno compilato il questionario in ogni sua parte sono 153 (11,4% del totale). Un valore così basso non è in ogni caso allarmante, bensì è dovuto al fatto che il questionario aveva alcune domande molto mirate, soprattutto nella parte riguardante i servizi, le cui risposte venivano probabilmente saltate perché non riguardanti la biblioteca in questione o considerate dai risponditori soltanto meno importanti ai fini dell'indagine.

Grafico 2.1 Istogramma delle frequenze relative alle mancate risposte delle variabili per biblioteca



Tuttavia la numerosità di mancate risposte per ogni singola biblioteca non è particolarmente elevata. La maggior parte dei questionari sono stati compilati saltando solo il 25% dei quesiti, mentre si può notare che le biblioteche che hanno risposto soltanto a qualche domanda sono in netta minoranza.
Anche se la qualità del censimento è risultata buona, si è scelto di operare, come descritto nel paragrafo 1.3, con l'analisi del dataset ridotto, comprendente 1112 biblioteche e 55 variabili, allo scopo di ridurre ulteriormente i dati mancanti per semplificare il successivo lavoro di imputazione.

2.2 Assunzioni sui dati mancati

La letteratura propone differenti metodi per trattare i dati mancanti, ma è comunemente accettato che nessun criterio è chiaramente superiore rispetto ad altri[1]. Un primo approccio per ottenere un *dataset* completo, è quello di basarsi solamente sulle unità osservate. Tale metodo ha il pregio di essere di semplice applicazione, ma può portare a risultati non soddisfacenti. Questo metodo è consigliato nel caso in cui i dati mancanti è limitato e i dati mancati sono MCAR, ossia *missing completely at random;* in altre parole il campione incompleto può essere considerato un sottocampione casuale del campione originario. Questa tecnica risulta molto semplice e sotto le condizioni MCAR le stime che si ottengono risultano consistenti. Però lo svantaggio è di scartare i casi incompleti, in special modo se questi sono numerosi. In tal modo si riduce la numerosità campionaria e con essa l'informazione.
Un alternativa valida, che consente di non rinunciare a nessun dato, è la tecnica chiamata *"Pairwise Delection"*, che tradotto letteralmente significa "omissione a coppie di due". Tale tecnica parte dal fatto che le principali analisi, quale quella fattoriale, quella cluster e i modelli di regressione possono essere effettuate attraverso le stime di media, varianza e covarianza, etc. provenienti solo dai dati completi. Così, per esempio, la covarianza tra due variabili X e Z può essere calcolata attraverso i valori presenti sia in X che in Z. Però, per applicare tale metodo, si deve verificare l'assunto di MCAR sui dati mancanti, altrimenti le stime dei parametri potrebbero essere inconsistenti e seriamente distorte.

[1] Si veda, ad esempio, Allison (2001).

Nel nostro caso, l'unico modo per verificare se i dati incompleti del questionario sono effettivamente prodotti da un meccanismo casuale, è controllare se il numero di dati mancanti è lo stesso nelle varie tipologie di biblioteche, che è l'unica informazione che si possiede per tutte le unità in esame.

Tabella 2.2 Tabella di contingenza per la distribuzione dei dati totalmente mancanti all'interno delle tipologie di biblioteche.

	Atenei e Interfacoltà	**Dipartimenti**	**Facoltà e Interdipartimenti**	**Istituti**	**Totale**
Unità rispondenti	87	588	256	82	1013
Unità mancanti	0	73	13	13	99
Totale	87	661	269	95	1112

La tabella mostra l'andamento delle mancate risposte totali all'interno delle tipologie di biblioteca in ordine sommario di grandezza. Dai valori che risultano apparire in tabella sembra che la dimensione delle biblioteche influisca sulle mancate risposte totali: infatti nelle strutture apparentemente più grandi, cioè quelle di Ateneo e Interfacoltà, tutti i questionari risultavano, almeno parzialmente, compilati.

Per confermare il fatto che vi è una certa dipendenza tra la dimensione delle biblioteche e la presenza di dati mancati si può eseguire il test chi-quadrato di Pearson. Questo test ha come ipotesi nulla l'indipendenza tra le due variabili in esame. In questo caso l'ipotesi nulla e l'ipotesi alternativa sono:

$$\begin{cases} H_0 \ : \ P[X = x_i, \ Y = y_j] \ = \ P[X = \ x_i]P[Y = \ y_j] \\ H_1 \ : \ P[X = x_i, \ Y = y_j] \ \neq \ P[X = \ x_i]P[Y = \ y_j] \end{cases} \qquad (2.3)$$

dove le variabili categoriali X e Y sono la dimensione della biblioteca e il tipo di risposta sul questionario.

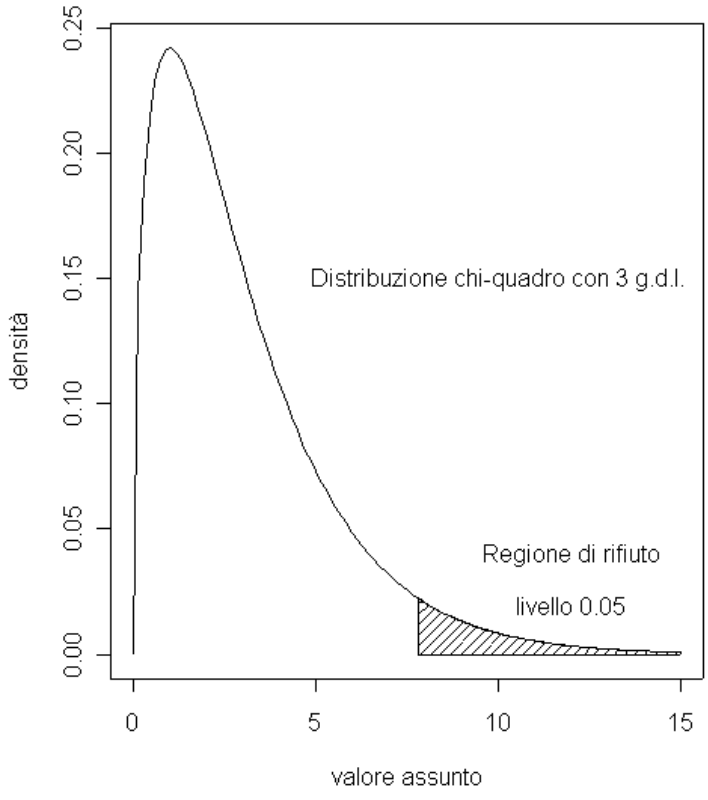
Per poter effettuare un test è necessario costruire una statistica test T . In generale si ha che sotto H_0 :

$$T = \sum_i \sum_j \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}} \sim \chi^2_{[I-1][J-1]} = \chi^2_{[3]} \qquad (2.4)$$

In questo caso :

- I rappresenta il numero di livelli della variabile X, grandezza presunta della biblioteca ($I = 4$);
- J rappresenta il numero di livelli della variabile Y tipo di risposta ($J = 2$);
- f e $\hat{f}$ sono le frequenze osservate e stimate sotto H_0 (ipotesi d'indipendenza).

Grafico 2.5 Distribuzione di una variabile chi-quadro con 3 gradi di libertà e regione di rifiuto per un livello di significatività del 0,05.



Il valore osservato della statistica test è 17,73 (*p-value* = 0,0005). Fissando un livello di significatività pari 0,05 si è propensi a rifiutare H_0 e ad affermare che c'è una dipendenza tra la presunta grandezza delle biblioteche e la numerosità delle mancate risposte totali. Infatti i questionari non compilati provengono soprattutto dalle biblioteche più piccole, come i dipartimenti. Pertanto, le assunzioni di dati mancanti MCAR non sono valide e risulta incorretta l'analisi dei dati completi perché potrebbe comportare delle consistenti distorsioni delle stime e falsare gli studi successivi.

2.3 Metodi di imputazione per la non risposta parziale

Non essendo possibile trattare i dati grezzi per effettuare il calcolo degli indicatori, si deve ricorrere a qualche metodo per effettuare l'imputazione: esso consiste nel sostituire al valore mancante un valore scelto in "modo opportuno". Tale imputazione risulta estremamente utile per poter ricostruire *dataset* rettangolari, sui quali possono essere applicate facilmente analisi statistiche. Inoltre l'imputazione è lo strumento più adatto per trattare la non risposta nelle indagini complesse come i censimenti, le indagini multiscopo e le ricerche di mercato. Esistono due principali tecniche per l'imputazione dei dati mancanti: *i criteri basati sui modelli* e *i metodi di tipo deterministico*. L'imputazione attraverso l'uso di modelli è ampiamente discussa da Little (1993), Rubin (1987) e Dempster (1987): essa implica l'uso di modelli predittivi per stimare il dato mancante attraverso una combinazione delle più significative variabili esplicative. Il secondo metodo raccoglie una serie di criteri più "*naive*", e in questo ambito ricordiamo:

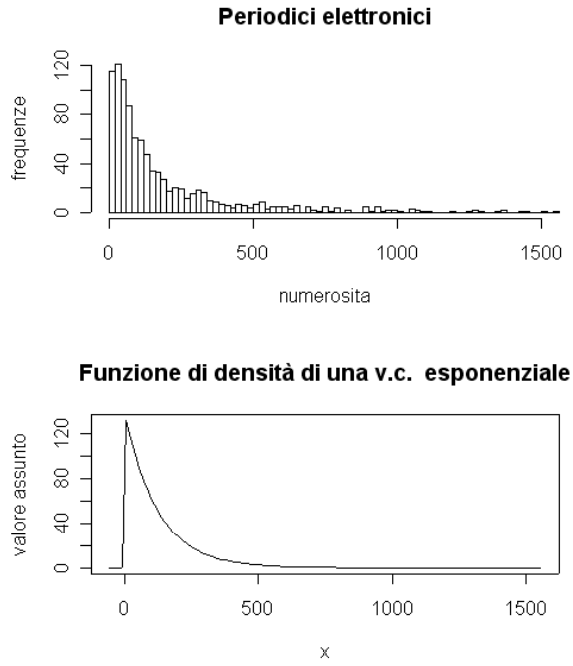
- imputazione deduttiva, in cui le mancanti risposte vengono dedotte dall'osservazione dei dati relativi alle altre variabili;
- imputazione "*cold-deck*", essa prevede che il valore mancante venga rimpiazzato da un valore proveniente da una fonte esterna, ad esempio il valore ricavato da una precedente indagine campionaria;
- imputazione di medie, totale e in classi, dove i valori mancanti per una certa variabile vengono rimpiazzati dalla media o mediana dei valori osservati. Tuttavia in questo modo si sottostima la varianza e si riducono le correlazioni esistenti tra le variabili.

2.3.1 Tecnica usata per l'imputazione dei dati numerici

Un criterio usato per effettuare l'imputazione per variabili numeriche si basa sull'utilizzo di modelli di regressione lineari. In questo modo, per una certa variabile risposta *Y*, si cerca di effettuare una modellazione della risposta attraverso l'uso di un certo numero di regressori. Una volta stimato il modello si cerca di risalire al valore mancante con i valori delle variabili esplicative relative a quel particolare caso. L'argomento riguardante i modelli lineari è ampiamente discusso in *Pace* e *Salvan (2001).* Gli assunti necessari per utilizzare un modello di regressione lineare sono:

- *linearità dell'associazione* fra le variabili;
- *normalità della distribuzione* dei valori che la variabile dipendente assume per ogni dato valore della variabile indipendente;
- *omogeneità delle varianze* relativa ai valori della variabile dipendente per ogni dato valore di quella indipendente.

Il maggior problema riscontrato nell'imputazione con questo metodo è la non normalità della distribuzione dei dati. Infatti, la quasi totalità delle variabili numeriche del *dataset* ha una distribuzione che assomiglia molto ad una variabile esponenziale. Ciò è dovuto al fatto che la maggior parte delle biblioteche è di piccole dimensioni e quindi non c'è una omogeneità tra le unità osservate.



Grafici 2.6 Esempio di distribuzione di una variabile quantitativa e paragone con una variabile casuale esponenziale di parametro opportuno.

La condizione di normalità è senza dubbio necessaria ai fini di ottenere un buon modello e che le assunzioni sopra citate siano rispettate. E' opportuno effettuare un qualche tipo di trasformazione dei dati. In questi caso la più comune trasformazione usata è quella logaritmica: ciò è dovuto essenzialmente alla struttura assunta dalle variabili da predire, che come mostrato dal grafico precedente, è di tipo esponenziale.

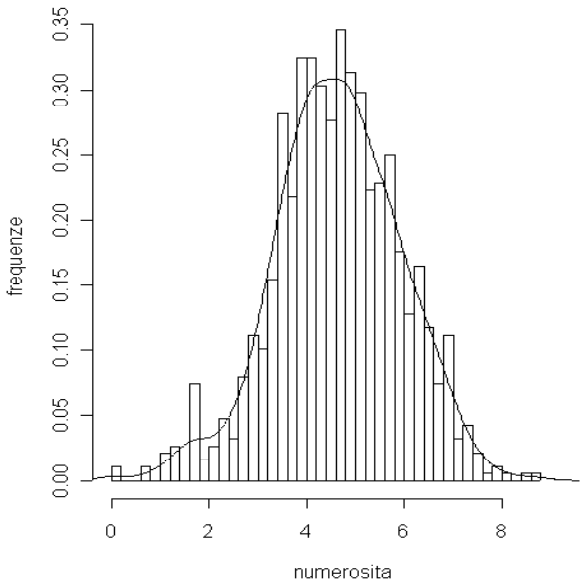


Grafico 2.7 Trasformazione logaritmica delle osservazioni relative ai periodici elettronici e stima della funzione di densità. Con tale operazione si cerca di normalizzare la risposta e di ottenere una migliore bontà di adattamento del modello.

La trasformazione logaritmica non può essere tuttavia applicata nelle variabili in cui si ha la presenza di zeri. Infatti la funzione logaritmica è definita solo per valori positivi.
Si rendeva allora necessaria una trasformazione più flessibile per cui ed è stata scelta la trasformazione di *Box – Cox*. Nel 1964, *Box* e *Cox* hanno proposto un metodo iterativo e concettualmente complesso, divenuto operativamente semplice e di vasta applicazione con l'uso dei computer, per individuare quale trasformazione dei dati può meglio normalizzare la loro distribuzione. Il metodo ricorre a una famiglia di trasformazioni di potenze così definita:

$$X' = \frac{X^{\lambda} - 1}{\lambda} \quad con \quad \lambda \neq 0 \qquad e \qquad X' = \log X \qquad se \ \lambda = 0. \tag{2.8}$$

Il valore di λ che meglio normalizza la distribuzione è quello che rende massima la funzione di log-verosimiglianza dei dati.

E' importante osservare che la trasformazione di Box – Cox definisce una famiglia di trasformazioni che variano la loro forma con il mutare del parametro λ: nel nostro caso la distribuzione originaria presenta una asimmetria positiva per cui si deve scegliere un opportuno valore di $\lambda < 1$ in modo da generare una "compressione" dei dati e quindi in grado di rendere la distribuzione degli stessi più simmetrica. Il valore di λ migliore viene scelto in modo da ottenere la migliore approssimazione alla normale dei dati e in tal modo avere un buon modello predittivo.

Dopo aver ottenuto un'approssimazione alla normale della variabile risposta, si cerca di modellare quest'ultima inserendo le variabili che sono più significative e che fanno aumentare la bontà della regressione del modello stesso. Questo lavoro di selezione e stima della regressione risulta particolarmente laborioso considerando che nel *dataset* la maggior parte delle variabili sono di tipo numerico.

Per valutare se il modello stimato era soddisfacente si è preso in considerazione l'indice di correlazione lineare multipla R^2 associato alla regressione (questo indicatore, che varia da 0 a 1, tiene conto della varianza spiegata dai regressori e tanto più l'R^2 è alto, maggiormente il modello specificato si adatta meglio ai dati). Durante il lavoro di stima si è cercato di tenere un valore di correlazione lineare della regressione abbastanza alto e raramente si è scesi sotto il 30% - 40%.

Il lavoro di modellazione è risultato abbastanza difficile per la presenza degli Atenei-monobiblioteca che, avendo valori nelle variabili mediamente più elevati, hanno prodotto degli *outliers* nettamente visibili nelle analisi dei residui.

In questi casi la cosa più semplice da fare è quella di escludere i valori anomali in modo da specificare un modello più centrato sui valori medi, diminuire la varianza e in questo modo migliorare la bontà di adattamento valutata dall'indice R^2.

L'ultima parte che riguarda il lavoro di imputazione è la stima dei valori mancanti con il modello precedentemente elaborato. Ogni dato mancante viene stimato con un numero diverso a seconda del valore delle variabili usate e quindi a seconda della biblioteca a cui il *"missing data"* appartiene.

2.3.2 Tecnica usata per l'imputazione dei dati dicotomici

La presenza nel dataset di variabili di tipo dicotomico ha reso più complicata l'imputazione dei dati mancanti. Per cercare di avere un modello predittivo accettabile per la mancante risposta si è ricorso ai modelli lineari generalizzati (GLM) ed in particolare alla regressione logistica. La trattazione dei modelli lineari generalizzati non è di difficile comprensione e per lo studio approfondito rimandiamo, come nei modelli lineari, alla visione di Pace e Salvan (2001).

Analogamente ai modelli lineari, si effettua una regressione, ma la formula che esprime la connessione tra la variabile risposta e le variabile esplicative non è più lineare, bensì è una trasformazione dei dati detta anche "funzione di legame". Nel caso della regressione binomiale in cui la variabile risposta assume valori dicotomici, si cerca di modellare la probabilità, con cui il dato considerato entra a far parte di un insieme o dell'altro a seconda dei valori assunti dalle variabile esplicative scelte.

Il modello specificato sarà il seguente

$$\mu_i = g(x_i^T \beta), \text{ con } E(Y_i) = \mu_i, \qquad (2.9)$$

dove la funzione g(·) è la funzione legame.

La scelta del modello migliore avviene verificando quali variabili nel *dataset* sono più significative e comportano una maggiore diminuzione di devianza relativa al modello.

2.3.3 Tecnica usata per l'imputazione dei dati categoriali

Nel dataset sono presenti alcune variabili di tipo qualitativo. Una tecnica predittiva che si basa sul valore assunto da altre variabili ausiliare, similmente ai modelli, può essere la regressione mediante modelli discriminanti. Scopo di questa tecnica è quello di costruire un "classificatore", ossia una regola per cui a partire da un vettore di variabili esplicative *X*, sia quantitative che qualitative, si possa associare un'etichetta alla variabile d'interesse. Tale criterio di allocazione detto *"classificazione ad albero"* è una particolare tecnica per costruire una regola di classificazione. Gli alberi di classificazione sono costruiti con ripetuti *splits* (divisioni) del gruppo originario in due sottoinsiemi discendenti, allo scopo di classificare le unità in gruppi omogenei al loro interno e quanto più possibile differenziati. E' possibile associare a tali modelli una rappresentazione grafica a forma di

albero in cui, partendo dal nodo radice si diramano una serie di nodi e rami. Le regole per la formazione e la verifica della bontà di una classificazione ad albero sono abbondantemente descritte in *Fabbris (1990).* La bontà della regola di allocazione ottenuta viene valutata attraverso vari metodi: da ricordare è la tecnica di *cross-validation* che permette calcolare la numerosità delle unità che vengono classificate in modo errato.
Grazie l'uso di questa tecnica non parametrica e di validi strumenti informatici, si è stimata la probabilità che ogni biblioteca aveva di entrare a far parte di ciascuno dei gruppi della variabile da predire e di conseguenza imputato il valore più probabile.

2.4 L'imputazione dei dati totalmente mancanti

Una volta terminata l'imputazione dei dati parzialmente mancanti, tutte le biblioteche che hanno risposto almeno in parte hanno dati completi per tutte le variabili. Tuttavia resta incompiuta l'imputazione dei dati relativi alle biblioteche che non hanno risposto al questionario (dati totalmente incompleti).
Pertanto predire valori per tali unità risulta difficile. Un criterio, in grado di predire valori abile a conservare le caratteristiche di media e varianza delle variabili è l'imputazione *hot-deck*. Questo processo sostituisce al valore mancante il valore osservato in un rispondente, detto donatore, scelto in maniera casuale.
Tale procedimento duplica i valori di una biblioteca e li sostituisce sulle righe che hanno solo valori mancanti nel dataset. L'aggettivo "*hot*" si riferisce al fatto che i valori imputati sono presi dall'indagine corrente, in contrapposizione a *"cold"* dell'imputazione *cold-deck* in cui i valori imputati sono tratti da precedenti indagini. Questa definizione di *hot-deck* è abbastanza generale anche perché, nella letteratura, non vi è definizione precisa comunemente accettata. Infatti, a questo proposito, bisogna sottolineare anche il fatto che non esiste una consolidata teoria per quanto riguarda l'*hot-deck,* ma la sua applicazione è soprattutto dettata dal buon senso che da un rigoroso sviluppo teorico.
L'*hot-deck* si basa sull'individuazione di una corrispondente tecnica deterministica tale che

$$\hat{y}_{is} = \hat{y}_i + \hat{e}_i, \qquad (2.8)$$

dove $\hat{y}_{is}$ è il valore imputato mediante tecnica stocastica, determinato aggiungendo un residuo $\hat{e}_i$ al valore imputato con tecnica deterministica, $\hat{e}_i$ è tale che, dato l'iniziale campione di osservazioni y_{obs}, il suo valore medio è nullo, e pertanto $E(\hat{y}_{is} \mid y_{obs}) = \hat{y}_i$.
Questi procedimenti sono molto usati nella pratica campionaria e possono assumere schemi molto elaborati per selezionare le unità per l'imputazione. La caratteristica comune a tutte le procedure *hot-deck* è quella di selezionare un donatore che abbia caratteristiche simili a quelle del non rispondente, allo scopo di ridurre la distorsione causata dalla non risposta. Questo viene fatto classificando tutte le unità del campione in gruppi, detti classi di imputazione (in inglese "*adjustment cell*") costruiti sulla base del valore delle variabili ausiliarie, in modo tale che le unità all'interno di essi siano omogenee. Nel nostro *dataset* l'unica variabile che può essere usata a tal scopo, dato che è presente in tutte le biblioteche, è la tipologia di biblioteca, in grado di fornirci un'informazione sulla grandezza della biblioteca. La scelta della variabile tipologia per definire le classi di imputazione sembra essere opportuna: come mostrato dal test chi-quadro del capitolo 2.2 la presenza di mancate risposte totali è in correlazione alla tipologia di biblioteca. Per definire se l'imputazione *hot-deck* in classi ha senso si deve analizzare se le biblioteche contenute nella classificazione sono diverse tra loro. Usualmente per verificare se due campioni statistici provengono dalla stessa popolazione si usano dei test statistici. Un test statistico non parametrico, sul quale non si deve presupporre una forma della distribuzione delle popolazioni è il *"test di Wilcoxon"*. Per approfondimenti sul test rimandiamo a Armitage - Berry (2001). Il test di *Wilcoxon* per campioni indipendenti è uno dei più potenti test non parametrici; corrisponde al test *t* di *Student* per campioni indipendenti. Questo test si basa sui ranghi e mette in confronto la distribuzione dei ranghi nei due campioni da confrontare. La statistica test utilizzata, per numerosità campionaria superiore a 20 nei due gruppi, si avvicina rapidamente alla distribuzione normale. Si è applicato il test di *Wilcoxon* a tutte le variabili numeriche e dicotomiche differenziando i gruppi tra le classi di appartenenza, cioè tra le tipologie di biblioteche. Le classi di imputazione che è ragionevole specificare sono 4 e coincidono con le seguenti tipologie di biblioteche:

- Ateneo e Interfacoltà;
- Facoltà e Interdipartimento;
- Dipartimento;
- Istituto.

Tabella 2.9 I risultati delle 6 combinazioni delle 4 classi di imputazione

Tipologia di biblioteche in confronto nel test	N° dei 2 gruppi	P-value osservato
Atenei e Interfacoltà --- Dipartimento	87 --- 587	0.01229699
Atenei e Interfacoltà --- Facoltà e Interdipartimento	87 --- 255	0.13597790
Atenei e Interfacoltà --- Istituto	87 --- 82	0.01106724
Dipartimento --- Facoltà e Interdipartimento	587 --- 255	0.03009766
Dipartimento --- Istituto	587 --- 82	0.08372351
Facoltà e Interdipartimento --- Istituto	255 --- 82	0.02858623

I test di *Wilcoxon* hanno confermato ciò che ci aspettavamo: in 4 dei 6 test abbiamo un p-*value* osservato minore di 0.05 per cui osserviamo una chiara differenza tra le variabili delle tipologie di biblioteche introdotte nei test. Come era prevedibile, i test condotti tra le tipologie di biblioteche simili, come tra Istituto e Dipartimento o tra Atenei e Facoltà, producono dei valori di p-*value* sostanzialmente elevato: questo è dovuto alla presenza di caratteristiche simili all'interno delle classi di imputazione. Tuttavia dalle prove condotte, si può affermare che la suddivisione delle classi risulta opportuna: osserviamo una sostanziale differenza tra le variabili di differenti tipologie di biblioteche.

