IL MERCATO ITALIANO: DIVERSIFICAZIONE REGIONALE DEI CONSUMI ALIMENTARI TRAMITE L'IMPIEGO DI ACP E ANALISI DEI CLUSTER

Dipartimento di Scienze Economico-Aziendali e Diritto per l'Economia

Corso di LM in Scienze Economico-Aziendali (Management)

Anno accademico 2018/19

Bonura Caterina

Matricola: 848818

Prof. Alessandro Zini

Indice

1. Introduzione

L'elaborato di seguito proposto mira ad analizzare i consumi alimentari delle diverse regioni italiane e a verificare se esistano similarità o, di contro, dissimilarità tra alcune di esse.

L'analisi è stata condotta con un approccio scientifico, presupponendo che il lettore disponesse delle conoscenze necessarie per la comprensione della stessa.

Ai fini dell'investigazione si è reso necessario individuare un dataset solido e di efficace lettura che consentisse di giungere ad esiti significativi; pertanto, quest'ultimo è stato estratto dalla banca dati Istat[1], nella sezione Salute e Sanità, alla voce Stili di Vita e Fattori di Rischio.

Trattasi di dati collezionati dall'Istat con cadenza annuale nell'ambito della più ampia indagine campionaria soprannominata "Multiscopo sulle Famiglie: Aspetti della vita quotidiana" che rileva le informazioni fondamentali circa gli individui e le famiglie.

La tesina è stata realizzata attraverso l'impiego del software SPSS, ricorrendo a due metodi: in primo luogo l'Analisi delle Componenti Principali (ACP) e successivamente l'Analisi Cluster.

In conclusione, si è proceduto al commento puntuale dell'output restituito.

[1] http://dati.istat.it/

2. Descrizione del dataset

Il dataset impiegato è l'esito di un'attività di combinazione di diversi micro-dataset reperibili presso la banca dati dell'Istat denominata I.Stat, nella sottovoce Stili Alimentari e Consumi di Cibi.

Il disegno di campionamento impiegato dall'Istat è di tipo complesso e si avvale di due differenti schemi: all'interno delle regioni, i comuni sono suddivisi in due sottoinsiemi sulla base della popolazione residente, giungendo a distinguerli in due sottoclassi: l'insieme dei comuni Auto rappresentativi (Ar) costituito dai comuni di maggiore dimensione demografica e l'insieme dei comuni Non auto rappresentativi (Nar), costituito dai rimanenti comuni. Nell'ambito dei comuni Nar viene adottato un disegno a due stadi con stratificazione delle unità primarie. I comuni vengono selezionati con probabilità proporzionali alla loro dimensione demografica e senza reimmissione.[2]

Nella varietà degli anni disponibili, si è scelto come anno di riferimento il 2016, in quanto esso rappresentava il miglior compromesso tra dati che fossero aggiornati e al tempo stesso completi (ad esempio i dati del 2017 presentavano voci con alcune mancanze).

Il dataset che si è venuto a definire è il seguente (riportato nella pagina successiva):

- n = 20, coincidenti con le regioni
- p = 15, variabili investigate, nello specifico:

salumi almeno qualche volta alla settimana
carni bianche almeno qualche volta alla settimana
carni bovine almeno qualche volta alla settimana
carni maiale almeno qualche volta alla settimana
uova almeno qualche volta alla settimana
pesce almeno qualche volta alla settimana
verdure almeno una volta
ortaggi almeno una volta
frutta almeno una volta
pane, pasta almeno una volta al giorno
latte almeno una volta al giorno
formaggio almeno una volta al giorno
legumi almeno qualche volta alla settimana
snack almeno qualche volta alla settimana
dolci almeno qualche volta alla settimana

