esercitazione 1

la densità di popolazione nelle sezioni censuarie di Feltre e i vari algoritmi di suddivisione in classi

per saperne di più su
- dove si concentra la popolazione di Feltre
- diversi metodi di suddividere i valori in classi

in che modo
- studio della distribuzione della variabile da rappresentare
- tematizzazione delle sezioni di censimento in base alla densità di popolazione per km2

facendo attenzione a
- scegliere tra i diversi algoritmi di suddivisione in classi
- è proprio vero che *"At best a choropleth map is an inappropriate representation of the spatial distribution, and, at worst, it contributes manifestly to the already difficult task of interpretation"* ?

bibliografia
- M. Monmonier, *How to lie with maps*, University Chicago Press, 1996
- J. Krygier, D. Wood, *Making maps: a visual guide to map design for GIS*, The Guilfor Press, 2005 – Google libri
  http://books.google.it/books?id=81aX0AIviv8C&printsec=frontcover&dq=choropleth map classification&source=gbs_slider_thumb - v=onepage&q=choropleth map classification&f=false

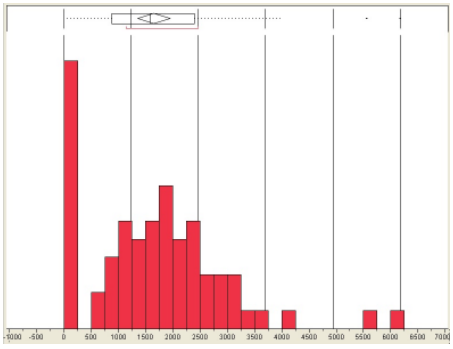
dati a disposizione
- tabelle: Feltre_Istat01.xls
- vettoriali: Feltre_comune_32632.shp, Feltre_strade_32632.shp, Feltre_sezcens_32632.shp
- raster: Feltre_raster_32632.tif
- sistema di riferimento: WGS 84 / UTM Zone 32 N - epsg 32632

software
- excel
- jmp – analisi statistica – trial 30 giorni
- gvsig – gis – free open source

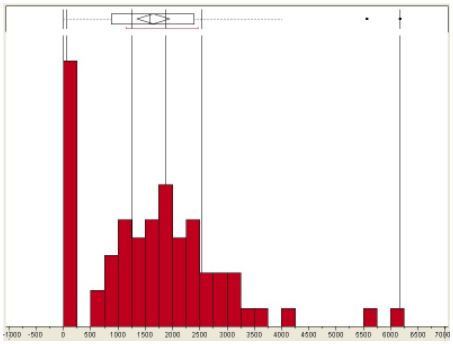
piccoli dettagli utili
- link in cui trovare i codici e i nomi dei comuni italiani http://www.istat.it/strumenti/definizioni/comuni/

gli algoritmi di classificazione

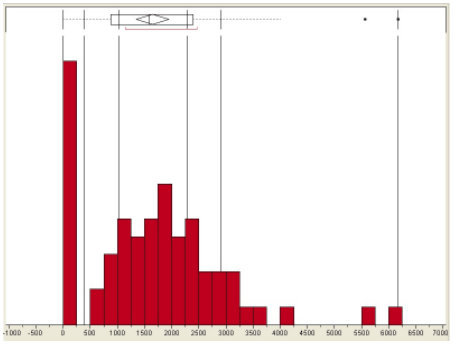
- **Equal interval**: i limiti delle classi sono individuati in modo tale che sia il più possibile simile l'ampiezza delle classi. L'ampiezza delle classi sarà uguale al range dei dati diviso il numero delle classi
  - è appropriato con una distribuzione "rettangolare", piuttosto rara
  - attenzione alle classi vuote



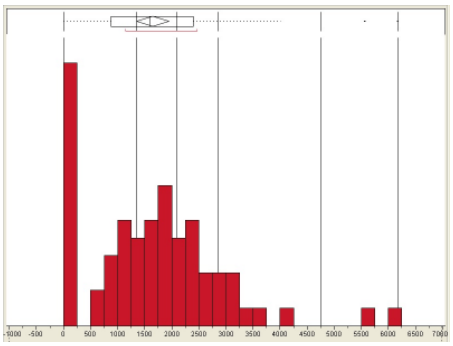
- **Quantile o Equal number**, cioè di uguale numerosità: i limiti delle classi sono definiti in modo da attribuire a ciascuna classe lo stesso numero di poligoni, cioè in modo da suddividere l'universo dei dati in classi di pari numero di unità
  - ogni classe è equamente rappresentata nella mappa
  - ci possono essere classi con ampiezza molto diversa



