RELAZIONE FORZA LAVORO

I dati in analisi riguardano le regione italiane e le caratteristiche analizzate sono le seguenti:
regione, zona d'Italia, numero abitanti, cerca, percentuale di forza lavoro, percentuali laureati, quanti abitanti hanno il diploma(%), lavoratori impiegati nel terziario.

Analisi Univariata/Bivariata:

Variabile	N	Media	Mediana	Coeff di variazione	Minimo	Massimo	Range	Quartile inferiore	Quartile superiore
abitanti	20	2878167.70	1866210.50	78.61	119610.00	8988951.00	8869341.00	1054467.50	4380298.50
cerca	20	12.45	9.90	55.95	3.90	25.50	21.60	6.30	17.90
fdl	20	40.33	40.10	10.29	33.80	46.60	12.80	36.50	43.90
laurea	20	9.20	9.00	18.41	6.10	14.00	7.90	7.95	10.20
diploma	20	29.26	29.15	10.49	22.80	36.50	13.70	27.75	30.80
terziario	20	61.98	61.75	10.28	53.20	75.70	22.50	56.25	66.65

Da una prima analisi descrittiva dalla tabella si osserva che l'unica caratteristica dove la media si discosta di molto dalla mediana sono il numero di abitanti (media: 2878167; mediana=1866210). Questo poiché ci sono regioni italiane con un alto numero di abitanti e altre regioni con un basso numero.

Infatti la Lombardia è la regione con il più alto numero di abitanti, quasi 9 milioni, mentre la seconda regione, ovvero la Campania ne ha quasi 6. Mentre quella con meno abitanti è la Valle d'Aosta con 119610 abitanti.

La percentuale di persone che cercano lavoro è in media del 12%, le percentuali più alte di abitanti che cercano lavoro sono in Campania e Calabria del 25%. Invece la regione con la più bassa percentuale è il Trentino col 4%.

Nel caso della forza lavoro, in percentuale, la media è del 40% ed è uguale alla mediana della distribuzione. La percentuale più elevata è nella Valle d'Aosta e Emilia Romagna, quella più bassa nella Sicilia e la Calabria.

Le percentuali media di diplomati e di laureati in Italia sono rispettivamente di circa il 30% e del 9%. Il Lazio ha la percentuale più elevata in Italia per entrambe i titoli di studio (36,5% di diplomati e 14% di laureati. Al contrario il Trentino ha le percentuali più basse (22,8% e 6,10%).

La percentuale media di lavoratori impiegati nel terziario ammonta al 61% circa e le due regioni ad avere tale percentuale superiore al 73% sono la Liguria e il Lazio. Invece in Veneto e nelle Marche la percentuale di impiegati nel terziario è al di sotto del 54%.

Tra i coefficienti di variazione, il più alto è quello del numero degli abitanti, un po' più basso quello della percentuale delle persone che cercano lavoro, mentre le altre caratteristiche hanno un coefficiente basso. Quindi tra le regioni italiane c'è omogeneità.

Analisi delle correlazioni:

Coefficienti di correlazione di Pearson, N = 20

	abitanti	cerca	fdl	laurea	diploma	terziario
abitanti	1.00000	0.07562	-0.02045	0.40130	0.05275	-0.02511
cerca	0.07562	1.00000	-0.92750	0.30011	0.12038	0.34004
fdl	-0.02045	-0.92750	1.00000	-0.33547	-0.18395	-0.25473
laurea	0.40130	0.30011	-0.33547	1.00000	0.78681	0.55909
diploma	0.05275	0.12038	-0.18395	0.78681	1.00000	0.45600
terziario	-0.02511	0.34004	-0.25473	0.55909	0.45600	1.00000

Tra la percentuale di persone in cerca di lavoro e quella di forza lavoro c'è una correlazione negativa quasi perfetta, quindi all'aumentare di una delle due, diminuisce l'altra in media. Anche Tra la percentuale di diplomati e laureati c'è una forte correlazione ma in questo caso positiva, quindi all'aumentare di una aumenta in media anche l'altra. Tutti gli altri legami invece sono trascurabili.