Inoltre, i risultati dei test sono corretti: infatti la numerosità campionaria è abbastanza elevata per garantire, attraverso l'uso del teorema del limite centrale, la normalità della distribuzione della statistica test utilizzata.

Dopo aver verificato che la classificazione nelle classi indicate è opportuna, abbiamo effettuato la procedura di *hot-deck* nel nostro *dataset.* Per ogni dato totalmente mancante, l'unico dato a cui si è a conoscenza è quello relativo alla tipologia della biblioteca, per cui risulta molto semplice scegliere in modo casuale dalla classe di imputazione afferente allo stesso modello di biblioteca e procurarsi una biblioteca "donatrice". Questo criterio di selezione e imputazione viene effettuato per tutte le biblioteche che non hanno risposto al questionario. Il pregio dell'utilizzo dell'imputazione *hot-deck*, al contrario di metodi più *"naive"* come la media per classi, è la non distorsione delle stime di media e varianza e inoltre non si attenua la correlazione esistente tra le variabili. La sintassi di una pseudo - funzione in linguaggio *R* per l'imputazione *hot-deck* è contenuta in Appendice.

CAPITOLO 3

GLI INDICATORI E LORO SINTESI

3.1 Presentazione degli indicatori

In questo capitolo si presenta la compilazione degli indicatori per i dati aggregati di ciascun Ateneo. Infatti le analisi successive non verranno effettuate con i dati grezzi, bensì con degli indici che possono fornire un'adeguata sintesi delle caratteristiche peculiari di ogni Ateneo. Questi indicatori provengono da varie fonti, alcuni sono stati presi dalla letteratura in materia o da indagini precedenti, altri sono stati creati "*ad hoc"* dal Gruppo Interbibliotecario per il Monitoraggio (GIM). Per ulteriori spiegazioni sull'uso degli indicatori come misure di performance nel comparto bibliotecario, rimandiamo all'Appendice di questo lavoro dove si possono trovare il modo in cui calcolarli e i riferimenti bibliografici.
Una panoramica sulle tappe che hanno portato alla definizione di una serie di indicatori, quali strumenti di misurazione di performance delle biblioteche sono:

- Negli Stati Uniti i primi sforzi in tal senso iniziano negli anni '70 e dopo solo 20 anni si è già in grado di definire standard di applicabilità internazionale;
- *EQLIPSE* sviluppa un sistema di gestione della qualità compatibile con le norme *ISO 9000 e ISO 11620*;
- CAMILE (1998-99), *Concerted Action on Management Information for Libraries in Europe*, divulga con seminari in tutta Europa i risultati di precedenti indagini;
- Con *EQUINOX* (1998-2000) si definisce un set di indicatori;
- *LIBECON* usa Internet come strumento per lo sviluppo e l'aggiornamento di un database di dati statistici sulle attività bibliotecarie e sui relativi costi;
- In Italia la costruzione degli indicatori informativi per le biblioteche è stata intrapresa da *Pilia, Tammaro, Solimine, IFLA / AIB, Osservatorio*.

Gli indicatori disponibili in letteratura e utilizzati nell'indagine GIM, possono essere divisi per area d'interesse:

- accessibilità;
- efficacia / fruibilità / innovazione;
- efficienza / produttività / economicità;
- vitalità del patrimonio / offerta delle risorse.

3.2 Calcolo degli indicatori

Il calcolo degli indicatori è stato effettuato agglomerando i valori relativi ad ogni singolo Ateneo. Infatti i dati per ogni singola biblioteca non possono essere analizzati in modo corretto perché le unità presenti sono troppo disomogenee tra loro: basti pensare che nello stesso *dataset* troviamo biblioteche afferenti ad Istituti o ad Atenei che hanno peculiarità non confrontabili tra loro. Ottenuti i valori delle biblioteche per ogni Ateneo, gli indicatori che si potevano calcolare erano veramente esigui. Quindi, per ottenere un buon numero di indicatori, si sono aggiunti al *dataset* i dati relativi all'indagine SBA sul Sistema Bibliotecario di Ateneo. Con un *dataset* così allargato si è potuto calcolare ben 28 indicatori afferenti alle quattro aree d'interesse introdotte nel paragrafo 3.1.

3.3 Analisi descrittiva degli indicatori

Effettuare un'analisi descrittiva di tutti gli indicatori risulta noioso e poco interessante al fine delle analisi successive. In questa parte mi soffermerò ad analizzare soltanto gli aspetti essenziali di ciò che emerge dalla distribuzione dei dati, utilizzando in special modo delle rappresentazioni grafiche che riescono a sintetizzare una notevole mole di informazioni. A tal scopo i diagrammi *Box-and-Whisker* (scatola e baffi), chiamati anche semplicemente *boxplot,* presentati in modo organico da Tukey (1977), sono un metodo grafico diffuso per la facilità con la quale possono essere costruiti.

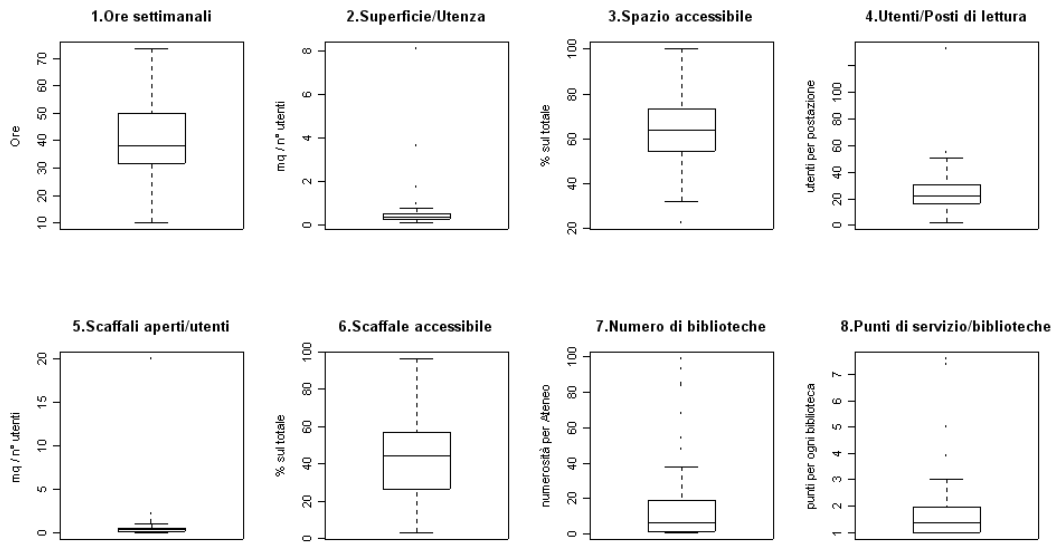
Essi sono in grado di rappresentare visivamente caratteristiche fondamentali di una distribuzione statistica:

- il grado di dispersione o variabilità dei dati, rispetto alla mediana;
- la simmetria;
- la presenza di valori anomali.

3.3.1 Accessibilità

Gli indicatori di accessibilità calcolati sono otto e forniscono una interpretazione di come il sistema bibliotecario di ogni Ateneo è organizzato: gli spazi e i tempi messi a disposizione dell'utenza sono diversi a seconda della politica di apertura al pubblico dell'Ateneo.

Grafici 3.1 Box - plot dei dati relativi agli indicatori di accessibilità.



Dai grafici sembra emergere che i dati relativi agli indicatori sono poco omogenei, con una componente significativa di *outliers* nella gran parte dei *box-plot*. Dalle informazioni relative ad ogni singolo indicatore emerge che:

1- La distribuzione delle medie di ore di apertura settimanali sembra essere equilibrata. Si osserva mediamente che le biblioteche degli Atenei hanno una apertura media sulle 40 ore settimanale, valore sicuramente adeguato. Da notare la grande variabilità di questo indicatore di accessibilità: esso implica

la presenza di Atenei con un forte valore di tempo di apertura settimanale, mentre vi sono strutture la cui scarsa apertura al pubblico è una conseguenza di una possibile adeguatezza delle risorse per garantire una buona copertura dell'apertura settimanale.

2- Il rapporto tra superficie totale e utenza potenziale serve a verificare la disponibilità degli spazi rispetto al pubblico. Ad una prima analisi del *box-plot* emerge una forte presenza di *outliers*: questi valori anomali sono riferiti a quegli Atenei di piccola-media dimensione che, avendo una utenza potenziale ristretta, hanno una particolare gestione degli spazi. Con una forte componente estrema, un indice robusto è sicuramente la mediana che si attesta a 35,6 cm quadrati per utente.

3- La percentuale di superficie accessibile al pubblico in rapporto a quella totale, indica come gli spazi sono organizzati e ciò è un indicatore della politica di accessibilità attuata dall'Ateneo. Dai dati si può evidenziare come la maggior parte delle biblioteche ha una percentuale di superficie accessibile superiore al 60% con la presenza di un solo *outlier* inferiore. Complessivamente questo mostra che le biblioteche sono organizzate con una consistente area di apertura verso al pubblico, rispetto agli spazi adibiti a magazzini.

4- Il numero di posti di lettura per ogni utente è determinante al fine di valutare il corretto dimensionamento della biblioteca rispetto all'utenza potenziale. Dal *box-plot* di tale indicatore emerge la presenza di due valori anomali positivi dovuti probabilmente ad una errata stima dell'utenza potenziale. La distribuzione di tale indicatore è simmetrica con un valore della mediana pari a 22,02 utenti per posto di lettura, valore che non si discosta dalla precedente indagine condotta dal GIM.

5- Il rapporto tra la misura dei metri lineari occupati da materiali negli scaffali e gli utenti potenziali indica la fruibilità del patrimonio cartaceo presente nella biblioteca, ma tale valore non tiene conto della logistica di disposizione del materiale nei scaffali aperti al pubblico. La presenza di un forte *outlier* positivo spinge verso il basso la distribuzione dei dati. La mediana risulta pari a 0,32 metri per utente, mentre lo scarto quadratico medio, con *outlier* escluso, pari a 0,36 indica una forte componente di variabilità all'interno delle biblioteche.

6- La percentuale di scaffale occupato sul totale dovrebbe misurare l'accessibilità diretta al patrimonio documentario e valutare la tipologia di organizzazione spaziale della biblioteca. La maggior parte dei valori si colloca al di sotto del 60% (3° quartile), per cui si può dedurre che la grande maggioranza delle biblioteche ha una consistente parte di materiale a scaffale chiuso, perciò non accessibile direttamente.

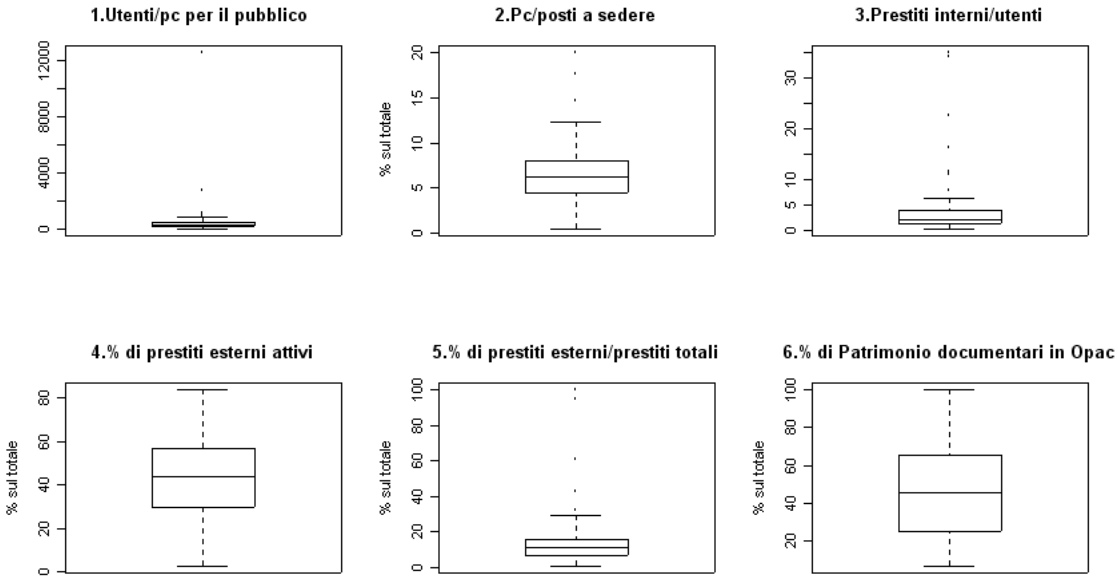
7- L'indicatore che tiene conto delle unità amministrative serve a quantificare il numero di biblioteche presenti per ciascun Ateneo e quindi il tipo di presenza posta nel territorio da ogni singola Università. Dalla distribuzione dei valori risulta che la maggior parte degli Atenei ha meno di 20 unità amministrative, mentre si collocano certamente fuori dallo schema le Università degli Studi di Milano e "La Sapienza" di Roma con più di 90 biblioteche, testimonianza che gli Atenei più grandi e di antica fondazione hanno una tradizione secolare di frammentazione strutturale.

8- Con il numero medio di punti di servizio per unità amministrativa si vuole quantificare l'indice di frammentazione delle biblioteche per ogni singolo Ateneo. La distribuzione dei valori indica che il numero medio di punti di servizio si colloca tra 1 e 2, mentre spiccano le Università del centro Italia, come l'Università degli Studi di Firenze e l'Università degli Studi di Urbino, che hanno un valore medio maggiore a 7 punti di servizio per unità amministrativa, sintomo di una organizzazione strutturale decentrata.

3.3.2 Efficacia – Fruibilità - Innovazione

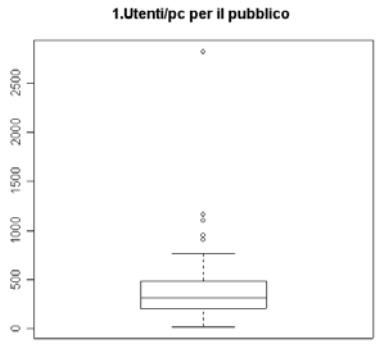
Questa serie di indicatori vogliono quantificare in ogni Ateneo, se le biblioteche sono adeguate dal punto di vista organizzativo. E' logico pensare che, dando ampio respiro allo sviluppo di risorse informatiche, si è in grado di migliorare l'efficacia dei servizi offerti.

Grafici 3.2 Box - plot dei dati relativi agli indicatori di efficacia – fruibilità - innovazione



Dai precedenti grafici, si osserva la minor presenza di *outliers* rispetto alla sezione precedente. Andando nello specifico, i *boxplot* denotano singolarmente quanto segue:

1- Nel primo grafico si evince la presenza di un importante valore anomalo che altera la lettura del grafico. Dopo averlo tolto si è in grado di poter estrarre una informazione significativa dal *boxplot.* Questo indicatore ci dovrebbe dare un'utile informazione sul livello di innovazione tecnologica delle biblioteche. Dal grafico qui a lato, si nota la presenza di qualche altro *outlier* positivo. Tuttavia la distribuzione sembra essere simmetrica intorno al valore della mediana pari a 316 utenti potenziali per ogni pc destinato al pubblico. Tale dato sembra



preoccupante dal punto di vista organizzativo, ma va commisurato al fatto che, all'interno degli Atenei, si tende a creare nuovi spazi, come laboratori e aule informatiche, in cui gli utenti possono usare le risorse informatiche.

2- La percentuale di postazioni informatiche sul totale dei posti a sedere è considerata un forte indice di innovazione tecnologica. Il valore mediano è pari a una percentuale di 6,32 pc sul totale di posti a sedere. Inoltre, la ridotta variabilità di questo indice (ben 35 Atenei hanno un valore compreso tra 4,5 e 8) fa pensare che questo indicatore non è in grado di discriminare in modo efficace un Ateneo dall'altro e che sono molto limitati gli Atenei che hanno un livello di attrezzature informatiche di livello superiore.

3- Il rapporto tra i prestiti interni, sia normali che interbibliotecari passivi, e l'utenza potenziale serve a valutare la capacità di una biblioteca di soddisfare le esigenze informative del pubblico. Dal *boxplot* relativo a questi dati si può notare la presenza di valori anomali positivi che esprimono il fatto che in certi Atenei, quali il Politecnico di Torino o la Libera Università degli Studi di Bolzano, vi è una forte attività di prestito interno. Considerando che la mediana si attesta intorno ai 3 volumi per utente potenziale, con pochissima variabilità, non si è in grado di effettuare una consistente discriminazione tra gli Atenei.

4- Il rapporto tra i prestiti interbibliotecari e *document delivery* attivi in rapporto a quelli totali indica la mole di volumi prestati all'esterno. Una alta percentuale può indicare che il patrimonio documentario della biblioteca è di buona qualità e quindi altamente richiesto, ma anche una mancata disponibilità di documenti per l'utenza interna. Per queste ragioni, può essere considerato un valore ottimale, il 50%. Dall'analisi dei dati e dal *boxplot,* mediamente siamo vicini a questo valore anche se c'è una grande variabilità, sintomo del fatto che non c'è una comune linea di tendenza tra gli Atenei.

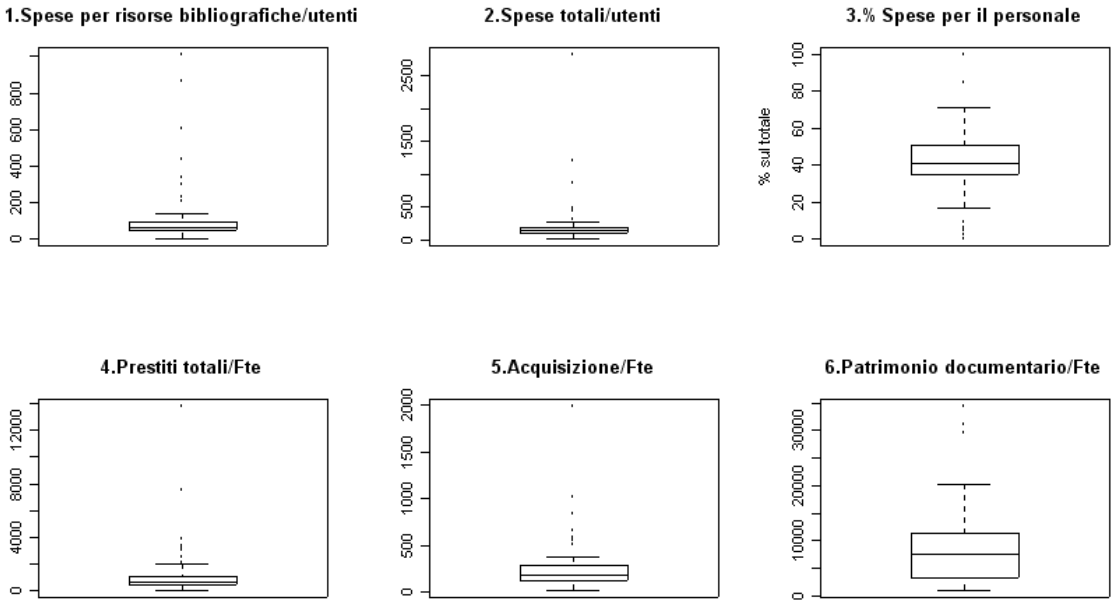
5- La percentuale di prestiti interbibliotecari e *document delivery* sul totale delle movimentazioni dovrebbe indicare le biblioteche più dinamiche e innovative, che non si limitano soltanto al prestito esterno. La presenza di molti *outliers* positivi indica che ci sono certi Atenei in cui le biblioteche hanno molti di questi prestiti in relazione al totale, per cui un'elevata apertura verso l'utenza esterna. Tuttavia la percentuale media per un Ateneo si attesta attorno all'11%.

6- Il numero di patrimonio documentario presente nell'inventario *OPAC* è un indicatore che esprime in modo corretto il livello di automazione nella ricerca di volumi all'interno delle biblioteche. In generale, il valore mediano pari al 46% è sicuramente un ottimo risultato, in confronto al fatto che sono pochi, solamente 9, gli Atenei che hanno biblioteche con una percentuale di volumi in OPAC minore del 20%.

3.3.3 Efficienza – Produttività – Economicità

Gli indicatori relativi a questa sezione danno una misura di come gli Atenei impiegano le risorse finanziarie e le risorse umane per le biblioteche. E' importante per ogni Ateneo riuscire fornire alle biblioteche un livello di risorse adeguato e raggiungere un compromesso tra le varie voci di spesa.

Grafici 3.3 Box - plot dei dati relativi agli indicatori di efficienza - produttività – economicità



Da un'analisi sommaria emergono molti *outliers*: infatti, in questa tipologia di indicatori, emergono le differenti organizzazioni degli Atenei a livello di gestione delle risorse e umane.

I valori più estremi sono relativi alle Università di piccola dimensione che offrono servizi specializzati. E' per questo motivo che certe Università come la Scuola Internazionale superiore ai studi avanzati (SISSA) di Trieste o la Scuola normale superiore di Pisa, offrono delle prestazioni rivolte ad una utenza di tipo diverso e perciò si differenziano in modo determinante dal resto degli Atenei. Analizzando i valori assunti da ogni singolo indicatore si nota quanto segue:

1. La spesa per le diverse categorie di materiali in rapporto all'utenza potenziale serve a misurare quanto gli Atenei dedicano le risorse finanziarie per incrementare il patrimonio documentario considerando l'utenza potenziale che può avvantaggiarsi da tali collezioni bibliografiche. Il valore della mediana dei dati, pari a 65,9 euro per utente potenziale, è leggermente maggiore della precedente indagine condotta dal GIM (60,54 euro).
2. Il rapporto tra le spese totali e gli utenti potenziali definisce, come nel caso precedente, quanto l'Ateneo investe per ogni singolo utente potenziale. In questo caso, trattandosi delle spese complessive, l'indicatore è in grado di fornire una vera stima del grado di produttività/economicità. Anche in questo caso vi sono dei valori estremi dovuti alle Università con situazioni organizzative particolari, in cui i servizi di questi Atenei sono rivolti ad un utenza limitata. Tralasciando questi singoli casi, non vi è una forte differenza nella spesa pro-capite tra gli Atenei italiani: la maggior parte di quest'ultimi assume valori compresi tra 100 e 200 euro per utente potenziale (mediana pari a 146 euro).
3. La percentuale di spesa per il personale soppesa come l'Ateneo è organizzato dal punto di vista gestionale, fornendo un'incidenza di questa spesa sul totale. Escludendo i valori estremi, sia in positivo che in negativo, la distribuzione dei dati relativi a questo indicatore sembra indicare che vi è una sostanziale omogeneità tra le Università. Infatti la percentuale di spesa media si attesta sul 41% e gran parte degli Atenei ha una spesa per il personale compresa tra il 30% e il 55% (ben 42 Atenei).

