

Università degli Studi di Napoli "Federico II"

Scuola Politecnica e delle Scienze di Base
Corso di Laurea Magistrale in Ingegneria Informatica

Giuseppe Francesco Di Cecio - M63001211
Nicola D'Ambra - M63001223
Emma Melluso - M63001176

Elaborato di *Impianti di Elaborazione*



Università degli Studi di Napoli *Federico II*

Napoli

A.A 2020/2021

Indice

1	Workload Characterization	1
1.1	Filtraggio	1
1.1.1	Colonne Identiche	1
1.1.2	Outlier	2
1.2	PCA	5
1.3	Clustering	6
1.3.1	4 Componenti Principali	6
1.3.2	5 Componenti Principali	7
1.3.3	6 Componenti Principali	7
1.3.4	Interpretazione	8
1.4	Workload Sintetico	8

Capitolo 1

Workload Characterization

Il dataset di partenza è composto da **3000 righe** e **24 colonne**, ciascuna delle quali rappresenta uno dei parametri del sistema oggetto di studio. Si tratta di parametri caratterizzanti l'esecuzione di vari Threads su un sistema operativo.

In particolar modo le colonne con prefisso *Vm* rappresentano informazioni sulla memoria virtuale occupata e utilizzata dai Threads, mentre le altre colonne rappresentano informazioni di carattere generale, come memoria libera, numero di threads, pagine inattive ecc.

1.1 Filtraggio

1.1.1 Colonne Identiche

Innanzitutto osservando il workload e effettuando un grafico delle distribuzioni ci rende conto della presenza di ben 4 colonne costanti:

- **Active**
- **AnonPages**
- **AvbLatency**
- **Error**

Essendo tali non spiegano varianza, dunque possono essere tranquillamente trascurate ai fini dell'analisi.

Osservando le distribuzioni dei parametri **WriteBack** e **MemFree** si sono notate alcune caratteristiche comuni. Per avere una maggiore chiarezza si è preferito calcolare la matrice delle correlazioni su questi due parametri.



	'MemFree'	'Writeback'
'MemFree'	1,0000	1,0000
'Writeback'	1,0000	1,0000

Figura 1.1: *Matrice di correlazione tra MemFree e WriteBack*

Osservando la matrice appare evidente che le due colonne sono esattamente identiche, fornendo quindi la stessa informazione. Per questo motivo si è deciso di trascurare una delle due, in particolare quella di WriteBack.

Le 24 colonne iniziali sono state ridotte a 19 colonne, riducendo il dataset di un numero di osservazioni pari a:

$$n_{dati} = 3.000 \times (24 - 19) = 15.000 \quad (1.1)$$

1.1.2 Outlier

Gli outliers sono valori che si discostano notevolmente dalle altre osservazioni. Non possono essere subito eliminati poiché influiscono sulle analisi statistiche in modo considerevole. Si possono però analizzare per capire se è possibile rimuoverli o meno. Osservando l'andamento dei seguenti parametri, attraverso box plot e grafici di distribuzione:

- **VmSize**, quanta memoria virtuale utilizza l'intero processo.
- **VmHWM**, di quanta RAM il processo necessita al massimo.
- **VmRSS**, quanta RAM il processo sta correntemente usando.
- **VmPTE**, quanta memoria Kernel è occupata dalle entries della tabella delle pagine.

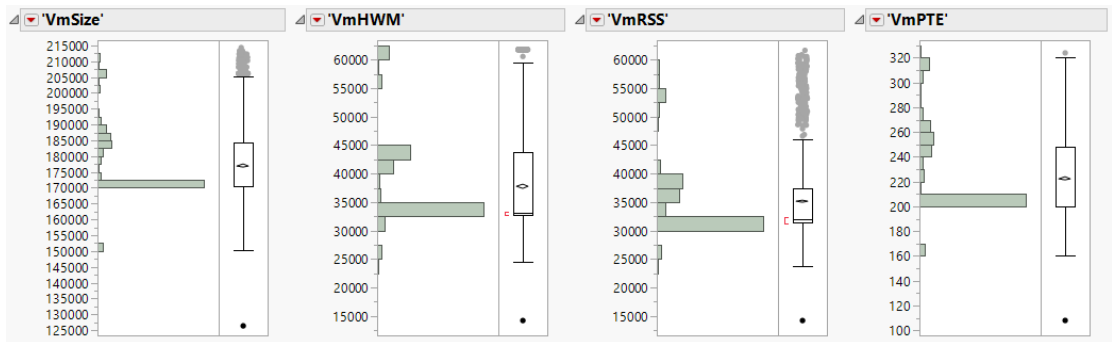
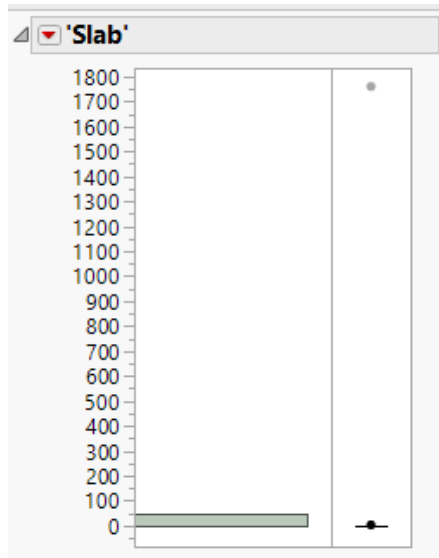