Analisi in Componenti Principali:

Con l'obiettivo di individuare eventuali relazioni lineare si procedere con un'analisi in componenti principali, che permette di sintetizzare le caratteristiche delle regioni italiane in alcuni indicatori.
Poiché le variabili sono espresse in diversa unità di misura e poiché presentano una, seppur leggera, differenza di variabilità, si procede con una standardizzazione dei dati, per una più adeguata applicazione di tale metodologia.

Per la regola dell'autovalore > 1 si scelgono 3 componenti principali.

Autovalori della matrice di correlazione: Totale = 6 Media = 1

	Autovalore	Differenza	Proporzione	Cumulativa
1	2.75220926	1.26895654	0.4587	0.4587
2	1.48325272	0.42711255	0.2472	0.7059
3	1.05614017	0.50940296	0.1760	0.8819
4	0.54673721	0.43703861	0.0911	0.9731
5	0.10969859	0.05773654	0.0183	0.9913
6	0.05196206		0.0087	1.0000

Le 3 componenti spiegano l'88% della variabilità totale e rappresentano bene le caratteristiche delle regioni

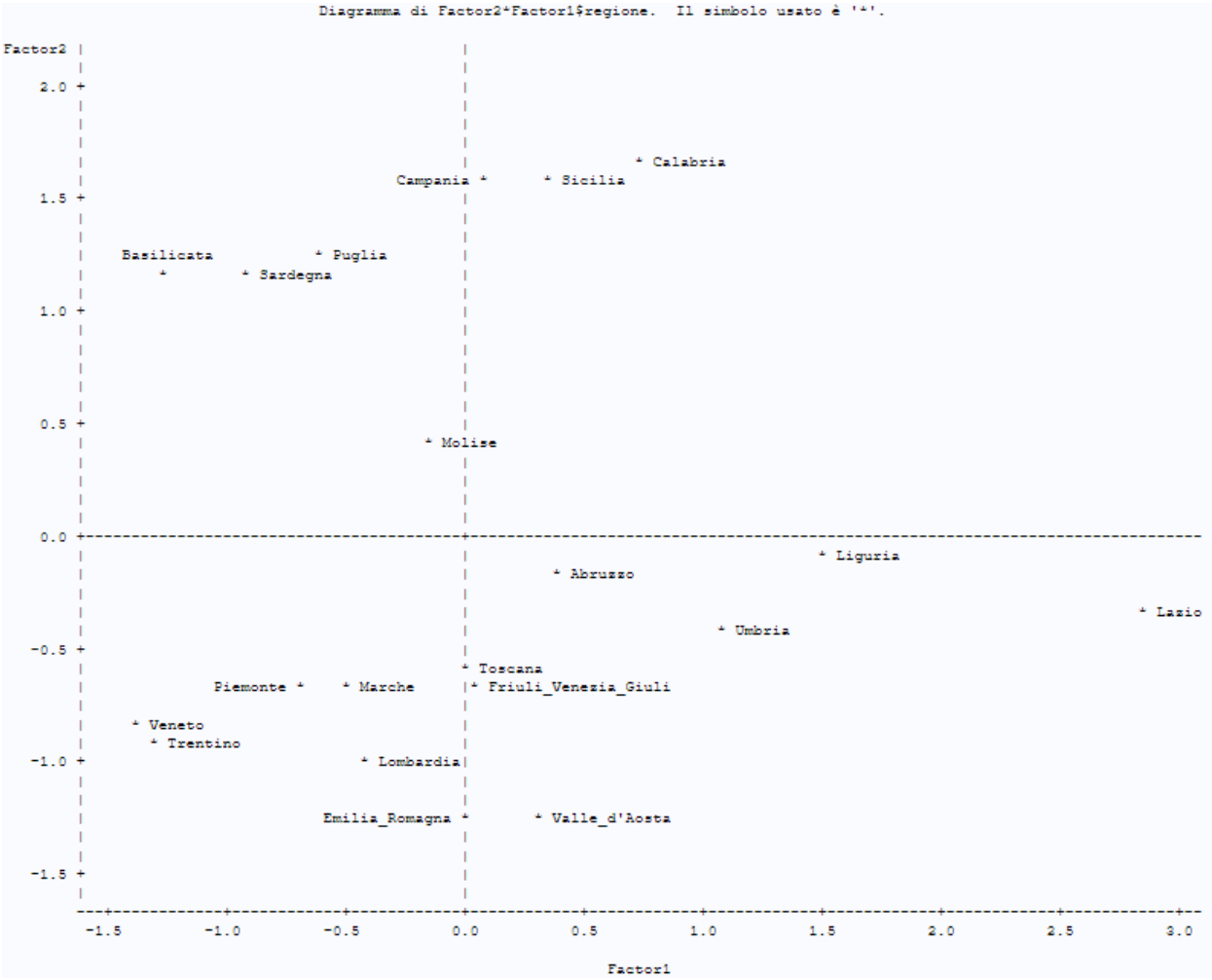
Stime di comunanza finali: Totale = 5.291602					
abitanti	**cerca**	**fdl**	**laurea**	**diploma**	**terziario**
0.96415233	0.96939884	0.94900947	0.93184440	0.81866514	0.65853196

