The method of increasing proportions

Ulisse Di Corpo[1]

Abstract

Predictions based on the classical method of multiple regressions, are based on the assumption that relations among variables are linear or can be translated in the linear form. But data show that this assumption is wrong and that in the real world relations are usually non-linear and do not satisfy the linear assumption at the base of multiple regression. In this paper a different methodology which allows to device highly predictive non linear models is proposed.

Forward

When we want to predict the values of geographical micro-zones (for example towns) starting from macro-zone data (for example regions) the method of multiple regressions is generally used. But when we verify the results, summing the quantitative data obtained at the local level into the macro-zone data, big differences are observed, and we discover that the multiple regression method does not allow to produce satisfactory prediction models.

[1] ulisse.dicorpo@gmail.com

This is the consequence of:

- The complex nature of variables: for example, in the Italian case, correlations in South Italy are usually totally different from those within North Italy. Linear correlation techniques are not able to handle this complexity.
- Often obvious intervening variables correlate with all the set of variables, for example the distinction between North and South Italy; the strong correlation observed among variables are often due to these intervening variables.
- When using few macro-units (for example the 20 Italian regions) statistical significance is very approximate and is not able to distinguish among predictive correlations and correlations due to an intervening variable (for example North/South).

The increasing proportion method

A different method was tested in order to study non linear relations and discriminate correlations on the basis of extremely high statistical significance.

This method functions in the following way:

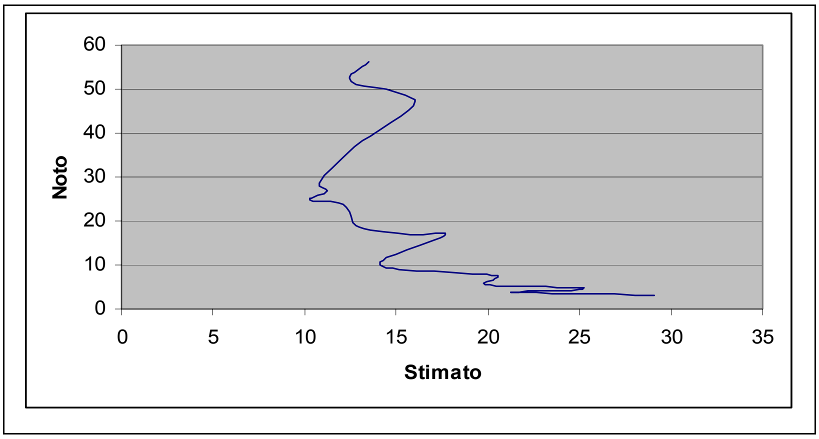
1. Each ratio is ordered from lower to higher values.
2. Proportions among the ratio and the ratio which has to be predicted are calculated.
3. Only those ratios which show increasing proportions are selected.

In the following example we see that, when the ratio increases also the proportion with the dependent ratio, which we want to predict, increases.

Tab. 1 – Increasing proportions

Ratio known only for macro-zones	Known ratio	Proportion
29.1	3.2209347572312	0.1106850432038
21.4	3.7472576341536	0.1751054969231
24.0	4.2306504791205	0.1762771032967
25.2	4.6657799308346	0.1851499972553
20.8	5.2952271608988	0.2545782288894
19.8	5.9854907244274	0.3022975113347
20.5	7.6421077174166	0.3727857423130
14.1	10.1008595988539	0.7163730211953
17.7	16.9204114240064	0.9559554476840
15.8	16.9357366771160	1.0718820681719
13.0	18.7267234701782	1.4405171900137
12.3	23.0345854828162	1.8727305270582
11.4	24.3379344343111	2.1349065293255
10.3	24.7761336866902	2.4054498724942
11.2	26.9976775283271	2.4105069221721
10.8	28.7036518210801	2.6577455389889
12.7	36.7435590173757	2.8931936234154
16.0	47.1063829787234	2.9441489361702
12.5	51.7574171029668	4.1405933682373
13.5	56.3550834597876	4.1744506266509

The probability that the proportion increases moving from line 2 to line 3 is ½, the probability that the proportion continues to increase moving from line 3 to line 4 is ¼. The total probability that the proportions increase moving from line 2 to line 4 is therefore ½ x ¼ ($1/2^{2!}$). The probability of 19 increasing proportions (as in this table) is therefore $1/2^{19!}$, which means 1 probability every 12,164,510 billion cases, practically impossible by pure chance.

