Statistica I

Unità I: dati qualitativi

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità I

Argomenti affrontati

- Moda
- Diagramma a barre, diagramma a torta
- Concetto di mutabilità
- Indice di Gini ed entropia di Shannon

Riferimenti al libro di testo

- §5.7
- **Nota**. Nel libro di testo sono discussi ulteriori indici di eterogeneità (Leti, Frosini), che non sono materia d'esame.

Descrizione del problema

- I dati delle elezioni comunali 2016 di Milano sono disponibili sul **sito del Comune di Milano**
- I dati si possono ottenere al link: `https://dati.comune.milano.it/dataset/ds1183_elezioni-comunali-2016--sindaco-voti-di-lista-per-sezione`
- Il **primo turno** delle elezioni si è svolto il 5 Giugno 2016 ed è stato eletto sindaco Giuseppe Sala.
- Sono noti i **voti** ricevuti da ciascun candidato in ogni **Municipio**.
- **Nota**. Ci sono alcune piccole differenze tra i dati ufficiali del comune e quelli riportati dai principali quotidiani.

I 9 municipi di Milano

Municipio	Quartieri
Municipio 1	Centro storico
Municipio 2	Stazione Centrale, Gorla, Turro, Greco, Crescenzago
Municipio 3	Città Studi, Lambrate, Porta Venezia
Municipio 4	Porta Vittoria, Forlanini
Municipio 5	Vigentino, Chiaravalle, Gratosoglio
Municipio 6	Barona, Lorenteggio
Municipio 7	Baggio, De Angeli, San Siro
Municipio 8	Fiera, Gallaratese, Quarto Oggiaro
Municipio 9	Porta Garibaldi, Niguarda

- Per chi fosse interessato: `https://it.wikipedia.org/wiki/Municipi_di_Milano`
- Università Milano-Bicocca si trova nel Municipio 9.

I dati grezzi

- I dati prendono la forma di una lunga tabella.

Elettore	Municipio	Voto
1	Municipio 1	Stefano Parisi
2	Municipio 4	Giuseppe Sala
3	Municipio 9	Stefano Parisi
⋮	⋮	⋮
537619	Municipio 7	Giuseppe Sala

- Per ogni elettore (unità statistica) vengono rilevate due variabili: il **municipio** di appartenenza ed il **voto**.
- Si tratta quindi di variabili **qualitative sconnesse**.
- I **voti validi** (numerosità campionaria) sono complessivamente $n = 537619$.

Frequenze assolute e relative

- La tabella della pagina precedente è poco "maneggevole".
- I dati possono essere rappresentati tramite seguente tabella di **frequenze assolute**.
- Ad esempio, 1073 è il numero di voti ricevuti da Marco Cappato nel Municipio 1, ovvero il centro storico di Milano.

	Municipio								
Candidato	1	2	3	4	5	6	7	8	9
Azzaretto, N.	91	212	235	229	180	256	427	283	302
Baldini, M. T.	81	127	144	140	121	112	124	144	148
Cappato, M.	1073	928	1375	1214	848	1143	1218	1152	1130
Corrado, G.	2023	5915	5324	6091	5284	6276	7300	7543	8349
Mardegan, N.	488	580	646	584	457	557	685	1266	759
Parisi, S.	17154	23508	23241	25595	19521	24156	28970	29636	27431
Rizzo, B. V.	1185	2032	2385	2251	1733	2085	2196	2652	2626
Sala, B.	18939	21448	26949	25626	20447	25508	28050	29765	27481
Santambrogio, L.	111	125	143	103	84	179	294	266	180