Figura 1.2: *Grafici di distribuzione di VmSize, VmHWM, VmRSS, VmPTE*

Si è notato che essi presentano un outlier isolato (in basso ad ogni grafico) in comune associato alla prima riga del dataset. Analizzando gli altri parametri (*MemFree*, *Dirty*, *PageTables*, *Buffer*, ...) è stato possibile evidenziare che anche per la maggior parte di essi lo è, ma non è un punto isolato.

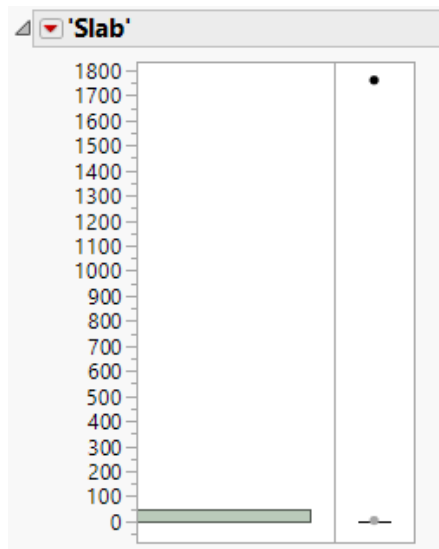
L'ipotesi fatta è che con molta probabilità le prime righe del dataset (da 0 a 100 circa), rappresentano la fase di avvio del processo e l'outlier oggetto di studio è la prima istanza di questa fase. Dato che l'obiettivo della caratterizzazione del workload è quello di analizzare le prestazioni a regime del sistema oggetto di studio (in questo caso), si è deciso di trascurare quel singolo outlier. E' bene notare che in ogni caso le informazioni riguardo questa fase di avvio non saranno del tutto perse dato che è stato rimosso un singolo punto e non tutti i punti che la rappresentano.

Un secondo outlier che può essere agevolmente rimosso è la riga 512 in cui il parametro **Slab** assume valore 4.

Figura 1.3: *Grafico di distribuzione di Slab*

Oltre ad avvicinarsi molto al valore medio assunto da Slab (zero), esso risulta essere un outlier solo per il parametro stesso dato che per gli altri è un valore compreso tra i quartili. Una sua rimozione quindi non influenza gli indici di caratterizzazione sintetica dei parametri del workload complessivo.

Un terzo outlier è il valore 1760 del parametro *Slab*, associato alla riga 90 del workload.

Figura 1.4: *Grafico di distribuzione di Slab*

Rispetto al precedente, tale outlier richiede un' analisi più approfondita visto che influenza significativamente l'andamento di parametri quali *Mapped* e *PageTables*. Per descrivere meglio la dipendenza tra questi parametri si può effettuare un grafico tra il numero dell'osservazione e il valore assunto da *Mapped*, analogo discorso con *PageTables*.