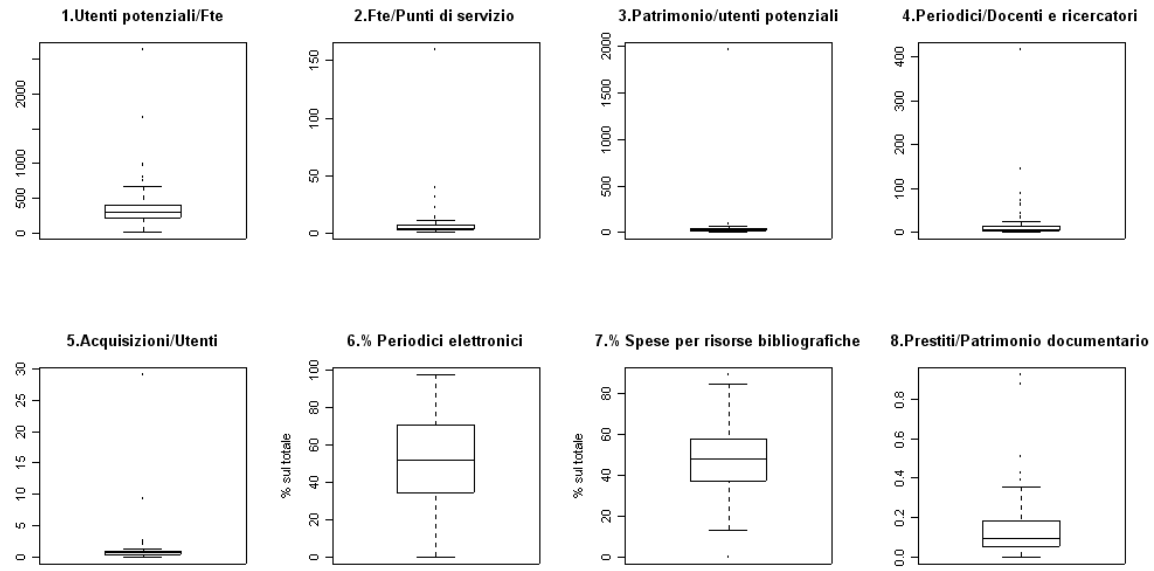
4. Il numero medio di prestiti totali in relazione al personale Fte[2] presente ci indica quanti prestiti vengono effettuati da ciascuna unità di personale, ed evidenziano l'entità dei cariche di lavoro per ciascun addetto. Escludendo i valori anomali, già discussi sugli altri indicatori, i dati sembrano avere una grande variabilità: lo scarto interquantile è pari a 650,6 quasi quanto la mediana (677,4). L'interpretazione di questi dati sembra alquanto contraddittoria: un valore piccolo significherebbe un basso indice di efficienza, al contrario un alto numero di prestiti per persona segnala pesanti carichi di lavoro.
5. Questo indicatore, come quello precedente, è in grado di fornire un'informazione sui carichi di lavoro del personale in biblioteca, soprattutto quelli inerenti all'attività di *back office*. Il comportamento dell'indicatore è simile a quello precedente: non considerando gli *outliers*, la distribuzione è abbastanza simmetrica intorno al valore mediano pari a 187,5 acquisti per persona Fte e la maggior parte degli Atenei si attesta intorno a 100 / 300 acquisti per addetto.
6. Il rapporto tra il totale del patrimonio documentario e il personale Fte mette in evidenza il carico di lavoro svolto nella gestione e movimentazione del materiale librario. Dai valori assunti dall'indicatore, la distribuzione dei dati è sufficientemente simmetrica attorno al valore mediano pari a 7620. Sia la deviazione standard (6971), sia un indice di variabilità più robusto come lo scarto interquantile (7940), indicano che i dati presentano una forte dispersione; sintomo che non c'è una omogeneità di gestione del patrimonio documentario da parte degli Atenei.

3.3.4 Vitalità del patrimonio - Offerta di risorse

In questa sezione si sono calcolati alcuni indici in grado di valutare la vitalità del patrimonio bibliografico, l'interazione delle biblioteche con gli utenti e misurare l'adeguatezza complessiva dell'offerta di servizi al pubblico.

[2] *Full time equivalent.*

Grafici 3.5 Box - plot dei dati relativi all'ultima sezione degli indicatori.



Da come si può notare dai *boxplot* nei primi cinque indici vi è la presenza di alcuni valori anomali positivi che portano a "schiacciare" la distribuzione dei dati e quindi a rendere quasi illeggibile ciò che vorrebbe trasmettere la rappresentazione grafica. Questo insieme di *outliers* è dovuto soltanto ad un ristretto gruppo di Atenei che riportano valori che sono sicuramente non confrontabili con quelli delle altre Università. La causa di questo va ricercata nella struttura di tali particolari Atenei o semplicemente in una errata compilazione del questionario. Tuttavia, scartando questi valori anomali si è in grado di poter analizzare in modo corretto questi indici.

1. L'indicatore che mette in rapporto l'utenza potenziale con l'indice Fte è in grado di valutare l'adeguatezza del proprio personale in rapporto al proprio bacino d'utenza. Dal *box-plot* precedente si può notare la presenza di 4/5 Atenei con valori anomali: ciò può essere dovuto al fatto che queste Università, di piccola dimensione, non hanno potuto fornire in maniera corretta né il dato riguardante l'utenza potenziale, né quello dell'effettivo numero di personale afferente alle biblioteche. Scartando dall'analisi questi valori, si nota che la distribuzione è simmetrica attorno al valore mediano di 292 utenti per addetto bibliotecario, con una consistente deviazione

standard (150). Questo valore indica che la gestione dell'organico bibliotecario risulta abbastanza disomogenea sul territorio nazionale.

2. La distribuzione delle risorse umane nelle strutture è evidenziata dal secondo indicatore di vitalità, che relaziona la numerosità dell'organico Fte con il numero di punti di servizio presenti sul territorio. Dal *box-plot* si evidenzia un *outlier*: sembra che tale valore sia dovuto ad una errata compilazione del questionario *on-line*. Trascurato questo dato, la distribuzione dell'indicatore non è simmetrica, con una concentrazione dei valori verso il basso. La mediana risulta pari a 4,155, quindi leggermente più alta rispetto alla precedente indagine svolta dal GIM.
3. Il numero di documenti bibliografici per utente è una misura della capacità delle biblioteche di soddisfare la richiesta formativa della propria utenza. Scartato, in partenza, il valore relativo alla Scuola normale superiore di Pisa, la distribuzione dei valori sembra abbastanza simmetrica con un valore centrale pari a 26,82, con valori molto diversi tra gli Atenei.
4. La quantità di periodici correnti, sia di titoli elettronici che di abbonamenti cartacei, commisurato al numero di docenti e ricercatori dà una valutazione dell'offerta di contenuti per la ricerca in rapporto al corpo accademico. Ciò che emerge è la grande varietà dei valori calcolati per ogni Ateneo e questo tende a sottolineare la grande promiscuità nell'ambito della ricerca nel territorio italiano. Basandosi soltanto sui dati, la mediana risulta pari a 7,9, un valore di poco superiore rispetto all'indagine GIM.
5. Il numero di acquisizioni per ogni singolo utente è un indicatore di quanto le biblioteche investono nell'aumento dell'offerta di materiale complessivo. Ciò che emerge dal *box-plot*, oltre alla presenza dei soliti *outliers*, è che la distribuzione dei dati è abbastanza simmetrica attorno al valore di 0,63 acquisti per utente. Questo valore può essere considerato molto negativo, ma va commisurato al fatto che l'utenza potenziale comprende una ampia gamma di figure, volendo essere, questa misura di utenza, la più inclusiva possibile.
6. La percentuale di periodici elettronici sul totale serve a verificare l'adeguamento delle biblioteche all'introduzione dei nuovi supporti informatici. Dalla rappresentazione grafica si nota la grande variabilità presente nella distribuzione

dei dati e, inoltre, non si è in grado nemmeno di ricavare un'informazione sommaria dalla mediana o dalla media, perché esse sono pari circa al 50%.

7. L'indicatore di vitalità 7 fornisce la percentuale delle uscite economiche dovuta alla spesa per le risorse bibliografiche delle biblioteche. Da come si può notare dal *box-plot* e da una breve analisi dei dati, la distribuzione dei valori è simmetrica attorno al valore della media-mediana pari al 50%, con poca variabilità, sintomo del fatto che nella maggior parte delle biblioteche accademiche si dà una importante rilevanza alle spese dovute alla gestione del proprio capitale bibliografico.
8. L'ultimo indicatore di questa sezione serve a stimare la vitalità del patrimonio. Infatti, il rapporto tra i prestiti totali attivi e il patrimonio documentario fornisce una valutazione di come una biblioteca risponde alle esigenze dell'utenza. Da una lettura degli indici, con la mediana pari 0,09 e la media pari a 0,14 prestiti per documento, l'indicatore denota una scarsa vitalità complessiva degli Atenei italiani. Indubbiamente queste indicazioni devono esser lette tenendo conto che, almeno negli Atenei più grandi, una consistente parte del patrimonio documentario dichiarato rimane in magazzino e quindi non facilmente visibile dalla clientela.

CAPITOLO 4

L'ANALISI FATTORIALE E DELLE COMPONENTI PRINCIPALI

4.1 L'Analisi Fattoriale

L'analisi fattoriale è un metodo statistico idoneo a ridurre un sistema complesso di correlazioni in un numero minore di dimensioni. Per uno studio approfondito dal punto di vista matriciale rimandiamo a *Fabbris (1990).* Questa tecnica ha avuto largo impiego nella psicologia come modello matematico per la formalizzazione di teorie nell'ambito degli studi sui test mentali e attitudinali e sul comportamento umano. Attualmente tale tecnica viene utilizzata in diversi campi, sociale, psicologico, economico, e gli impieghi più ricorrenti sono:

- ridurre la complessità di una matrice di dati, riducendo il numero delle variabili;
- semplificare la lettura di un fenomeno;
- verificare ipotesi sulla struttura delle variabili, in termini di numero di fattori significativi, sui loro legami, sulle cause comuni che agiscono sulle loro manifestazioni;
- misurare costrutti non direttamente osservabili a partire da indicatori osservabili ad essi correlati.

4.1.1 Modello dell'analisi fattoriale

In generale la correlazione tra due variabili aleatorie X_1 e X_2 può risultare dall'associazione di entrambe con una terza variabile F. Se il coefficiente di correlazione parziale tra X e Y rispetto a F ha un valore prossimo allo 0, allora F spiega quasi completamente la relazione tra X_1 e X_2. A livello multivariato, partendo da una variabile aleatoria X k-dimensionale, se esiste un insieme F di variabili (il meno numeroso possibile) tale che tutte le correlazioni parziali tra gli elementi di X per determinati valori degli elementi di F sono significativamente nulle, allora gli elementi di F spiegano completamente l'interdipendenza tra gli elementi di X. L'incorrelazione condizionata è una condizione necessaria affinché l'insieme F di variabili condizionanti offra una spiegazione adeguata della correlazione tra le componenti della X. L'obiettivo è quello di spiegare

l'interdipendenza esistente all'interno di un insieme numeroso di variabili tramite un numero esiguo di *fattori* non osservabili sottostanti, incorrelati tra loro. In questo senso l'analisi fattoriale costituisce un superamento dell'analisi delle componenti principali in quanto, piuttosto che nella semplice trasformazione sintetica delle variabili osservate, consiste nella stima di un modello che riproduca la struttura della covarianza tra le stesse. Lo studio della relazione tra k variabili tramite m *fattori comuni* a tutte le variabili e k *fattori specifici* di ciascuna variabile si è sviluppato a partire dalle idee di Galton (1898) e Pearson, e grazie alle prime applicazioni in ambito psicometrico (Spearman,1904). Più tardi con il calcolo delle stime di massima verosimiglianza dei fattori (Lawley, 1940), la metodologia fu completamente formalizzata dal punto di vista inferenziale. Nell'analisi fattoriale ciascuna variabile viene espressa come funzione lineare di un certo numero m di fattori comuni, responsabili della correlazione con le altre variabili, e di un solo fattore specifico, responsabile della variabilità della variabile stessa:

$$X_j = a_{j1}F_1 + \ldots + a_{jq}F_q + u_j c_j \qquad \text{con } (j = 1,\ldots,p). \qquad (4.1)$$

Questo modello somiglia solo apparentemente a quello di regressione multipla, infatti i fattori $F_1,\ldots, F_p$ non sono osservabili: tutto ciò che giace a destra dell'uguaglianza è dunque incognito.

4.1.2 Definizione del modello di analisi fattoriale

Si supponga di aver osservato un insieme di p variabili quantitative o dicotomiche presso n unità statistiche (con n abbastanza elevato rispetto a p), di aver ordinato le osservazioni nella matrice X il cui elemento generico x_{hj} denota il valore della variabile x_j osservato presso l'unità h, e di aver successivamente standardizzato i dati (le variabili hanno media nulla e varianza unitaria). Il modello di analisi fattoriale si esprime con l'equazione dove i deponenti relativi alle unità statistiche sono stati soppressi per semplificare l'esposizione,

$$X_j = a_{j1} + F_1 + \ldots + a_{jq}F_q + u_j c_j = \sum_{i}^{q} a_{ji}F_i + u_j c_j \qquad \text{con } (j = 1,\ldots,p), \qquad (4.2)$$

con F_i $(i = 1,\ldots,q)$ fattore comune i-esimo *(variabile latente);* a_{ji} coefficiente che lega il fattore F_i alla variabile x_j ,ed è detto *peso fattoriale (factor loading);* c_j fattore specifico

di x_j e u_j è il suo coefficiente. Nella notazione matriciale, il modello consiste nella scomposizione della matrice di dati in matrici di fattori comuni e specifici:

$$X = F\, A_q{}^T + E \qquad (4.3)$$

dove F è la matrice n x q di fattori, A_q è una matrice di pesi fattoriali di ordine p x q, mentre $E = C\, U$ è una matrice n x p di fattori specifici e U è la matrice diagonale di coefficienti dei fattori specifici $c_1, c_2, \ldots, c_p$.

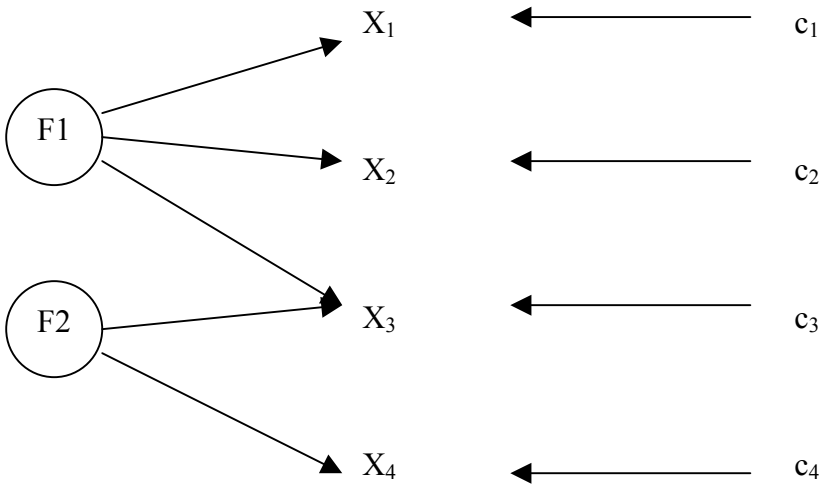
Nel modello fattoriale vengono fatte le seguenti ipotesi:

- $Corr\,(F_i, F_j) = 0 \qquad \forall\ i, j\ ;$
- $Corr\,(c_i, c_j) = 0 \qquad \forall\ i, j\ ;$
- $Corr\,(c_i, F_j) = 0 \qquad \forall\ i, j\ .$

Il fattore F_i si dice comune perché è presente in tutte le p possibili equazioni; se ha coefficienti non nulli con tutte le variabili, si dice generale; c_j si dice specifico perché appartiene solo alla variabile x_j. Ogni fattore comune è combinazione di tutte le variabili osservate:

$$F_i = \sum_j^p w_{ji} x_j \quad con\ (i=1,\ldots,q)\ ,$$

dove w_{ji} è il coefficiente fattoriale *(factor score coefficient)* della variabile x_j nella combinazione F_i. Adottando il modello di analisi fattoriale si assumono dunque relazioni lineari ed additive tra le variabili osservate. Graficamente un modello fattoriale può essere così rappresentato: esso è formato da due fattori ortogonali F, quattro variabili x e quattro fattori unici c.



(4.4)

4.1.3 Comunanza e unicità dei fattori

La formula (4.2) ha la forma di un'equazione di regressione dove x_j è la variabile dipendente, i fattori sono le esplicative e c_j il termine residuale. Per analogia con l'analisi di regressione, se una variabile è esprimibile in funzione di fattori comuni e di un fattore specifico, anche la sua varianza è scomponibile in due parti: la varianza comune *(comunanza)* e la varianza unica *(unicità)*. Se i fattori sono incorrelati tra loro e con quello specifico, per ogni x_j vale l'identità:

$$\sigma_j^2 = 1 = Var\left[\sum_i^q a_{ij} F_i + u_j c_j\right] \quad con\ (j = 1,\ldots,p). \qquad (4.5)$$

La comunanza h_j^2 è la frazione di varianza di x_j spiegata dall'insieme dei fattori comuni. Essendo il coefficiente di correlazione tra la variabile x_j e il fattore f_j uguale al peso fattoriale, $r_{ij} = a_{ij}$, la comunanza, data dalla somma del quadrato dei coefficienti di correlazione con i singoli fattori comuni, è anche ottenibile sommando il quadrato dei pesi fattoriali:

$$h_j^2 = \sum_i^q r_{ji}^2 = \sum_i^q a_{ji}^2 \qquad con\ (j = 1,\ldots,p). \qquad (4.6)$$

Proprio per essere fattori comuni a tutte le variabili, si può dire che la comunanza di una variabile è la parte di varianza che questa condivide con le altre variabili fattorizzate.

4.2 Il metodo di analisi delle componenti principali

L'analisi delle componenti principali è un metodo di trasformazione matematica di un insieme di variabili in uno nuovo di variabili composite (componenti principali) ortogonali tra loro e che spiegano la totalità della variabilità del fenomeno. Per un'analisi approfondita di questo argomento è rimandiamo a *Saporta e Bouroche (1978)* e *Bolasco (1999)*. Essa si distingue dall'analisi fattoriale in quanto vengono considerate tutte le componenti principali, anche se solo alcune saranno poi utilizzate a fini interpretativi. A differenza della analisi fattoriale, le componenti principali sono ottenute per via algebrica, non fissando alcun vincolo per la costruzione delle stesse. Le componenti principali, sono delle combinazioni lineari delle variabili di partenza, che nell'insieme ricostituiscono la variabilità originaria. Il calcolo matriciale delle

componenti principali è discusso in maniera dettagliata e comprensibile da Fabbris (1997). In generale la generica variabile $x_{j.}$ è funzione lineare di tutte le possibili componenti principali estraibili (pari al rango della matrice di correlazione):

$$x_j = \sum_i^r a_{ji} F_i \qquad (j = 1,...,p) \qquad (4.7)$$

che in notazione matriciale diventa $X=FA^T$, dove la matrice F comprende tutte le componenti F_i e la matrice A i pesi fattoriali.

Le componenti principali si ricavano identificando in sequenza la combinazione lineare delle variabili osservate che estrae la quota massima di variabilità man mano depurata della variabilità delle componenti principali estratte. La prima componente sarà quella a varianza maggiore, generalmente indicata con λ_1 e chiamata autovalore, per cui valgono le seguenti relazioni:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \quad \text{e} \quad \sum_i \lambda_i = \sum_i \text{var}(x_i) \qquad (4.8)$$

Applicando l'analisi delle componenti principali si assiste pertanto ad una ridistribuzione della varianza totale, con una forte concentrazione nelle prime componenti principali. Questa tecnica può essere vista come una generalizzazione della dell'analisi fattoriale classica, dovuto al fatto che la conduzione dell'analisi è molto simile e porta spesso a risultati comuni.

4.3 Il procedimento dell'analisi fattoriale e delle componenti principali

Nello svolgere un'analisi fattoriale vanno prese alcune decisioni. In genere si deve:

1. Identificare la matrice sulla quale si svolgerà l'analisi. Solitamente si considera la matrice di correlazione R o la matrice di varianze e covarianze.
2. Stabilire il numero di fattori da estrarre. Il numero massimo di fattori che possono essere considerati è r (rango della matrice di correlazione) anche se solitamente ne vengono utilizzati un numero inferiore. Tale scelta deve essere coerente con i principi di parsimonia della soluzione finale, ossia il numero di fattori deve essere inferiore a quello delle variabili, e di partecipazione di ogni fattore all'interpretazione della variabilità dei fenomeni osservati.

3. Definire il criterio di estrazione dei fattori. Solitamente inizialmente vengono estratte tutte le componenti principali e viene trasformata la matrice delle informazioni in funzione di queste. Quindi si pongono sulla diagonale della matrice *R* le relative comunanze in modo da ottenere una soluzione fattoriale unica.
4. Determinare il criterio di rotazione degli assi ortogonali trovati. Le rotazioni, che possono essere ortogonali od oblique, modificano i fattori in modo da rendere più realistici e semplici i fattori e facile l'interpretazione finale dell'analisi.
5. Calcolare i punteggi fattoriali, ossia il valore che una unità statistica ha sul fattore, valutare e interpretare i fattori.

4.3.1 Criteri per determinare il numero dei fattori

Solitamente il numero dei fattori non è noto a priori per cui l'analista inizia con quello che crede sia il numero più probabile di fattori e poi, per approssimazioni successive, trova la soluzione più congruente con gli obiettivi della ricerca. I criteri maggiormente utilizzati per la determinazione del numero dei fattori sono due: uno basato sulla varianza spiegata dai fattori e uno sulla rappresentazione grafica degli autovalori.

- *Varianza spiegata dai fattori*. Questo criterio consiste nell'estrarre un numero di fattori tale per cui venga spiegata una certa quota di varianza. Ricordiamo che l'autovalore λ_i del fattore *i-esimo* è la sua varianza e la somma degli autovalori è uguale alla somma delle varianze se l'analisi è condotta su una matrice di varianze – covarianze e a *p,* numero di variabili, se è condotta su una matrice di correlazione. Se l'analisi è condotta su una matrice di varianze – covarianze, la quota di varianza estratta dal fattore *i* è:

$$\lambda_i \Big/ \sum_k^r \lambda_k = \lambda_i \Big/ \sum_k^p s_k^2 \qquad (4.9)$$

  Se l'analisi è condotta sulla matrice di correlazione, la varianza spiegata da ogni singolo fattore sarà

$$\lambda_i \Big/ \sum_k^r \lambda_k = \lambda_i \Big/ p \; . \qquad (4.10)$$

Una percentuale di varianza del 75% è considerata un buon traguardo, anche se spesso si tollerano percentuali inferiori a questo valore. La frazione di varianza complessivamente estratta si valuta in funzione del numero di variabili inserite nell'analisi e dal tipo di impiego che si farà delle nuove variabili latenti costruite.

- Rappresentazione grafica degli autovalori. La rappresentazione grafica degli autovalori λ_i in relazione all'ordine di estrazione i permette di individuare gli autovalori importanti. Rappresentando i punti $(i, \lambda_i) (i = 1,..., r)$ sul piano cartesiano e collegandoli con segmenti, si ottiene una spezzata: se questa mostra due tendenze (una forte inclinazione all'altezza dei primi fattori e un successivo appiattimento che la porta ad essere quasi parallela all'asse delle ascisse) i fattori che appartengono a quest'ultima parte della spezzata possono essere ignorati. Si considerano rilevanti per l'analisi solo i fattori il cui autovalore, stando più in alto del flesso, descritto dalle due tendenze, si stacca visibilmente dagli altri. Inoltre, per motivi che ora non analizziamo, vengono considerati di rilevante importanza solo gli autovalori superiori all'unità. Se non ci sono fattori che prevalgono nettamente sugli altri allora significa che l'analisi fattoriale non è un metodo adatto per l'analisi di quei dati.

4.3.2 Rotazione dei fattori

I pesi fattoriali a_{ji} *(factor loadings)* coincidono con i coefficienti di correlazione tra le variabili iniziali e i fattori ed indicano quanto la variabile sia determinante per il fattore. Dall'analisi della matrice dei pesi fattoriali è possibile riuscire a comprendere quali variabili contribuiscono maggiormente alla definizione del fattore e quindi alla sua interpretazione ed essa inizialmente viene prodotta senza essere sottoposta ad alcuna rotazione. La rotazione dei fattori, o degli assi, è pertanto un cambiamento di posizione delle dimensioni estratte nella prima fase dell'analisi che facilita la comprensione del significato dei fattori stessi.

La rotazione procede alla riduzione del valore dei pesi fattoriali marginali, ossia quelli che nella costruzione originaria dei fattori risultano essere relativamente piccoli, e nell'incremento, in valore assoluto, dei pesi più significativi. La soluzione ideale, ai fini dell'interpretabilità dei fattori, è quella in cui tutti i pesi fattoriali siano prossimi a *0* o a *1*. La rotazione comporta per

tanto una ridistribuzione delle comunanze delle variabili e della varianza spiegata dai fattori.
I principali criteri di rotazione ortogonali e non sono Varimax, Quartimax, Equamax, Promax.

Varimax:

La rotazione con il metodo Varimax tende a minimizzare il numero di variabili con cui ciascun fattore ha coefficienti di correlazione elevati. Tale criterio è raccomandabile se si vuole ottenere una netta separazione tra i fattori e se la rotazione è effettuata senza precisi criteri di riferimento.

Quartimax:

Tale criterio semplifica le righe della matrice dei pesi fattoriali, cercando di stabilire la corrispondenza tra la variabile sulla riga e uno o pochissimi fattori. Tale criterio è adatto per identificare i fattori che governano la variabilità delle caratteristiche osservate e dà risultati migliori del metodo precedente quando si vuole semplificare il primo fattore estratto, che tende ad essere un fattore generale.

Equamax:

E' un compromesso tra i due criteri precedenti in quanto tenta di realizzare la semplificazione simultanea di righe e colonne della matrice dei pesi fattoriali. Non si adatta efficacemente a strutture semplici.