Frequenze assolute e relative

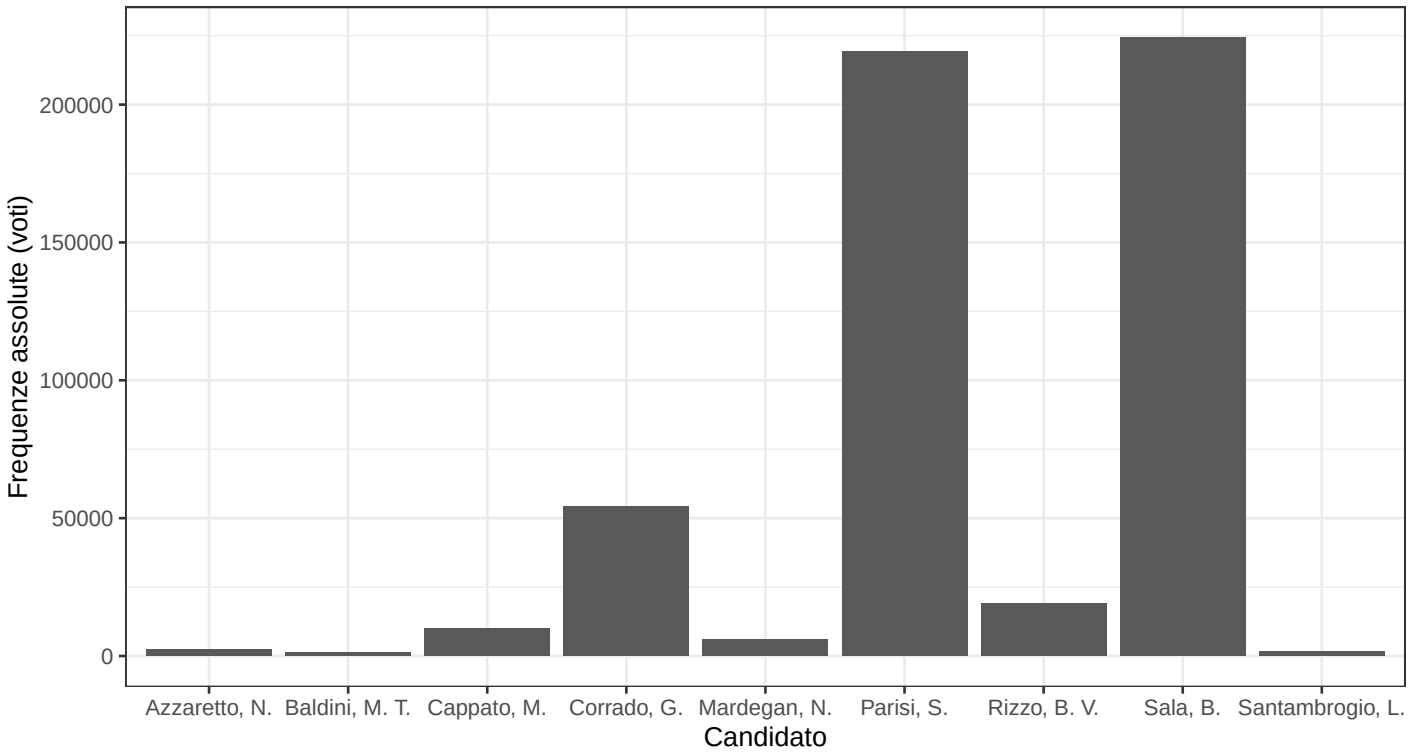
- La variabile **Voto** ha la seguente distribuzione di frequenze.
- I candidati Giuseppe Sala e Stefano Parisi hanno quindi ricevuto la maggior parte dei voti e pertanto sono stati ammessi al ballottaggio.

Candidato	Frequenze assolute (Voti)	Frequenze relative
Azzaretto, N.	2215	0.00
Baldini, M. T.	1141	0.00
Cappato, M.	10081	0.02
Corrado, G.	54105	0.10
Mardegan, N.	6022	0.01
Parisi, S.	219212	0.41
Rizzo, B. V.	19145	0.04
Sala, B.	224213	0.42
Santambrogio, L.	1485	0.00

Commento ai dati

- La natura di questi dati è diversa da quelli visti in precedenza.
- Nei precedenti esempi sono stati considerati **dati numerici**.
- Viceversa, in questo caso le variabili sono nomi e luoghi. Sono pertanto dei **dati qualitativi** o categoriali.
- Questo cambia (di molto!) quello che possiamo e non possiamo fare.
- **Nota importante**. Non ha senso chiederci quanto valga la media aritmetica o la varianza ad esempio della variabile Municipio.
- Pertanto, dobbiamo costruire delle rappresentazioni grafiche, indici di posizione, di variabilità che siano opportuni per questa tipologia di dati.

