Clustering interval data: an application in the epidemiologic field

Luisa Canal, Rocco Micciolo

Department of Cognitive Sciences and Education,
University of Trento, Italy

Corresponding Author:
Luisa Canal
Department of Cognitive Sciences and Education,
University of Trento
Corso Bettini, 31 - 38068 Rovereto, Italy
e-mail: luisa.canal@unitn.it

Summary
Information collected in biological and medical sciences usually consists of single-valued numerical variables. However many phenomena are measured by intervals and statistical approaches have been proposed to deal with this kind of data. In this paper an algebra for interval data based on the extension to an interval of real differentiable functions is employed. Using two scalar indices associated with an interval, involving a free parameter which tunes up the uncertainty in the data, it is possible to obtain a ranking of a set of intervals. A cluster analysis on intervals based on the incidence rates of cutaneous melanoma in the districts of an Italian province during a 15-year period is proposed.

KEY WORDS: *clustering; interval data; uncertainty*

Introduction

Empirical studies usually analyze single-valued numerical variables. However interval-data naturally arise in different circumstances. Sometimes intervals are generated in the framework of repeated measurements when one variable can be coded in an interval using the lowest and the highest registered measure (for example in air pollution data). In other cases intervals arise from a couple of variables complementary with respect to a given concept (expected and perceived healthiness). Imprecise measurements also generate interval data through the error associated with a physical measure or the uncertainty associated with sampling from a given population. Historically, an algebra for interval data was introduced to deal with computer representation of real numbers with floating point numbers and more recently some statistical approaches to deal with interval data have been proposed (Diday, 1996; Hickey et al., 2001; Lauro e Palumbo, 2000). In this paper an alternative approach will be presented, based on a numerical structure $X = \langle x, \varepsilon \rangle$ where x is the centre and ε is the spread.

Theoretical framework

The real interval $[x - \varepsilon, x + \varepsilon]$ can be represented as $X = \langle x, \varepsilon \rangle$ where x is the centre and ε is the spread. From such intervals a generalized numerical space arises whose algebraic structure is based on the extension to X of the real differentiable functions $f(x)$, $f(x,y)$. The approach followed by Canal and Marques Pereira (1998) will be briefly summarized.
The extended function $f(X)$ is constructed on the basis of the linear approximation of $f(x)$. The interval algebra constitutes a first order approximation of *interval analysis* (Moore, 1966).

The linear approximation of the function *f* for the interval X is defined as

$$f(x+\phi) \approx f(x) + f'(x)\phi \qquad \phi \in [-\varepsilon, \varepsilon]$$

and is graphically displayed in figure 1.

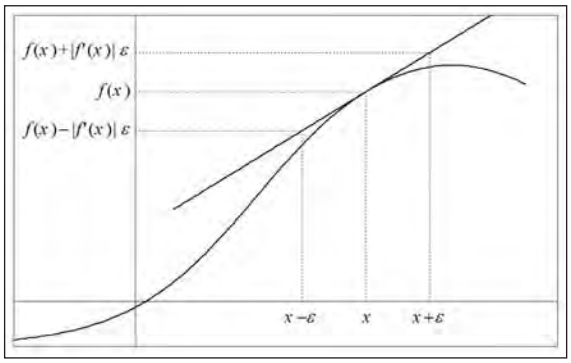

Figure 1. Linear approximation for the interval X.

The function f of the interval X, i.e. the set of the images of X obtained employing the linear approximation of the function *f*, is

$$f(X) = \left\langle f(x), \max_{\phi}\left(f'(x)\phi\right)\right\rangle \qquad \phi \in [-\varepsilon, \varepsilon]$$

Therefore $f(X) = \left\langle f(x), \left|f'(x)\right|\varepsilon \right\rangle$.

In such a way it is possible to define the product between a real number λ and the interval X

$$\lambda X = \langle \lambda x, |\lambda|\varepsilon \rangle$$

and the *n*-th power of the interval X

$$X^n = \langle x^n, n|x^{n-1}|\varepsilon \rangle$$

The extension to a vector $\mathbf{X}$ of intervals is straightforward. In particular, given two intervals $X = \langle x,\varepsilon \rangle$ and $Y = \langle y, \vartheta \rangle$, the definitions of sum, difference, product and ratio are:

$$X + Y = \langle x + y, \varepsilon + \vartheta \rangle;$$
$$X - Y = \langle x - y, \varepsilon + \vartheta \rangle;$$
$$XY = \langle xy, |y|\varepsilon + |x|\vartheta \rangle;$$
$$X/Y = \left\langle x/y, (1/|y|)\varepsilon + \left(|x|/y^2\right)\vartheta \right\rangle.$$

It can be seen that neither an additive inverse nor a multiplicative inverse exist.