Promax:

A differenza delle precedenti, il metodo Promax è un procedimento iterativo di rotazione non ortogonale degli assi. Tale tecnica parte con una rotazione ortogonale Varimax dei pesi originari. Poi cerca una trasformazione dei pesi ruotati che incrementi i pesi già grandi in assoluto e riduca quelli più piccoli. L'angolo tra gli assi varia secondo passi predeterminati fino a raggiungere una posizione ottima. In questo caso gli assi possono essere molto ravvicinati e si perde la condizione di ortogonalità dei fattori e la correlazione può risultare elevata

4.4 Analisi dei dati

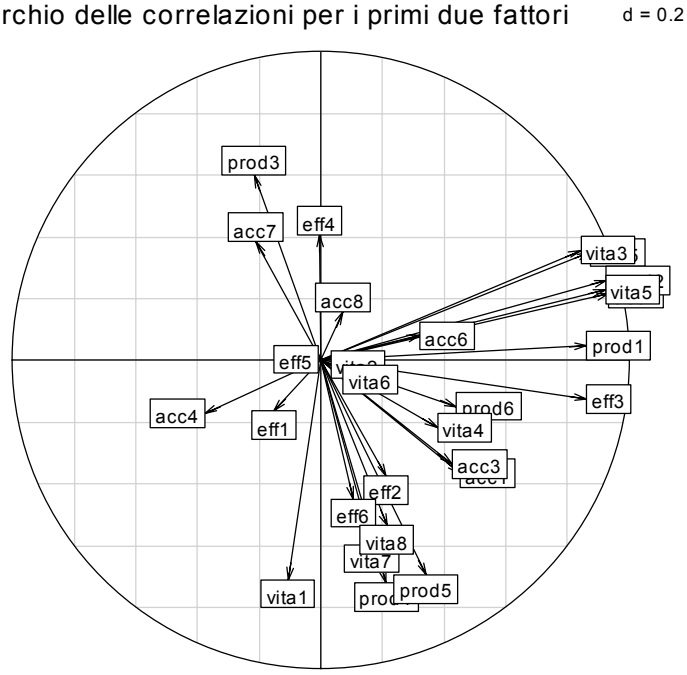
Usare un modello di analisi fattoriale per cercare di spiegare l'interdipendenza esistente all'interno di un insieme numeroso di variabili è un metodo alquanto oneroso dal punto di vista temporale, non tanto per la complessità matematica, ma per ricercare una giusta spiegazione a ciò che emerge dall'analisi.

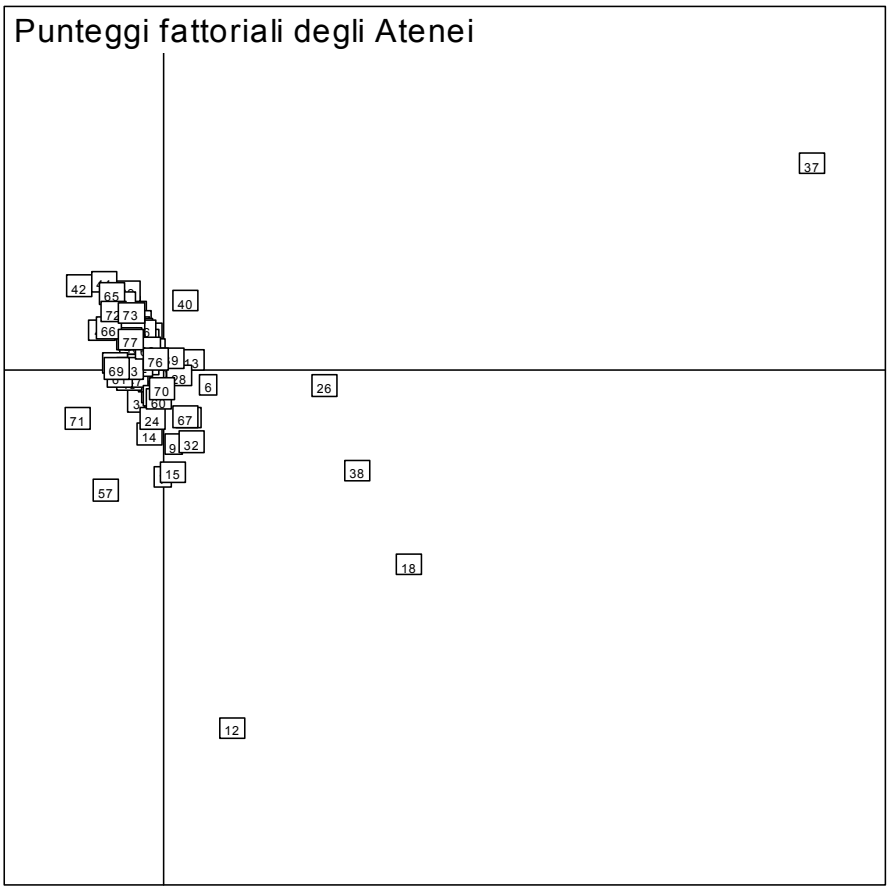
Nel pacchetto statistico *R* l'analisi fattoriale viene eseguita usando la funzione *"factanal()"* contenuta nella libreria *"stats"*, che stima i fattori massimizzando la funzione di verosimiglianza basandosi sulla matrice di covarianza dei dati. Sempre nell'ambiente R, l'estrazione delle componenti principali è stata effettuata con la funzione *"dudi.pca"* incorporata nella libreria *"ade4"*, e con la funzione *"princomp()"* della libreria *"stats"*.

Il *dataset* di 28 indicatori presentati nel Capitolo 3, sono raggruppati in 4 aree tematiche di appartenenza. Da una prima analisi delle correlazioni tra i dati, non è presente alcun insieme di dati correlati in modo significativo tra loro. Questo è dovuto essenzialmente al fatto che ogni indicatore ha una scala di misura diversa e quindi, le relazioni, se esistenti, tra le variabili risultano estremamente indebolite. Per cercare di avere degli indici "simili" tra loro si è sono standardizzati i valori degli indicatori, utilizzando la funzione *"scale()"*di R. Anche dopo la standardizzazione, vi è una scarsa struttura di correlazione tra gli indicatori scelti per l'analisi, e in più ci si pone il problema di come comportarsi alla costante presenza di valori *outliers.* Come suggerito da Bolasco (1999, cap. 6), quando esistono modalità sistematicamente ridondanti, come nel caso di risposte non dovute, queste devono essere eliminate, dal momento che produrrebbero, come risultato, degli assi falsati dalla loro ridondanza. Dalle analisi descrittive si nota come ci siano alcuni Atenei che hanno ripetuti valori anomali in gran parte degli indicatori, per cui si decide che è opportuno togliere tali unità al fine di non alterare le analisi.

Sia l'analisi fattoriale che l'analisi delle componenti principali sono tecniche in grado di scovare fattori latenti all'interno dei dati. Dopo alcune prove a livello iniziale, la scelta della tecnica migliore è ricaduta sull'analisi delle componenti principali: quest'ultima, non ipotizzando nessuna struttura sottostante per la stima dei fattori, sembra cogliere in modo leggermente migliore la variabilità degli indicatori, che come visto in precedenza appaiono scarsamente correlati tra loro.

Grafici 4.11 - 4.12 Cerchio delle correlazioni dei pesi fattoriali attribuiti agli indicatori dalle prime due componenti principali e rappresentazione grafica dei punteggi fattoriali ottenuti dagli Atenei.





L'analisi in componenti principali di tutti gli indicatori non riesce ad essere sufficientemente esplicativa. Infatti i 2 fattori ricavati spiegano soltanto il 39% della variabilità complessiva. Dal grafico 4.11, si può notare come siano veramente pochi gli indicatori che hanno una correlazione abbastanza consistente con uno dei due fattori. Inoltre, confrontando l'andamento dei pesi con il grafico dei punteggi fattoriali, si nota che l'analisi è falsata dalla presenza di valori anomali: infatti gli Atenei con valori caratterizzati da *outliers* introducono nel calcolo dei pesi fattoriali una distorsione che comporta un maggior peso dato a quegli indicatori che presentano valori estremi in queste unità. Tale effetto è dimostrato dall'effetto grafico di "compressione" esercitato dall'Ateneo con id 37 (Scuola normale superiore di Pisa) nella distribuzione dei punteggi fattoriali nei due assi.

Uno studio successivo è stato effettuato cercando di eliminare dallo studio quegli Atenei con valori così anomali: l'esito non è stato confortante per la presenza di un folto numero di valori estremi. Inoltre non si sono avuti significativi aumenti della varianza spiegata dalle componenti principali calcolate e l'interpretazione degli *output* grafici non è migliorata in modo apprezzabile.

Partendo dal fatto che la matrice dei dati originaria contenga valori molto eterogenei tra di loro, si è cercato di effettuare un'analisi delle componenti principali degli indicatori dividendoli per ogni area d'interesse. Teoricamente gli indicatori facenti capo ad ogni sezione dovrebbero essere molto correlati tra loro e quindi la varianza dovrebbe essere spiegata grazie un numero esiguo di fattori.

4.4.1 Analisi fattoriale degli indicatori di Accessibilità

Con questa analisi si vuole cercare di ricavare un esiguo numero di fattori in grado di spiegare in maniera adeguata i livelli di accessibilità riscontrati dai dati raccolti dal questionario GIM. Potendo contare su una matrice di analisi di grandezza minore rispetto alla precedente si è potuti analizzare in maniera più accurata i valori degli indicatori, facendo leva sui *box-plot* dei dati del paragrafo 3.4.2. Da subito si è cercato di eliminare i valori di quegli Atenei che potrebbero sfalsare l'analisi. Quindi sono state tolte dall'analisi i valori di queste Strutture: Scuola internazionale superiore degli studi avanzati di Trieste, Università degli studi di Pisa, Scuola normale superiore di Pisa, Libera Università di Bolzano e Università degli Studi di Catanzaro Magna Grecia, per un totale di 5 Atenei.

Grafico 4.13 Rappresentazione della correlazione esistente tra gli indicatori di accessibilità.

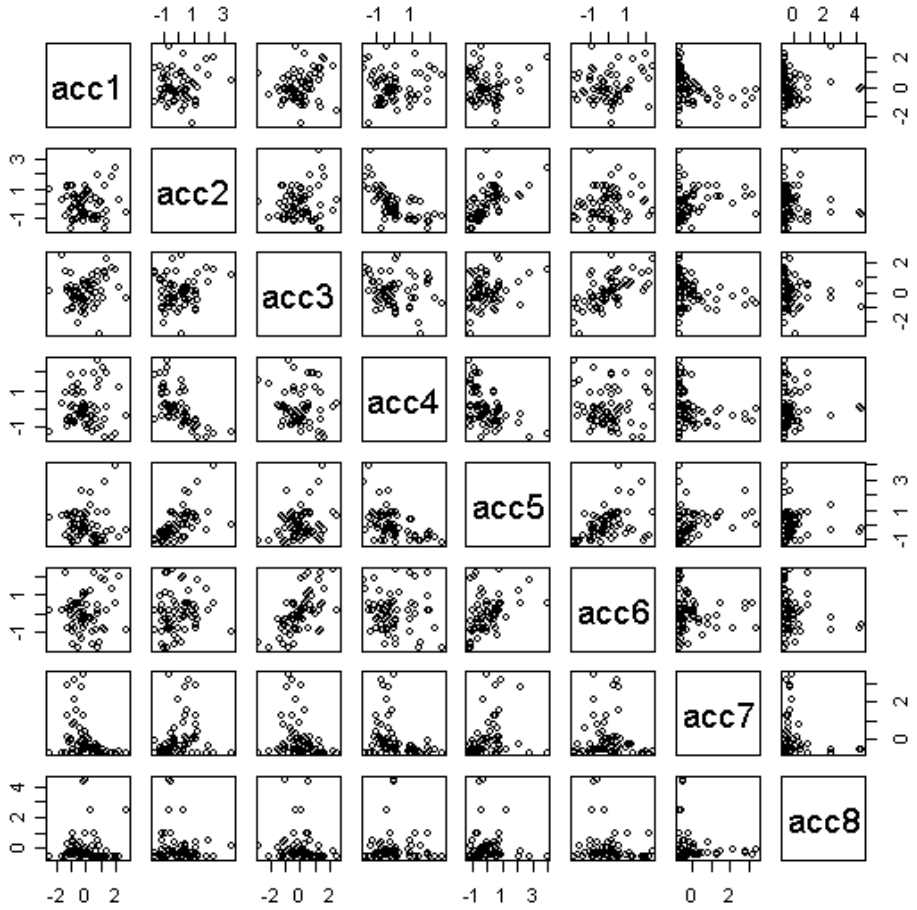


Tabella 4.14 Correlazione esistente tra gli indicatori di accessibilità.

	acc1	acc2	acc3	acc4	acc5	acc6	acc7	acc8
acc1	1,000	0,039	0,202	0,040	0,007	0,131	-0,350	0,048
acc2	0,039	1,000	0,135	-0,670	0,597	0,110	0,186	-0,105
acc3	0,202	0,135	1,000	-0,160	0,287	0,616	-0,232	-0,110
acc4	0,040	-0,670	-0,160	1,000	-0,516	-0,081	-0,241	-0,005
acc5	0,007	0,597	0,287	-0,516	1,000	0,343	0,272	-0,021
acc6	0,131	0,110	0,616	-0,081	0,343	1,000	-0,067	-0,129
acc7	-0,350	0,186	-0,232	-0,241	0,272	-0,067	1,000	-0,121
acc8	0,048	-0,105	-0,110	-0,005	-0,021	-0,129	-0,121	1,000

Dal grafico precedente si può notare come non esistano forti legami tra gli indicatori introdotti nell'analisi, ma sussistano medio-deboli relazioni lineari.

Una media relazione lineare positiva esiste tra gli indicatori acc2 e acc5 e tra acc3 e acc6, mentre sono correlati negativamente gli indicatori acc2 e acc4.

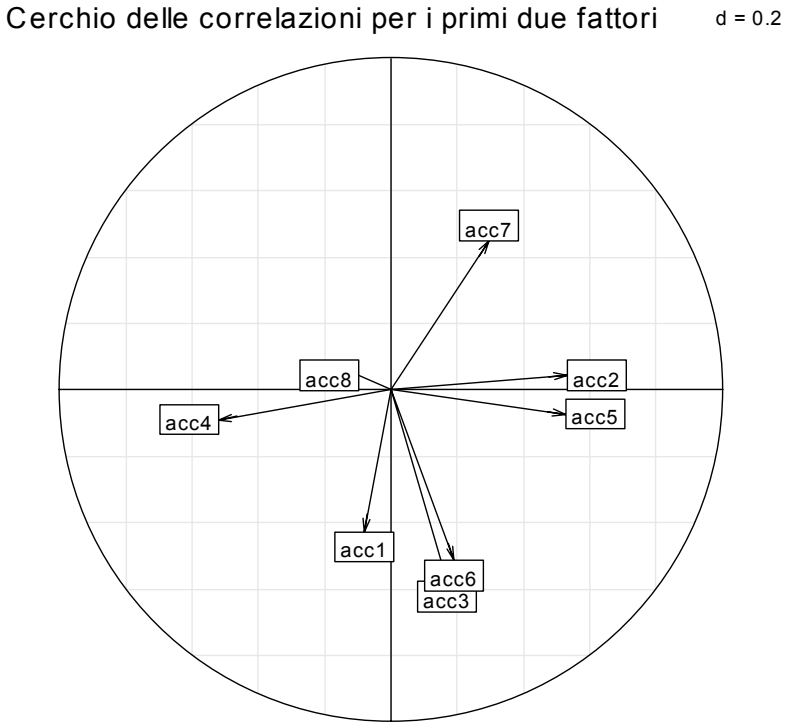
Con l'obiettivo di trovare due fattori nascosti, si è effettuata un'analisi delle componenti principali degli indicatori di accessibilità. Dalla rappresentazione dei pesi fattoriali distribuiti sui 2 fattori emerge che il primo fattore è correlato positivamente con l'indicatore 2 e 5 e negativamente con l'indicatore 4, mentre il secondo fattore è in relazione negativa con l'indicatore acc1, acc3 e acc6

Tabella 4.15 Pesi fattoriali ottenuti dai due fattori con rotazione varimax.

	Acc1	Acc2	Acc3	Acc4	Acc5	Acc6	Acc7	Acc8
Fattore 1		0.532	0.170	-0.521	0.527	0.191	0.297	
Fattore 2	-0.427		-0.579			-0.513	0.450	

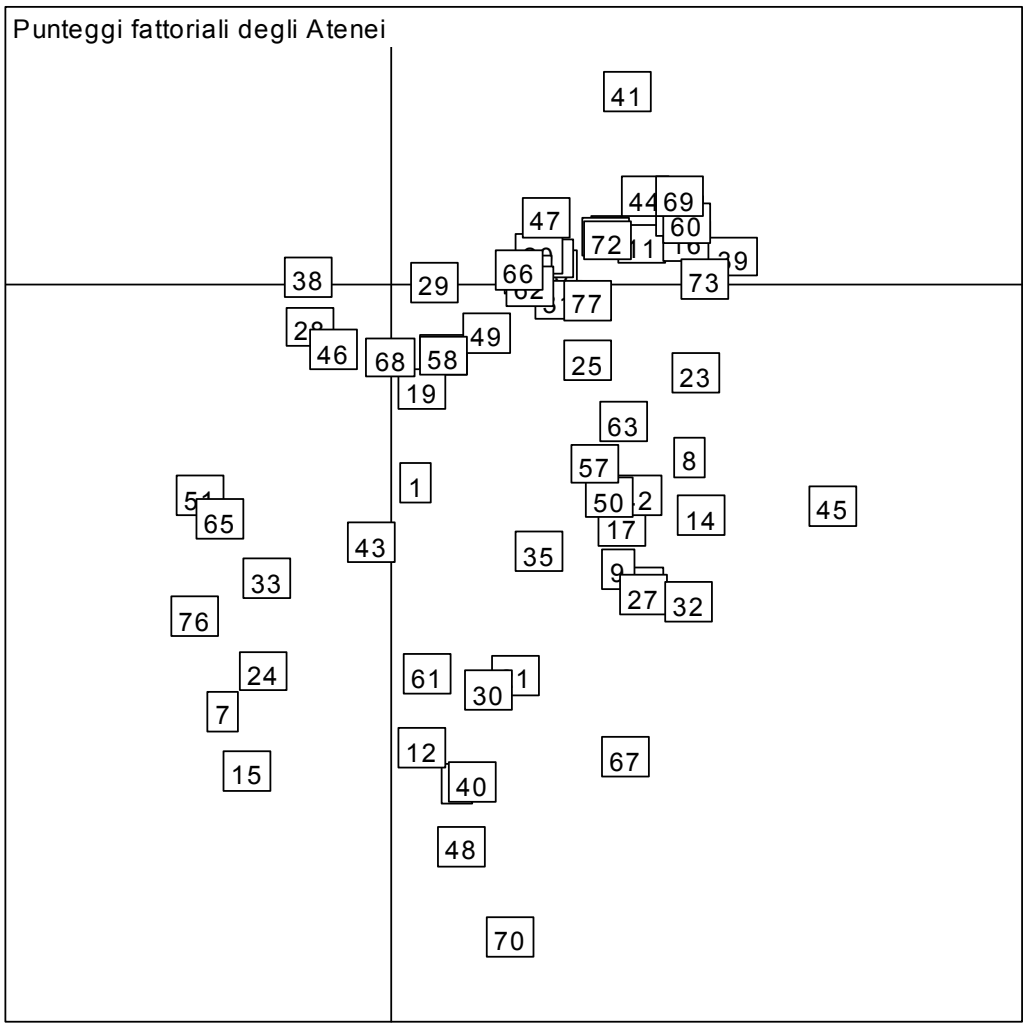
La percentuale di varianza spiegata dai fattori non è molta (53.7%), però i risultati della rappresentazione grafica dei pesi e dei punteggi fattoriali sembra essere buona.

Grafico 4.16 Rappresentazione del cerchio delle correlazioni.



L'interpretazione dei fattori viene effettuata andando a controllare cosa esprimono gli indicatori. Gli indicatori acc2 e acc5 sono correlati positivamente al 1° fattore (mentre lo è negativamente acc4), esprimono come sono gestiti gli spazi e le attrezzature in rapporto all'utenza potenziale, indicando la superficie, gli scaffali e i posti a sedere dedicati ad ogni singolo utente, quindi è una misura del livello generale dell'importanza dell'utenza potenziale da parte di ogni Università. Il secondo fattore ha una media correlazione negativa agli indicatori acc1,acc3 e acc6, e positiva con acc7, quindi risultano con punteggi bassi gli Atenei che hanno una alta percentuale di superficie accessibile e di scaffalatura dedicata al pubblico: la seconda componente principale è un indice dell'accessibilità complessiva delle strutture bibliotecarie. Da queste considerazioni sono da considerare "più accessibili" quelle Università che hanno un valore alto nel primo fattore e basso nel secondo fattore.

Grafico 4.17 Rappresentazione dei punteggi fattoriali degli Atenei (indicatori di accessibilità)



Dal grafico dei punteggi fattoriali non si intravede un suddivisione netta degli Atenei, bensì abbiamo dei gruppi di unità simili. Gli Atenei che hanno strutture bibliotecarie più accessibili in rapporto all'utenza sono stanziati un piccolo gruppo in basso a destra. In questa zona risaltano le organizzazioni bibliotecarie afferenti a strutture come Università degli Studi di Milano (8), Università degli Studi di Roma - "Tor Vergata" (45), Università degli Studi di Padova (23), Università degli Studi di Firenze (35).
Per un'informazione più completa sulla codifica degli indici nel grafico 4.17 rimandiamo alla tabella in appendice.

4.4.2 Analisi fattoriale degli indicatori di Efficacia / Fruibilità / Innovazione

Gli indicatori facente capo a questa area di interesse sono 6. Dall'analisi descrittiva emerge che nel primo indicatore è presente un consistente *outliers.* Si è deciso, quindi, di eliminare i valori dell'Istituto Universitario Suor Orsola Benincasa di Napoli dall'analisi.

Grafico 4.18 Rappresentazione della correlazione esistente tra gli indicatori di questa sezione.

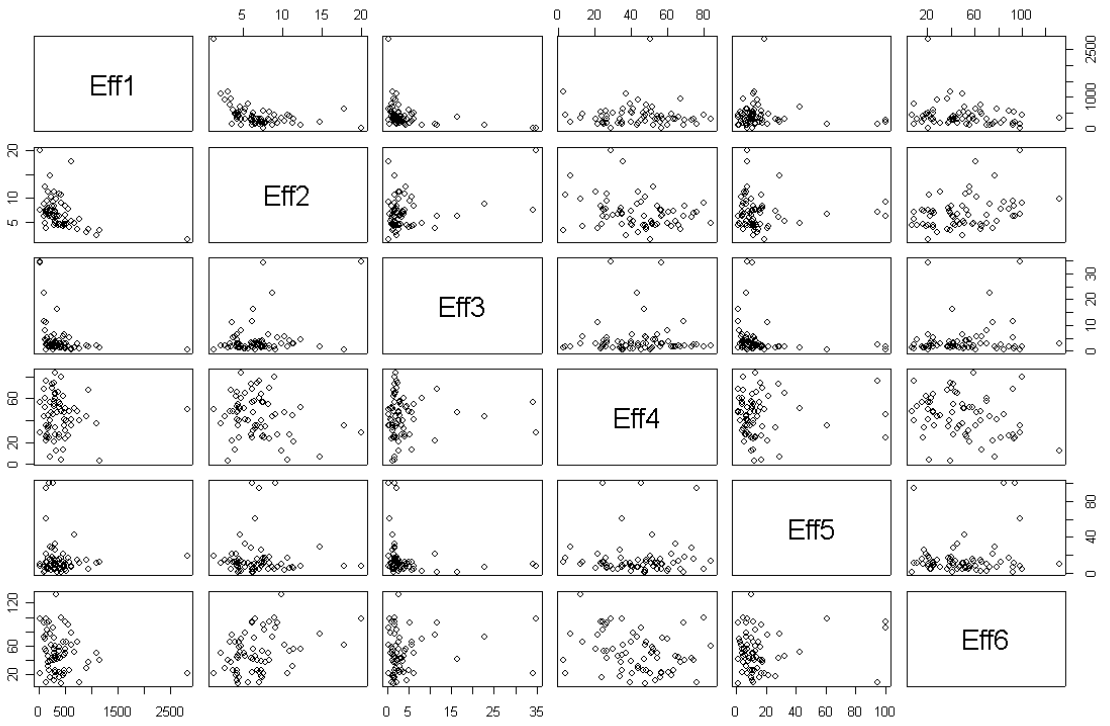


Tabella 4.19 Matrice di correlazione degli indicatori di efficacia.

