Statistica I

Unità I: dati qualitativi

Tommaso Rigon

Università Milano-Bicocca

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 del Parlamento Europeo 2024 di Milano sono disponibili sul **sito del Comune di Milano**
- I dati si possono ottenere al link: `https://dati.comune.milano.it/dataset/ds2747_elezioni-del-parlamento-europeo-2024-voti-di-lista-per-sezi`
- Le elezioni si è svolte nei giorni 8-9 Giugno 2024.
- Sono noti i **voti** ricevuti da ciascuna lista in ogni **Municipio**.
- **Nota**. Ci potrebbero 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	Partito Democratico
2	Municipio 4	Partito Democratico
3	Municipio 9	Fratelli d'Italia
⋮	⋮	⋮
505075	Municipio 7	Fratelli d'Italia

- 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 $n = 505075$.

Frequenze assolute e relative

- La tabella della pagina precedente è poco "maneggevole".
- I dati possono essere rappresentati tramite la seguente tabella di **frequenze assolute**.
- Ad esempio, 2898 è il numero di voti ricevuti da Alleanza Verdi e Sinistra nel Municipio 1, ovvero il centro storico di Milano.

	Municipio								
Lista	1	2	3	4	5	6	7	8	9
Alleanza Verdi e Sinistra	2898	6687	6916	5781	4841	6064	5812	6779	7359
Alternativa Popolare	79	109	106	107	87	122	157	161	142
Azione - Siamo Europei	4282	2757	4220	3593	2582	3540	4208	4199	3252
F.I. - Noi Moderati - PPE	4992	4464	5124	4927	3749	4502	5789	5868	5410
Fratelli D'Italia	9194	11034	11897	12458	9737	11808	14787	15313	13509
Lega Salvini Premier	1787	4151	2831	3598	2646	3326	4093	4337	4235
Libertà	120	270	246	335	273	304	374	332	328
Movimento 5 Stelle	939	2885	2553	2994	2637	3194	3463	3766	4141
Pace Terra Dignità	737	1295	1474	1518	1154	1430	1521	1664	1660
Partito Democratico	11832	16882	21067	18868	14072	17655	18536	20236	19358
Rassemblement Valdotain	18	35	51	44	33	37	68	64	50
Stati Uniti D'Europa	4723	2773	4064	3569	2663	3316	3758	4040	3250