[2] http://dati.istat.it/OECDStat_Metadata/ShowMetadata

Territorio: regioni	Variabili rilevate: salumi almeno qualche volta alla settimana	carni bianche almeno qualche volta alla settimana	carni bovine almeno qualche volta alla settimana	carni maiale almeno qualche volta alla settimana	uova almeno qualche volta alla settimana	pesce almeno qualche volta alla settimana	verdure almeno una volta	ortaggi almeno una volta	frutta almeno una volta	pane, pasta almeno una volta al giorno	latte almeno una volta al giorno	formaggio almeno una volta al giorno	legumi almeno qualche volta alla settimana	snack almeno qualche volta alla settimana	dolci almeno qualche volta alla settimana
Piemonte	55,4	82,2	64	38	63,4	55,5	60,1	53,1	76,5	77,3	50,3	27,9	45,3	25,5	50,7
Valle d'Aosta / Vallée d'Aoste	56,4	80,2	60,8	34,5	62,9	52	57,7	54,3	68,7	79,6	49,1	35,6	39,7	20,2	46,8
Liguria	55,3	78,1	54,6	34,1	62,3	57,9	51,4	47,6	76	76,1	56,8	22,9	44	21,4	50,2
Lombardia	58,8	79	57,2	35,1	54,5	58	54	46,2	71	75,1	51	27,2	41,7	30,8	54,7
Trentino Alto Adige / Südtirol	60	63,1	45,8	35,3	56,4	40,7	58,9	49,2	66,2	73,5	55,4	35	34,9	21,5	46,6
Veneto	51,7	79,4	58,2	39,9	54,4	55,7	59,1	50,6	69,6	76,7	50,7	24,7	39,9	28,4	54,1
Friuli-Venezia Giulia	56,3	79,3	53,5	40,5	55,1	54,3	63,3	46,5	73,4	76,8	52,4	26,2	40,9	24,1	53,5
Emilia-Romagna	61,3	81,3	58,3	48,9	56,8	57	63,6	58,3	77,2	84,1	49,9	20,2	48,4	27,2	57,2
Toscana	57,9	84	64,4	47,1	56,6	57,7	55,6	50,7	75	83,7	56,3	18,7	50,5	21,4	47,1
Umbria	63,3	86,2	68,2	58,1	59,9	61,3	57,9	51	79,8	87,3	56,7	18,4	59,6	21,2	48,4
Marche	61,9	81,4	61,9	45,2	55	65,4	57,3	47,6	75,3	85,7	49,9	15,9	49,3	21,9	49,3
Lazio	47,2	78,2	63,9	39,8	60,4	65,4	61,3	52,8	75	79,4	60,7	15	52,3	22,3	41,3
Abruzzo	63,3	84,3	60,4	45,3	64,5	61	43,1	39,1	76	86,2	52,4	12,4	62,5	26,6	49
Molise	67,5	85,5	68,4	57,7	68,1	64,7	41,3	36,2	73,3	81,1	54,6	15,7	70,4	29,8	48,9
Campania	55,7	80,4	64,9	49,3	60,3	70,5	41,6	35,9	71,1	78,2	50,6	8,2	72,3	30,8	39,8
Puglia	56,1	74,4	58,9	36,9	53,1	61	36,9	34,6	77,1	82	58,5	14	63,6	26,1	43,7
Basilicata	64,7	81,9	64,6	52,2	65,2	59,8	40,1	31,3	73,8	88,4	49,9	14,6	70,4	29,3	41,3
Calabria	57,2	77	59,7	45,1	63,8	68,4	42,6	38,8	75,3	85,4	51,6	17,4	70,8	25,8	40,1
Sicilia	56,7	79,6	69,4	46	66,6	62	43,6	36,8	80,4	86,6	51,1	19,1	55,9	29,5	42,2
Sardegna	52	78,4	62,6	52	55,4	56,8	51,9	45,2	79,9	76,1	52,8	27,1	41,5	25,9	52

I dati riportati nelle celle sono delle frequenze e la misura delle stesse è data da 100 persone con le stesse caratteristiche (persone dai 3 anni in su per consumo di alcuni cibi).

3. Analisi delle Componenti Principali (ACP)

Cenni Teorici

L'analisi delle Componenti Principali è un metodo di analisi multivariata che consente una riduzione delle variabili investigate.

A partire dalle variabili originarie si definiscono delle variabili nuove denominate Componenti Principali, che sono combinazioni lineari delle prime.

Ciascuna di esse consente di raggruppare all'interno di sé tutta una serie di variabili, ponderandole tramite l'assegnazione di pesi diversificati.

Il pregio di tale metodo è quello di ridurre la numerosità delle variabili e di evitare l'indesiderata duplicazione di variabili dall'andamento simile.

La scelta del numero di Componenti Principali è effettuata sulla base della Varianza Cumulativa da queste spiegata, che non deve essere inferiore ad un certo livello soglia.

Tale scelta risponde ad un compromesso; da un lato la volontà di estrarre il minor numero di componenti principali, dall'altro la volontà di conservare quanta più significatività possibile, ossia senza alterare il dato disaggregato di partenza.

Applicazione pratica in SPSS per passaggi

PASSAGGIO 1: MATRICE DI CORRELAZIONE

Attraverso la matrice di correlazione si valuta la desiderabilità di effettuare l'Analisi delle Componenti Principali.

Essa consente di osservare se tra le variabili esista un certo grado di correlazione.

Tale correlazione può essere positiva o negativa, a seconda che le due variabili investigate si modifichino nello stesso verso o in verso opposto, giungendo agli estremi ad assumere i valori rispettivamente di 1 e -1.

Dalla matrice (riportata alla pagina seguente) si osserva come esista un certo grado di correlazione tra le diverse tipologie di consumi alimentari; tuttavia, le dimensioni relativamente ampie di tale matrice rendono difficoltosa una sua efficace analisi.