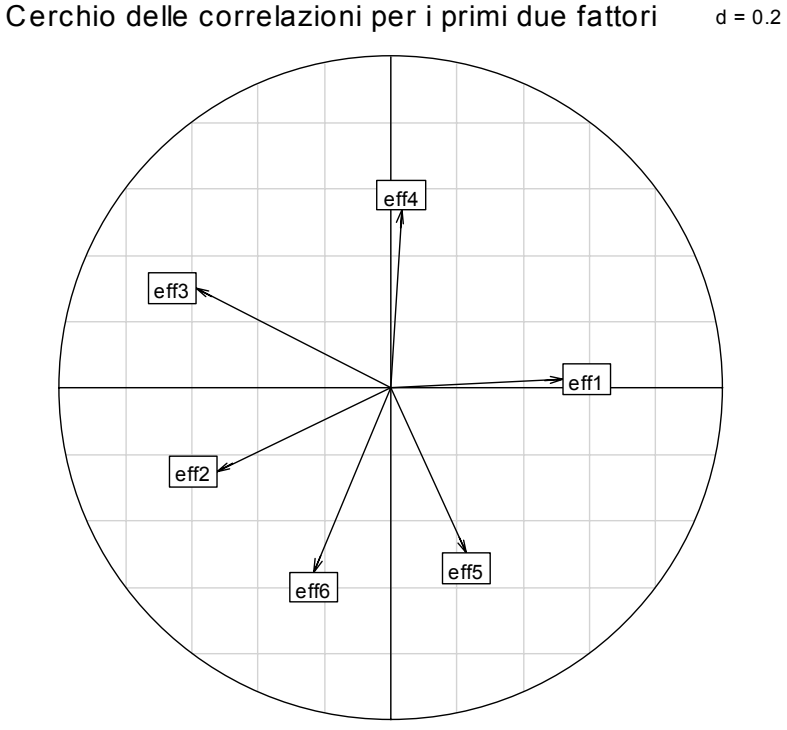
	eff1	eff2	eff3	eff4	eff5	eff6
eff1	1,000	-0,425	-0,301	-0,035	-0,082	-0,211
eff2	-0,425	1,000	0,315	-0,246	-0,011	0,343
eff3	-0,301	0,315	1,000	0,004	-0,186	0,125
eff4	-0,035	-0,246	0,004	1,000	0,029	-0,262
eff5	-0,082	-0,011	-0,186	0,029	1,000	0,119
eff6	-0,211	0,343	0,125	-0,262	0,119	1,000

Dal grafico precedente non si intravedono forti strutture di correlazione. Deboli relazioni lineari si intravedono soltanto tra l'indicatore 2 e 6 e tra 1 e 2. Di conseguenza le prime due componenti spiegano soltanto il 53% della variabilità complessiva, ricordando che gli indicatori raccolti nell'analisi sono soltanto 6.

Tabella 4.20 Pesi fattoriali ottenuti con rotazione varimax.

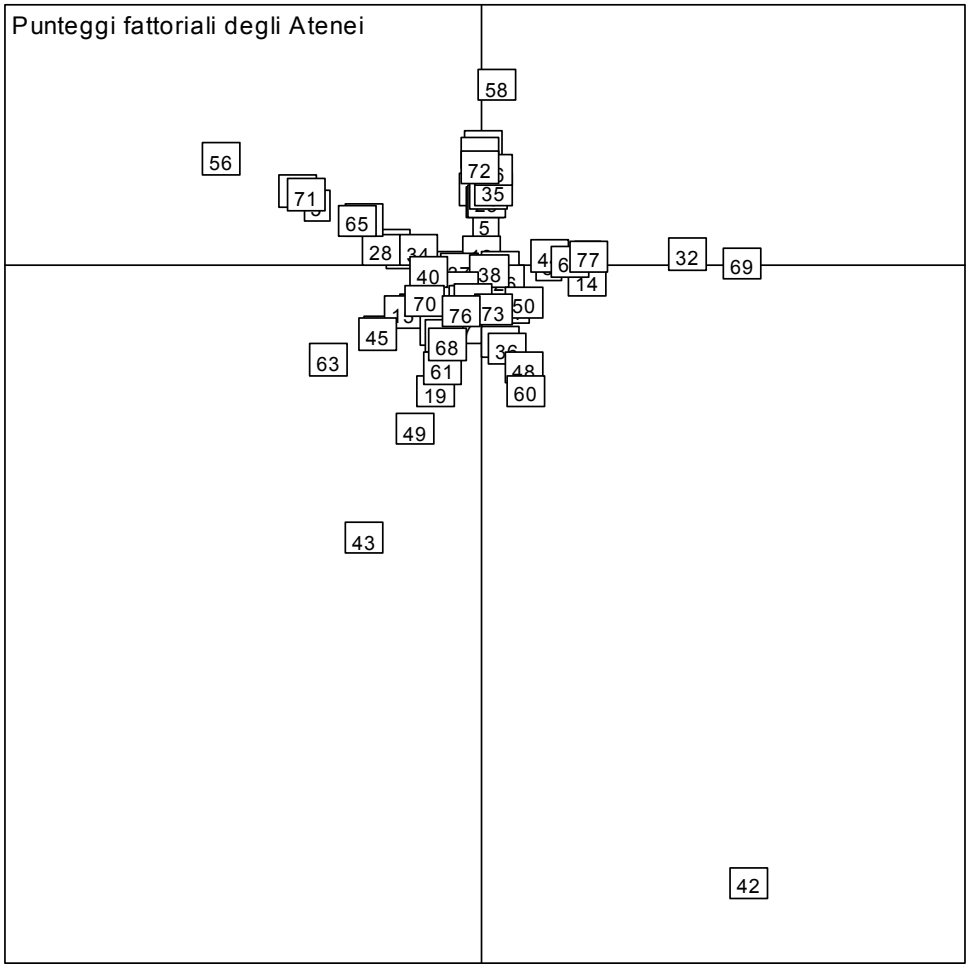
	Eff1	Eff2	Eff3	Eff4	Eff5	Eff6
Fattore 1	0.520	-0.525	-0.589		0.229	-0.233
Fattore 2		-0.251	0.303	0.538	-0.498	-0.554

Grafico 4.21 Cerchio delle correlazioni per gli indicatori di efficacia.



La struttura di correlazione tra i due fattori e gli indici risulta quindi blanda e risulta difficile l'interpretazione pratica dei 2 fattori. Tuttavia il primo fattore, correlato positivamente al primo indicatore e negativamente al secondo e al terzo, sembra fornire una blanda informazione sul livello di fruibilità, da parte dell'utenza, della sezione informatica. Il fattore 2, essendo debolmente correlato positivamente al quarto indicatore, penalizza chi ha un elevato indice di prestiti interbibliotecari attivi, sintomo di mancata disponibilità di documenti per l'utenza interna.

Grafico 4.22 Rappresentazione dei punteggi fattoriali nell'analisi degli indicatori di efficacia



Come era previsto, il grafico dei punteggi fattoriali ci mostra come gli indicatori di questa sezione siano poco correlati. Le componenti principali calcolate sono difficilmente interpretabili: infatti gli Atenei sembrano diramarsi lungo gli assi formati dal cerchio delle correlazioni (5.21). A questo punto sembra inutile proporre un'ulteriore analisi eliminando altri Atenei *outliers*.

4.4.3 Analisi fattoriale degli indicatori di Efficienza / Produttività / Economicità

In questa sezione centrale degli indicatori troviamo 6 indicatori. Dall'analisi descrittiva emergono diversi *outliers*: per evitare che dati anomali alterino in modo considerevole l'analisi vengono eliminati i valori di Libera Università di Lingue e Comunicazione (12), Scuola internazionale superiore degli Studi avanzati di Trieste (26) e Università degli Studi di Pisa (36).

Grafico 4.23 Rappresentazione della correlazione esistente tra gli indicatori di questa sezione.

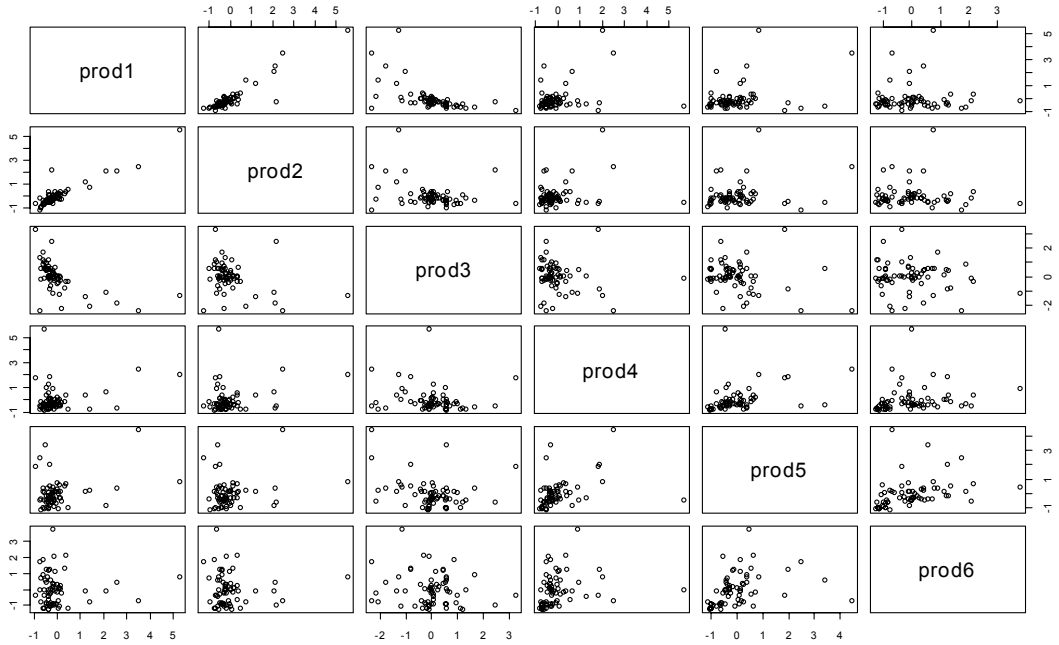


Tabella 4.24 Matrice di correlazione degli indicatori di efficienza.

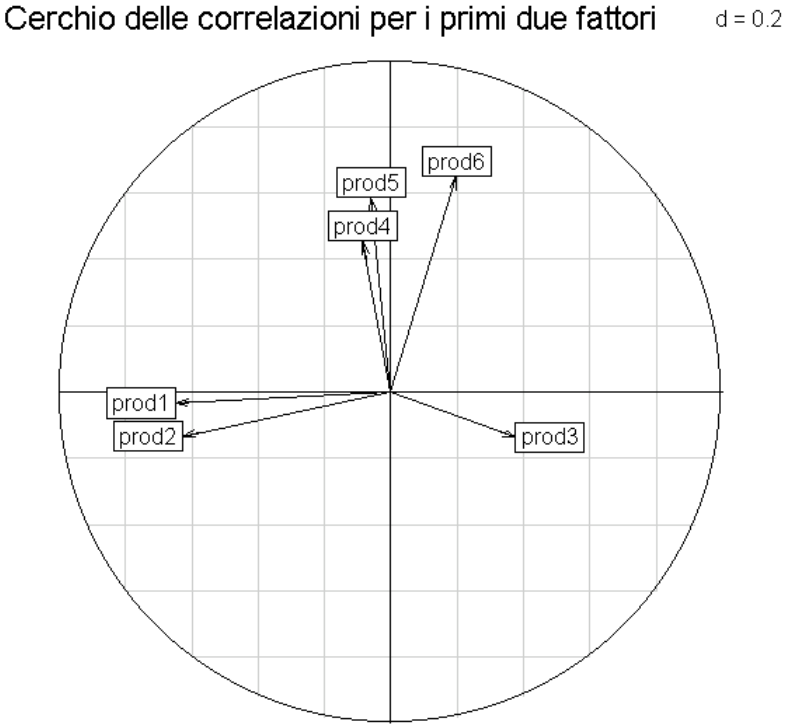
	prod1	prod2	prod3	prod4	prod5	prod6
prod1	1,000	0,916	-0,568	0,265	0,306	0,022
prod2	0,916	1,000	-0,305	0,224	0,188	-0,037
prod3	-0,568	-0,305	1,000	-0,133	-0,268	-0,144
prod4	0,265	0,224	-0,133	1,000	0,361	0,192
prod5	0,306	0,188	-0,268	0,361	1,000	0,335
prod6	0,022	-0,037	-0,144	0,192	0,335	1,000

Dal grafico del cerchio delle correlazioni si evidenzia come l'indicatore prod1 sia fortemente correlato positivamente all'indicatore prod2 (0,916), e in correlazione negativa all'indicatore prod3 (-0,568). Nella tabella emergono altre blande relazioni tra variabili (intorno a 0,3), ma meno significanti delle precedenti.

Tabella 4.25 Pesi fattoriali ottenuti con rotazione promax.

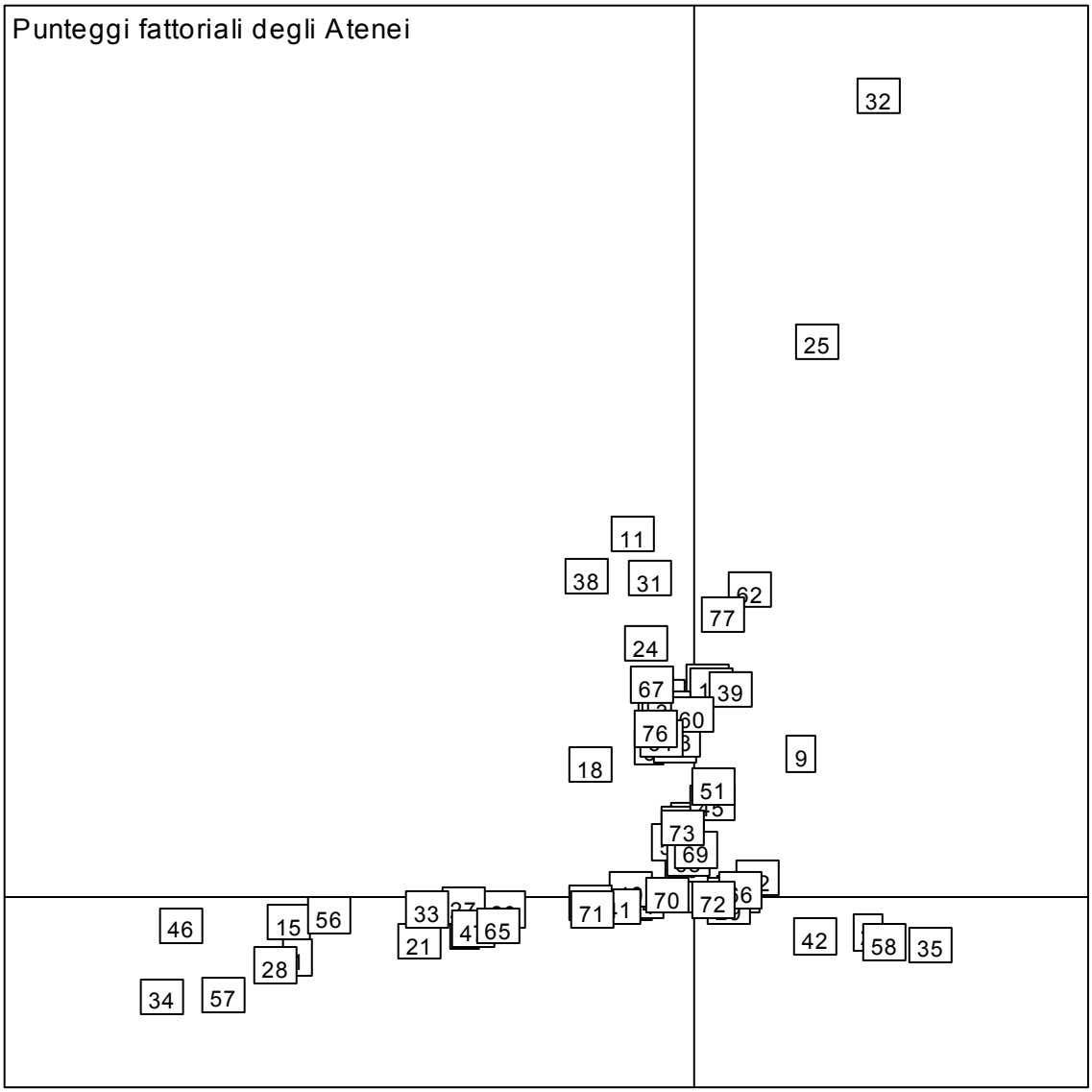
	Prod1	Prod2	Prod3	Prod4	Prod5	Prod6
Fattore 1	-0.650	-0.630	0.377			0.198
Fattore 2		-0.134	-0.138	0.456	0.586	0.650

Grafico 4.26 Cerchio delle correlazioni per gli indicatori di efficacia.



In questo caso l'analisi delle componenti principali è abbastanza buona: infatti la varianza spiegata dai primi due fattori è pari al 65,1%, risultato inaspettato se confrontato alle analisi delle sezioni precedenti. Inoltre le due componenti principali sembrano maggiormente interpretabili: la prima componente è correlata positivamente all'indicatore prod3 (percentuale di spese per il personale) e negativamente con le variabili prod1 e prod2 che indicano la spesa per le risorse bibliografiche e la spesa complessiva in rapporto all'utenza potenziale; il secondo fattore è correlato positivamente agli altri tre indici restanti che definiscono il numero prestiti, acquisizioni e patrimonio documentario in rapporto al personale Fte. Quindi il primo fattore definisce l'organizzazione delle spese all'interno delle strutture bibliotecarie di ogni Ateneo, mentre il secondo fattore è una misura della produttività e della efficienza del personale Fte.

Grafico 4.27 Rappresentazione dei punteggi fattoriali nell'analisi degli indicatori di efficienza - produttività - economicità



L'analisi dei due fattori ci fa pensare che un Ateneo molto efficiente dovrebbe avere una organizzazione che gli permetta di spendere in modo adeguato all'utenza, senza trascurare la retribuzione del personale Fte, e che, la produttività del personale Fte si attesti su livelli medi (sono da considerare negative le situazioni che indicano troppi carichi di lavoro per il personale). Quindi gli Atenei migliori devono avere una situazione di equilibrio nel primo fattore ed ottenere un punteggio medio nel secondo fattore.

Dall'analisi dei punteggi fattoriali si può notare la presenza di un paio di outliers (Atenei 25 e 32). Come ci aspettavamo dall'analisi delle due componenti principali, gli Atenei migliori sotto il punto di vista dell'efficienza – produttività – economicità si collocano ad una altezza media lungo l'asse delle "ascisse". In questa regione dello spazio possiamo trovare sia Atenei più blasonati come l'Università degli Studi di Milano (8) , il Politecnico di Torino (2), l'Università degli Studi Padova (23), sia istituzioni accademiche più piccole come Università degli Studi del Molise (63) o Libera Università di Bolzano (18).

4.4.4 Analisi fattoriale degli indicatori di Vitalità del patrimonio - Offerta risorse

In questa sezione si trovano gli indicatori che forniscono una misura complessiva dello stato dell'offerta di risorse, sia umane che materiali, da parte delle strutture bibliotecarie di Ateneo e il tipo di gestione del patrimonio. Dando "un'occhiata" alle statistiche descrittive, si nota la numerosa presenza di diversi Atenei *outliers*. Non potendo escluderli tutti dalla analisi (sono circa una decina), si opta nell'eliminare quelli che si distinguono particolarmente. Si procede senza i valori relativi alla Libera Università di Lingue Comunicazione (12), alla Libera Università di Bolzano (18), all'Università degli Studi di Pisa (36), alla Scuola normale di Pisa e all'Università degli Studi della Calabria, per un totale di 5 Atenei esclusi.

Grafico 4.28 Rappresentazione della correlazione esistente tra gli indicatori di questa sezione

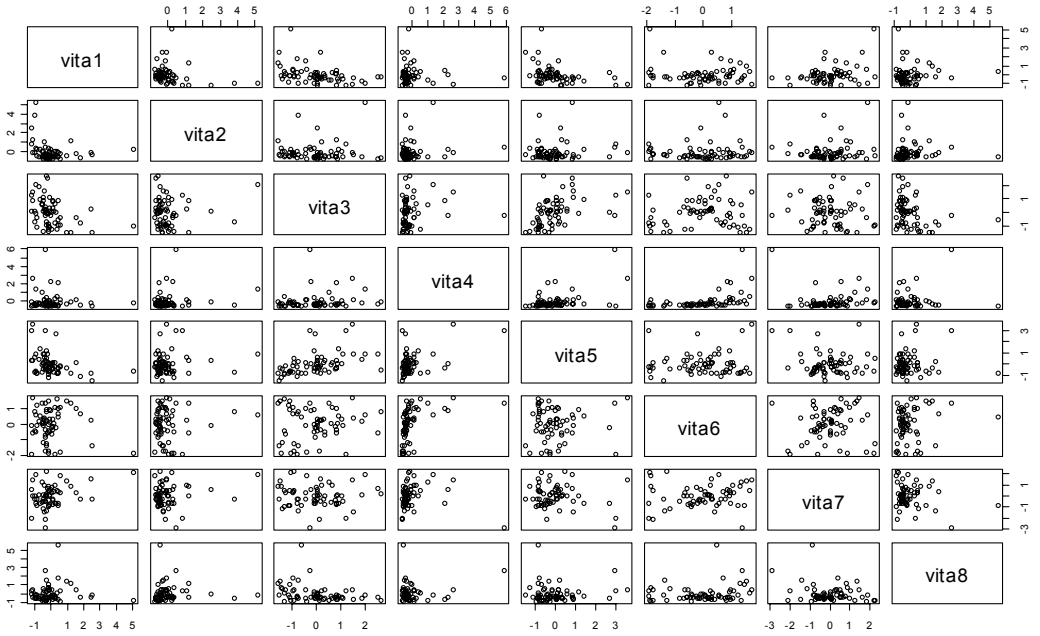


Tabella 4.29 Matrice di correlazione degli indicatori di vitalità.

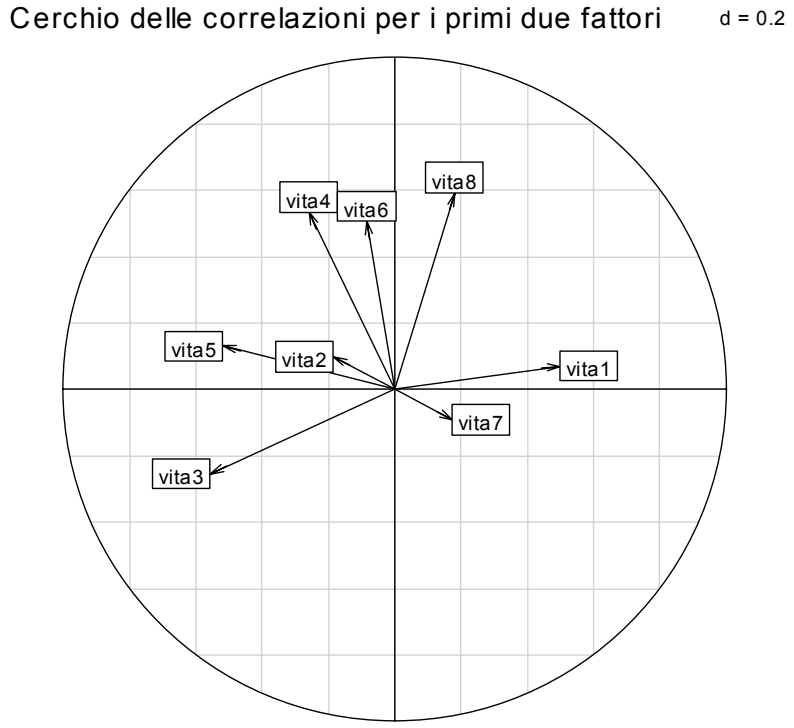
	vita1	vita2	vita3	vita4	vita5	vita6	vita7	vita8
vita1	1,000	-0,217	-0,370	-0,088	-0,357	-0,098	0,330	0,037
vita2	-0,217	1,000	0,077	0,165	0,111	0,076	0,159	0,022
vita3	-0,370	0,077	1,000	0,144	0,456	0,095	-0,038	-0,312
vita4	-0,088	0,165	0,144	1,000	0,469	0,442	-0,068	0,197
vita5	-0,357	0,111	0,456	0,469	1,000	0,063	-0,160	0,063
vita6	-0,098	0,076	0,095	0,442	0,063	1,000	0,080	0,186
vita7	0,330	0,159	-0,038	-0,068	-0,160	0,080	1,000	-0,216
vita8	0,037	0,022	-0,312	0,197	0,063	0,186	-0,216	1,000

Né dalla figura 4.28, né dalla matrice di correlazioni, sono presenti forti relazioni tra le variabili. Tuttavia si osservano numerose correlazioni dell'ordine del 0,4. Da notare che gli indicatori vita2, vita7 e vita8 sono molto poco correlati nei confronti delle altre variabili.