Frequenze assolute e relative

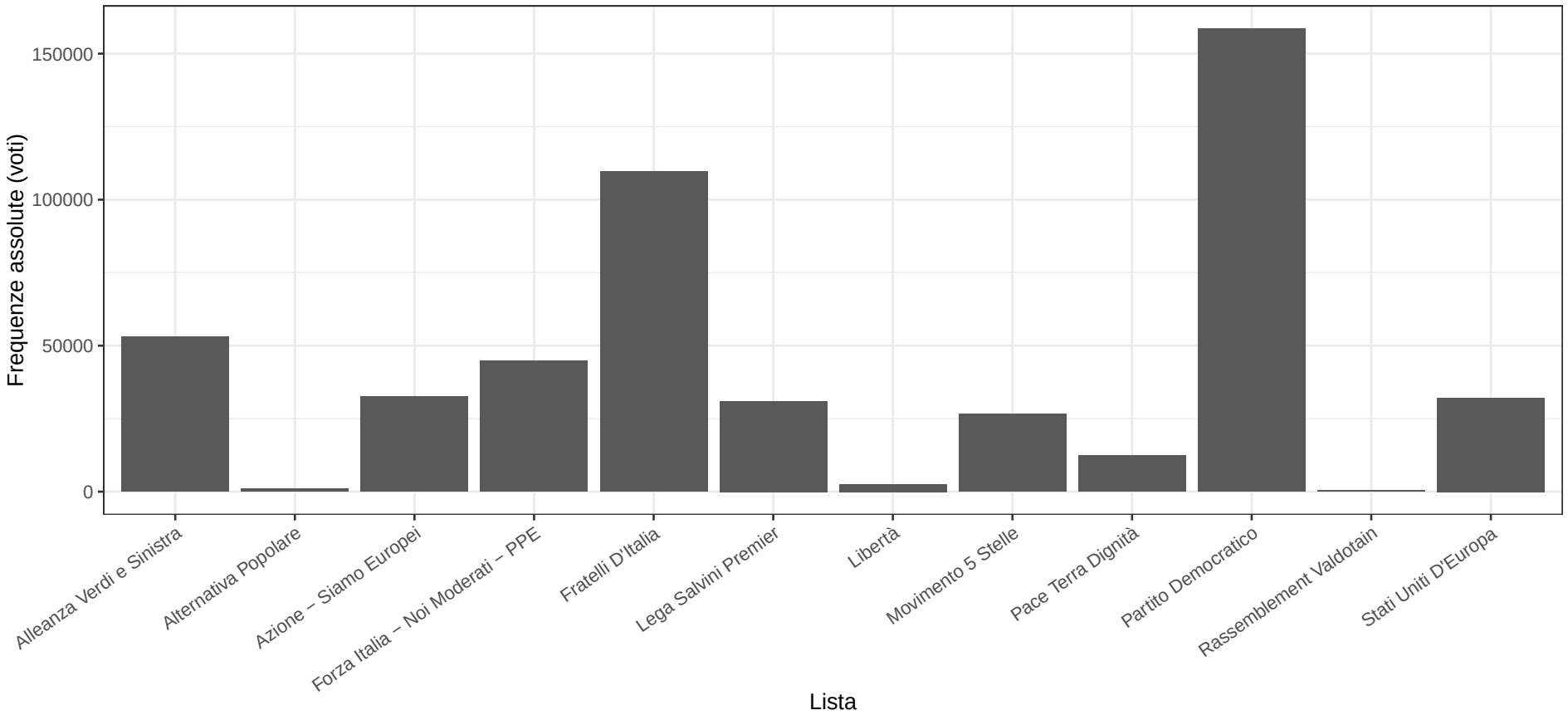
- La variabile **Voto** ha la seguente distribuzione di frequenze.
- Il Partito Democratico e Fratelli d'Italia hanno quindi ricevuto la maggior parte dei voti.

Lista	**Frequenze assolute (Voti)**	**Frequenze relative**
Alleanza Verdi e Sinistra	53137	0.11
Alternativa Popolare	1070	0.00
Azione - Siamo Europei	32633	0.06
Forza Italia - Noi Moderati - PPE	44825	0.09
Fratelli D'Italia	109737	0.22
Lega Salvini Premier	31004	0.06
Libertà	2582	0.01
Movimento 5 Stelle	26572	0.05
Pace Terra Dignità	12453	0.02
Partito Democratico	158506	0.31
Rassemblement Valdotain	400	0.00
Stati Uniti D'Europa	32156	0.06

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



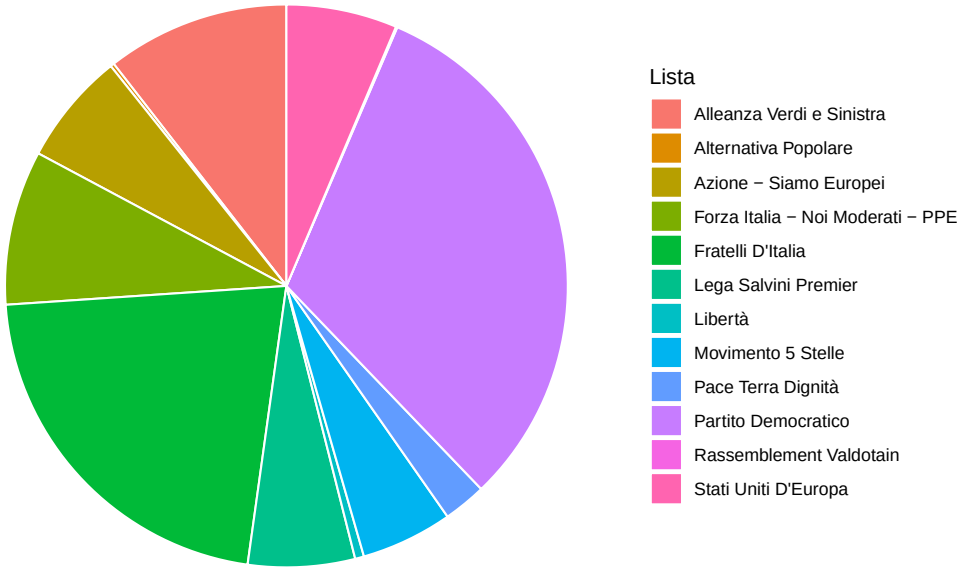
- La rappresentazione grafica più utilizzata è il **diagramma a barre**: ogni modalità è rappresentata da una barra di altezza pari alla frequenza (assoluta o relativa).