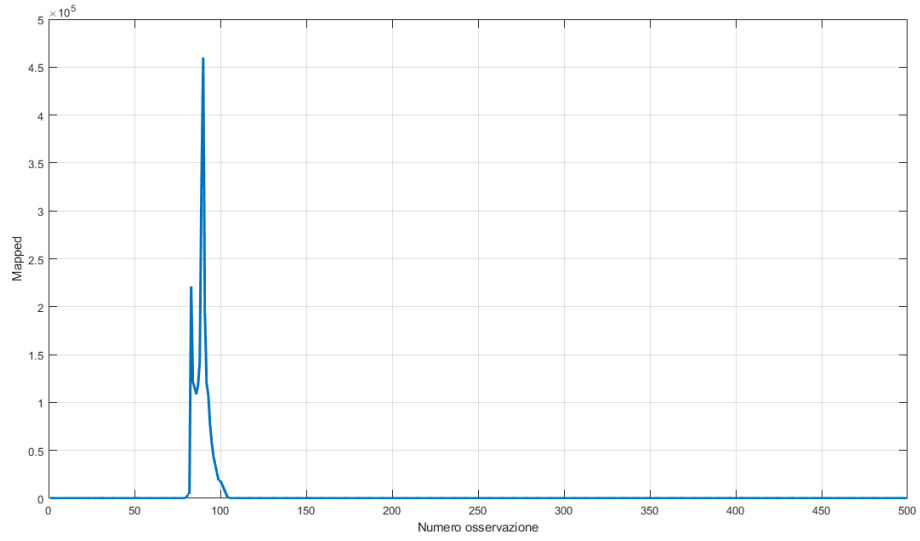


Figura 1.5: *Grafico tra numero di osservazione e valore assunto dal parametro Mapped*

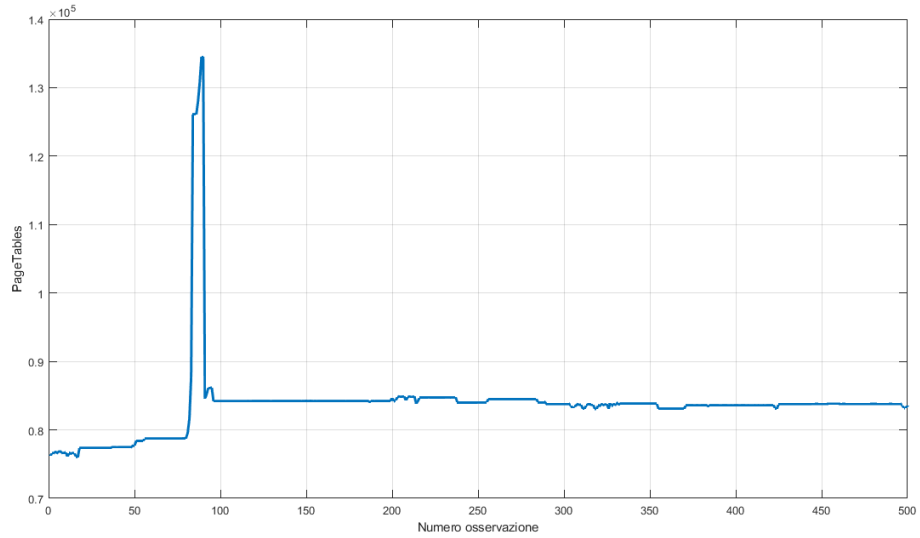


Figura 1.6: *Grafico tra numero di osservazione e valore assunto dal parametro PageTables*

Si nota che in corrispondenza (in realtà nell'osservazione appena precedente) dell'outlier del parametro *Slab* i due parametri sopra indicati hanno un picco, durante la fase di avvio del sistema.

Lo *Slab* si riferisce ad un particolare meccanismo di allocazione/deallocazione della memoria nel Kernel. Dato che influenza in particolar modo altri parametri si è preferito di non trascurarlo.

In conclusione sono stati eliminati dal dataset solo 2 outlier che corrispondono a 38 campioni. Quindi il dataset è stato ridotto in totale di 15.038 elementi, provocando una diminuzione dei dati iniziali di poco più del 20%.

1.2 PCA

A seguito del filtraggio il dataset risulta ridotto grazie alla rimozione di alcune colonne e righe che rappresentano outlier.

Sul questo dataset si possono quindi iniziare a fare le prime considerazioni.