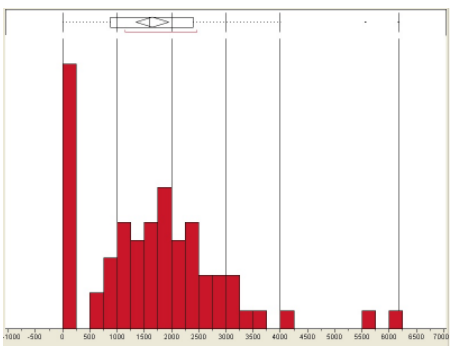
- **Standard deviation**: con questo strumento si analizza la distanza del valore di un oggetto dalla media dei valori. I limiti delle classi sono calcolati a partire dalla media sommando e sottraendo un valore pari alla deviazione standard, al 50% della deviazione standard o al 25% della deviazione standard
  - tiene in considerazione la dispersione dei dati intorno alla media
  - evidenzia gli scostamenti dalla media
  - dovrebbe essere usato quando l'istogramma mostra una distribuzione approssimativamente normale



- **Natural break o interruzioni naturali**: utilizza un algoritmo (jeank's) per minimizzare la somma della varianza interna a ciascuna classe
  - minimizza le differenze interne alle classi
  - enfatizza le differenza tra le classi
  - può essere fatta manualmente, osservando l'istogramma e cercando gli "avvallamenti"



- **Manuale**: seguendo criteri soggettivi, numeri tondi, valori significativi, limiti normativi, …



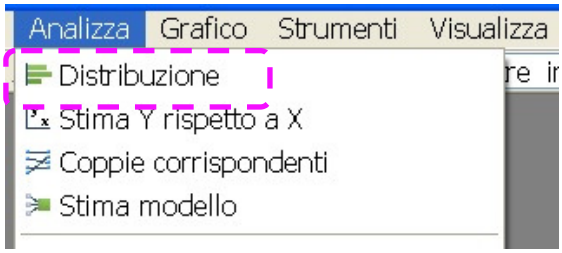
parte 1
studiare la distribuzione della densità abitativa

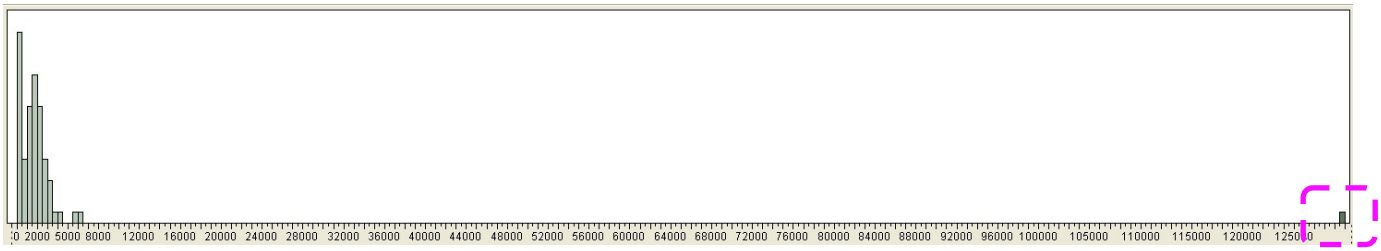
il file **Feltre_sezcens2001_denspop.xls** contiene la popolazione per km2 per sezione di censimento

	A	B	C	D	E
1	area	perimeter	residenti	denspop	sezione
2	23924779,4	27834,0643	0	0	93
3	5229599,13	24926,8759	44	8,41365	101
4	78727,9355	2855,64737	156	1981,50756	54
5	5049158,36	21140,2448	18	3,56495	100
6	81098,6371	2541,88659	179	2207,18875	53
7	13610,2	566,55181	13	955,16598	52
8	94271,5988	2036,42232	244	2588,26627	51
9	154400,903	3159,74511	378	2448,17221	55

apritelo con il programma

calcolate le statistiche sulla distribuzione della variabile **denspop**





Quantili

100,0%	massimo	129820
99,5%		129820
97,5%		18537
90,0%		3123,54
75,0%	quartile	2393,51
50,0%	mediana	1527,67
25,0%	quartile	654,385
10,0%		11,9384
2,5%		0
0,5%		0
0,0%	minimo	0