Pertanto, si procede con il successivo Test di Sfericità di Bartlett, per comprendere se si possa proseguire con l'implementazione del metodo.

		matrice di correlazione														
		salumi almeno qualche volta alla settimana	carni bianche almeno qualche volta alla settimana	carni bovine almeno qualche volta alla settimana	carni maiale almeno qualche volta alla settimana	uova almeno qualche volta alla settimana	pesce almeno qualche volta alla settimana	verdure almeno una volta al giorno	ortaggi almeno una volta al giorno	frutta almeno una volta al giorno	pane, pasta almeno una volta al giorno	latte almeno una volta al giorno	formaggio almeno una volta al giorno	legumi almeno qualche volta alla settimana	snack almeno qualche volta alla settimana	dolci almeno qualche volta alla settimana
Correlazione	salumi almeno qualche volta alla settimana	1,000	0,314	0,144	0,505	0,331	0,030	-0,304	-0,314	0,006	0,512	-0,240	-0,214	0,407	0,167	0,083
	carni bianche almeno qualche volta alla settimana	0,314	1,000	0,755	0,580	0,361	0,538	-0,075	-0,024	0,447	0,521	-0,204	-0,442	0,384	0,180	0,150
	carni bovine almeno qualche volta alla settimana	0,144	0,755	1,000	0,666	0,505	0,664	-0,341	-0,271	0,546	0,583	-0,051	-0,536	0,577	0,280	-0,335
	carni maiale almeno qualche volta alla settimana	0,505	0,580	0,666	1,000	0,311	0,465	-0,286	-0,309	0,454	0,577	-0,059	-0,528	0,596	0,282	-0,113
	uova almeno qualche volta alla settimana	0,331	0,361	0,505	0,311	1,000	0,307	-0,439	-0,369	0,158	0,397	-0,130	-0,236	0,522	0,189	-0,443
	pesce almeno qualche volta alla settimana	0,030	0,538	0,664	0,465	0,307	1,000	-0,506	-0,462	0,427	0,497	0,027	-0,847	0,768	0,368	-0,445
	verdure almeno una volta al giorno	-0,304	-0,075	-0,341	-0,286	-0,439	-0,506	1,000	0,932	-0,175	-0,378	-0,014	0,585	-0,785	-0,526	0,593
	ortaggi almeno una volta al giorno	-0,314	-0,024	-0,271	-0,309	-0,369	-0,462	0,932	1,000	-0,114	-0,327	0,037	0,567	-0,738	-0,584	0,569
	frutta almeno una volta al giorno	0,006	0,447	0,546	0,454	0,158	0,427	-0,175	-0,114	1,000	0,517	0,187	-0,403	0,282	0,021	-0,013
	pane, pasta almeno una volta al giorno	0,512	0,521	0,583	0,577	0,397	0,497	-0,378	-0,327	0,517	1,000	-0,104	-0,606	0,642	0,053	-0,361
	latte almeno una volta al giorno	-0,240	-0,204	-0,051	-0,059	-0,130	0,027	-0,014	0,037	0,187	-0,104	1,000	-0,189	0,057	-0,411	-0,231
	formaggio almeno una volta al giorno	-0,214	-0,442	-0,536	-0,528	-0,236	-0,847	0,585	0,567	-0,403	-0,606	-0,189	1,000	-0,844	-0,370	0,464
	legumi almeno qualche volta alla settimana	0,407	0,384	0,577	0,596	0,522	0,768	-0,785	-0,738	0,282	0,642	0,057	-0,844	1,000	0,440	-0,637
	snack almeno qualche volta alla settimana	0,167	0,180	0,280	0,282	0,189	0,368	-0,526	-0,584	0,021	0,053	-0,411	-0,370	0,440	1,000	-0,022
	dolci almeno qualche volta alla settimana	0,083	0,150	-0,335	-0,113	-0,443	-0,445	0,593	0,569	-0,013	-0,361	-0,231	0,464	-0,637	-0,022	1,000

PASSAGGIO 2: TEST DI SFERICITÀ DI BARTLETT

Tale test opera implicitamente un confronto tra la matrice di correlazione sopra riportata e la matrice di perfetta incorrelazione, vale a dire la matrice identità.

Se il test restituisce un ρ value inferiore a 0,05, si desume che è opportuno procedere con l'analisi.

Test di Bartlett		
Test della sfericità di Bartlett	**Appross. Chi-quadrato**	226,542
	gl	105
	Sign.	0,000

PASSAGGIO 3: TABELLA DI COMUNALITA'

Comunalità		
	Iniziale	Estrazione
salumi almeno qualche volta alla settimana	1,000	0,811
carni bianche almeno qualche volta alla settimana	1,000	0,834
carni bovine almeno qualche volta alla settimana	1,000	0,774
carni maiale almeno qualche volta alla settimana	1,000	0,679
uova almeno qualche volta alla settimana	1,000	0,502
pesce almeno qualche volta alla settimana	1,000	0,841
verdure almeno una volta al giorno	1,000	0,873
ortaggi almeno una volta al giorno	1,000	0,880
frutta almeno una volta al giorno	1,000	0,611
pane, pasta almeno una volta al giorno	1,000	0,763
latte almeno una volta al giorno	1,000	0,696
formaggio almeno una volta al giorno	1,000	0,804
legumi almeno qualche volta alla settimana	1,000	0,928
snack almeno qualche volta alla settimana	1,000	0,893
dolci almeno qualche volta alla settimana	1,000	0,805

Tale tabella riporta le percentuali di varianza di ciascuna variabile spiegate dalle componenti principali estratte, dato che emerge dalla tabella a fianco riportata.