Il numero degli abitanti, le percentuali delle persone che cercano lavoro, della forza lavoro e dei laureati sono quasi perfettamente spiegate (93%). La percentuale dei diplomati è spiegata all'82%, mentre la variabilità della percentuale delle persone che lavoro del settore terziario è la meno spiegata al 66% ma comunque un valore elevato.

Per facilitare l'interpretazione delle 3 componenti principali si fa una rotazione con il metodo VARIMAX.

Pattern fattoriale ruotato			
	Factor1	**Factor2**	**Factor3**
abitanti	0.05156	0.02476	0.98025
cerca	0.13064	0.97511	0.03869
fdl	-0.14772	-0.96279	-0.01473
laurea	0.85795	0.19158	0.39884
diploma	0.90326	-0.00404	0.05270
terziario	0.75475	0.24280	-0.17302

La prima componente è indicatore crescente della percentuale di laureati, diplomati e dei lavoratori del terziario, indica il lv di istruzione e impiego nel terziario. La seconda componente è indicatore della percentuale di abitanti che cercano lavoro ed è correlata inversamente con la percentuale della forza lavoro. Quest'ultima può essere definita come indicatore della disoccupazione.
Infine la terza componente è indicatore crescente del numero degli abitanti.



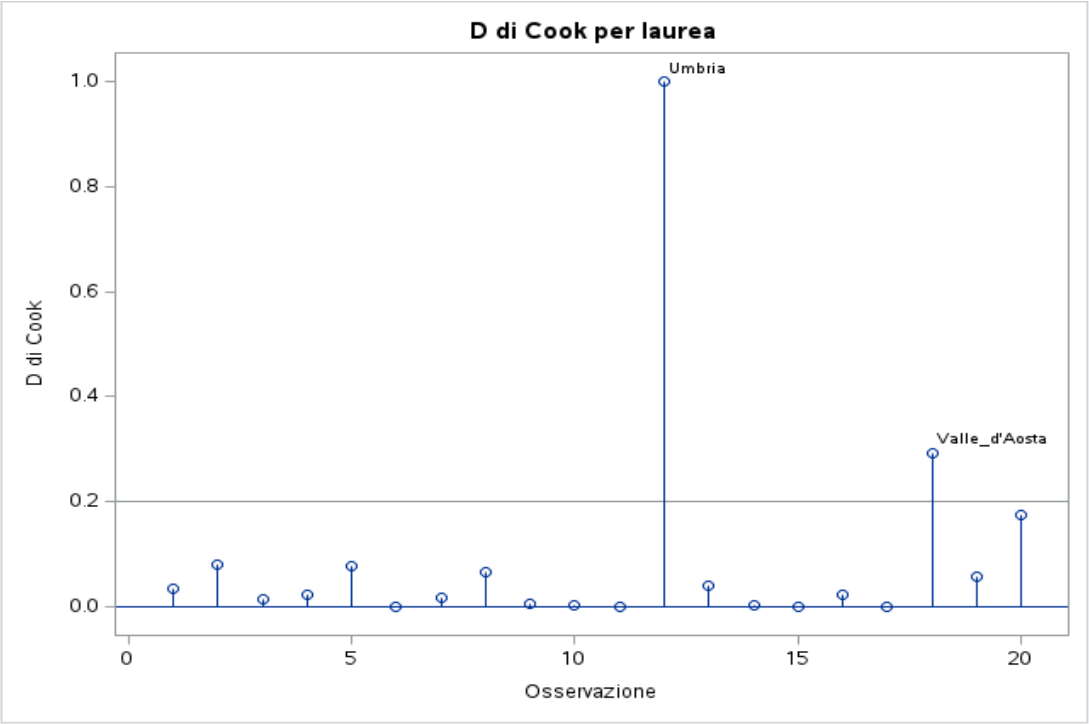
Il diagramma, delle prime due componenti, rappresenta le regioni italiani sul piano principale. Guardando da sinistra a destra si passa dalle regioni con il più basso lv di istruzione e impiego nel terziario (come Basilicata e Veneto) a quello più alto (Lazio).