Momenti

Media	3297,9948
Deviazione standard	14861,227
Errore std della media	1716,0267
LSC della media al 95%	6717,2526
LIC della media al 95%	-121,2629
N	75

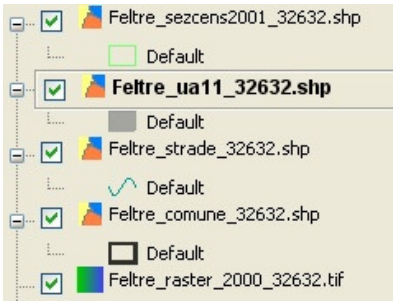
perché c'è un valore così terribilmente anomalo?

che fare con gli zeri?

clickando sopra la barra dell'istogramma all'estrema destra, nella tavola dei dati si evidenzia l'osservazione incriminata, la sezione 888888 che ha un'area di soli 177 metri quadrati, un perimetro di 61 metri e in cui risiedono ben 23 persone … è un errore?

Col 5/3 / Righe 75/1	area	perimeter	residenti	denspop	sezione
59	177,1684	61,2937	23	129819,986	888888

aprite in gvSIG



definite una nuova vista con epsg 32632 e aggiungete i layer

- Feltre_comune_32632.shp confini comunali
- Feltre_sezcens2001_32632.shp confini delle sezioni di censimento
- Feltre_strade_32632.shp rete stradale
- Feltre_ua11_32632.dbf edificato residenziale
- Feltre_raster_2000_32632.tif immagine del 2000

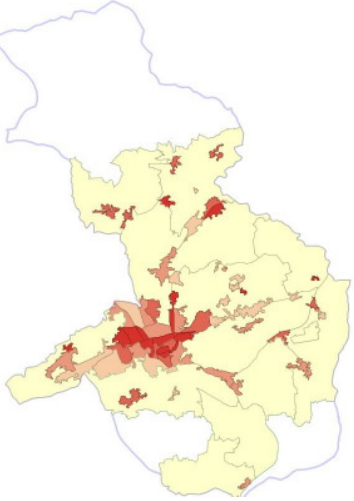
visualizzate la sezione 888888

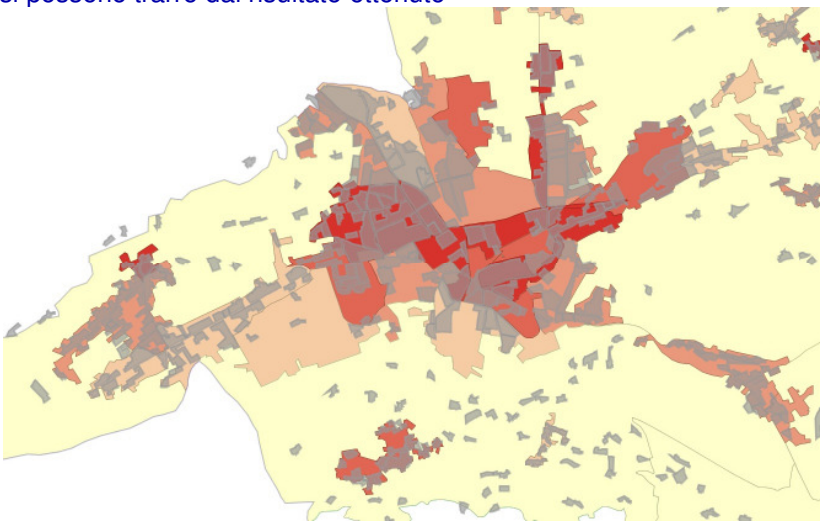
si tratta di persone iscritte all'anagrafe del comune, ma senza abitazione (di solito apolidi o nomadi), che risiedono per convenzione presso la casa comunale

ricalcolate istogramma e statistiche dopo aver escluso la sezione 888888 e valutato cosa fare con i valori pari a zero

scegliete quale potrebbe essere l'algoritmo appropriato per la suddivisione in classi e provatelo