Tabella 4.30 Pesi fattoriali ottenuti con rotazione varimax.

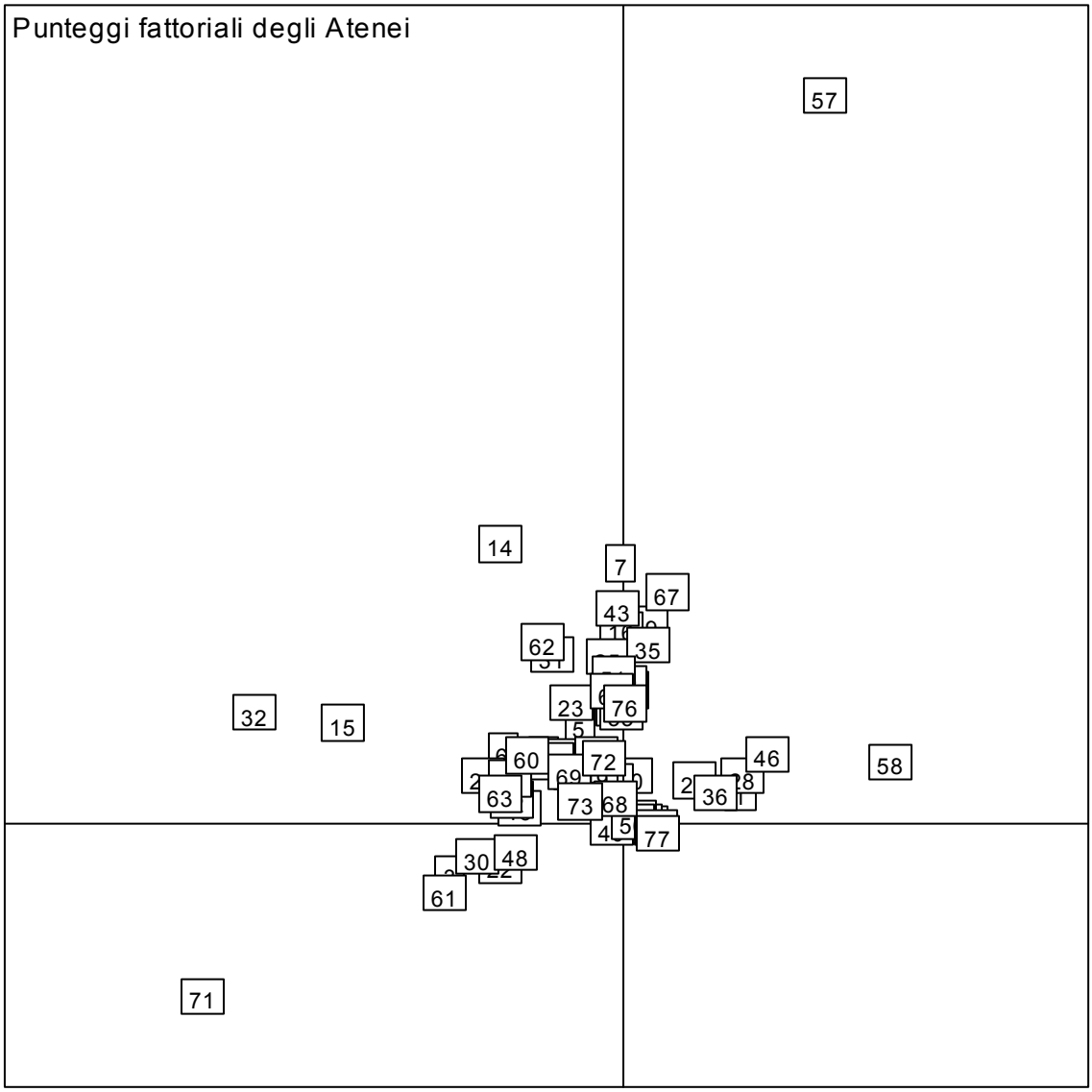
	Vita1	Vita2	Vita3	Vita4	Vita5	Vita6	Vita7	Vita8
Fattore 1	0.500	-0.185	-0.556	-0.259	-0.519		0.175	0.180
Fattore 2		0.101	-0.253	0.534	0.131	0.507		0.594

Grafico 4.31 Cerchio delle correlazioni per gli indicatori di vitalità.



Le prime due componenti della analisi spiegano soltanto discretamente la variabilità complessiva degli indicatori (46,1%). Dal cerchio delle correlazioni si può vedere che il primo fattore è correlato positivamente con l'indicatore vita1 e negativamente con vita3 e vita5. Dai valori misurati da questi indicatori, gli Atenei che presentano punteggi alti nel primo fattore sono quelli che hanno un offerta di risorse inadeguata al numerosità dell'utenza potenziale. Il secondo fattore, correlato in maniera positiva agli indicatori vita4, vita6 e vita8, sembra valutare la vitalità e l'adeguatezza generale del patrimonio documentario. Quindi gli Atenei che esprimono una migliore situazione sulla base di questi indicatori, dovrebbero collocarsi nel quadrante in alto a sinistra.

Grafico 4.32 Rappresentazione dei punteggi fattoriali nell'analisi degli indicatori di vitalità



Dal grafico dei punteggi fattoriali non vi è una netta separazione delle unità: infatti, gli Atenei sono molto concentrati e si nota la presenza, come preventivato inizialmente, di alcuni *outliers*. Nel quadrante in alto a sinistra dovrebbero collocarsi gli Atenei che offrono il miglior livello di risorse e buona qualità del patrimonio. Situati in questa zona spiccano i nomi dell'Università di Bologna (29), l'Università di Milano – Bicocca (14), l'Università di Ferrara (30), l'Università di Bergamo (15), l'Università di Padova (23) e la Libera Università degli Studi San Pio V di Roma (51) .

CAPITOLO 5
CLUSTER ANALYSIS

5.1 La *Cluster Analysis*

Sotto il termine generale di analisi di raggruppamento o *cluster analysis* si accorpano varie tecniche operanti su dati di tipo quantitativo volte a classificare l'insieme delle unità dell'analisi in gruppi, *cluster*, non definiti a priori, in base alle caratteristiche possedute. Per approfondimenti si veda Zani (2000) e Fabbris (1990). I gruppi vengono formati cercando di massimizzare l'omogeneità interna e le differenze tra i vari *clusters*. Per stabilire la similarità tra le varie unità campionate, vengono calcolate delle distanze: questo comporta il dover scegliere una metrica che sia in grado di esprimere al meglio la distanza tra gli elementi considerati. In sintesi, l'*input* dell'analisi è costituito da una matrice di dati quantitativi che riporta, per ciascuna unità statistica, il valore delle variabili rispetto alle quali si vuole operare la classificazione, mentre l'output è una nuova variabile categoriale le cui modalità rappresentano il *cluster* di appartenenza a cui ciascun elemento è assegnato in modo univoco. Fondamentalmente, esistono due differenti tipi di algoritmi di classificazione: quelli gerarchici, suddivisi in scissori e agglomerativi, e quelli non gerarchici.

5.1.1 Algoritmi gerarchici

Ogni gruppo fa parte di un gruppo più ampio, il quale è contenuto a sua volta in uno di ampiezza maggiore e così in progressione fino al gruppo che contiene l'intero insieme di unità analizzate. Gli algoritmi gerarchici si suddividono in:

- scissori: quando l'insieme delle *n* unità, in *n-1* passi, si ripartisce in gruppi che sono, ad ogni passo dell'analisi, sottoinsieme di un gruppo formato allo stadio precedente, e che termina con la situazione in cui ogni gruppo è composto da una unità;
- aggregativi: se procedono a una successione di fusioni delle *n* unità, a partire dalla situazione di base nella quale ogni unità costituisce un gruppo a sé stante e fino allo stadio *n-1* nel quale si forma un gruppo che le contiene tutte (questi sono maggiormente usati in quanto richiedono un minor tempo di elaborazione).

5.1.2 Algoritmi non gerarchici

In questo caso è necessario conoscere a priori il numero di *cluster* che si vogliono ottenere ed i centroidi iniziali di tali *cluster*. L'algoritmo procede in maniera iterativa cercando di ottenere la migliore classificazione degli elementi secondo il numero di classi prestabilito: ad ogni iterazione dispari vengono accorpati i due *cluster* più vicini mentre ad ogni iterazione pari viene separato il *cluster* più disomogeneo. Si procede poi al calcolo dei centroidi fino a quando lo spostamento dei centroidi da un'iterazione all'altra diventa infinitesimale. Le procedure di analisi non gerarchica si suddividono in due categorie a seconda che generino partizioni, ossia classi mutuamente esclusive, o classi sovrapposte, per le quali si ammette la possibilità che un elemento appartenga contemporaneamente a più *cluster*.

5.2 Il percorso di analisi

Per effettuare una *cluster analysis* si devono prendere diverse decisioni:

1. identificare le variabili di classificazione. In questo caso consideri la matrice di dati $X = \{x_{hj}\}$ $(h = 1,\ldots,n \; ; \; j = 1,\ldots,p)$ relativa ad n osservazioni su p variabili (nell'analisi si possono considerare le variabili osservate o una loro opportuna trasformazione);
2. selezione della misura di prossimità tra le unità da raggruppare. Se l'obiettivo dell'analisi è la classificazione delle unità si userà una matrice simmetrica di ordine n (solitamente matrice di varianze e covarianze o matrice di correlazione), se invece è la classificazione delle variabili una matrice di ordine p;
3. selezione della tecnica di raggruppamento delle entità. Le tecniche di raggruppamento proposte in letteratura sono numerose e diverse, tanto che risulta difficile riuscire a capire quale si adatti meglio agli obiettivi di ogni singola analisi. Ricordiamo le gerarchiche, agglomerative e scissorie e le non gerarchiche, che generano partizioni o classi sovrapposte;
4. identificazione del numero di gruppi entro i quali ripartire le entità. Questo problema risulta similare a quello per la scelta del numero di fattori nell'analisi fattoriale vista precedentemente;
5. completamento dell'analisi e interpretazione dei risultati dell'analisi.

5.2.1 Selezione della misura di prossimità tra le variabili

Dopo aver deciso se operare l'analisi sulla matrice iniziale di dati X_{hj} o su una loro trasformazione lineare ottenuta tramite una standardizzazione dei dati, si determina la matrice delle distanze, ossia quella matrice quadrata il cui elemento generico D_{hk} è una misura di distanza tra le unità h e k.

Tra le misure più utilizzate per la *cluster analysis* vi sono:

- distanza euclidea. La distanza calcolata tra le entità h e k basata sulla distanza euclidea viene calcolata nel seguente modo:

$$d_{hk} = \left\{ \sum_{j}^{p} (x_{hj} - x_{hj})^{2} \right\}^{1/2} \qquad (h,k = 1,\ldots,n). \qquad (5.1)$$

La distanza tra x_{hj} e x_{kj} non varia al variare dell'origine o al ruotare degli assi.

- distanza media assoluta. La distanza media assoluta d_{hk} tra le unità statistiche h e k nello spazio *p-dimensionale* definito dalle p variabili osservate è data da:

$$d_{hk} = \sum_{j}^{p} | x_{hj} - x_{kj} | \qquad (h,k = 1,\ldots,n,), \qquad (5.2)$$

ed è particolarmente appropriata quando le variabili sono su scala ordinale. Rispetto la distanza euclidea, la distanza media assoluta non è invariante rispetto a traslazioni o rotazioni degli assi coordinati.

- distanza di Lagrange – Tchebychev. La distanza di *Lagrange – Tchebychev* tra due unità statistiche h e k è lo scostamento massimo, in valore assoluto, tra tutti gli scostamenti tra le singole variabili osservate e le unità h e k:

$$d_{hk} = Max | x_{hj} - x_{kj} | \qquad (h,k = 1,\ldots,n), \qquad (5.3)$$

dove il valore massimo è calcolato in relazione alle p variabili osservate.

5.2.2 Selezione di un algoritmo di classificazione

Le tecniche di analisi dei gruppi possono essere divise in gerarchiche, aggregative e scissorie, e non gerarchiche. Gli algoritmi gerarchici non necessitano della definizione a priori del numero di *cluster* che si vuole ottenere e risultano molto onerosi e poco efficienti dal punto di vista computazionale. Inoltre, sono fortemente influenzati dalla presenza di *outliers*. Nel caso di *dataset* di elevate dimensioni, gli algoritmi non gerarchici risultano estremamente più efficienti e meno influenzati da valori anomali inoltre, essendo non monotoni, permettono che un'unità statistica, inizialmente inserita in un *cluster*, possa modificare il proprio gruppo di appartenenza durante il processo iterativo

5.2.3 Tecniche gerarchiche aggregative

Date tre unità h, k e l di numerosità rispettivamente n_h, n_k, n_l, le tecniche di analisi gerarchica aggregative prevedono di utilizzare la matrice delle distanze per trovare la coppia di elementi h e k che sono più vicine e formare così il primo *cluster*. Successivamente si ricalca la matrice delle distanze sostituendo le righe e le colonne relative ai gruppi h e k con una riga e una colonna di distanze tra il gruppo (h, k) e il gruppo l. L'individuazione delle unità più prossime e il ricalco delle distanze si ripetono per $n-1$ volte finché tutte le unità fanno parte di un gruppo unico. Il calcolo della distanza d_{hk} tra l'entità l e il gruppo (h, k) può essere effettuato mediante vari criteri:

- metodo della media di gruppo. La distanza tra l'elemento l ed il gruppo formato dalla fusione di h e k è data dalla media aritmetica delle distanze d_{hl} e d_{kl} ponderate con la numerosità degli individui appartenenti ai gruppi h e k:

$$d_{l(h,k)} = \alpha_h d_{hl} + \alpha_k d_{kl} \quad con\ h \neq k \neq l = 1,\ldots,n \quad , \tag{5.4}$$

dove $\alpha_h = n_h / (n_h + n_k)$ e $\alpha_k = n_k / (n_h + n_k)$ e d_{hl} e d_{kl} sono due misure qualsiasi di dissomiglianza, calcolate come mostrato precedentemente.

- metodo del centroide. Operando con il metodo del centroide (vettore delle medie di una distribuzione multivariata), la distanza tra due gruppi è la distanza euclidea tra i centroidi dei gruppi.

La distanza tra l'unità *l* e il gruppo formato dalla fusione di *h* e *k* è data da:

$$d_{l(h,k)} = \{(\alpha h_2 d_{hl}^2 + \alpha_{k2} d_{kl}^2 + \alpha_h \alpha_{k2} d_{hk}^2)\}^{1/2} \qquad con\ h \neq k \neq l = 1,...,n, \qquad (5.5)$$

dove d_{hk} indica la distanza euclidea tra due punti *h* e *k* qualsiasi e α_l è il peso relativo del gruppo *l* ($\alpha_l = n_l / (n_l + n_k)$).

- metodo del legame singolo. Con la strategia del legame singolo la distanza tra l'unità *l* e la fusione (*h, k*) è la distanza minore tra *l* e le due unità aggregate:

$$d_{l(h,k)} = \min\{d_{hl}, d_{kl}\} \qquad con\ h \neq k \neq l = 1,...,n. \qquad (5.6)$$

- metodo del legame completo. Il criterio del legame completo si contrappone, come logica e come risultati, a quello del legame singolo. Tra l'elemento *l* e il gruppo (*h, k*), la distanza è infatti data dal valore più elevato tra d_{hl} e d_{kl}:

$$d_{l(h,k)} = \max\{d_{hl}, d_{kl}\} \qquad con\ h \neq k \neq l = 1,...,n. \qquad (5.7)$$

A differenza del metodo del legame singolo, con il metodo del legame completo, poiché si ottengono gruppi di forma circolare caratterizzati da notevole somiglianza interna, è possibile eseguire una ricerca dei gruppi omogenei.

- metodo di Ward. Con il metodo di *Ward*, la scelta della coppia di unità da aggregare si basa sulla minimizzazione della devianza tra i centroidi dei possibili gruppi. La devianza ha un minimo pari a 0 quando tutti gli elementi sono isolati e un massimo pari alla somma delle devianze delle variabili di classificazione quando tutte le unità appartengono a un unico gruppo. La distanza euclidea tra l'elemento *l* e il *cluster (h,k)* è data da:

$$\sqrt{\frac{n_l\, n_{(h,k)} d_{l(h,k)}^{\ 2}}{n_l + n_{(h,k)}}}\ , \qquad (5.8)$$

dove n_l è il numero di unità che compongono il gruppo *l* e $n_{(h,k)} = n_h + n_k$.

Per ogni livello gerarchico dell'algoritmo di classificazione si ottengono indicatori statistici che possono aiutarci nella scelta del numero ottimale di *cluster*.

Tali indicatori si basano sulla scomposizione della variabilità tra e dentro i *cluster*: la variabilità tra i gruppi misura il livello di eterogeneità tra un *cluster* e l'altro (tanto più elevata è la variabilità, tanto più differenziati sono i gruppi di clienti a cui ci riferiamo); la variabilità entro i *cluster* misura il livello di omogeneità all'interno del gruppo (tanto più bassa è la variabilità, tanto più in ciascun *cluster* i comportamenti degli atenei sono simili).

5.2.4 Metodi gerarchici scissori o divisivi

Il procedimento di suddivisione è concettualmente opposto a quello della aggregazione progressiva delle unità. Si parte infatti dalla situazione nella quale le *n* unità fanno parte di un unico gruppo e in *n-1* passi si perviene alla situazione nella quale ogni unità fa gruppo a sé stante. Tra i metodi divisori, uno dei più utilizzati è il *K-Means* basato sulla distanza tra i centroidi, che prevede di effettuare una prima suddivisione in due gruppi sulla base della combinazione delle unità che minimizza la devianza interna ai gruppi. Ad ogni passo successivo, individuato il gruppo che ha la massima devianza interna (devianza di ogni elemento dal centroide), la suddivisione dicotomica delle *n* unità del gruppo si effettua provando tutte le possibili combinazioni con 1 e *n-1* unità, 2 e *n-2* unità e così via, individuando quella che minimizza la funzione:

$$D = \sum_{g}^{G} \sum_{h}^{n_g} \sum_{i}^{p} ({}_g x_{hi} - {}_g \bar{x}_{xi})^2 , \qquad (5.9)$$

dove ${}_g x_{hi}$ ($g = 1,2;\ h = 1,\ldots,n_g;\ i = 1,\ldots,p$) è il valore della variabile x_i osservato presso l'unità statistica h appartenente al sottogruppo g e ${}_g \bar{x}_{xi}$ è il valore medio della variabile i nel sottogruppo g. Il metodo di analisi *K-Means* consiste nella suddivisione ad ogni passo del campione sulla base di un numero qualsiasi ma opportuno di suddivisioni.

5.2.5 Criteri che generano partizioni non gerarchiche

La maggior parte di questi criteri consiste nell'eseguire una successione, anche iterata, di tre procedure volte ad avviare il processo classificatorio, individuando una soluzione provvisoria; ad assegnare le unità ai gruppi individuati nella prima fase; ad assegnare gli elementi a gruppi diversi da quelli precedentemente individuati, ottimizzando una funzione obiettivo.

Per quanto riguarda le procedure di avvio dell'analisi, si sfruttano le informazioni sui gruppi o ottenute da altre analisi, anche gerarchiche, oppure da un'analisi *K-Means* non gerarchica vista precedentemente.
Se non sono disponibili queste informazioni, si può utilizzare la tecnica proposta da Beale (1969) che considera un numero elevato di centroidi casuali ed assegna le unità statistiche ai diversi gruppi in base alla minima distanza euclidea dai centroidi; quindi iterativamente vengono spaccati i *cluster* meno omogenei, fornendo i due nuovi gruppi e ricalcolando i centroidi, fino a quando gli spostamenti tra questi diventano irrilevanti.

5.2.6 Tecniche non gerarchiche con sovrapposizione

Le tecniche di raggruppamento con sovrapposizione ammettono che, per un dato numero di gruppi, le unità appartengano a più insiemi disgiunti. Tra le varie tecniche ricordiamo brevemente:

- ricerca di insiemi sfuocati: in tale tecnica i gruppi risultano compenetrati e le unità hanno un grado più o meno elevato di appartenenza ai gruppi. Appartengono al cluster gli elementi che si trovano entro un raggio fissato dal centro del gruppo, per cui un elemento può avere un livello di appartenenza non nullo su più gruppi;
- analisi di miscugli di distribuzione: si ipotizza una certa distribuzione delle frequenze delle n unità osservate, si identificano i gruppi e quindi si stabilisce la probabilità di appartenenza delle singole unità ai gruppi individuati;
- analisi Fattoriale *Q*: l'analisi fattoriale condotta sulla trasposta della matrice dei dati, dopo una standardizzazione che rende uniforme la scala di misura delle variabili. La matrice fattorizzata è pertanto una matrice di similarità tra individui e i fattori sono combinazioni lineari di unità (non di variabili); la rappresentazione grafica degli elementi sugli assi definiti dai fattori, solitamente i primi due, è essenziale per decidere a quale gruppo assegnare le unità.

5.3 Scelta tra i metodi di analisi

La qualità di una tecnica di raggruppamento può essere valutata in base a vari criteri:

- l'oggettività data dal fatto che se diversi ricercatori conducono la stessa analisi separatamente, questi devono giungere alla stessa conclusione;
- la stabilità dei risultati della classificazione operando su campioni equivalenti, ossia vogliamo che i risultati dati dai metodi di analisi non risentano significativamente di piccole variazioni del campione di riferimento;
- l'informazione ottenuta dal risultato intermedio e finale;
- la semplicità dell'algoritmo e la rapidità di esecuzione.

Vi è una sostanziale differenza tra le tecniche gerarchiche e quelle non gerarchiche. Infatti le tecniche non gerarchiche sono in genere più informative delle gerarchiche perché danno anche risultati intermedi e indici relativi alla qualità dei risultati, mentre i metodi gerarchici risentono della presenza di errori di misura o di altre fonti di variabilità presenti nelle misure di prossimità e i dati anomali creano alcuni problemi. Se si cercano gruppi caratterizzati da forte omogeneità interna, le tecniche gerarchiche sono in genere meno efficaci di quelle non gerarchiche.
Il calcolo delle soluzioni gerarchiche, in particolare quelle agglomerative, è più rapido degli altri. Tuttavia uno svantaggio delle tecniche gerarchiche è la rigidità della soluzione: una aggregazione impropria effettuata nei primi stadi dell'analisi si trascina fino alla fine e può rendere i risultati artificiosi. D'altra parte, se un procedimento di analisi non gerarchica è avviato senza una adeguata conoscenza a priori, i risultati sono modesti. La soluzione più conveniente può essere allora quella di far precedere l'analisi non gerarchica da una gerarchica.

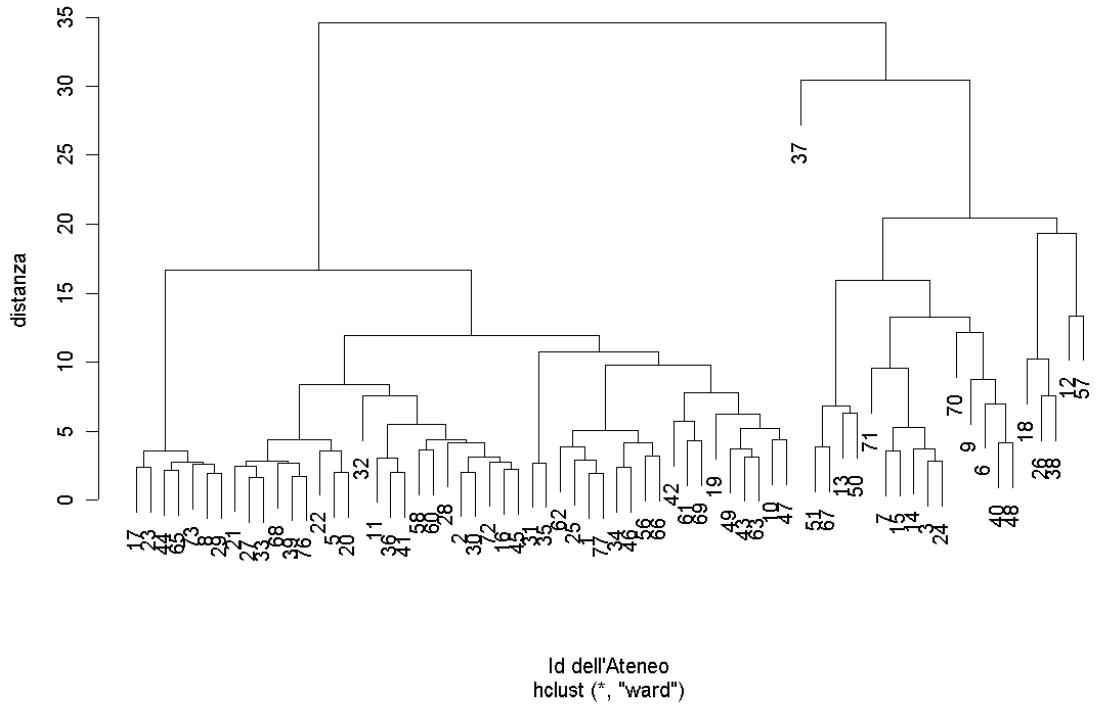
5.4 Analisi dei *dataset* con metodi gerarchici agglomerativi

L'analisi della matrice degli indicatori standardizzati con metodi gerarchici agglomerativi ci consente, partendo dalle singole unità accademiche, di effettuare una operazione di "fusione" tra gli Atenei che sono più vicini. Tale tecnica ha il pregio di ottenere, con un basso costo computazionale, un buon risultato dal punto di vista grafico ed informativo.
Nel linguaggio R troviamo implementata nella libreria *"cluster"*, la funzione "*hclust()"* che consente di ottenere una clusterizzazione partendo dalla matrice di distanza delle unità che si può calcolare con la funzione *"dist()"*.