Utilizzando la tecnica della *Principal Component Analysis* il dataset può essere estremamente ridotto, sfruttando solo le *Componenti Principali* che mantengono più varianza. Il risultato della PCA è quindi:

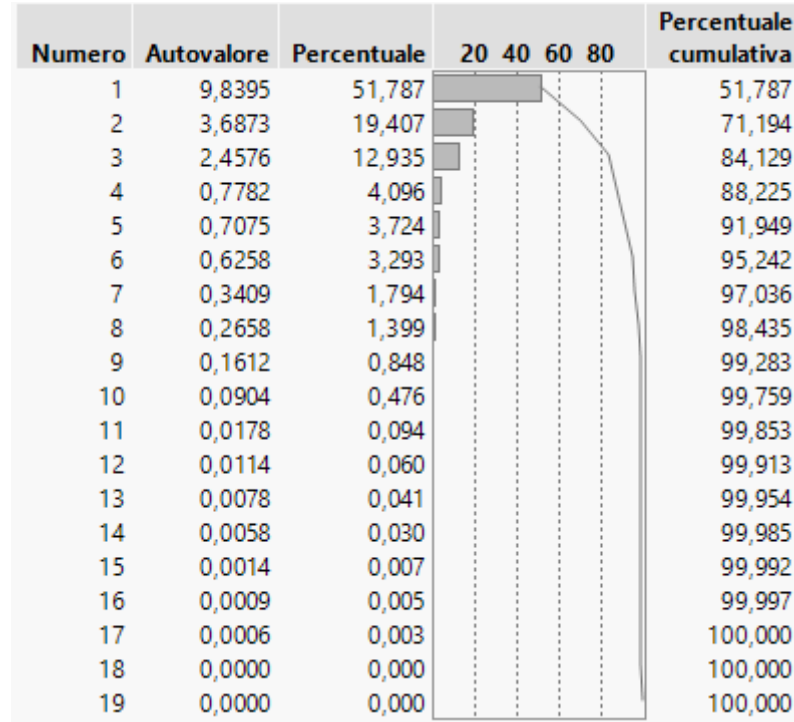


Figura 1.7: PCA applicata al dataset filtrato

La scelta del numero di componenti principali ricade in particolar modo sulla devianza che quelle componenti mantengono rispetto al dataset reale. Inoltre essa dipende anche dal tipo di osservazioni ed esperimento che è stato effettuato.

La scelta migliore ricade nel prendere 5 componenti principali poiché rappresentano il 92% della devianza totale. Esso rappresenta un valore anche abbastanza elevato, ma è stato scelto per mantenersi in una regione di tolleranza durante la clusterizzazione.

Per completezza si possono riportare le analisi considerando tre set di componenti principali nel seguente modo:

- Prendere 4 PC

$$DEV_{PCA-MANTENUTA} \approx 88\%$$

- Prendere 5 PC

$$DEV_{PCA-MANTENUTA} \approx 92\%$$

- Prendere 6 PC

$$DEV_{PCA-MANTENUTA} \approx 95\%$$

Sulla base di queste tre alternative, viene effettuata la procedura di clustering per ognuna di esse.

1.3 Clustering

Il clustering è una tecnica che consiste nel raggruppare osservazioni "simili" tra loro. La similitudine tra un elemento e un cluster, o tra un cluster e un altro cluster, può essere calcolata secondo varie tecniche. In questa analisi si è preferito utilizzare il **metodo di Ward**, il quale pesa la distanza tra due cluster in relazione al numero di elementi che li compongono.

Dati due cluster P e Q (un elemento non appartenente ad un cluster, può essere visto come un cluster di dimensione 1), sia $|P|$ la cardinalità di P , analogo con $|Q|$, e sia \bar{x}_p il centroide di P , analogo con \bar{x}_q , la distanza tra P e Q viene calcolata come:

$$d(P, Q) = 2 \frac{|P| |Q|}{|P| + |Q|} \|\bar{x}_p - \bar{x}_q\|^2$$

Per ogni raggruppamento viene poi scelto una singola osservazione che la rappresenta, riducendo quindi il dataset pari al numero di cluster scelti durante l'analisi.

A tal proposito il numero di cluster da scegliere può dipendere da vari fattori:

- Omogeneità dei cluster: i cluster devono raggruppare un numero di osservazioni quanto il più possibile omogeneo rispetto agli altri cluster. Avere un cluster con un numero di elementi di vari ordini di grandezza rispetto ad un altro cluster non sempre può portare a buoni risultati (in termini di devianza).
- Devianza mantenuta: a seguito della PCA parte della devianza nei dati viene persa. Dato che il clustering viene effettuato sulle *Componenti Principali* allora esso produce un'ulteriore perdita di devianza nel risultato finale.

Devianza Persa

Per effettuare il calcolo della devianza totale persa bisogna prima calcolare la devianza intra-cluster (la somma delle devianze per ogni cluster) e sulla base di questa si può calcolare la quantità richiesta.