Nove variabili sulle 15 complessive presentano una percentuale di varianza spiegata superiore all'80%, soglia decisamente elevata.

Le rimanenti variabili presentano comunque una soglia di varianza spiegata non inferiore al 60%, ad eccezione di un'unica variabile (consumo di uova) che si attesta al 50%.

PASSAGGIO 4: TABELLA DI VARIANZA SPIEGATA

Varianza totale spiegata						
Componente	Autovalori iniziali			Caricamenti somme dei quadrati di estrazione		
	Totale	% di varianza	% cumulativa	Totale	% di varianza	% cumulativa
1	6,511	43,404	43,404	6,511	43,404	43,404
2	2,235	14,900	58,304	2,235	14,900	58,304
3	1,761	11,742	70,046	1,761	11,742	70,046
4	1,187	7,912	77,958	1,187	7,912	77,958
5	0,958	6,384	84,342			
6	0,643	4,284	88,626			
7	0,569	3,792	92,418			
8	0,414	2,763	95,182			
9	0,251	1,674	96,855			
10	0,189	1,263	98,118			
11	0,106	0,707	98,825			
12	0,071	0,471	99,296			
13	0,051	0,341	99,636			
14	0,045	0,302	99,938			
15	0,009	0,062	100,000			

L'analisi delle Componenti Principali restituisce 15 nuove variabili (le Componenti principali appunto), tante quante erano le variabili originarie.

Qualora tutte le Componenti Principali fossero estratte, esse infatti spiegherebbero il 100% della varianza.

Tra esse si estraggono le prime 4 componenti (quelle

associate ai maggiori autovalori), che spiegano cumulativamente circa il 78% della varianza.

La prima componente spiega il 43% di essa, segue la seconda componente che spiega un ulteriore 15% circa, quindi la terza componente rende conto di un aggiuntivo 12% di varianza; infine, la quarta e ultima componente estratta rappresenta un altro 8% di varianza, giungendo così complessivamente al dato prima evidenziato.

PASSAGGIO 5: MATRICE DEI COMPONENTI

La Matrice dei Componenti riporta sulle colonne le 4 componenti principali estratte e sulle righe le 15 variabili originarie; nelle celle sono riportate dei valori, positivi o negativi, che indicano la correlazione di ciascuna delle componenti estratte con le variabili originarie (consumo dei vari beni alimentari).

Ad esempio, il dato contenuto nella prima cella (0,417) indica che la prima componente principale è positivamente correlata con la prima variabile (consumo di salumi).

Invece, il valore negativo contenuto nella cella riferita alla variabile "consumo di verdure" all'interno della prima componente principale indica come essa sia negativamente correlata con questa (-0,746).

Con riferimento alla prima componente principale (prima colonna), le variabili consumo di "carni bovine", "carni maiale", "pesce" e "legumi" presentano un elevato grado di correlazione positiva.

Pertanto, si procede a denominare tale componente principale "**Componente proteica di tipo animale e vegetale**".

Nella seconda componente una correlazione elevata è osservata con riguardo alle variabili consumo di "carni bianche", "verdure", "frutta" e "ortaggi", che portano a definirla come "**Componente salutista**".

Le variabili consumo di "salumi", "snack" e "dolci", e seppure in misura minore anche la variabile "formaggio", risultano positivamente correlate con la terza componente principale, individuando così la "**Componente calorica**".

Infine, all'interno dell'ultima componente assumono un ruolo centrale le variabili consumo di "salumi", "uova" e "pane"; si tratta di cibi accomunati dall'essere alimenti di veloce consumo e di natura tradizionalmente povera. Si denomina così tale combinazione come "**Componente povera/veloce**".