Invece dal basso verso l'alto si hanno le regioni con la disoccupazione più bassa come l'Emilia Romagna e la Valle d'Aosta a quella più alta come la Campania, la Sicilia e la Calabria.
Le regioni più in prossimità dell'origine sono il Molise e l'Abruzzo: queste regioni presentano un livello d'istruzione/di sviluppo del terziario e una disoccupazione simili a quelli medi nazionali.

Dalle contribuzioni relative le unità statistiche poco rappresentate sono l'Abruzzo (26%) e il Molise (47%), l'Umbria e le Marche sono sopra il 50%, mentre tutte le altre sopra l'80%. Il Lazio, il Veneto e la Lombardia sono quasi perfettamente rappresentate (quasi 100%).

Regressione:

Con l'obiettivo di studiare se la percentuale di laureati è influenzata dalle altre caratteristiche si costruisce un modello di regressione.
Inizialmente sono state selezionate come determinante con il metodo Stepwise alcune caratteristiche: gli abitanti, la percentuale di diplomati e quella dei lavoratori del terziario. Di seguito è stata esclusa la regione Umbria in quanto presentava un valore elevato, al di sopra della soglia, nell'indice di D-Cook, il quale la individuava come regione influente sul modello.



A questo punto dal test t risulta che la variabile terziario non abbia più un contributo significativo nello spiegare la percentuale di laureati e quindi la eliminiamo dalla regressione.

Stime dei parametri

Variabile	DF	Stima dei parametri	Errore standard	Valore t	Pr > |t|	Inflazione varianza	Limiti di confidenza al 95%	
Intercept	1	-8.08789	1.68049	-4.81	0.0002	0	-11.66976	-4.50601
abitanti	1	2.14958E-7	6.573772E-8	3.27	0.0052	1.04802	7.484141E-8	3.550747E-7
diploma	1	0.47095	0.06208	7.59	<.0001	1.40727	0.33864	0.60326
terziario	1	0.04885	0.02607	1.87	0.0806	1.36590	-0.00672	0.10441

Il modello definitivo è stato quindi stimato selezionando come determinanti il numero di abitanti, la percentuale di diplomati e considerando l'intercetta.

Essendo tutti i VIF < 10 non si evidenziano problemi di multicollinearità.

Stime dei parametri								
Variabile	**DF**	**Stima dei parametri**	**Errore standard**	**Valore t**	**Pr > |t|**	**Inflazione varianza**	**Limiti di confidenza al 95%**	
Intercept	1	-6.75532	1.63769	-4.12	0.0008	0	-10.22707	-3.28358
abitanti	1	1.991563E-7	7.012346E-8	2.84	0.0118	1.03077	5.050117E-8	3.478113E-7
diploma	1	0.53111	0.05714	9.29	<.0001	1.03077	0.40997	0.65225