Diagramma a barre

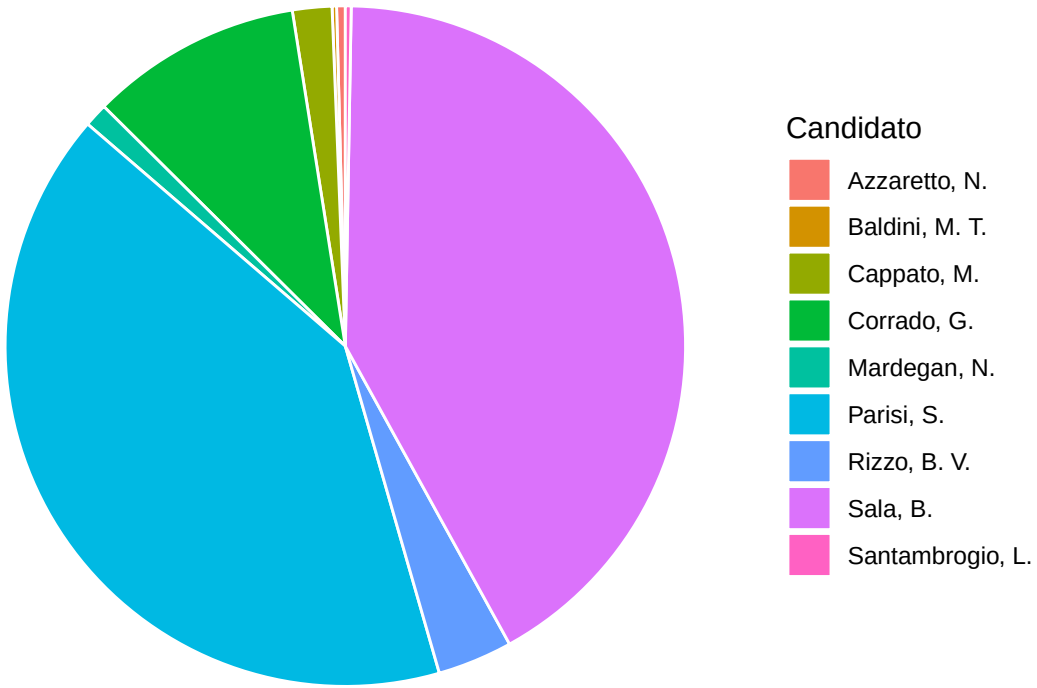


- La rappresentazione grafica più utilizzata è il **diagramma a barre**: ogni modalità è rappresentata da una barra di altezza pari alla frequenza (assoluta o relativa).
- I rettangoli, contrariamente al caso di un istogramma, sono disegnati **staccati**.
- Se la variabile non è ordinale, l'**ordine** delle modalità è **arbitrario**.

Diagramma a torta

- Una diversa rappresentazione grafica per variabili qualitative è il **diagramma a torta**.
- Ogni modalità è rappresentata da una fetta di torta proporzionale alla sua frequenza relativa, ovvero

$$(\text{Angolo in gradi}) = 360^\circ \times (\text{frequenza relativa}).$$



La moda

- Volendo sintetizzare una variabile qualitativa tramite un unico valore si può usare un **indice di posizione** chiamato moda, che caratterizza la modalità più frequente.

- **La moda**. La moda dei dati è la modalità cui corrisponde la massima frequenza assoluta.

- La moda della variabile **Voto** è Mo = Giuseppe Sala. Infatti, Giuseppe Sala ha ricevuto 224213 voti al primo turno.

- **Nota**. Attenzione a non confondersi: la moda è Giuseppe Sala e **NON** la sua frequenza 224213.

- **Esercizio - proprietà**. Dimostrare che la moda coincide con la modalità avente la più alta frequenza relativa.

La moda e le variabili numeriche

- La moda può essere usata per qualsiasi distribuzione di frequenza, incluse quelle delle unità precedenti basate su dati numerici.
- In caso di variabili **numeriche discrete**, la moda si calcola come nel caso di variabili qualitative, ovvero considerando la modalità associata alla frequenza più alta.
- In caso di variabili **numeriche continue**, la moda non esiste. Infatti, se i dati sono tutti diversi tra loro, allora necessariamente le modalità hanno frequenza assoluta pari a 1.
- In caso di variabili **numeriche (discrete e continue) raggruppate in classi**, allora si parla di **classe modale**.
- **Nota**. La classe modale è quella con **densità di frequenza** più elevata e NON quella avente frequenza più alta. Si veda l'Unità G per la definizione di densità.