Matrice dei componenti[a]				
	Componente			
	1	2	3	4
salumi almeno qualche volta alla settimana	0,417	0,175	0,551	0,550
carni bianche almeno qualche volta alla settimana	0,577	0,691	0,083	-0,126
carni bovine almeno qualche volta alla settimana	0,769	0,389	-0,134	-0,118
carni maiale almeno qualche volta alla settimana	0,706	0,405	0,098	0,084
uova almeno qualche volta alla settimana	0,577	-0,031	0,153	0,381
pesce almeno qualche volta alla settimana	0,809	0,027	-0,245	-0,355
verdure almeno una volta al giorno	-0,746	0,551	-0,118	-0,004
ortaggi almeno una volta al giorno	-0,707	0,584	-0,193	0,037
frutta almeno una volta al giorno	0,480	0,458	-0,393	-0,128
pane, pasta almeno una volta al giorno	0,743	0,294	-0,048	0,349
latte almeno una volta al giorno	-0,039	-0,174	-0,797	0,170
formaggio almeno una volta al giorno	-0,841	0,068	0,244	0,180
legumi almeno qualche volta alla settimana	0,935	-0,223	-0,027	0,066
snack almeno qualche volta alla settimana	0,472	-0,234	0,550	-0,559
dolci almeno qualche volta alla settimana	-0,533	0,574	0,365	-0,240

PASSAGGIO 6: MATRICE DEI COEFFICIENTI DI PUNTEGGI DEI COMPONENTI

La Matrice dei Coefficienti di Punteggi dei Componenti riporta sulle colonne le 4 componenti principali estratte e sulle righe le 15 variabili originarie; nelle celle sono riportati i pesi assegnati a ciascuna delle variabili originarie (consumo dei vari beni alimentari) all'interno delle varie componenti estratte.

Ad esempio, il dato contenuto nella prima cella (0,064) indica che la prima variabile (consumo di salumi) assume all'interno della prima componente principale un peso di pari entità. Tale valore si discosta significativamente dal valore massimo assumibile 1 (corrispondente al caso in cui la componente principale coincide con quella sola variabile) e mostra come il peso rivestito da questa nella combinazione lineare non sia preponderante.

Invece, i pesi assunti all'interno della medesima componente dalle variabili consumo di "carni bianche", "carni bovine", "carni di maiale", "pesce" e "legumi", sono più elevati ed indicano che essi svolgono un ruolo cardine nella determinazione della stessa.

Matrice dei coefficienti di punteggi dei componenti				
	Componente			
	1	2	3	4
salumi almeno qualche volta alla settimana	0,064	0,078	0,313	0,463
carni bianche almeno qualche volta alla settimana	0,089	0,309	0,047	-0,106
carni bovine almeno qualche volta alla settimana	0,118	0,174	-0,076	-0,099
carni maiale almeno qualche volta alla settimana	0,108	0,181	0,056	0,071
uova almeno qualche volta alla settimana	0,089	-0,014	0,087	0,321
pesce almeno qualche volta alla settimana	0,124	0,012	-0,139	-0,299
verdure almeno una volta al giorno	-0,115	0,246	-0,067	-0,003
ortaggi almeno una volta al giorno	-0,109	0,261	-0,109	0,031
frutta almeno una volta al giorno	0,074	0,205	-0,223	-0,108
pane, pasta almeno una volta al giorno	0,114	0,132	-0,027	0,294
latte almeno una volta al giorno	-0,006	-0,078	-0,453	0,144
formaggio almeno una volta al giorno	-0,129	0,030	0,138	0,152
legumi almeno qualche volta alla settimana	0,144	-0,100	-0,015	0,056
snack almeno qualche volta alla settimana	0,072	-0,105	0,312	-0,471
dolci almeno qualche volta alla settimana	-0,082	0,257	0,207	-0,202

Le componenti principali individuate sono state di qui in poi assunte come variabili; pertanto, la matrice dei dati originariamente di dimensioni 20 × 15, diviene ora una matrice di dimensioni 20 ×4, qui riportata.

Territorio: regioni	PC1	PC2	PC3	PC4
Piemonte	37,12802	104,8805	-13,9022	24,60769
Valle d'Aosta / Vallée d'Aoste	33,75451	101,9781	-12,2522	31,78129
Liguria	37,01054	97,24015	-17,7284	26,28182
Lombardia	35,8849	98,63438	-9,17888	19,63278
Trentino Alto Adige / Südtirol	27,80272	91,89557	-11,0409	36,39953
Veneto	34,76761	101,7749	-12,4913	18,66591
Friuli-Venezia Giulia	34,38463	102,2347	-13,1394	24,11449
Emilia-Romagna	38,43089	110,4013	-12,0444	25,10934
Toscana	41,44491	104,5149	-18,8278	27,73619
Umbria	46,0569	109,7817	-18,4367	31,27526
Marche	42,26008	104,3376	-15,5827	25,52667
Lazio	40,56606	99,68217	-27,5068	21,54349
Abruzzo	48,09203	97,21972	-12,1475	28,65187
Molise	52,21014	97,90667	-9,39593	29,07291
Campania	51,2953	90,06713	-15,2636	18,79892
Puglia	45,76885	87,09548	-19,2459	23,28193
Basilicata	52,09661	92,81766	-9,48203	31,56287
Calabria	48,97475	91,15925	-16,372	27,33493
Sicilia	48,22506	96,31172	-14,7001	25,66684
Sardegna	38,75418	102,7674	-15,2403	20,2461

4. Analisi dei Cluster

L'analisi dei cluster è un metodo di analisi multivariata dei dati che consente di raggruppare le varie unità statistiche oggetto di indagine in una serie di gruppi (detti appunto cluster).