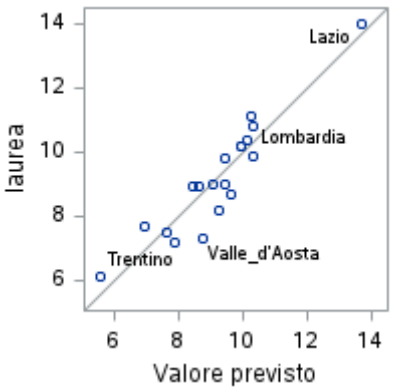
Radice MSE	0.68361	**R-quadro**	0.8697
Media dip.	9.20526	**R-quadro corr**	0.8534
Coeff var	7.42633		

Tale modello è caratterizzato da un'elevata bontà di adattamento, in quanto il coefficiente di determinazione R^2 è pari a 0.87.

Le assunzioni di normalità, media nulla e varianza costante, riguardo la distribuzione degli errori risultano verosimili, sulla base dell'analisi del q-q plot del grafico di dispersione dei residui. Gli errori risultano casuali e non sistematici.

All'aumentare di un punto percentuale di diplomati, la percentuale di laureati aumenta in media di circa metà punto percentuale (o con probabilità al 95% aumenta tra 0.4 e 0.65 punti percentuali), a parità di numero di abitanti.

Stime dei parametri								
Variabile	**DF**	**Stima dei parametri**	**Errore standard**	**Valore t**	**Pr > |t|**	**Inflazione varianza**	**Limiti di confidenza al 95%**	
Intercept	1	-6.75532	1.63769	-4.12	0.0008	0	-10.22707	-3.28358
abitanti	1	1.991563E-7	7.012346E-8	2.84	0.0118	1.03077	5.050117E-8	3.478113E-7
diploma	1	0.53111	0.05714	9.29	<.0001	1.03077	0.40997	0.65225



Nel grafico “valore previsto-percentuale di laureati” i punti sono vicini alla retta, quindi i valori osservati sono simili a quelli previsti. Dunque il modello è buono.

La regione che ha la percentuale di laureati prevista più simili a quella reale è la Liguria.

Cluster Analysis:

Si procede con la Cluster Analysis in modo da individuare gruppi di regioni in base alle loro caratteristiche, tali che quelle che appartengono allo stesso gruppo siano simili, mentre quelle di gruppi diversi siano eterogenei.

Poiché le variabili sono espresse in unità di misura differente si effettua la standardizzazione dei dati, per una più adeguata analisi.