Esempio di calcolo della classe modale

- Supponiamo di avere i seguenti dati **raggruppati**

Classi	$(0,1]$	$(1,2]$	$(2,5]$	$(5,7]$	$(7,10]$
Frequenze assolute	1	4	5	2	1

- In primo luogo, otteniamo le densità per ciascuna classe, pari a

Classi	$(0,1]$	$(1,2]$	$(2,5]$	$(5,7]$	$(7,10]$
Densità	$1/1$	$4/1$	$5/3$	$2/2$	$1/3$

- La classe modale è quindi $(1,2]$, ovvero la classe avente la più alta densità, pari a 4.
- **Nota**. La classe modale **NON** è $(2,5]$, nonostante questa abbia la frequenza più alta.

La mediana per dati ordinali

- Nel caso in cui i dati siano qualitativi **ordinali** è possibile utilizzare la mediana, la cui definizione deve essere leggermente adattata. Infatti, in questo contesto non è possibile considerare semi-somme.

- I seguenti dati sono i voti ricevuti da una classe di $n = 26$ persone.

Modalità	Sufficiente	Buono	Distinto	Ottimo
Frequenze assolute	3	5	10	8

Ovviamente si ha che: Sufficiente $<$ Buono $<$ Distinto $<$ Ottimo.

- Poichè $n = 26$ è pari, la mediana coinciderà con il valore centrale $x_{(13)}$ oppure con $x_{(14)}$.

- In questo caso si ha che $x_{(13)} = x_{(14)} =$ Distinto, e pertanto concludiamo che

$$\text{Me} = \text{Distinto}.$$

- Inoltre, in questo caso si ha anche che Me $=$ Mo $=$ Distinto.

La mutabilità

- La **mutabilità**, **eterogeneità** o **diversità** è l'analogo della variabilità per dati qualitativi.

Minima mutabilità

- La minima mutabilità si osserva se le unità statistiche sono tutte uguali. Le unità statistiche sono perfettamente **omogenee** rispetto al fenomeno considerato.

- Si osservi che in questo caso la distribuzione delle frequenze relative si presenta come

Modalità	c_1	...	c_j	...	c_k
Frequenze relative	0	...	1	...	0

Massima mutabilità

- La massima mutabilità si osserva se le unità statistiche si ripartiscono eugualmente.

- Si osservi che in questo caso la distribuzione delle frequenze relative si presenta come

Modalità	c_1	...	c_j	...	c_k
Frequenze relative	1/k	...	1/k	...	1/k

La mutabilità, esempi applicativi

- Nelle analisi delle preferenze elettorali, i risultati possono oscillare tra un estremo di **indecisione assoluta** (tutti i candidati ricevono gli stessi voti), ed **estrema polarizzazione** (uno o due candidati ricevono la maggior parte dei voti).
- In questo contesto specifico, gli indici di mutabilità rappresentano degli **indici di polarizzazione** del consenso elettorale.
- In ecologia, la problematica dell'eterogeneità è connessa alla diversità delle specie animali e vegetali presenti nel territorio.
- Infatti, più le specie sono **diversificate** maggiore sarà il patrimonio genetico. Di conseguenza, il sistema sarà maggiormente capace di adattarsi a cambiamenti di qualsiasi origine. Viceversa, un territorio popolato da una sola specie è fragile.

Indice di Gini

Modalità	c_1	$\dots$	c_j	$\dots$	c_k
Frequenze relative	f_1	$\dots$	f_j	$\dots$	f_k

- **Indice di mutabilità Gini**. L'indice di Gini dei dati aventi frequenze relative $f_1, \dots, f_k$ è

$$G = \sum_{j=1}^{k} f_j(1 - f_j) = 1 - \sum_{j=1}^{k} f_j^2.$$

- In condizioni di **minima mutabilità** l'indice di Gini è pari a zero. Infatti

$$G = 1 - \sum_{j=1}^{k} f_j^2 = 1 - (0^2 + \cdots + 1^2 + \cdots + 0^2) = 1 - 1 = 0.$$

- In condizioni di **massima mutabilità** l'indice di Gini è invece pari a:

$$G = 1 - \sum_{j=1}^{k} \frac{1}{k^2} = 1 - \frac{k}{k^2} = 1 - \frac{1}{k} = \frac{k-1}{k}.$$

Indice di Gini, definizione alternativa

- In maniera analoga alla varianza (si veda l'Unità F), l'indice di Gini si può derivare come la media delle **distanze tra tutte le osservazioni**.