5.4.1 Analisi del *dataset* completo

Con l'analisi di tutti i 28 indicatori si cerca di raggruppare fra loro gli Atenei che avevano una forte similarità in tutte le aree di interesse. In questo modo si possono cercare strutture accademiche che condividono la stessa gestione bibliotecaria in tutti negli ambiti accessibilità, efficienza produttività e vitalità del patrimonio.

Grafico 5.10 Dendrogramma degli Atenei utilizzando tutti gli indicatori



Dopo aver effettuato numerose prove la soluzione grafica migliore è stata ottenuta scegliendo come misura della distanza la distanza euclidea, mentre come tecnica

aggregativa la scelta è ricaduta sul metodo di Ward. Quest'ultimo è basato sulla minimizzazione della variabilità all'interno dei gruppi: obiettivo della partizione è, infatti, minimizzare la quota di variabilità interna, massimizzando, nel contempo, la variabilità fra i gruppi, così da ottenere classi omogenee e ben separate le une dalle altre (vedi paragrafo 5.2.1).

In questo caso l'interpretazione del risultato sembra più immediata di quanto non lo sia stato fatto con l'analisi fattoriale. Infatti dal dendrogramma (Figura 5.10) si può vedere come gli Atenei si dividano inizialmente in due gruppi. Il gruppo a destra sembra contenere le istituzioni accademiche che hanno caratteristiche "particolari" nel panorama universitario italiano. Infatti, in tale partizione troviamo scuole per studi avanzati, istituti specializzati e Atenei con scarso bacino d'utenza. Invece, nella partizione di destra la restante parte di Università hanno misura di distanza molto vicine tra loro per cui esprimono realtà organizzative comuni. Addentrandoci in questo secondo gruppo possiamo notare altri tre grappoli in cui le unità sono molto vicine tra loro. Nel caso dell'Università degli Studi di Padova (8), essa sembra essere molto vicina dalle misure degli indicatori a realtà come l'Università degli Studi di Bologna (29) e l'Università di Pavia (17) .

In appendice si può trovare l'allegato contenente la codifica dell'Id con il nome effettivo dell'Ateneo.

5.4.2 Analisi gerarchica degli indicatori di accessibilità

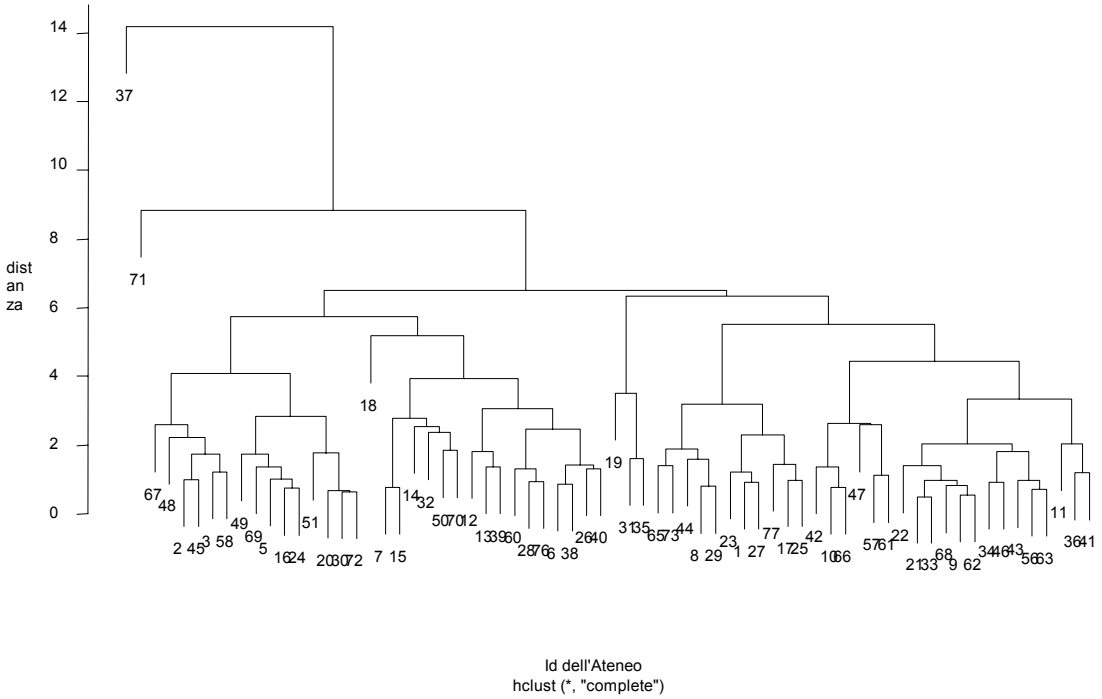
Il motivo di questa analisi è di identificare gli Atenei simili sotto il profilo dell'accessibilità. In questo modo gli Atenei che condividono la stessa organizzazione degli spazi e lo stesso livello di apertura verso l'utenza.

Dopo varie prove, per l'analisi gerarchica si è deciso di utilizzare, come misura di prossimità nella procedura *"hclust()"*, il metodo del legame completo, perché sembra migliorare graficamente le distinzioni tra le Università.

Nel dendrogramma (Figura 5.11) si può notare come gli Atenei si dividano in 6 sottogruppi, non tenendo conto della presenza di 2 *outliers*. Dalla rappresentazione grafica non si evince una particolare distribuzione degli Atenei, anche se analizzando i vari gruppi, le componenti che tendono ad avvicinare due Atenei sono quelli di localizzazione geografica, di grandezza fisica o di tipologia di utenza.

Infatti si nota come siano poco distanti, anche a livello chilometrico, Atenei come l'Università degli Studi di Pisa (36) e l'Università degli Studi di Perugia (41). Inoltre, come è logico pensare, gli indicatori di accessibilità tendono ad accostare strutture bibliotecarie con organizzazione simili come l'Università degli Studi di Milano con l'Università degli Studi di Bologna. Tuttavia, ulteriori confronti possono essere fatti, senza però riuscire a far emergere aspetti latenti più importanti.

5.11 Dendrogramma basato sugli indicatori di accessibilità



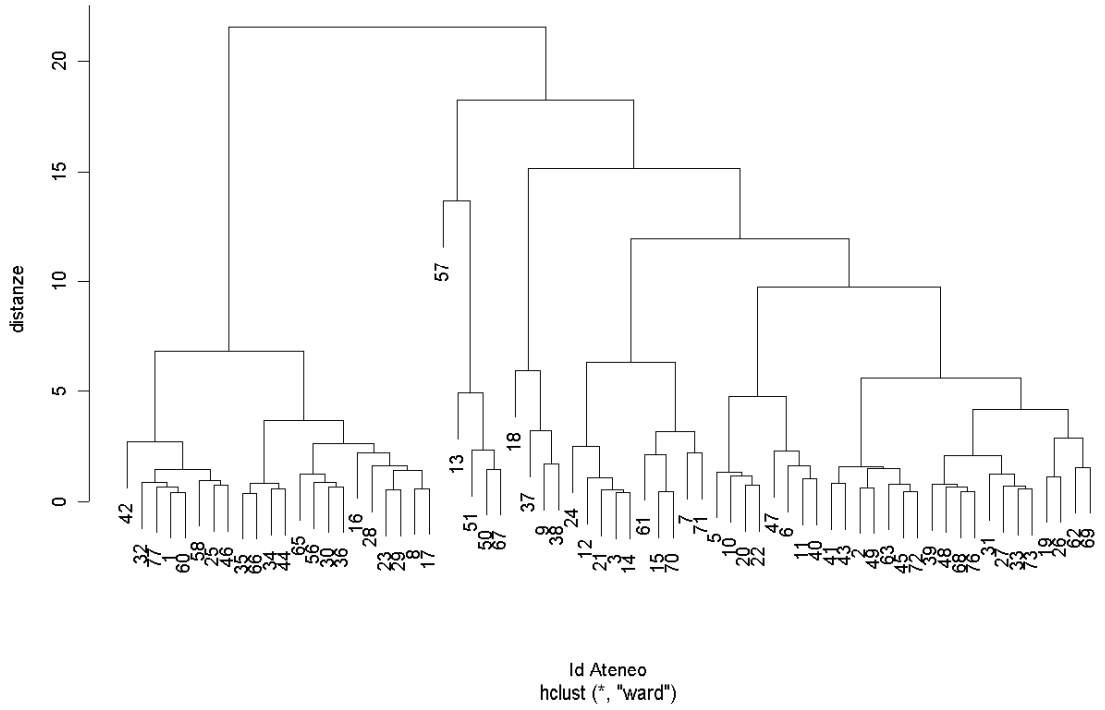
Soltanto un ristretto gruppo di Atenei *outliers* sembra avere caratteristiche di accessibilità diverse dal resto delle altre Istituzioni Accademiche, come la Scuola normale superiore di Pisa e l'Università degli Studi di Catanzaro Magna Grecia.

5.4.3 Analisi gerarchica degli indicatori di efficacia / fruibilità / innovazione

I sei indicatori contenuti in questa area qualificano, in un sistema bibliotecario, l'efficacia e la fruibilità dei servizi offerti e il grado innovazione presente nella struttura informativa. Quindi lo scopo di questa analisi *cluster* è trovare comportamenti simili tra gli Atenei osservati. Dal dendrogramma 5.12, dove la misura di prossimità è definita dal legame completo, si possono definire inizialmente due gruppi che appaiono molto distanti tra loro. Inoltre si può notare come questa divisione definisce una blanda classificazione dei due pattern: nel ramo di sinistra si trovano raggruppati strettamente gli Atenei di medio-grande

di dimensione, mentre nel secondo gruppo, quello di destra, le strutture accademiche hanno una minore dimensione. Inoltre, è possibile notare in questo gruppo, come le distanze siano molto alte e si tende ad effettuare dei piccoli "grappoli" di 5-6 Atenei che hanno, quindi, valori indicatori molto simili tra loro.

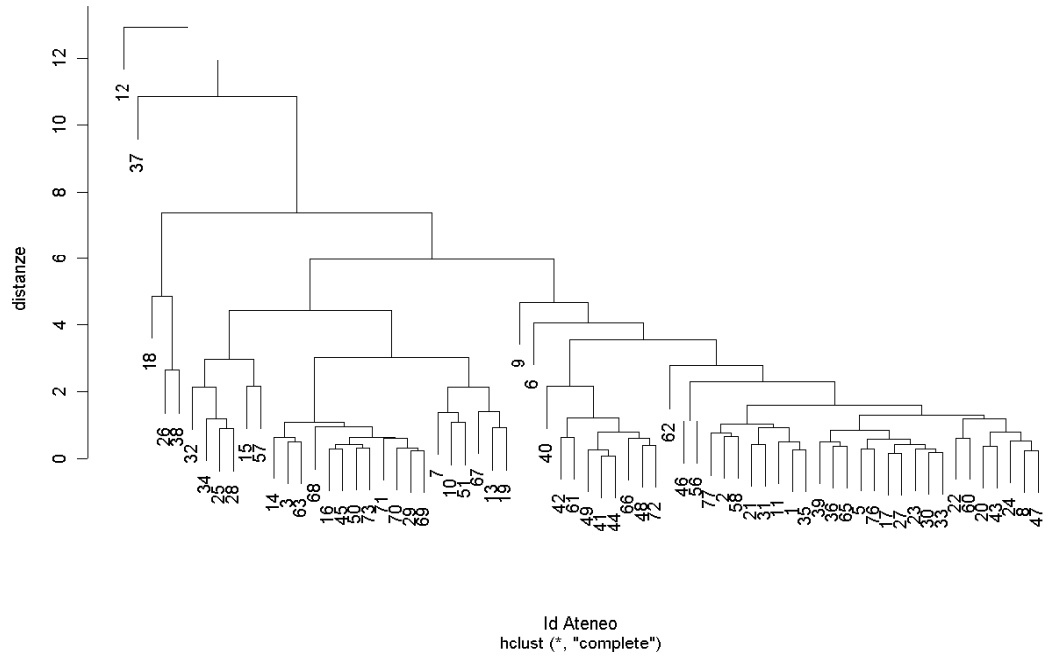
5.12 Dendrogramma degli Atenei basato sugli indicatori di efficacia



5.4.4 Analisi gerarchica degli indicatori di efficienza / produttività / economicità

In questa area la classificazione delle strutture bibliotecarie degli Atenei dovrebbe accomunare biblioteche con misure di efficienza ed economicità simili. Ciò che ci si può aspettare da un analisi del genere che gli Atenei più grandi, fornendo servizi e prestazioni su larga scala, abbiano un valore di questi indicatori molto differente da quelle strutture che, contando in un'utenza ristretta, hanno una diversa gestione delle spese. Tuttavia dal dendrogramma ottenuto basandosi col metodo del legame completo, si può verificare che questa distinzione non è netta, e che sono raggruppati insieme, Atenei molto distanti dal punto di vista organizzativo. Trascurando gli *outliers,* si possono scorgere due gruppi principali, anche se, a prima vista, non c'è una grossa distinzione tra i due *pattern* di Atenei.

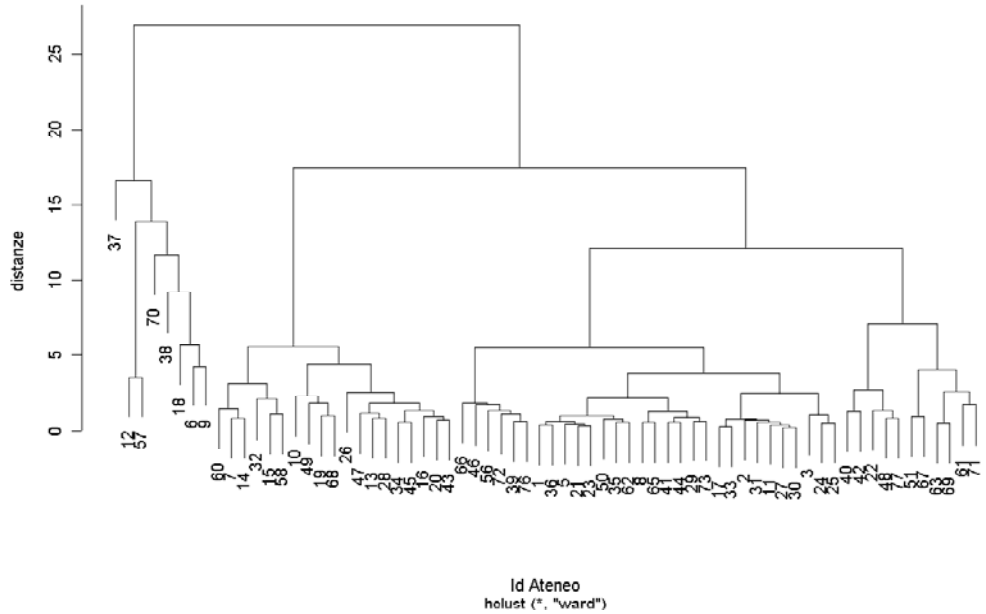
5.13 Dendrogramma degli Atenei basato sugli indicatori di efficienza.



5.4.5 Analisi gerarchica degli indicatori di vitalità del patrimonio / offerta risorse

Il dendrogramma ottenuto con questa tipologia di indicatori porta a definire quattro gruppi ben distinti. A parte il primo pattern a sinistra che sembra contenere Atenei eterogenei tra di loro (lo indicano le distanze), gli altri tre gruppi sono molto vicini e hanno valori di indicatori molto simili.

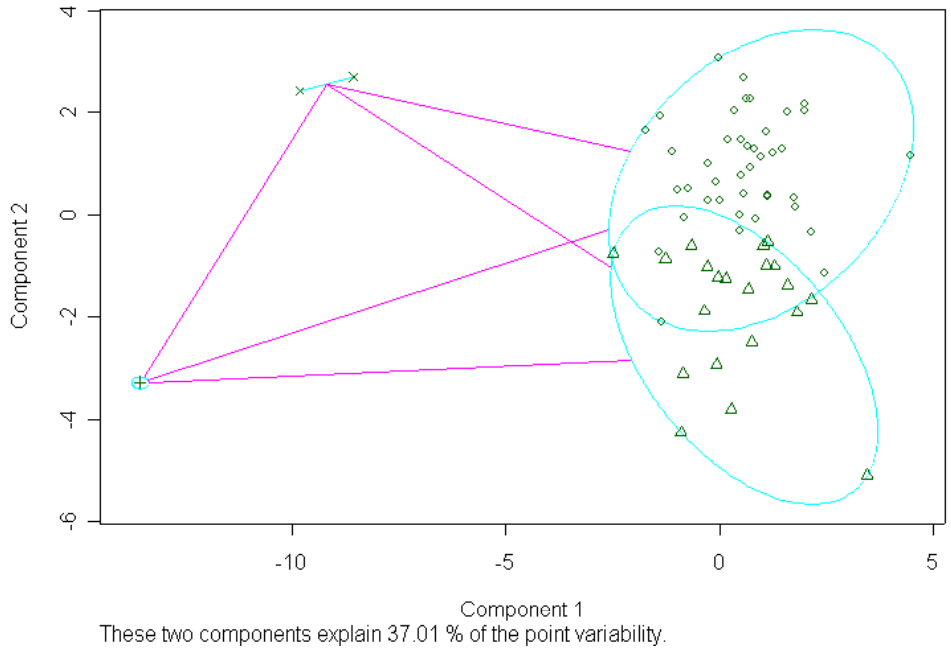
5.14 Dendrogramma degli Atenei basato sugli indicatori di vitalità.



5.5 Analisi del *dataset* completo con metodi con gerarchici divisivi

Per effettuare questa analisi in R, utilizzo l'algoritmo PAM[3], la cui funzione è contenuta nella funzione *"pam()"* della libreria *"cluster()"*. Questa tecnica si fonda sulla ricerca di *k* punti rappresentativi, detti medoidi, tra quelli osservati; le restanti unità sono allocate ai medoidi in ragione della distanza più piccola (vedi paragrafo 5.3.2). La media delle distanze dal medoide più vicino misura la bontà della soluzione ottenuta. L'obiettivo finale quello di giungere ad una partizione che minimizza la somma delle distanze entro i gruppi.
Questa tecnica risente molto degli *ouliers* per cui si è deciso di escludere dal *dataset* quegli Atenei che sono stati individuati anomali. Per questo motivo, già noto dall'analisi descrittiva, vengono lasciati fuori dall'analisi gli "Id" 12, 37 e 57 afferenti alla Libera Università degli Studi e Comunicazione, la Scuola normale superiore di Pisa e l'Istituto Universitario Suor Orsola Benincasa di Napoli.
Per utilizzare questo algoritmo devo prefissare il numero di partizioni in cui devo dividere tutte le unità. In una prima analisi si possono fissare quattro gruppi e verificare come converge l'algoritmo e controllare la bontà della soluzione.

Grafico 5.15 Suddivisione degli Atenei con l'uso dell'algoritmo PAM



[3] *Partitioning around Medoids*

Come si può vedere dalla figura precedente, l'algoritmo PAM tende ad isolare i tre *outliers* in due gruppi e a dividere le restanti osservazioni negli altri due partizioni principali. Se si cerca di aumentare la numerosità dei gruppi inizializzata nell'algoritmo PAM, non si riesce ad ottenere un risultato migliore, nemmeno escludendo gli *outliers* emersi. I gruppi definiti da questo metodo gerarchico divisivo sono così composti.

Tabella 5.16 Università del 1° Cluster (41 Atenei).

Università degli Studi di Torino	Università degli Studi di Parma	Libera Università degli Studi degli Studi Maria SS. Assunta - (LUMSA) di Roma
Politecnico di Torino	Università degli Studi di Modena e Reggio Emilia	Libera Univeristà Internazionale di Studi Sociali Guido Carli - (LUISS) di Roma
Università degli Studi di Genova	Università degli Studi di Bologna	Università degli Studi Roma Tre
Università degli Studi di Milano	Università degli Studi di Ferrara	Istituto Universitario Orientale di Napoli
Politecnico di Milano	Università degli Studi di Urbino	Seconda Università degli Studi di Napoli
Università Commerciale Luigi Bocconi di Milano	Università degli Studi di Macerata	Università degli Studi Gabriele D'Annunzio di Chieti
Università Cattolica del Sacro Cuore di Milano	Università degli Studi di Camerino	Università degli Studi di Bari
Libera Università Vita Salute San Raffaele di Milano	Università degli Studi di Firenze	Politecnico di Bari
Università degli Studi di Pavia	Università degli Studi di Pisa	Università degli Studi di Lecce
Università degli Studi di Trento	Università degli Studi di Siena	Università degli Studi di Reggio Calabria
Università degli Studi di Verona	Università per gli stranieri di Siena	Università degli Studi di Palermo
Università degli Studi Cà Foscari di Venezia	Università degli Studi di Perugia	Università degli Studi di Sassari
Università degli Studi di Padova	Università per gli stranieri di Perugia	Università degli Studi di Cagliari
Università degli Studi di Trieste	Università degli Studi di Roma - La Sapienza	

Tabella 5.17 Università del 2° Cluster (21 Atenei).

Università degli Studi del Piemonte Orientale Amedeo Avogadro	Università degli Studi di Udine	Università degli Studi di L'Aquila
Libero istituto Universitario Carlo Cattaneo di Castellanza	Università Politecnica delle Marche	Università degli Studi di Teramo
Università degli Studi dell'Insubria	Università degli Studi della Tuscia	Università degli Studi del Molise
Università degli Studi di Milano - Bicocca	Università degli Studi di Roma - Tor Vergata	Libera Università Mediterranea J ean Monnet - Casamassima
Università degli Studi di Bergamo	Istituto Universitario di Scienze Motorie di Roma	Università degli Studi della Basilicata
Università degli Studi di Brescia	Università Campus Bio-Medico di Roma	Università degli Studi della Calabria
Istituto Universitario di Architettura di Venezia	Libera Università degli Studi San Pio V di Roma	Università degli Studi di Catanzaro Magna Grecia

Tabella 5.18 Università outliers del 3° Cluster e 4° cluster (1 e 2 Atenei).

Libera Università di Bolzano	Scuola Superiore di studi avanzati e di perfezionamento S.Anna di Pisa
	Scuola Internazionale superiore di Studi avanzati (SISSA) di Trieste

Tabella 5.19 Medie degli indicatori per i due gruppi principali.