- I rettangoli, contrariamente al caso di un istogramma, sono disegnati **staccati**. Inoltre, 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 = Partito Democratico. Infatti, il Partito Democratico ha ricevuto 158506 voti.

- **Nota**. Attenzione a non confondersi: la moda è il Partito Democratico e **NON** la sua frequenza 158506.

- **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

$$G = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(x_i \neq x_j) = 1 - \sum_{j=1}^{k} f_j^2.$$

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

- In primo luogo, si noti che la **media** aritmetica delle frequenze relative è

$$\bar{f} = \frac{1}{k}\sum_{j=1}^{k} f_j = \frac{1}{k}.$$

- Si noti inoltre che la funzione $g(x) = x(1-x)$ è **concava**.

- Pertanto, grazie alla **disuguaglianza di Jensen**, otteniamo che

$$\frac{1}{k}G = \frac{1}{k}\sum_{j=1}^{k} f_j(1-f_j) \leq \bar{f}(1-\bar{f}) = \frac{1}{k}\left(1-\frac{1}{k}\right).$$

- Il risultato segue moltiplicando per k entrambi i lati della precedente disuguaglianza.

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 $G = 0.82$ e che $G_{\text{norm}} = 0.8945$.

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

- **Esercizio - proprietà**. Si dimostri 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 $H = 1.9628$ e che $H_{\text{norm}} = 0.7899$.

Le elezioni del Parlamento Europeo del 2024

- Frequenze relative di voto, all'interno dello stesso municipio.

Lista	Municipio 1	2	3	4	5	6	7	8	9
Alleanza Verdi e Sinistra	0.070	0.125	0.114	0.100	0.109	0.110	0.093	0.102	0.117
Alternativa Popolare	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002	0.002
Azione - Siamo Europei	0.103	0.052	0.070	0.062	0.058	0.064	0.067	0.063	0.052
F.I. - Noi Moderati - PPE	0.120	0.084	0.085	0.085	0.084	0.081	0.093	0.088	0.086
Fratelli D'Italia	0.221	0.207	0.196	0.216	0.219	0.214	0.236	0.229	0.215
Lega Salvini Premier	0.043	0.078	0.047	0.062	0.059	0.060	0.065	0.065	0.068
Libertà	0.003	0.005	0.004	0.006	0.006	0.005	0.006	0.005	0.005
Movimento 5 Stelle	0.023	0.054	0.042	0.052	0.059	0.058	0.055	0.056	0.066
Pace Terra Dignità	0.018	0.024	0.024	0.026	0.026	0.026	0.024	0.025	0.026
Partito Democratico	0.284	0.316	0.348	0.326	0.316	0.319	0.296	0.303	0.309
Rassemblement Valdotain	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Stati Uniti D'Europa	0.114	0.052	0.067	0.062	0.060	0.060	0.060	0.061	0.052

Le elezioni del Parlamento Europeo del 2024

Municipio	G	G_{norm}	H	H_{norm}
Municipio 1	0.825	0.900	1.937	0.779
Municipio 2	0.819	0.894	1.956	0.787
Municipio 3	0.806	0.879	1.917	0.772
Municipio 4	0.815	0.889	1.950	0.785
Municipio 5	0.818	0.893	1.959	0.789
Municipio 6	0.818	0.893	1.961	0.789
Municipio 7	0.823	0.898	1.975	0.795
Municipio 8	0.822	0.897	1.969	0.792
Municipio 9	0.822	0.897	1.968	0.792

- Il Municipio 1, ovvero il centro storico, risulta maggiormente eterogeneo rispetto agli altri, ovvero meno polarizzato.
- Viceversa, il Municipio 3 presenta un comportamento leggermente più polarizzato.