TECNICHE DI CLASSIFICAZIONE

La tecnica di classificazione più conosciuta è la **cluster analysis**, che ha l'obiettivo di identificare gruppi di soggetti (o oggetti) omogenei al loro interno ed eterogenei tra di loro rispetto a un set di variabili (e.g., attributi dell'offerta)

Esistono diversi **algoritmi** di cluster analysis:

- **gerarchici:** si tratta di algoritmi che prevedono l'aggregazione sequenziale (i.e., gerarchica) dei soggetti in gruppi; tali algoritmi possono analizzare variabili solo qualitative (e.g., dummy) o solo quantitative, ma tendono ad essere poco efficienti e a richiedere molte decisioni soggettive da parte del ricercatore
- **non gerarchici:** si tratta di algoritmi che prevedono l'aggregazione dei soggetti in gruppi sulla base di punteggi rispetto a variabili quantitative; tendono ad essere molto efficienti, ma non permettono l'utilizzo di variabili qualitative

CLUSTER ANALYSIS

In questa sede verrà approfondita l'applicazione di un algoritmo non-gerarchico di cluster analysis, ovvero il metodo **k-medie,** che permette di identificare gruppi di soggetti sulla base delle medie di un set di variabili di classificazione quantitative (preferibilmente standardizzate, se misurate con un'unità di misura diversa o se percettive)

La cluster analysis k-medie può essere applicata in "tandem" con altre tecniche; in particolare, in:

- **analisi di segmentazione classica (factor + cluster):** applicazione della cluster k-medie sui punteggi fattoriali, interpretazione dei cluster sulla base delle medie rispetto ai fattori
- **analisi di segmentazione flessibile (conjoint + cluster):** applicazione della cluster k-medie sui punteggi di utilità (output conjoint), successiva applicazione e interpretazione dei risultati della conjoint analysis sui singoli cluster

SEGMENTAZIONE CLASSICA

La segmentazione classica prevede che il ricercatore, dalla precedente applicazione della factor analysis, abbia individuato dei fattori descrittivi dell'offerta e ne abbia salvato i punteggi standardizzati

È possibile, quindi, applicare **la cluster analysis k-medie sui punteggi fattoriali**

L'algoritmo non gerarchico k-medie prevede la suddivisione di un campione in un numero **predefinito** di cluster, in modo da massimizzare, rispetto alle variabili di classificazione, **il rapporto tra varianza esterna (tra i gruppi) e varianza interna (nei gruppi)**

Il metodo k-medie è sensibile agli outlier: verifica preliminare dei punteggi standardizzati maggiori di |3|

IL METODO K-MEDIE

L'applicazione del metodo k-medie richiede la definizione a priori

- del **numero dei cluster:** il ricercatore prova le soluzioni da 2 a n cluster e individua, sulla base di particolari output, la soluzione migliore
- dei **centri iniziali dei cluster**, rispetto ai quali effettuare la partizione (SPSS offre i centri iniziali in automatico, sulla base delle osservazioni più discriminanti rispetto alle variabili di classificazione)

Il metodo k-medie segue poi un processo iterativo

Dopo aver identificato i centri iniziali e creato la partizione di partenza, i soggetti vengono ri-allocati al cluster con medie (centri) più vicine ai loro punteggi, in modo da massimizzare la varianza esterna e minimizzare la varianza interna

IL METODO K-MEDIE

Il ricercatore dovrà scegliere la soluzione a k cluster sulla base di una serie di criteri

- **i valori dei test F** su ogni variabile di classificazione, riportati nella tabella ANOVA, e i relativi valori del **p-value**
- l'interpretabilità dei cluster, sulla base dei **centri finali dei cluster** (medie dei cluster rispetto alle variabili di classificazione)
- **numerosità dei cluster** (situazione preferita: omogeneità)

Identificata la soluzione ottimale a k gruppi, è necessario interpretare (assegnandogli dei nomi specifici) i cluster rispetto ai centri finali, che indicano le medie dei cluster rispetto alle variabili di classificazione

I cluster possono essere descritti, inoltre, rispetto ad altre variabili terze (es.: socio-demo, comportamenti, percezioni)

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

Dalla precedente applicazione della factor analysis sono stati salvati i punteggi fattoriali; per prima cosa è opportuno verificare la presenza di potenziali outlier, calcolando le statistiche descrittive dei punteggi fattoriali

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Rapporto personale	160	-2.56063	2.75148	.0000000	1.00000000
Accessibilità dell'offerta	160	-3.10076	2.97955	.0000000	1.00000000
Struttura	160	-2.93176	2.39654	.0000000	1.00000000
Aspetti distributivi	160	-2.90730	2.10473	.0000000	1.00000000
Parking	160	-2.65617	3.05512	.0000000	1.00000000
Valid N (listwise)	160				

Sebbene siano evidenti alcuni valori vicini a |3|, non sembrano presenti casi eclatanti di outlier

È quindi possibile procedere con le successive analisi

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

Applicando la cluster k-medie, è possibile verificare che le soluzioni da 2 a 6 cluster mostrano tutti F-test significativi su tutte le variabili di classificazione, oltre che una numerosità dei cluster abbastanza omogenea

In questi casi, la scelta della soluzione ottimale avviene sulla base dell'interpretabilità della soluzione e delle esigenze manageriali (più cluster: più precisione, ma anche più complessità)

Scegliamo la soluzione a 4 cluster

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Rapporto personale	5,690	3	,910	156	6,254	,000
Accessibilità	9,712	3	,832	156	11,667	,000
Struttura	25,086	3	,537	156	46,733	,000
Aspetti distributivi	15,637	3	,719	156	21,762	,000
Parcheggi	25,791	3	,523	156	49,290	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Come detto, tutti gli F-test sono significativi: le variabili di classificazione discriminano i cluster in termini di medie