Two scalar indices, the *rank* and the *modulus*, are associated with X, both involving a free parameter β which tunes up the uncertainty in the data. The *rank* of X is defined as

$$r(X) \equiv \frac{1}{1 + \exp\left(-\beta\left(x^2 - \varepsilon^2\right)\right)}; \quad \beta \geq 0$$

and is a scalar measure (between 0 and 1) of the extent to which the interval X can be considered separate from the real number zero.

If zero coincides with one of the interval extremes then the rank is 0.5; if zero lies in X then the rank is lower than 0.5; if zero does not lie in X then the rank is higher than 0.5.

For high values of the parameter β the rank polarizes towards the values zero (if zero lies in X) and one (if zero does not lie in X).

The *modulus* is defined via its square, which is a linear combination of the squares of the centre and of the spread:

$$\lfloor X \rfloor^2 \equiv r(X) \times x^2 + (1 - r(X)) \times \varepsilon^2$$

The modulus is a scalar measure of the difference between the interval X and the *null interval* (0,0). It plays a similar rôle to that of the absolute value for standard real numbers. In the modulus, x^2 is dominant if zero does not lie in X and ε^2 is dominant if zero lies in X. For high values of the parameter β, the square of the modulus coincides with x^2 if $\varepsilon^2 < x^2$ and with ε^2 if $\varepsilon^2 > x^2$.

An interesting result connects the parameter β and the amount of uncertainty in the data: the modulus of X is monotonically increasing in $|x|$ and monotonically decreasing in ε if $\beta \leq 1/\varepsilon^2$.

When considering a vector $\mathbf{X}$ of intervals, a key rôle is played by the *norm*, which is defined (via its square) as

$$\|\mathbf{X}\|^2 \equiv \left\|\langle \mathbf{x}, \boldsymbol{\varepsilon} \rangle\right\|^2 = \left\langle \|\mathbf{x}\|^2, 2\left|\mathbf{x}^T\right|\boldsymbol{\varepsilon} \right\rangle$$

The (square of the) norm of the difference between two vectors $\mathbf{X}$ and $\mathbf{Y}$ of intervals is the extension of the Euclidean distance

$$\|\mathbf{X} - \mathbf{Y}\|^2 \equiv \left\langle \|\mathbf{x} - \mathbf{y}\|^2, 2\left|\mathbf{x} - \mathbf{y}\right|^{\mathrm{T}} (\boldsymbol{\varepsilon} + \boldsymbol{\vartheta}) \right\rangle.$$

Data application from epidemiologic field

The analysis of the incidence of different illnesses is a major topic in the epidemiologic field in particular with reference to geographic areas.

In this paper we analyze data on incidence cases of cutaneous melanoma retrieved from the Skin Cancer Registry of the Province of Trento (Italy) between 1992 and 2006. The description of the criteria for collection, registration and analysis fol-

lowed by the registry has been presented elsewhere (Boi et al., 2003). The province of Trento has about 470,000 inhabitants (density 76 inhabitants per km^2) subdivided into 11 districts. Amongst them, 2 districts have together a population that is about one half that of the entire Province and the population of the remaining 9 districts accounts for between 2% and 9% of the total.

In this paper we analyze the distribution of incidence rates for cutaneous melanoma searching for a suitable aggregation of districts with similar rates. Starting with year, sex and age specific incidence rates a direct standardized figure was calculated for each district using the yearly mid-year population of the entire Province as standard.

Indicating year (1992-2006), sex and age (20 classes: 0-4, 5-9, …, 90-94, 95 and over) respectively with k (1,2,…,15), i (1,2) and j (1,2,…,20), the corresponding incidence rate for the district l was ${}_l\lambda_{kij}$. The corresponding proportion of subjects in the standard population was w_{kij}, so that the (year, sex and age–)standardized incidence rate is given by

$$ {}_l\Lambda = \sum_{k=1}^{15}\sum_{i=1}^{2}\sum_{j=1}^{20} {}_l\lambda_{kij} w_{kij} . $$

Following Breslow & Day (1987), the standard error of the directly standardized rate was estimated by

$$ \mathrm{SE}({}_l\Lambda) = \left(\sum_{k=1}^{15}\sum_{i=1}^{2}\sum_{j=1}^{20} \frac{{}_l\lambda_{kij} w_{kij}}{{}_l n_{kij}} \right)^{1/2} , $$

where ${}_l n_{kij}$ refers to the mid-year population living in the district l.