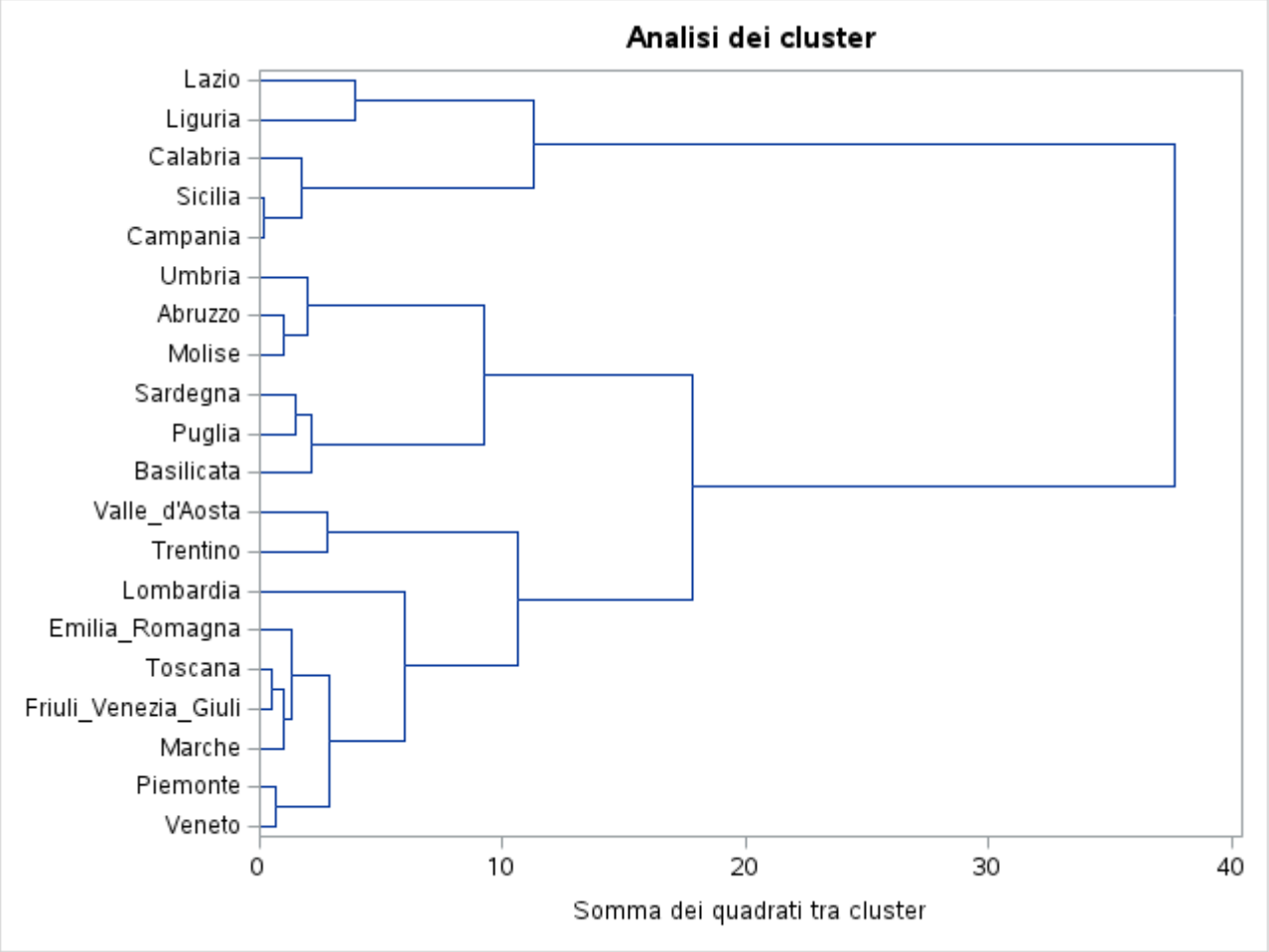
Si effettua un’analisi gerarchica aggregativa scegliendo il metodo di WARD, che permette la formazione di gruppi di regioni con una bassa variabilità interna.

Si scelgono 5 gruppi giustificato dall’indice R^2 e dall’indice RMSSTD. L’andamento del primo presenta una brusca diminuzione del suo valore in corrispondenza della partizione successiva in 4 gruppi (da 0.585 in giù) che indica un netto peggioramento della qualità del gruppo. Invece l’RMSSTD assume valore minimo (nell’intervallo tra 2 e 5 gruppi) indicando che il gruppo formato a quel livello della gerarchia ha una bassa variabilità interna.

Cronologia dei cluster

Numero di cluster	Cluster uniti		Freq	Dev std RMS nuovo cluster	R-quadro semiparziale	R-quadro	R-quadro atteso approssimato	Criterio di clusterizzazione cubica	Statistica pseudo F	Pseudo t-quadro	Between Cluster Sum of Squares	Legame
19	Campania	Sicilia	2	0.1739	0.0016	.998	.	.	34.8	.	0.1815	
18	Friuli_Venezia_Giuli	Toscana	2	0.3016	0.0048	.994	.	.	18.3	.	0.5457	
17	Veneto	Piemonte	2	0.3313	0.0058	.988	.	.	15.2	.	0.6587	
16	Marche	CL18	3	0.3542	0.0084	.979	.	.	12.7	1.8	0.9596	
15	Molise	Abruzzo	2	0.4043	0.0086	.971	.	.	11.9	.	0.9807	
14	CL16	Emilia_Romagna	4	0.3942	0.0113	.959	.	.	10.9	1.7	1.2915	
13	Puglia	Sardegna	2	0.4946	0.0129	.947	.	.	10.3	.	1.4677	
12	CL19	Calabria	3	0.3969	0.0150	.932	.	.	9.9	9.4	1.7084	
11	CL15	Umbria	3	0.4936	0.0170	.915	.	.	9.6	2.0	1.9428	
10	Basilicata	CL13	3	0.5483	0.0188	.896	.	.	9.6	1.5	2.1403	
9	Trentino	Valle_d'Aosta	2	0.6824	0.0245	.871	.	.	9.3	.	2.7943	
8	CL17	CL14	6	0.4594	0.0252	.846	.	.	9.4	3.3	2.8768	
7	Liguria	Lazio	2	0.8127	0.0348	.811	.	.	9.3	.	3.963	
6	CL8	Lombardia	7	0.5839	0.0521	.759	.	.	8.8	4.7	5.9421	
5	CL10	CL11	6	0.7245	0.0808	.678	.	.	7.9	5.6	9.2145	
4	CL6	CL9	9	0.7319	0.0934	.585	.683	-2.3	7.5	4.9	10.644	
3	CL12	CL7	5	0.8452	0.0991	.486	.579	-1.7	8.0	5.8	11.294	
2	CL4	CL5	15	0.8399	0.1562	.330	.403	-1.1	8.9	5.6	17.803	
1	CL2	CL3	20	1.0000	0.3298	.000	.000	0.00	.	8.9	37.592	