L'obiettivo è quello di pervenire ad una classificazione in gruppi che risponda a criteri di omogeneità interna e di eterogeneità esterna.

La scelta ottimale del numero di gruppi rappresenta un compromesso tra due esigenze opposte; da un lato la volontà di giungere ad un numero il più ristretto possibile di gruppi, dall'altro la necessità di non racchiudere all'interno di uno stesso gruppo entità con caratteristiche profondamente distinte.

Cenni Teorici

Esistono diverse tipologie di analisi cluster che possono essere condotte sulla base della scelta di alcuni elementi.

Tra di essi rientrano la scelta circa la misura di distanza tra cluster da adottare e la tipologia di metodo aggregativo.

Si è scelto di impiegare la distanza euclidea quadratica (impostazione di default all'interno del software SPSS), dal momento che essa risulta essere più equilibrata rispetto alle altre varianti disponibili.

Per quanto riguarda la scelta del metodo di aggregazione, si è deciso di procedere tanto con l'implementazione di un metodo gerarchico, quanto poi di uno partitivo (k-means).

Metodo Gerarchico

Tale tipologia di clustering consente di ottenere una panoramica completa di tutti i vari raggruppamenti intervenuti per passare dall'articolazione in n gruppi a quella in un solo gruppo (metodo gerarchico di tipo Top-Down), o viceversa, nel passaggio da un unico gruppo allo smembramento in n gruppi (metodo gerarchico di tipo Bottom-Up).

Metodo delle k-means

Trattasi di un metodo di tipo partitivo in cui, a differenza dei metodi gerarchici, il numero di cluster finali in cui sarà suddiviso il campione è prestabilito.

Applicazione pratica in SPSS

Si è partiti dall'implementazione del metodo gerarchico, al fine di individuare il numero ottimale di gruppi finali in cui ripartire la popolazione campionaria.

Questo stesso numero sarà impiegato per l'applicazione del metodo delle k-means.

METODO GERARCHICO:

Riepilogo elaborazione casi[a,b]

Casi					
Valido		Mancante		Totale	
N	Percentuale	N	Percentuale	N	Percentuale
20	100,0	0	0,0	20	100,0

Il software ha impiegato correttamente tutti i valori, di modo che ciascun caso è stato assegnato ad uno dei cluster individuati.

La seguente tabella riporta i vari step intervenuti nel raggruppamento delle differenti unità. Si è partiti da un numero di cluster pari a 20, uno per ciascuna regione, quindi progressivamente il software ha provveduto ad aggregarli giungendo in ultima fase ad ottenere un unico gruppo.