Matematicamente, definita $DEV_{PCA-PERSA}$ la devianza persa (in termini percentuali) a causa della PCA, viceversa $DEV_{PCA-MANTENUTA}$ la devianza mantenuta dalla PCA, e DEV_{INTRA} la devianza intra-cluster, la devianza totale persa percentuale vale:

$$DEV_{PCA-LOST} + DEV_{INTRA} \times DEV_{PCA-MANTENUTA}$$

Essa può essere calcolata in MATLAB passando ad uno script i cluster, le componenti principali e il dataset iniziale.

Per dare valore ai fattori sopra citati il clustering viene effettuato scegliendo un numero di cluster che varia da 6 a 16 (per ogni gruppo di componenti principali) su cui poi viene calcolata la devianza persa.

1.3.1 4 Componenti Principali

Utilizzando le prime quattro PC si ha:

Riepilogo cluster						
Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	
1	82	-10.920	8.451	-2.554	-0.241	
2	7	0.178	12.283	20.284	4.779	
3	1	4.824	19.063	52.299	-20.679	
4	1013	-1.560	-0.981	0.393	0.650	
5	1631	0.474	-0.295	-0.085	-0.514	
6	264	6.427	2.564	-0.923	0.711	

Riepilogo cluster						
Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	
1	82	-10.920	8.451	-2.554	-0.241	
2	7	0.178	12.283	20.284	4.779	
3	1	4.824	19.063	52.299	-20.679	
4	1013	-1.560	-0.981	0.393	0.650	
5	786	-1.282	-0.852	0.100	-0.551	
6	227	1.973	0.234	-0.467	-1.486	
7	618	2.158	0.220	-0.181	-0.110	
8	264	6.427	2.564	-0.923	0.711	

Riepilogo cluster						
Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	
1	82	-10.920	8.451	-2.554	-0.241	
2	6	-0.740	12.018	17.492	4.744	
3	1	5.681	13.876	37.033	4.986	
4	1	4.824	19.063	52.299	-20.679	
5	74	-1.585	-0.915	1.004	1.692	
6	939	-1.558	-0.987	0.345	0.567	
7	786	-1.282	-0.852	0.100	-0.551	
8	227	1.973	0.234	-0.467	-1.486	
9	501	2.229	0.269	-0.222	-0.316	
10	117	1.852	0.009622	-0.00249	0.769	
11	77	6.308	2.508	-0.539	1.549	
12	187	6.476	2.550	-1.081	0.366	

Riepilogo cluster						
Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	
1	82	-10.920	8.451	-2.554	-0.241	
2	6	-0.740	12.018	17.492	4.744	
3	1	5.681	13.876	37.033	4.986	
4	1	4.824	19.063	52.299	-20.679	
5	74	-1.585	-0.915	1.004	1.692	
6	939	-1.558	-0.987	0.345	0.567	
7	786	-1.282	-0.852	0.100	-0.551	
8	227	1.973	0.234	-0.467	-1.486	
9	501	2.229	0.269	-0.222	-0.316	
10	117	1.852	0.009622	-0.00249	0.769	
11	77	6.308	2.508	-0.539	1.549	
12	187	6.476	2.550	-1.081	0.366	

Riepilogo cluster						
Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	
1	82	-10.920	8.451	-2.554	-0.241	
2	6	-0.740	12.018	17.492	4.744	
3	1	5.681	13.876	37.033	4.986	
4	1	4.824	19.063	52.299	-20.679	
5	74	-1.585	-0.915	1.004	1.692	
6	939	-1.558	-0.987	0.345	0.567	
7	786	-1.282	-0.852	0.100	-0.551	
8	227	1.973	0.234	-0.467	-1.486	
9	501	2.229	0.269	-0.222	-0.316	
10	117	1.852	0.009622	-0.00249	0.769	
11	77	6.308	2.508	-0.539	1.549	
12	187	6.476	2.550	-1.081	0.366	

Figura 1.8: Numero di cluster e dimensione per diversi valori

La devianza persa durante la clusterizzazione varia in relazione al numero di cluster scelti. Si possono racchiudere le informazioni in un'unica tabella:

6 Cluster	8 Cluster	10 Cluster	12 Cluster	14 Cluster	16 Cluster
26%	17%	16%	15%	14.5%	14%

1.3.2 5 Componenti Principali

Utilizzando le prime quattro PC si ha:

Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	Principale 5
1	82	-10.920	8.451	-2.554	-0.241	0.200
2	7	0.178	12.283	20.284	4.779	-8.536
3	1	4.824	19.063	52.299	-20.679	29.023
4	1013	-1.560	-0.981	0.393	0.650	0.328
5	786	-1.282	-0.852	0.100	-0.551	-0.265
6	227	1.973	0.234	-0.467	-1.486	-0.707
7	618	2.158	0.220	-0.181	-0.110	-0.252
8	264	6.427	2.564	-0.923	0.711	29.023

Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	Principale 5
1	82	-10.920	8.451	-2.554	-0.241	0.200
2	6	-0.740	12.018	17.492	4.744	-7.536
3	1	5.681	13.876	37.033	4.986	-14.538
4	1	4.824	19.063	52.299	-20.679	29.023
5	74	-1.585	-0.915	1.004	1.692	0.328
6	939	-1.558	-0.987	0.345	0.567	0.328
7	786	-1.282	-0.852	0.100	-0.551	-0.265
8	227	1.973	0.234	-0.467	-1.486	-0.707
9	501	2.229	0.269	-0.222	-0.316	-0.252
10	117	1.852	0.009622	-0.00249	0.769	0.473
11	77	6.308	2.508	-0.539	1.549	0.328
12	187	6.476	2.550	-1.081	0.366	0.473

Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	Principale 5
1	82	-10.920	8.451	-2.554	-0.241	0.200
2	6	-0.740	12.018	17.492	4.744	-7.536
3	1	5.681	13.876	37.033	4.986	-14.538
4	1	4.824	19.063	52.299	-20.679	29.023
5	74	-1.585	-0.915	1.004	1.692	0.328
6	939	-1.558	-0.987	0.345	0.567	0.328
7	786	-1.282	-0.852	0.100	-0.551	-0.265
8	227	1.973	0.234	-0.467	-1.486	-0.707
9	501	2.229	0.269	-0.222	-0.316	-0.252
10	117	1.852	0.009622	-0.00249	0.769	0.473
11	77	6.308	2.508	-0.539	1.549	0.328
12	187	6.476	2.550	-1.081	0.366	0.473

Medie dei cluster						
Cluster	Conteggio	Principale 1	Principale 2	Principale 3	Principale 4	Principale 5
1	82	-10.920	8.451	-2.554	-0.241	0.200
2	6	-0.740	12.018	17.492	4.744	-7.536
3	1	5.681	13.876	37.033	4.986	-14.538
4	1	4.824	19.063	52.299	-20.679	29.023
5	74	-1.585	-0.915	1.004	1.692	0.328
6	939	-1.558	-0.987	0.345	0.567	0.328
7	786	-1.282	-0.852	0.100	-0.551	-0.265
8	227	1.973	0.234	-0.467	-1.486	-0.707
9	501	2.229	0.269	-0.222	-0.316	-0.252
10	117	1.852	0.009622	-0.00249	0.769	0.473
11	77	6.308	2.508	-0.539	1.549	0.328
12	187	6.476	2.550	-1.081	0.366	0.473

Figura 1.9: Numero di cluster e dimensione per diversi valori

La devianza persa durante la clusterizzazione varia in relazione al numero di cluster scelti. Si possono racchiudere le informazioni in un'unica tabella:

6 Cluster	8 Cluster	10 Cluster	12 Cluster	14 Cluster	16 Cluster
25.5%	16%	15%	12%	11%	10%

1.3.3 6 Componenti Principali

Utilizzando le prime quattro PC si ha:

Riepilogo cluster							
Medie dei cluster							
Cluster	Conteggio	Principale1	Principale2	Principale3	Principale4	Principale5	Principale6
1	82	-10.920	8.451	-2.554	-0.241	0.200	0.291
2	7	0.178	12.283	20.284	4.779	-8.536	-0.748
3	863	2.980	0.7229	-0.369	-0.444	-0.2988	0.419
4	116	6.040	2.283	-0.919	1.019	1.102	-2.367
5	1929	-1.241	-0.7444	0.238	0.141	0.0733	-0.037
6	1	4.824	19.083	52.299	-20.679	29.023	3.044