Analizzando il dendogramma (e la cronologia dei cluster riportati poco più su), si osserva che le prime regioni ad aggregarsi sono la Campania con la Sicilia, il Friuli Venezia Giulia con la Toscana, il Piemonte con il Veneto, ovvero quelle che presentano valori più simili tra loro. L'ultima invece ad aggregarsi è la Lombardia: ciò vuol dire che ha valori diversi dalle altre regioni rispetto alle caratteristiche analizzate.



CLUSTER=1

Variabile	Media	Coeff di variazione	N
abitanti	4325319.33	45.8336578	3
cerca	24.7333333	4.3608328	3
fdl	34.2000000	2.0257904	3
laurea	10.4666667	2.9188380	3
diploma	30.1000000	3.5935727	3
terziario	67.0333333	1.2861777	3

Il primo gruppo è caratterizzato dalla più alta percentuale media di persone in cerca di lavoro.

CLUSTER=2

Variabile	Media	Coeff di variazione	N
abitanti	3979926.57	64.7472686	7
cerca	6.8000000	17.3587561	7
fdl	43.7000000	3.1375994	7
laurea	8.8571429	10.5078304	7
diploma	28.4285714	4.2440749	7
terziario	56.7714286	5.7427162	7

Il secondo gruppo, con 7 regioni, è il gruppo con la minor percentuale media di persone impiegate nel terziario.

CLUSTER=3

Variabile	Media	Coeff di variazione	N
abitanti	1466579.17	93.4183214	6
cerca	15.5166667	29.6383030	6
fdl	37.5833333	5.9601283	6
laurea	8.6833333	11.3869898	6
diploma	29.3500000	12.9000307	6
terziario	59.8833333	6.3740995	6

Il terzo gruppo presenta tutte le caratteristiche medie comprese tra quelle degli altri gruppi.

CLUSTER=4

Variabile	Media	Coeff di variazione	N
abitanti	521945.50	109.0129756	2
cerca	4.7500000	25.3069795	2
fdl	46.2000000	1.2244273	2
laurea	6.7000000	12.6645991	2
diploma	25.9500000	17.1667542	2
terziario	66.1000000	6.6324691	2

Il quarto gruppo, costituito da Valle d'Aosta e dal Trentino, presenta il valore maggiore di percentuale media di forza lavoro e il valore minimo della media degli abitanti, dei laureati e dei diplomati.

CLUSTER=5

Variabile	Media	Coeff di variazione	N
abitanti	3442272.00	73.9686586	2
cerca	12.2500000	6.3495303	2
fdl	40.0500000	0.5296680	2
laurea	12.5500000	16.3395192	2
diploma	33.9500000	10.6222226	2
terziario	74.7500000	1.7973283	2

Infine il quinto gruppo costituito dalla Liguria e dal Lazio presenta il più alto valore della percentuale media dei laureati, diplomati e impiegati nel settore terziario.