Pianificazione di agglomerazione

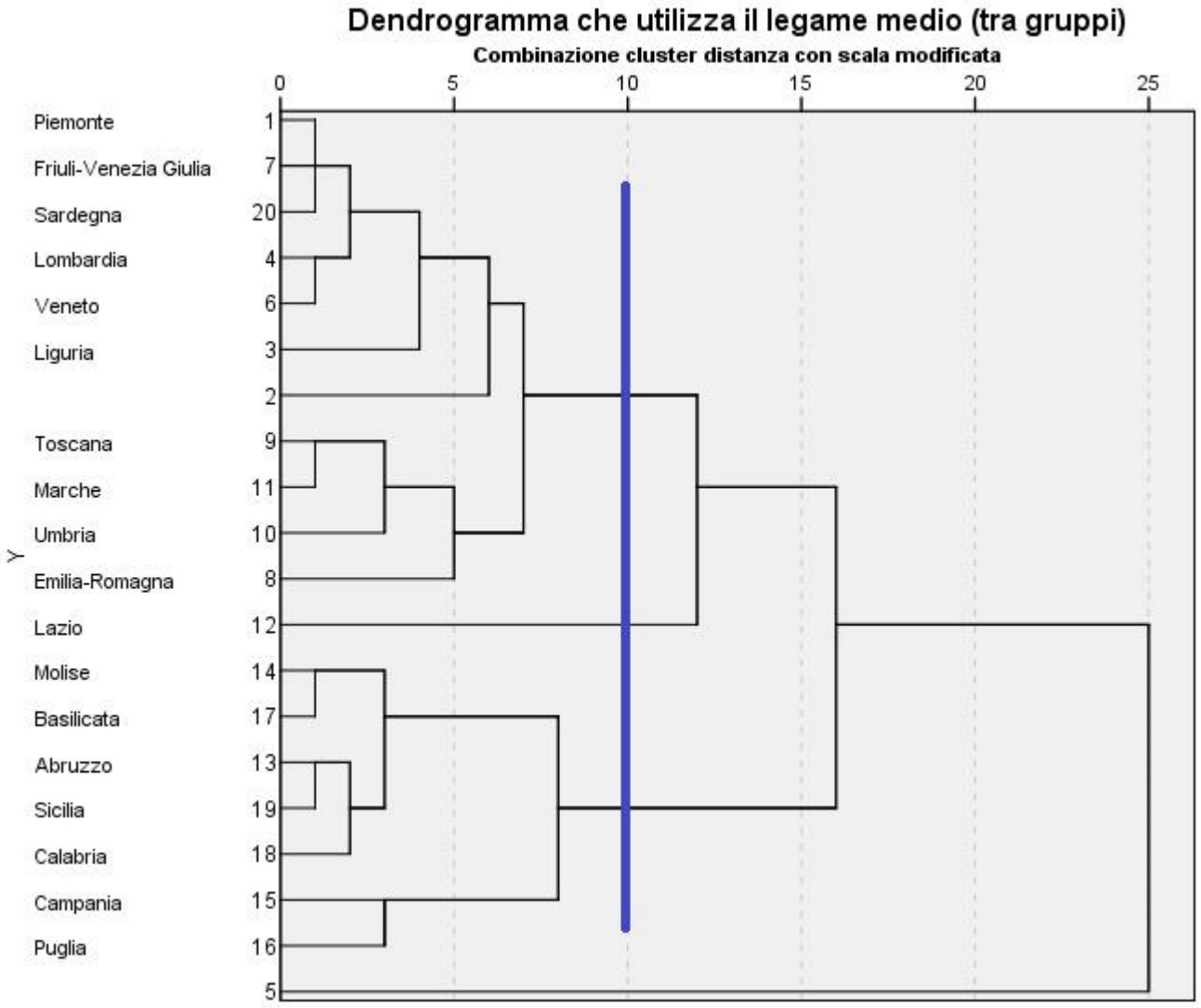
Stadio	Combinato in cluster		Coefficienti	Stadio prima apparizione cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	7	15,351	0	0	6
2	9	11	16,109	0	0	10
3	13	19	16,268	0	0	7
4	4	6	23,018	0	0	8
5	14	17	32,118	0	0	9
6	1	20	33,340	1	0	8
7	13	18	44,889	3	0	9
8	1	4	51,824	6	4	12
9	13	14	68,174	7	5	16
10	9	10	73,467	2	0	13
11	15	16	75,329	0	0	16
12	1	3	88,235	8	0	14
13	8	9	99,399	0	10	15
14	1	2	123,466	12	0	15
15	1	8	155,632	14	13	17
16	13	15	163,316	9	11	18
17	1	12	256,089	15	0	18
18	1	13	339,388	17	16	19
19	1	5	522,074	18	0	0

Come si evince dalla prima riga, al primo stadio il cluster 1 è stato accorpato al cluster 7, costituendo il nuovo cluster 1.

Tale cluster è stato poi oggetto di una successiva aggregazione allo stadio 6, ove è stato unito al cluster 20, gruppo costituito da una sola regione, non essendo ancora questo stato accorpato in precedenza ad alcun altro cluster.

L'osservazione della quarta colonna consente di cercare il più grande gap nei valori di distanza tra oggetti, intervenuti in questo caso agli stadi 17 e 19.

Il taglio va operato prima delle aggregazioni corrispondenti a salti rilevanti dell'indice di distanza.



Come si osserva dal dendogramma, esso è effettuato prima del valore 10, cui corrisponde il gap maggiore nel coefficiente riscalato. Ciò corrisponde ad identificare 4 cluster.

METODO DELLE K-MEANS:

Si impiega tale numero di cluster e lo si va ad inserire all'interno dell'applicazione del metodo delle k-means.

Il metodo delle k-means, qualificandosi come metodo non gerarchico, inizia da alcuni centri di partenza indicati con G.

Centri cluster iniziali

	Cluster			
	1	2	3	4
PC1	38,43089	27,80272	40,56606	52,09661
PC2	110,40132	91,89557	99,68217	92,81766
PC3	-12,04437	-11,04092	-27,50678	-9,48203
PC4	25,10934	36,39953	21,54349	31,56287

Tali centri sono progressivamente modificati man mano che l'algoritmo opera onde aggiustarli alle modifiche intervenute nella composizione dei gruppi.

Cronologia delle iterazioni[a]

Iterazione	Modifica nei centri del cluster			
	1	2	3	4
1	6,608	0,000	5,845	7,262
2	0,000	0,000	0,000	0,000

In questo caso l'algoritmo opera due sole iterazioni, giungendo a definire i nuovi centri G sotto riportati.