Indicatori	Medie del 1° gruppo	Medie del 2° gruppo		Indicatori	Medie del 1° gruppo	Medie del 2° gruppo
acc1	41,591	40,718		prod1	134,209	80,269
acc2	0,658	0,344		prod2	267,237	155,732
acc3	63,76	66,5		prod3	39,124	43,269
acc4	24,213	28,5		prod4	1327,375	1020,648
acc5	0,894	0,271		prod5	259,404	248,332
acc6	47,028	42,853		prod6	9183,859	7506,898
acc7	18,442	9,591		vita1	359,675	415,805
acc8	1,778	1,723		vita2	9,653	6,837
eff1	395,923	1003,598		vita3	74,491	21,369
eff2	7,174	6,256		vita4	23,568	19,443
eff3	5,111	2,605		vita5	1,634	0,654
eff4	43,44	41,953		vita6	47,922	53,523
eff5	14,438	18,306		vita7	49,362	47,636
eff6	51,008	50,007		vita8	0,143	0,155

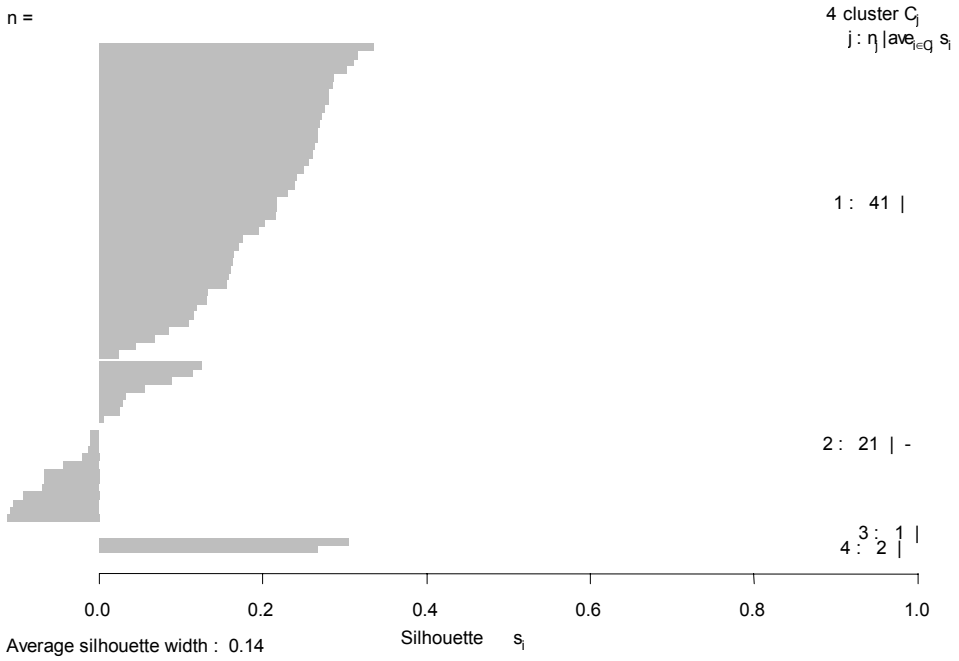
Si nota che le maggiori differenze tra gli indicatori dei due gruppi non sono molte, ma risultano significative per comprendere la tipologia di Ateneo inclusa nelle due partizioni. Dai valori assunti dalla sezione riguardante l'accessibilità, si osserva che il primo *cluster* di Atenei ottiene dei valori negli indicatori acc2, acc5 e acc7 nettamente superiori: ciò sta a significare che al primo gruppo appartengono strutture che hanno una maggiore superficie in rapporto all'utenza, e più decentrate (l'indicatore acc7 nel primo *pattern* di Atenei e quasi il doppio rispetto al secondo).

Nella seconda parte degli indicatori, quella riguardante l'efficacia, troviamo evidenti differenze nella variabile eff1 ed eff3: essi indicano che nel primo *cluster* si collocano gli Atenei che hanno un livello di fruibilità migliore, e ciò si traduce nell'avere un più alto numero di pc destinati al pubblico e maggiori prestiti per ogni singolo utente.

Andando ad analizzare gli indicatori di efficienza, non si notano eclatanti differenze, ma in generale i valori assunti dal primo pattern di Atenei, sembrano essere leggermente migliori rispetto al secondo gruppo.

Nell'ultima sezione riguardante le vitalità del patrimonio, le differenze che si riscontrano tra i due gruppi riguardano gli indicatori vita2, vita3 e vita5. Essi esprimono che, in media, le strutture bibliotecari del secondo gruppo hanno meno personale per punti di servizio rispetto al primo, hanno un patrimonio documentario più ridotto ed effettuano meno rapporto all'utenza.

Grafico 5.19 Silhoutte plot ottenuto dal partizionamento precedente.



Le prestazioni dell'algoritmo PAM vengono generalmente valutate utilizzando un particolare tipo di grafico, detto *"silhouette" plot*, che, per ogni medoide, associa ad ogni pattern un punteggio cha va da –1 (*pattern* mal associato) a +1 (*pattern* ben associato).
I pattern con punteggio intermedio (punteggio zero) non appartengono in definitiva a nessun medoide (dati intermedi). In figura (5.19) riportiamo il plot ottenuto con i nostri dati di partenza. La media dei punteggi ottenuti è di 0.14, perciò l'associazione tra medoidi e *patterns* non è perciò del tutto soddisfacente, come ci aspettavamo, del resto, data la natura dei dati. Nel grafico vengono disegnati, a partire dall'alto verso il basso i vari punteggi dei *patterns* (ordinati a decrescere) per il cluster 1 fino al cluster 4. Tuttavia vi è un consistente numero di punteggi negativi: ci indica, in special modo, che nel secondo gruppo ci sono alcuni Atenei mal associati.

Conclusioni

A chiusura di tutto il lavoro svolto, l'analisi dei dati proveniente dall'indagine GIM ha prodotto alcuni elementi di valutazione: essi meritano di essere considerati allo scopo di capire lo stato attuale dell'offerta bibliotecaria degli Atenei italiani.

Rispetto all'indagine precedente del GIM, l'imputazione dei *missing data* con tecniche che mirano a conservare la natura dei dati, ha comportato un leggero aumento dei valori calcolati per gli indicatori. Ciò fa ritenere che, l'imputazione dei dati mancati con metodi più *"naive"* come media e mediana, in genere comporta una sottostima del valori originari.

Come si può notare dal capitolo 3, la forte componente composta da una serie di unità *outliers*, ha comportato un abbassamento della qualità generale dello studio. Questo fa pensare che alcuni indicatori non sono in grado di valutare in maniera efficace gli Atenei e le Istituzioni Accademiche più piccole, che rivolgono la loro attenzione su un'utenza di minor numerosità, o quelli mono - biblioteca che sicuramente spiccano per una diversa organizzazione bibliotecaria. Per questo motivo, l'analisi *cluster* e quella fattoriale sono risultate solo discretamente significative conducendo, in taluni casi, a risultati di dubbia interpretazione. Inoltre, a causa della scarsa correlazione tra gli indicatori, non si sono potuti estrarre dal *dataset* indici complessivi di performance, che contando su un elevata numerosità campionaria, potevano essere usati per successive analisi.

L'uso di dati scarsamente misurabili, come l'utenza potenziale, sono labili a diverse interpretazioni da parte di chi compila il questionario, comportando una distorsione dei risultati: una possibile soluzione può essere cercata introducendo nell'analisi il valore di utenza effettiva in modo di ottenere una informazione concreta dei servizi realmente offerti. L'analisi svolta è riuscita a far emergere delle informazioni generali di estreme interesse sul mondo bibliotecario accademico:

- non emergono valori estremamente negativi dal quadro generale dei valori assunti dagli indicatori per ogni singolo Ateneo;
- vale la pena sottolineare che, dall'analisi in componenti principali, in ogni area di interesse degli indicatori, è possibile notare un ristretto gruppo di Atenei che si collocano nelle posizioni migliori.

Pur non ottenendo significative discriminazioni, questi Atenei hanno punteggi talmente vicini da ritenere che vi sia una comune visione dell'organizzazione e della gestione bibliotecaria;

- la *cluster analysis* ci permette di raggruppare gli Atenei simili sul piano della organizzazione bibliotecaria, fornendo elementi di uno studio per analisi successive che possono mirare alla selezione di un particolare sottogruppo di Atenei simili.

Il questionario *on–line*, proposto dal GIM, ha ottenuto ottimi risultati sul piano della raccolta delle informazioni e fornendo un modello per successive indagini in tal senso. Tuttavia, considerando che il lavoro di censimento elaborato dal GIM includeva Atenei e biblioteche molto eterogenei tra loro, era difficile aspettarsi un risultato migliore. Una possibile soluzione, in tal senso, potrebbe essere l'analisi degli Atenei con organizzazione bibliotecaria simile, ma i risultati ottenibili non coglierebbero la reale situazione riscontrata con l'indagine corrente.

APPENDICE

Allegato A

Codifica degli ID assegnati agli Atenei

ID	NOME	REGIONE
1	Università degli Studi di Torino	PIEMONTE
2	Politecnico di Torino	PIEMONTE
3	Università degli Studi del Piemonte Orientale Amedeo Avogadro	PIEMONTE
4	Università della Valle d'Aosta	VALLE D'AOSTA
5	Università degli Studi di Genova	LIGURIA
6	Libero Istituto Universitario Carlo Cattaneo di Castellanza	LOMBARDIA
7	Università degli Studi dell' Insubria	LOMBARDIA
8	Università degli Studi di Milano	LOMBARDIA
9	Politecnico di Milano	LOMBARDIA
10	Università Commerciale Luigi Bocconi di Milano	LOMBARDIA
11	Università Cattolica del Sacro Cuore di Milano	LOMBARDIA
12	Libera Università di Lingue e Comunicazione (IULM)	LOMBARDIA
13	Libera Università Vita Salute San Raffaele di Milano	LOMBARDIA
14	Università degli Studi di Milano - Bicocca	LOMBARDIA
15	Università degli Studi di Bergamo	LOMBARDIA
16	Università degli Studi di Brescia	LOMBARDIA
17	Università degli Studi di Pavia	LOMBARDIA
18	Libera Università di Bolzano	TRENTINO ALTO
19	Università degli Studi di Trento	TRENTINO ALTO
20	Università degli Studi di Verona	VENETO
21	Università degli Studi Ca' Foscari di Venezia	VENETO
22	Istituto Universitario di Architettura di Venezia	VENETO
23	Università degli Studi di Padova	VENETO
24	Università degli Studi di Udine	FRIULI VENEZIA
25	Università degli Studi di Trieste	FRIULI VENEZIA
26	Scuola internazionale superiore di Studi avanzati (SISSA) di Trieste	FRIULI VENEZIA
27	Università degli Studi di Parma	EMILIA
28	Università degli Studi di Modena e Reggio Emilia	EMILIA
29	Università degli Studi di Bologna	EMILIA
30	Università degli Studi di Ferrara	EMILIA
31	Università degli Studi di Urbino	MARCHE
32	Università Politecnica delle Marche	MARCHE
33	Università degli Studi di Macerata	MARCHE
34	Università degli Studi di Camerino	MARCHE

ID	NOME	REGIONE
35	Università degli Studi di Firenze	TOSCANA
36	Università degli Studi di Pisa	TOSCANA
37	Scuola normale superiore di Pisa	TOSCANA
38	Scuola superiore di Studi avanzati e di perfezionamento S. Anna di Pisa	TOSCANA
39	Università degli Studi di Siena	TOSCANA
40	Università per stranieri di Siena	TOSCANA
41	Università degli Studi di Perugia	UMBRIA
42	Università per stranieri di Perugia	UMBRIA
43	Università degli Studi della Tuscia	LAZIO
44	Università degli Studi di Roma La Sapienza	LAZIO
45	Università degli Studi di Roma Tor Vergata	LAZIO
46	Libera Università degli Studi Maria SS.Assunta - (LUMSA) di Roma	LAZIO
47	Libera Università Internazionale di Studi Sociali Guido Carli - (LUISS)	LAZIO
48	Istituto Universitario di Scienze Motorie di Roma	LAZIO
49	Università degli Studi Roma Tre	LAZIO
50	Università Campus Bio-Medico di Roma	LAZIO
51	Libera Università degli Studi San Pio V di Roma	LAZIO
52	Università degli Studi di Cassino	LAZIO
53	Università degli Studi del Sannio	CAMPANIA
54	Università degli Studi di Napoli Federico II	CAMPANIA
55	Istituto Universitario Navale di Napoli (Parthenope)	CAMPANIA
56	Istituto Università rio Orientale di Napoli	CAMPANIA
57	Istituto Universitario Suor Orsola Benincasa di Napoli	CAMPANIA
58	Seconda Università degli Studi di Napoli	CAMPANIA
59	Università degli Studi di Salerno	CAMPANIA
60	Università degli Studi di L'Aquila	ABRUZZO
61	Università degli Studi di Teramo	ABRUZZO
62	Università degli Studi Gabriele D'Annunzio di Chieti	ABRUZZO
63	Università degli Studi del Molise	MOLISE
64	Università degli Studi di Foggia	PUGLIA
65	Università degli Studi di Bari	PUGLIA
66	Politecnico di Bari	PUGLIA
67	Libera Università Mediterranea Jean Monnet - Casamassima	PUGLIA
68	Università degli Studi di Lecce	PUGLIA
69	Università degli Studi della Basilicata	BASILICATA
70	Università degli Studi della Calabria	CALABRIA
71	Università degli Studi di Catanzaro Magna Grecia	CALABRIA
72	Università degli Studi di Reggio Calabria	CALABRIA
73	Università degli Studi di Palermo	SICILIA
74	Università degli Studi di Messina	SICILIA
75	Università degli Studi di Catania	SICILIA
76	Università degli Studi di Sassari	SARDEGNA
77	Università degli Studi di Cagliari	SARDEGNA

Allegato B

Indicatori

Nr	Fonte bibl.	*Formula di calcolo*	nome	Area d'interesse
1	Eqlipse, Osservatorio	Media delle ore di apertura settimanale	Acc1	accessibilità
2	Eqlipse; CE; SCONUL; CRUI; Osservatorio	Superficie totale / utenti potenziali	Acc2	accessibilità
3	GIM	Superficie accessibile al pubblico / superficie totale *100	Acc3	accessibilità
4	Eqlipse; SCONUL; Osservatorio	Utenti potenziali / posti di lettura	Acc4	accessibilità
5	SCONUL	Metri lineari a scaffale aperto occupati dai materiali / utenti potenziali	Acc5	accessibilità
6	SCONUL	Metri lineari totali a scaffale aperto / metri lineari totali di scaffalatura * 100	Acc6	accessibilità
7	Osservatorio	Unità amministrative	Acc7	accessibilità
8	Osservatorio	Punti di servizio / unità amministrative	Acc8	accessibilità
9	Eqlipse; CE; Osservatorio	Utenti potenziali / personal computer destinati al pubblico	Eff1	efficacia/fruibilità/innovazione
10	GIM	Personal computer destinati al pubblico / posti di lettura + personal computer destinati al pubblico * 100	Eff2	efficacia/fruibilità/innovazione
11	Eqlipse; ISO; CE; SCONUL; Osservatorio	Prestiti + prestiti interbibliotecari passivi + document delivery passivi / utenti potenziali	Eff3	efficacia/fruibilità/innovazione
12	ARL; SCONUL; Osservatorio	Prestiti interbibliotecari attivi + document delivery attivi / prestiti interbibliotecari totali + document delivery totali * 100	Eff4	efficacia/fruibilità/innovazione
13	Eqlipse; CE; Osservatorio	Prestiti interbibliotecari totali + document delivery totali / prestiti + prestiti interbibliotecari totali + document delivery totali* 100	Eff5	efficacia/fruibilità/innovazione
14	GIM	Inventari in OPAC / patrimonio documentario * 100	Eff6	efficacia/fruibilità/innovazione
15	Eqlipse; CE; SCONUL; Osservatorio	Spese della biblioteca per risorse bibliografiche utenti potenziali	Prod1	efficienza/produttività/economicità
16	Eqlipse; SCONUL; Osservatorio	Spese totali della biblioteca / utenti potenziali	Prod2	efficienza/produttività/economicità
17	ARL; SCONUL	SpSpese della biblioteca per il personale / spese totali della biblioteca * 100	Prod3	efficienza/produttività/economicità
18	Eqlipse; EAL	Prestiti + prestiti interbibliotecari totali + document delivery totali / personale FTE	Prod4	efficienza/produttività/economicità
19	Eqlipse; EAL	Acquisizioni / personale FTE	Prod5	efficienza/produttività/economicità
20	Eqlipse; EAL	Patrimonio documentario / personale FTE	Prod6	efficienza/produttività/economicità
21	Eqlipse; EAL; SCONUL; ARL	Utenti potenziali / personale FTE	Vita1	vitalità del patrimonio/offerta risorse
22	CRUI; Osservatorio	Personale FTE / punti di servizio	Vita2	vitalità del patrimonio/offerta risorse
23	Eqlipse; CE; SCONUL; Osservatorio	Patrimonio documentario / utenti potenziali	Vita3	vitalità del patrimonio/offerta risorse
24	ARL; Osservatorio [per studente]	Periodici elettronici + Periodici cartacei: abbonamenti / docenti e ricercatori	Vita4	vitalità del patrimonio/offerta risorse
25	Eqlipse; CE; SCONUL; Osservatorio	Acquisizioni / utenti potenziali	Vita5	vitalità del patrimonio/offerta risorse
26	ARL	Periodici elettronici / periodici totali correnti (elettronici + abbonamenti cartacei) * 100	Vita6	vitalità del patrimonio/offerta risorse
27	ARL; SCONUL; Osservatorio	Spese della biblioteca per risorse bibliografiche / spese totali della biblioteca * 100	Vita7	vitalità del patrimonio/offerta risorse
28	Eqlipse; IFLA	Prestiti + prestiti interbibliotecari attivi + document delivery attivi / patrimonio documentario * 100	Vita8	vitalità del patrimonio/offerta risorse

Allegato C

Comandi in ambiente R

Per una sintassi più completa dei comandi utilizzati si rimanda all'aiuto in linea, alla funzione help di R e a Masarotto G. Iacus S.M., (2003), "Laboratorio di statistica con R".

 # Importazione dati in ambiente R

```
dati<-read.table(file.choose(),header=T)       # importo in R dati di un file testo separate da
                                               # tabulazione con la prima riga di nomi delle variabili
dati<-red.csv(file.choose(),header=T,sep=";")  # importo in R dati in formato *.csv
```

 # Ecco alcuni comandi per condurre l'analisi descrittiva sui dati

```
data(iris)                  # carichiamo il dataset iris nell'ambiente R
boxplot(iris[,1])           # grafico boxplot della prima variabile
hist(iris[,2])              # istogramma della seconda variabile
plot(iris[,1:4])            # grafico di dispersione tra le variabili
cor(iris[,1:4])             # matrice di correlazione dei dati
stand<-scale(iris[,1:4])    # standardizza le variabili (media zero e varianza unitaria)
```

 # Modelli lineari

```
data(cars)                          # dataset cars
names(cars)
fit<-lm(dist ~speed,data=cars)
#oppure
fit<-lm(cars$dist~cars$speed)       # stima di un modello lineare
summary(fit)                        # stima dei parametri e test sulla bontà del modello
anova(fit)                          # analisi della varianza del modello lineare
new<-data.frame(speed=c(,24,30))
predict(fit,newdata=new)            # valori predetti dal modello
# analogo procedimento per la stima di modelli lineari generalizzati con la funzione glm()
```

 # Analisi componenti principali e analisi fattoriale

```
data(iris)
fit2<-princomp(iris[,1:4],cor=T)        # calcolo delle componenti principali
summary(fit2)                           # valori degli autovalori e varianza spiegata
names(fit2)                             # oggetti dell'elemento fit3
loadings(fit2)                          # pesi delle componenti principali
plot(fit2)                              # istogramma degli autovalori
biplot(fit2)                            # plot dei punteggi fattoriali ottenuti dai dati
plot(fit2scores[,1],fit2$scores[,2])    # punteggi fattoriali sulle prime 2 componenti

library(ade4)
```

per informazioni più dettagliate sulla libreria consultare il manuale di riferimento presso http://cran.r-project.org/doc/packages/ade4.pdf.

```
fit3<-dudi.pca(iris[,1:4])              # calcolo delle componenti principali
names(fit3)                             # oggetti dell'elemento fit3
s.corcircle(fit3$li)                    # cerchi delle correlazioni dei pesi fattoriali
rotate<-varimax(as.matrix(fit3$co))     # rotazione degli assi delle componenti principali
names(rotate)                           # oggetti dell'elemento rotate
s.corcircle(rotate$loadings[,1:2])      # cerchi delle correlazioni dei pesi fattoriali
ruotati
s.class(fit3$li,fac=iris[,5])           # grafico dei punteggi fattoriali delle 2 componenti

data(mtcars)                                        # dataset mtcars
fit4<-factanal(mtcars,2,rotation="promax")          # calcolo di 2 fattori
fit4                                                # pesi fattoriali e varianza spiegata
scores<- factanal(mtcars,2,rotation="promax",scores="Bartlett")$scores
                                                    # calcolo dei punteggi fattoriali di Bartlett
plot(scores[,1],scores[,2])                         # plot dei punteggi fattoriali
abline(v=0,lty=2)                                   # aggiungo linea asse verticale
abline(h=0,lty=2)                                   #aggiungo linea asse orizzontale
```

#Analisi Cluster

```
library(cluster)                                    # carico libreria cluster
distanze<-dist(mtcars,method="euclidean")           # calcolo della matrice delle distanze
fit5<-hclust(distanze,method="ward")                # clusterizzazione gerarchica
plot(fit5)                                          # dendrogramma basato sulle distanze di fit5
fit6<- pam(mtcars,4)                                # modello di partizioni basato sull'algoritmo PAM
par(mfrow=c(1,2))                                   # inizializzo l'output grafico dividendolo in 2 parti
plot(fit6)          # grafico della partizioni e silhoutte plot per valutare la bontà del modello
```

#Pseudo-codice di imputazione hot-deck

```
hotdeck<-function(dati){
n<-trova1(dati)                          # funzione trova1() che ricava le righe di dati mancanti
l<-length(n)
dati2<-dati[-n,]                              # sottomatrice di dati completi
                    # creo le classi di imputazioni differenziate per tipologia di biblioteca
ateneoei<-dati2[dati2$tipologi=="ateneo" | dati2$tipologi=="interfacolta",,]
dipartimento<-dati2[dati2$tipologi=="dipartimento",,]
facoltaei<-dati2[dati2$tipologi=="facolta" | dati2$tipologi=="interdipartimentale",,]
istituto<-dati2[dati2$tipologi=="istituto",,]
for(i in 1:l){
if(dati$tipologi[n[i]]=="ateneo" | dati$tipologi[n[i]]=="interfacolta"){
q<-dim(ateneoei)[1]
a<-sample(q,1)                      # scelgo un campione casuale dalla classe di imputazione
dati[n[i],8:51]<-dati2[a,8:51]
}
if(dati$tipologi[n[i]]=="dipartimento"){
q<-dim(dipartimento)[1]
a<-sample(q,1)
dati[n[i],8:51]<-dati2[a,8:51]
}
if(dati$tipologi[n[i]]=="facolta" | dati$tipologi[n[i]]=="interdipartimentale"){
q<-dim(facoltaei)[1]
a<-sample(q,1)
dati[n[i],8:51]<-dati2[a,8:51]
}
if(dati$tipologi[n[i]]=="istituto"){
q<-dim(istituto)[1]
a<-sample(q,1)
dati[n[i],8:51]<-dati2[a,8:51]
}}
dati
}
```

Riferimenti Bibliografici

Bolasco S., (1999), *"Analisi multidimensionale dei dati"*, Carrocci Editore.

Fabbris L., (1990), *"Analisi esplorative di dati multidimensionali"* , Cleup Editore.

Masarotto G. Iacus S.M., (2003), *"Laboratorio di statistica con R"* , McGraw-Hill.

Hair, Anderson, Tatham, Black (1998), *"Multivariate data analysis"* , Fifth Edition, Prentice-Hall.

Saporta G., Bouroche J.M., (2002),*"Analisi dei dati"* , Clu, Napoli.

Venables V.N., Ripley B.D., (*2002) ,"Modern Applied Statistics with {S}",* Springer.

Little, R.J.A., and Schenker N., (1994), "*Missing data. In: Handbook for Statistical Modeling in the Social and Behavioral Sciences",* Plenum Press.

Allison P.D. , (2001), *"Missing data",* Sage Publications Ltd.

Armitage P., Berry G. (1996), *"Statistica medica: metodi statistici per la ricerca in medicina"*, McGraw-Hill.

Pace L., Salvan A., (2001), *"Introduzione alla Statistica – II. Inferenza, Verosimiglianza, Modelli"*, Cedam..

Bortot P., Ventura L. e Salvan A., (2000), *"Inferenza Statistica: Applicazioni con S-Plus e R",* Cedam.

Little R.J.A., Rubin D.B., (1987), *"Statistical Analysis with Missing Data"*, Wiley.

Giraldo A., (1995), *"Imputazione Multipla per Meccanismi di Non Risposta Non Ignorabili"*, Tesi di Dottorato, Università degli Studi di Firenze.

Diana G., lucidi delle lezioni di *"Analisi dei dati multidimensionali",* Facoltà di Scienze Statistiche, Padova.

Box G., Cox D., (1964), "An analysis of transformations", Journal of the Royal Statistical Society.

L'indagine del 2002 svolta dal gruppo GIM è consultabile *on-line* all'indirizzo http://www.biblio.unive.it/sba/statistiche.