To each district l an interval $X_l = \langle x_l, \varepsilon_l \rangle$ was associated, where x_l was the standardized incidence rate (${}_l\Lambda$) and ε_l was the corresponding standard error (SE(${}_l\Lambda$)) calculated as described above.

Districts were aggregated two by two employing the following two-steps procedure:

Step 1 – for each couple of districts l and m were calculated (i) the squared norm of the difference between the corresponding intervals

$$ \|X_l - X_m\|^2 \equiv \left\langle |x_l - x_m|^2 , \; 2|x_l - x_m|(\varepsilon_l + \varepsilon_m) \right\rangle $$

and (ii) the modulus of the squared norm

$$ M_{lm} = \left\lfloor \|X_l - X_m\|^2 \right\rfloor $$

employing for the parameter β the reciprocal of the squared spread of $\|X_l - X_m\|^2$.

Step 2 – The two districts with the lowest modulus were aggregated, i.e. the two districts were considered as a unique district (summing up the corresponding population and the number of incident cases of cutaneous melanoma) for which (year, sex and age) standardized incidence rates were calculated.

These two steps were repeated 10 times until the 11 districts were aggregated all together. All the analyses were performed employing R (R Development Core Team, 2008).

Results

During the 15 year period considered a total of 943 cases of cutaneous melanoma were diagnosed (ranging from 46 to 83 yearly cases without any evident trend). Since the average population of the province of Trento was 474,223 inhabitants, a raw incidence figure was 13.3 yearly cases/100,000 inhabitants. Table 1 shows the standardized incidence rates within the 11 districts together with the corresponding standard errors.

Table 1. Standardized incidence rates (per 100,000 inhabitants) of cutaneous melanoma in the 11 districts of the Trento Province (1992-2006).

District	Directly Standardized Rate	Standard Error
1	7.02	0.314
2	11.95	0.410
3	13.45	0.435
4	11.92	0.409
5	15.31	0.464
6	10.65	0.387
7	8.47	0.345
8	9.57	0.367
9	13.48	0.435
10	14.96	0.459
11	10.66	0.387

For each district an interval was defined where the standardized incidence rate was the centre and the standard error was the spread.

When the aggregation procedure was run the results summarized in table 2 were found.

Step by step results are reported in the rows of table 2. The districts 6 and 11 were aggregated firstly (see the first row of table 2). The intervals were $X_6 = \langle 10.65, 0.387 \rangle$ and $X_{11} = \langle 10.66, 0.387 \rangle$; the square of the norm of their difference was

Table 2. Results of the aggregation procedure of the 11 districts of the Trento Province.

Step number	Districts aggregated		Squared Norm Center	Spread	Beta	Rank	Modulus
1	6	11	1.011E-04	1.557E-02	4.126E+03	0.2689	0.115
2	3	9	7.262E-04	4.690E-02	4.546E+02	0.2690	0.200
3	2	4	1.006E-03	5.197E-02	3.702E+02	0.2690	0.211
4	5	10	1.190E-01	6.365E-01	2.469E+00	0.2759	0.738
5	7	8	1.224E+00	1.575E+00	4.031E-01	0.4022	1.202
6	(2,4)	(6,11)	1.566E+00	1.998E+00	2.504E-01	0.4047	1.355
7	(3,9)	(5,10)	2.649E+00	2.929E+00	1.166E-01	0.4547	1.675
8	(7,8)	(2,4,6,11)	4.775E+00	3.325E+00	9.045E-02	0.7432	2.109
9	1	(2,4,6,7,8,11)	1.335E+01	5.126E+00	3.805E-02	0.9969	3.651
10	(3,9,5,10)	(1,2,4,6,7,8,11)	2.121E+01	7.716E+00	2.120E+01	0.9986	4.604

$||X_6 - X_{11}||^2 = \langle 1.01 \cdot 10^{-4}, 1.56 \cdot 10^{-2} \rangle$ (see columns 4 and 5 of table 2). These two districts were aggregated since $||X_6 - X_{11}||^2$ had the lowest modulus (see column 8 of table 2); the rank employed to calculate this modulus (see column 7 of table 2) gives much more importance to the spread with respect to the centre and is approaching the limiting value $1/(1+e)$ implied by the choice for β. Quite similar values for the rank were found in the first four steps. Figure 2 shows the dendrogram reporting the results of the aggregation procedure. There is some evidence for the presence of three clusters with different incidence rates. The cluster with the highest incidence (14.9; SE: 0.46) is an aggregation of 4 districts (3, 5, 9, 10); another cluster of six districts (2, 4, 6, 7, 8, 11) shows an intermediate incidence (10.7; SE: 0.39). The remaining district which showed the lowest incidence rate among all the 11 districts (7.0; SE: 0.31) appears separate from the other two aggregations. It