- In questo caso utilizziamo la cosiddetta **distanza di Hamming**, che è semplicemente

$$(\text{distanza di Hamming tra } x_i \text{ e } x_j) = \mathbb{1}(x_i \neq x_j) = \begin{cases} 0, & \text{se } x_i = x_j \\ 1, & \text{se } x_i \neq x_j. \end{cases}$$

- La distanza di Hamming quindi misura se due quantità sono uguali o diverse. Si noti che per dati qualitativi questa è sostanzialmente l'unica misura coerente di distanza.

Teorema

L'indice di mutabilità di Gini G dei dati $x_1, \ldots, x_n$ aventi modalità $c_1, \ldots, c_k$ e frequenze assolute $n_1, \ldots, n_k$ è pari a

$$G = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(x_i \neq x_j) = \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{k} n_i n_j \mathbb{1}(c_i \neq c_j).$$

Dimostrazione

- In primo luogo si noti che

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}(x_i \neq x_j) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{k} n_j \mathbb{1}(x_i \neq c_j) = \frac{1}{n^2}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j \mathbb{1}(c_i \neq c_j).$$

- La dimostrazione quindi segue con qualche manipolazione algebrica

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j \mathbb{1}(c_i \neq c_j) &= \frac{1}{n^2}(0 \times n_1 n_1 + n_1 n_2 + \cdots + 1 \times n_1 n_k + \\ &\qquad + n_2 n_1 + 0 \times n_2 n_2 + \cdots + n_2 n_k + \cdots + 0 \times n_k n_k) \\ &= \frac{1}{n^2}[n_1(n-n_1) + n_2(n-n_2) + \cdots + n_k(n-n_k)] \\ &= \frac{1}{n^2}\sum_{j=1}^{k} n_j(n-n_j) = 1 - \frac{1}{n^2}\sum_{j=1}^{k} n_j^2 = 1 - \sum_{j=1}^{k} f_j^2.\end{aligned}$$

Proprietà dell'indice di Gini

- **Proprietà**. L'indice di Gini si può anche scrivere come

$$G = 1 - \frac{1}{n^2} \sum_{j=1}^{k} n_j^2.$$

Per convincersene, si veda l'ultima riga della dimostrazione precedente.

Teorema

L'indice di Gini dei dati $x_1, \ldots, x_n$ con k modalità è tale che

$$G \leq \left(1 - \frac{1}{k}\right),$$

ed è pari al valore massimo $G = 1 - 1/k$ **se e solo se** le frequenze relative assumono il valore $f_j = 1/k$, per ogni $j = 1, \ldots, k$.

- In altri termini, l'indice di Gini raggiunge il valore massimo $1 - 1/k$ solo in situazione di massima mutabilità.

Dimostrazione (facoltativa)

- Per costruzione delle frequenze relative, si ha che $f_k = 1 - \sum_{j=1}^{k-1} f_j$. Pertanto, possiamo scrivere l'indice di Gini come segue

$$G = \sum_{j=1}^{k-1} f_j(1-f_j) + f_k(1-f_k) = \sum_{j=1}^{k-1} f_j(1-f_j) + \left(1 - \sum_{j=1}^{k-1} f_j\right)\left(\sum_{j=1}^{k-1} f_j\right)$$

- Il massimo va cercato tra i punti che annullano le derivate prime. Quindi otteniamo

$$\frac{\partial G}{\partial f_j} = 1 - 2f_j - \sum_{j=1}^{k-1} f_j + 1 - \sum_{j=1}^{k-1} f_j = 2 - 2f_j - 2\sum_{j=1}^{k-1} f_j,$$

per ogni $j = 1, \ldots, k-1$.