Final Cluster Centers

	Cluster			
	1	2	3	4
Rapporto personale	-,02033	-,61684	,21064	,27173
Accessibilità	-,05245	,62908	,08897	-,69472
Struttura	-1,08883	,00125	,70206	,14848
Aspetti distributivi	-,42929	,63474	,38966	-,74458
Parcheggi	-,42284	,86327	-,67606	,80219

Number of Cases in each Cluster

Cluster	1	40,000
	2	32,000
	3	55,000
	4	33,000
Valid		160,000
Missing		,000

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

A fini descrittivi, è possibile applicare opportune analisi bivariate tra il cluster di appartenenza dei soggetti e le variabili qualitative "genere" e "frequenza d'uso del servizio"

Crosstab

Count

		Sesso dell'intervistato		
		maschio	femmina	Total
Cluster Number of Case	1	25	15	40
	2	18	14	32
	3	28	27	55
	4	15	18	33
Total		86	74	160

Chi-quadrato: non significativo

Crosstab

Count

		Frequenza servizi PT					
		3 volte a settimana	2 volte a settimana	1 volta a settimana	2 volte al mese	1 volta al mese	Total
Cluster Number of Case	1	1	7	9	12	11	40
	2	1		6	5	20	32
	3		7	11	17	20	55
	4	3	3	2	11	14	33
Total		5	17	28	45	65	160

Chi-quadrato: significativo

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

A fini descrittivi, è possibile applicare opportune analisi bivariate tra il cluster di appartenenza dei soggetti e le variabili quantitative "età" e "soddisfazione generale"

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Età	1	40	43,7000	14,9944	2,3708	38,9046	48,4954	21,00	70,00
	2	32	40,8750	13,8977	2,4568	35,8644	45,8856	19,00	69,00
	3	55	45,2545	15,3227	2,0661	41,1122	49,3969	19,00	70,00
	4	33	51,5758	16,0274	2,7900	45,8927	57,2588	22,00	70,00
	Total	160	45,2938	15,3948	1,2171	42,8900	47,6975	19,00	70,00
Soddisfazione servizio pt della zona	1	40	4,1000	1,2362	,1955	3,7046	4,4954	1,00	7,00
	2	32	4,6250	1,3137	,2322	4,1514	5,0986	1,00	6,00
	3	55	4,7273	1,0084	,1360	4,4547	4,9999	1,00	6,00
	4	33	4,1212	1,4949	,2602	3,5911	4,6513	1,00	7,00
	Total	160	4,4250	1,2617	9,97E-02	4,2280	4,6220	1,00	7,00

In entrambi i casi: F-test significativi

+ post-hoc t-test

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

Il **primo cluster (N = 40)** mostra medie tutte negative; i punteggi di rapporto personale e accessibilità sono in realtà prossimi alla media, mentre è molto negativo (quindi molto inferiore alla media) il punteggio di struttura; questo cluster sembra essere formato dai clienti (tendenzialmente maschi non soddisfatti, che usano abbastanza spesso il servizio) più indifferenti, particolarmente insensibili alla struttura: **gli indifferenti**

Il **secondo cluster (N = 32)** mostra medie molto alte su accessibilità, aspetti distributivi e parcheggi, mentre la media su rapporto personale è negativa; questi soggetti (i più giovani, abbastanza soddisfatti, ma che usano poco il servizio) sembrano interessati ad elementi molto funzionali e poco all'interazione sociale: **i pratici**

Il **terzo cluster (N = 55)** mostra una media molto alta su struttura, e medie positive su rapporto personale e aspetti distributivi; la media su parcheggi è decisamente negativa; questi soggetti (i più soddisfatti) sembrano interessati alla fruizione del servizio sia in filiale (andando a piedi o con mezzi pubblici) che a casa: **i vicini contenti**

Il **quarto cluster (N = 33)** mostra medie molto basse su accessibilità e aspetti distributivi, mentre le medie su rapporto personale, struttura e soprattutto parcheggi sono positive; questi soggetti (i più anziani e poco soddisfatti) sembrano interessati ad aspetti legati ai servizi in filiale da raggiungere, però, in auto: **i nonni socievoli lontani**

CLUSTER ANALYSIS (K-MEDIE): UN ESEMPIO

Passo successivo: proposta di azioni manageriali da indirizzare ai cluster, considerandone le caratteristiche distintive

Esempi:

- vantaggi sui parcheggi ai soggetti del cluster 4

- promozioni e bundling ai soggetti del cluster 2

SEGMENTAZIONE CLASSICA: UNA SEQUENZA LOGICA

1. Applicare la **factor analysis** sulle variabili osservate (valutazioni d'importanza su attributi dell'offerta) e salvare i punteggi fattoriali
2. Applicare la **cluster analysis k-medie** sui fattori e salvare la soluzione a k cluster ideale
3. Applicare **ulteriori analisi bivariate**, incrociando la cluster membership e altre variabili disponibili (es.: socio-demo, comportamentali, percezioni), per identificare ulteriori tratti descrittivi dei cluster
4. Interpretare i cluster, **fornendo un nome e una breve descrizione** per ognuno di essi, sulla base dei centri finali della cluster analysis e delle eventuali evidenze significative dalle ulteriori analisi bivariate
5. Definire le **implicazioni manageriali**, identificando il o i migliori segmenti target per l'impresa e proponendo azioni di marketing specifiche per ogni target