Centri finali del cluster

	Cluster			
	1	2	3	4
PC1	38,28666	27,80272	38,78830	49,52325
PC2	104,13054	91,89557	98,46116	93,22538
PC3	-14,10959	-11,04092	-22,61758	-13,80099
PC4	24,86957	36,39953	23,91266	26,33861

L’algoritmo restituisce 4 cluster così costituiti, che presentano una numerosità rispettivamente pari a:

Numero di casi in ciascun cluster		
Cluster	1	10,000
	2	1,000
	3	2,000
	4	7,000
Valido		20,000
Mancante		0,000

- 10 unità/regioni all’interno del cluster 1;
- 1 sola regione costituisce il cluster 2;
- 2 regioni sono unite nel cluster 3;
- le rimanenti 7 regioni confluiscono nel cluster 4.

Appartenenza cluster			
Numero di caso	Regioni	Cluster	Distanza
1	Piemonte	1	1,420
2	Valle d'Aosta / Valle d'Aoste	1	8,740
3	Liguria	3	5,845
4	Lombardia	1	9,365
5	Trentino Alto Adige / Sudtirol	2	0,000
6	Veneto	1	7,684
7	Friuli-Venezia Giulia	1	4,509
8	Emilia-Romagna	1	6,608
9	Toscana	1	6,372
10	Umbria	1	12,332
11	Marche	1	4,293
12	Lazio	3	5,845
13	Abruzzo	4	5,108
14	Molise	4	7,484
15	Campania	4	8,491
16	Puglia	4	9,522
17	Basilicata	4	7,262
18	Calabria	4	3,489
19	Sicilia	4	3,531
20	Sardegna	1	4,973

Il test ANOVA indica quali variabili hanno maggiormente contribuito all'individuazione dei cluster; in questo caso si osserva come la Componente Principale 1, che si ricorda essere la "Componente Proteica di natura Animale e Vegetale" sia quella che abbia influenzato significativamente la composizione finale dei gruppi.

Segue per importanza la Componente Principale 2, "Componente Salutista" con un valore della F di Fisher pari a 13, 256.

Le Componenti Principali 3 e 4 hanno invece contribuito solo in misura residuale.

ANOVA

	Cluster		Errore		F	Sign.
	Media quadratica	gl	Media quadratica	gl		
PC1	251,660	3	11,442	16	21,994	0,000
PC2	182,446	3	13,764	16	13,256	0,000
PC3	49,303	3	12,992	16	3,795	0,031
PC4	43,360	3	19,026	16	2,279	0,119

5. Conclusioni

L'indagine condotta sull'analisi dei consumi regionali ha riscontrato esiti positivi attraverso l'applicazione del metodo delle Componenti Principali, individuando le 4 componenti che secondariamente sono state assunte come variabili e implementate nell'analisi dei cluster.

Quest'ultima è stata articolata in due sottofasi, la prima delle quali prevede l'applicazione di un metodo di clustering gerarchico, la seconda l'implementazione di un metodo di clustering partitivo.

La prima ha permesso di individuare il numero ottimale dei cluster, pari a k=4. Tale valore è stato poi inserito nel metodo di clustering delle k-means; quest'ultimo, a differenza dei metodi gerarchici, ha il pregio di consentire un'eventuale riallocazione dei casi ai cluster man mano che l'algoritmo opera.

L'assegnazione infatti di un caso ad un gruppo non è irrevocabile, rendendolo un metodo più flessibile e dinamico.

Il suo usuale svantaggio, consistente nella previsione di una decisione arbitraria del numero di cluster finali che il processo deve restituire, è stato ovviato tramite l'aggiustamento sopra descritto.

Il metodo delle k-means ha evidenziato i seguenti risultati:

- La suddivisione in cluster corrisponde anche ad una suddivisine di tipo territoriale. Il cluster 1 è costituito dalla Sardegna e dalle regioni del Nord, ad esclusione del Trentino-Alto Adige che costituisce cluster a sé stante (cluster 2). Formano il cluster 4 le regioni del basso Centro e del Sud Italia, ivi compresa la Sicilia. Infine, le regioni Liguria e Lazio costituiscono il cluster 3;
- La ragione Trentino-Alto Adige risulta separata dalle altre a causa del valore assunto dalla Componente Principale 1, ossia la "Componente proteica", che risulta essere nettamente inferiore rispetto ai valori registrati dagli altri cluster. Contribuisce anche a questa differenziazione la Componente Principale 4 (componente povera/veloce) che assume un valore superiore rispetto agli altri gruppi;

- Lazio e Liguria (cluster 3), pur presentando valori molto simili alle regioni del Nord e alto-Centro (cluster 1) in merito alle variabili componente protetica salutista e veloce/povera, differiscono da queste per via della componente calorica nettamente superiore (PC 3);
- I cluster 1 e 4 (regioni del Nord e del Sud rispettivamente) sono accumunate dal valore delle variabili componente calorica e veloce/povera, mentre differiscono a causa delle variabili Componente Proteica (maggiore nelle regioni del Sud) e Componente Salutista (maggiore nelle regioni del Nord).