Riepilogo cluster							
Medie dei cluster							
Cluster	Conteggio	Principale1	Principale2	Principale3	Principale4	Principale5	Principale6
1	82	-10.920	8.451	-2.554	-0.241	0.200	0.291
2	7	0.178	12.283	20.284	4.779	-8.536	-0.748
3	132	6.518	2.801	-0.862	0.505	0.121	0.639
4	711	2.239	0.279	-0.209	-0.466	-0.360	0.373
5	116	6.040	2.283	-0.919	1.019	1.102	-2.367
6	1126	-1.207	-0.864	0.325	0.058	0.338	0.252
7	803	-1.290	-0.847	0.082	-0.584	-0.279	-0.489
8	1	4.824	19.083	52.299	-20.679	29.023	3.044

Riepilogo cluster							
Medie dei cluster							
Cluster	Conteggio	Principale1	Principale2	Principale3	Principale4	Principale5	Principale6
1	82	-10.920	8.451	-2.554	-0.241	0.200	0.291
2	6	-0.740	12.018	17.482	4.744	-7.536	-0.298
3	1	5.681	13.876	37.033	4.986	-14.538	-3.445
4	15	6.154	4.666	2.755	1.375	-2.095	-0.750
5	137	6.558	2.597	-1.070	0.409	0.364	0.789
6	497	2.190	0.258	-0.203	-0.310	-0.258	0.548
7	214	2.354	0.326	-0.533	-1.427	-0.695	-0.034
8	116	6.040	2.283	-0.919	1.019	1.102	-2.367
9	639	-1.718	-0.946	0.329	0.639	0.363	-0.111
10	487	-0.536	-0.626	0.330	0.682	0.281	0.727
11	803	-1.290	-0.847	0.082	-0.584	-0.279	-0.489
12	1	4.824	19.083	52.299	-20.679	29.023	3.044

Riepilogo cluster							
Medie dei cluster							
Cluster	Conteggio	Principale1	Principale2	Principale3	Principale4	Principale5	Principale6
1	82	-10.920	8.451	-2.554	-0.241	0.200	0.291
2	6	-0.740	12.018	17.482	4.744	-7.536	-0.298
3	1	5.681	13.876	37.033	4.986	-14.538	-3.445
4	15	6.154	4.666	2.755	1.375	-2.095	-0.750
5	137	6.558	2.597	-1.070	0.409	0.364	0.789
6	497	2.190	0.258	-0.203	-0.310	-0.258	0.548
7	214	2.354	0.326	-0.533	-1.427	-0.695	-0.034
8	116	6.040	2.283	-0.919	1.019	1.102	-2.367
9	639	-1.718	-0.946	0.329	0.639	0.363	-0.111
10	487	-0.536	-0.626	0.330	0.682	0.281	0.727
11	803	-1.290	-0.847	0.082	-0.584	-0.279	-0.489
12	1	4.824	19.083	52.299	-20.679	29.023	3.044

Figura 1.10: Numero di cluster e dimensione per diversi valori

6 Cluster	8 Cluster	10 Cluster	12 Cluster	14 Cluster	16 Cluster
22%	15%	13%	12%	11%	9%

1.3.4 Interpretazione

All'inizio dell'analisi sono state effettuate delle ipotesi che hanno trovato riscontro nella procedura di clustering:

1. La fase iniziale del sistema (descritto dalle prime righe) trova riscontro con il cluster numero uno qualunque siano le componenti principali e qualunque sia il numero di cluster scelto. Questo quindi prova l'ipotesi definita inizialmente
2. L'ultimo cluster contiene sempre un elemento singolo. Questo accade a causa del fatto che è stato identificato un picco nella fase iniziale delle misure. Esso inoltre è stato definito grazie all'outlier nel parametro Slab che non è stato eliminato, di conseguenza il picco viene racchiuso in un unico cluster in tutte le situazioni.

In conclusione si può costruire una tabella che racchiude le informazioni riguardanti le PCA e la clusterizzazione in termini di percentuale di devianza persa.

	6 Cluster	8 Cluster	10 Cluster	12 Cluster	14 Cluster	16 Cluster
4 PC	26%	17%	16%	15%	14.5%	14%
5 PC	25.5%	16%	15%	12%	11%	10%
6 PC	22%	15%	13%	12%	11%	9%

1.4 Workload Sintetico

Come già deciso nel paragrafo precedente, sono state scelte 5 PC. A tal proposito per non perdere troppa varianza, e non sfociare in un numero di cluster molto elevato, si è scelto di considerare 10 cluster. In tal caso la perdita di devianza con PCA e Cluster è di circa del 15%.