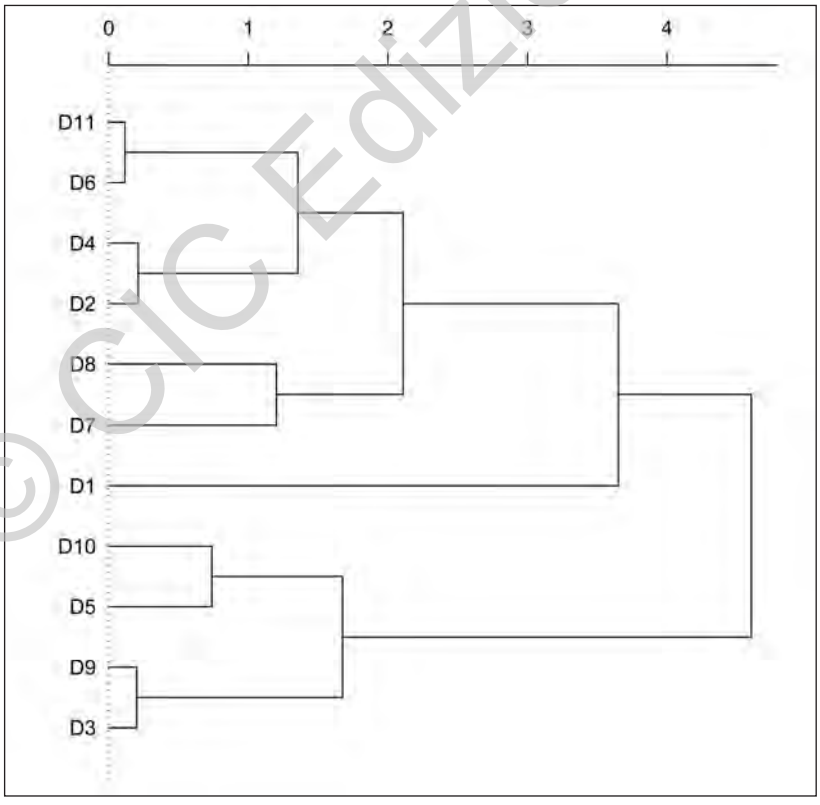

Figure 2. Dendrogram showing the aggregation of the 11 districts considered (D1, …, D11) together with the values of the modulus of the squared norms of the difference between standardized incidence rates for cutaneous melanoma of the two districts to be aggregated.

is interesting to note that while the aggregation proceeds, the rank increases, giving a progressively higher weight to the centre.

Conclusions

Interval data naturally arise in many contexts where uncertainty is present. Here we concentrated on cases where uncertainty is linked to a precision measure of a point estimate. The novelty of this approach consists in the definition of two scalar indices associated with an interval which permits a ranking of the intervals.

These characteristics were used to perform an analysis of incidence rates of cutaneous melanoma in the districts of an Italian province. We were able to take into account both standardized incidence rates as well as their standard errors in evaluating the similarities between districts. Uncertainty was more important in the first steps of the procedure while the last steps were based almost only on point estimates. Such a result is to be expected, since aggregation creates clusters of increasing size, but the approach presented makes it numerically evident by quantifying the "importance" of the uncertainty.

References

1. Boi, S., Cristofolini, M., Micciolo, R., Polla, E., Dalla Palma, P., 2003. Epidemiology of Skin Tumors: Data from the Cutaneous Cancer Registry in Trentino, Italy. J. Cutan. Med. Surg. 7, 300-305.
2. Breslow, N.E., Day, N.E., 1987. Statistical Methods in Cancer Research. Volume II. The Design and Analysis of Cohort Studies. IARC, Lyon.
3. Canal, L., Marques Pereira, R.A., 1998. Towards statistical indices for numeroid data. Proceedings "NTTS'98. International Seminar on New Techniques and Technologies for Statistics". Napoli: Studio Idea, 97–102.
4. Diday, E., 1996. Une introduction á l'analyse des données symboliques. SFC, Vannes, France.
5. Hickey, T., Ju, Q., Van Emden, M.H., 2001. Interval arithmetic: from principles to implementation. Journal of the ACM. 48, 1038-1068.
6. Lauro, C.N., Palumbo, F., 2000. Principal component analysis of interval data: a symbolic data analysis approach. Comput. Stat., 15, 73-87.
7. Moore, R.E., 1966. Interval analysis. Prentice-Hall, Englewood Cliffs NJ.
8. R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-044.