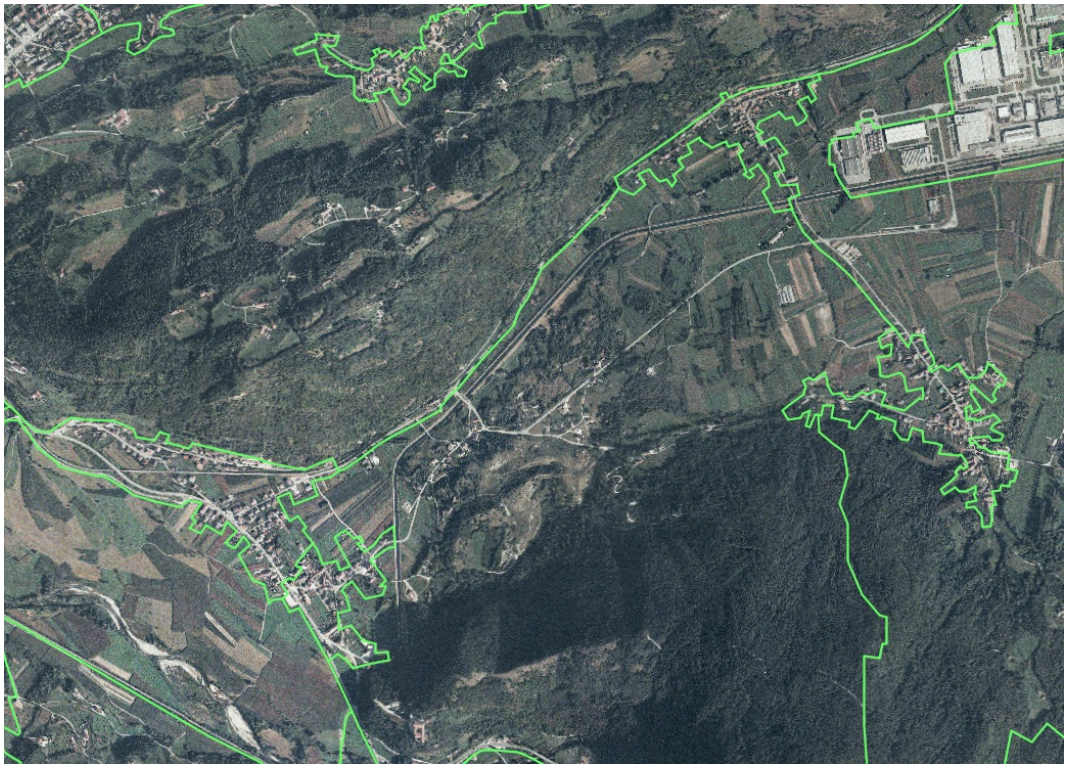
poi ragionerete su quali informazioni / conoscenze si possono trarre dal risultato ottenuto





è convincente la tematizzazione? che cosa può minare la sua significatività?
osservate il rapporto tra le strade e i confini delle sezioni; aiuta a definire sezioni omogenee al loro interno?

osservate il legame tra la copertura del suolo con edificato residenziale e i confini delle sezioni:
aiuta a definire sezioni che ha senso confrontare in termini socio-economici?



che ruolo giocano l'area della sezione e la sua densità abitativa?
una diversa suddivisione del comune avrebbe portato a risultati diversi?

http://jratcliffe.net/

The ecological fallacy

The Ecological Fallacy is a situation that can occur when a researcher or analyst makes an **inference about an individual based on aggregate data for a group**. For example, a researcher might examine the aggregate data on income for a neighbourhood of a city, and discoverer that the average household income for the residents of that area is $30,000.

To state that the average income for residents of that area is $30,000 is true and accurate. No problem there. The ecological fallacy can occur when the researcher then states, based on this data, that people living in the area earn about $30,000. This may not be true at all, and may be an ecological fallacy.

Close examination of the neighbourhood might discover that the neighbourhood is actually composed of two housing estates, one of a lower socio-economic group of residents, and one of a higher socio-economic group. The poorer part of town residents earn on average $10,000 while the more affluent citizens can average $50,000. When the researcher stating that individuals who live in the area earn $30,000 (the mean rate) this did not account for the fact that the average in this example is constructed of two disparate groups, and it is likely that not one person earns $30,000.