- Ponendo pari a zero i termini precedenti, otteniamo che per ogni $j = 1, \ldots, k-1$

$$2 - 2f_j - 2\sum_{j=1}^{k-1} f_j = 0 \quad \Longleftrightarrow \quad f_j = 1 - \sum_{j=1}^{k-1} f_j = f_k \quad \Longleftrightarrow \quad f_1 = \cdots = f_k,$$

da cui segue che $f_1 = \cdots = f_k = 1/k$.

Dimostrazione (facoltativa)

- Abbiamo quindi dimostrato che il punto $f_1 = \cdots = f_k = 1/k$ è l'unico che annulla tutte le derivate prime.

- Concludiamo quindi che $f_j = 1/k$ per $j = 1, \ldots, k$ è il punto di massimo assoluto e quindi necessariamente si avrà che

$$G \leq 1 - \sum_{j=1}^{k} \frac{1}{k^2} = 1 - \frac{1}{k}.$$

- **Nota matematica**. La dimostrazione non è completa: bisogna infatti verificare che tale punto sia effettivamente un punto di massimo e non ad esempio un punto stazionario.

- Per fare ciò, bisognerebbe ottenere le derivate seconde a verificare che la matrice ottenuta sia definita negativa. Tale verifica è lasciata come (difficile!) esercizio.

Indice di Gini normalizzato

- In pratica spesso viene utilizzato l'**indice di Gini normalizzato**, definito come

$$G_{\text{norm}} = \frac{G}{(\text{massimo valore di } G)} = \frac{k}{k-1} G.$$

- L'indice normalizzato pertanto è tale che $0 \leq G_{\text{norm}} \leq 1$, ovvero varia tra 0 e 1.
- In particolare, assume il valore 0 in presenza di minima mutabilità e valore 1 in presenza di massima mutabilità.
- Per la variabile **Voto** si ha che $G = 0.6479$ e che $G_{\text{norm}} = 0.7289$.

Entropia di Shannon

- **Entropia di Shannon**. L'entropia di Shannon dei dati aventi frequenze relative $f_1, \ldots, f_k$ è

$$H = -\sum_{j=1}^{k} f_j \log f_j,$$

in cui se $f_j = 0$ per convenzione poniamo $f_j \log f_j = 0$.

- Proviene dalla **teoria dell'informazione**, dove viene utilizzata per misurare la complessità di un messaggio.

- In condizioni di **minima mutabilità** l'entropia di Shannon è pari a zero.

- In condizioni di **massima mutabilità** l'entropia di Shannon è invece pari a:

$$H = -\sum_{j=1}^{k} \frac{1}{k} \log\left(\frac{1}{k}\right) = -\log\left(\frac{1}{k}\right) = \log k.$$

Proprietà dell'entropia di Shannon

- È possibile dimostrare che anche questo indice assume valore massimo nelle situazioni di massima mutabilità, ovvero

$$H \leq \log k,$$

e si ottiene $H = \log k$ se e solo se $f_1 = \cdots f_k = 1/k$.

- Spesso viene definita anche l'**entropia di Shannon normalizzata**, ovvero

$$H_{\text{norm}} = \frac{H}{(\text{massimo valore di } H)} = H/\log k.$$

- Per la variabile **Voto** si ha che $H = 1.2572$ e che $H_{\text{norm}} = 0.5722$.

Le elezioni comunali del 2016

Municipio	G	G_{norm}	H	H_{norm}
Municipio 1	0.610	0.687	1.162	0.529
Municipio 2	0.650	0.732	1.259	0.573
Municipio 3	0.643	0.724	1.254	0.571
Municipio 4	0.645	0.726	1.245	0.567
Municipio 5	0.649	0.730	1.252	0.570
Municipio 6	0.648	0.729	1.253	0.570
Municipio 7	0.648	0.730	1.260	0.573
Municipio 8	0.654	0.735	1.278	0.582
Municipio 9	0.661	0.744	1.285	0.585

- Il Municipio 1, ovvero il centro storico, risulta leggermente meno eterogeneo rispetto agli altri, ovvero più polarizzato.
- Viceversa, il Municipio 9 presenta un comportamento leggermente più eterogeneo.