Assumptions made about individuals based on aggregate data are vulnerable to the ecological fallacy.

This does not mean that identifying associations between aggregate figures is necessarily defective, and it doesn't necessarily mean that any inferences drawn about associations between the characteristics of an aggregate population and the characteristics of sub-units within the population are absolutely wrong either. What it does say is that the process of aggregating or disaggregating data may conceal the variations that are not visible at the larger aggregate level, and researchers, analysts and crime mappers should be careful.

http://www.geog.ubc.ca/courses/geog570/talks_2001/scale_maup.html

Shifting Boundaries, Shifting Results: The Modifiable Areal Unit Problem (MAUP)

Lisa Oliver Geog 516 2001

Openshaw and Alvanides (1999) warn that GIS users need to seriously consider how the zones of analysis effect results. If relations between variables change with the selection of different areal units, the reliability of results is called into question. The effect of the selection of areal units on analysis, is termed the modifiable areal unit problem (MAUP), it is formally defined as: "**a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns** (1998 Heywood)."

The MAUP had been most prominent in the analysis of socio-economic and epidemiological data (see 1999 Wong, Lasus, Falk; 2000 Nakaya; 1999 Openshaw and Alvandies). Such areal data cannot be measured at a single point, but must be contained within a boundary to be meaningful. For example, it is not possible to measure the percent of low birth-weight babies at a single point, this percentage must be calculated within a defined area. It is the selection of these artificial boundaries and their use in analysis that produces the MAUP.

The effects of the MAUP can be divided into two components: the scale effect and the zonation effect (1995 Armhein). The scale effect is the variation in numerical results that occurs due to the number of zones used in an analysis. For example, the difference in numerical results between mortality rates by municipality and health area in British Columbia is a scale effect. The zonation effect is the variation in numerical results arising from the grouping of small areas into larger units. For example, using Canada census data, numerical differences in employment rates between a census tract data and its enumeration area would be a zonation effect.

It is necessary to understand the ways in which the MAUP effects the results of statistical analysis. Caution, however, is required, as there is a random aspect to the effects of the MAUP. It may be difficult to generalise how different data sets with different spatial units are effected by the MAUP. This caution aside, the use of small areal units has a tendency to provide unreliable rates because the population used to calculate the rate is smaller. On the other hand, using larger areal units will provide more stable rates but may mask meaningful geographic variation evident with smaller areal units (Nakaya 2000). Choosing between the scale of zones depends upon the particular use and requirements of the data.

Armheim (1995) in his study of a region, aggregated data into three scales of analysis and calculated mean, variance, regression and correlation coefficients. This study found that regression correlation coefficients increase as data is aggregated into larger spatial units. Intuitively, this makes sense because unstable rates tend to be averaged when aggregated into larger units, thereby increasing correlation coefficients. According to Armheim (1995) mean and variance are more stable as areal units of analysis change. Example 1 shows the change in means that occurs when smaller units are aggregated into larger units. Given the susceptibility of statistical results to change at different scales of analysis ecological fallacy needs to be considered. For example, it is unlikely that an increase in correlation with larger areal units reflects stronger correlation at the individual level.

These few paragraphs have discussed the MAUP and its effects on analysis. It has, however, avoided discussing solutions to the problem. There are two reasons for this omission: (1) researchers have only begun to unpack the effects of the MAUP on analysis and; (2) few generic and practical solutions exist. The weighing areal units by population, as well as complex statistical procedures are currently being researched to address the MAUP. A simple strategy to deal with the problem, is to undertaking analysis at multiple scales or zones. Despite the lack of solutions, **being cognisant of the fact that analysis results may be dependent on the zones used to aggregate data is an important step**.

Per approfondire:

P. Longley, M. Batty Ed., Spatial analysis: modelling in a GIS environment, Wiley, 1996

Part 1 – Analysis of spatial distributions

Google books

http://books.google.it/books?id=1DCq7fg_0QcC&printsec=frontcover&dq=longley+professor&source=bl&ots=og5NB23Jbs&sig=WU_nUAovM6XD0XaFZlGastpkdxA&hl=it&ei=h2YfTNrYB4XfsAa6l6H1DQ&sa=X&oi=book_result&ct=result&resnum=2&ved=0CBoQ6AEwAQ - v=onepage&q=longley pro