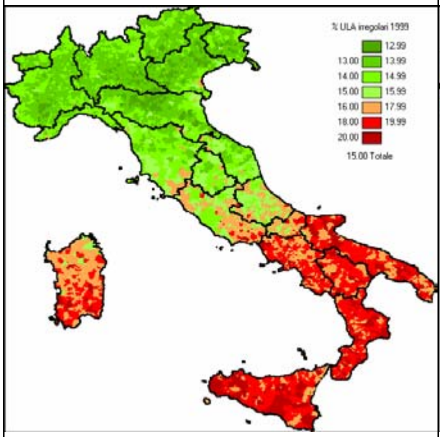
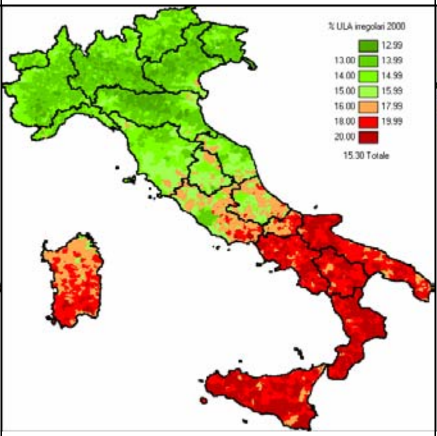
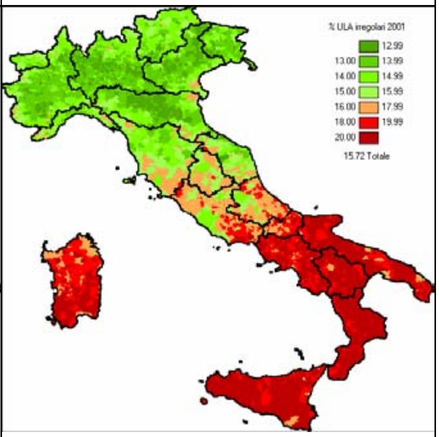


If we plot the values of the table (known ratio and dependent ratio which we want to predict), we observe that the relation is non-linear. The relation corresponds therefore to a function of which we have only a limited number of points. Points which can be used to assess the intermediate values of the function.

An example

The method of the increasing proportions has been used to predict irregular work at the local level starting from data which is known only at the regional level. Using the method of the increasing proportions, 13 indicators produced from the census data obtained 19 increasing proportions. These 13 indicators have been used to predict the local value of irregular work. In this example the quantitative value of irregular work was obtained and it was then summed to obtain the regional values already known. The difference observed is 14.54% in 1999, 13.00% in 2000 and 12,54% in 2001.

In order to reduce this error, local values were corrected according to the regional known values, in such a way that the sum of local values coincided with the known regional values. The result is the following:

Stima su base comunale del lavoro irregolare		
1999	**2000**	**2001**
		

Anlysis produced using sintropia-AS http://www.sintropia.it/ricerca/landstat/landstat.htm

Comparison with a known variable (Active population 1991)

In order to verify this method a variable known at the local and macro level was used (Active Population).

	Prediction	Real data	Difference
Piemonte	1.810.256	1.915.651	- 5.50%
Valle d'Aosta	51.118	52.712	- 3.02%
Lombardia	3.879.296	4.020.360	- 3.51%
Trentino Alto Adige	397.782	392.729	- 1.29%
Veneto	1.927.250	1.936.915	- 0.50%
Friuli Venezia Giulia	505.731	509.894	- 0.82%
Liguria	673.259	673.315	- 0.01%
Emilia Romagna	1.707.527	1.814.770	- 5.91%
Toscana	1.488.400	1.543.354	- 3.56%
Umbria	347.504	336.412	+ 3.30%
Marche	620.276	626.172	- 0.94%
Lazio	2.183.445	2.168.728	+ 0.68%
Abruzzo	525.869	502.429	+ 4.67%
Molise	132.778	132.390	+ 0.29%
Campania	2.353.818	2.197.869	+ 7.10%
Puglia	1.648.401	1.562.468	+ 5.50%
Basilicata	245.075	245.622	- 0.22%
Calabria	823.986	800.200	+ 2.97%
Sicilia	2.041.341	1.829.059	+11.61%
Sardegna	675.876	663.589	+ 1.85%

A big part of these differences can be explained by the fact that extreme values are predicted using the last known point of the function. This choice cuts extreme values (high and low values) and the effect is that positive differences are observed in those territories where the value is low (South Italy) and negative differences are observed in those territories where the value is high (North Italy).

Comparison of towns with low and high values: the example of Liguria

The region of Liguria was chosen for this comparison, as it obtained the lowest differences in the previous table. In the following table it is possible to observe the contraction of extreme values towards the central values:

	Real data	Predicted data
Genova	270.602	262.771
La Spezia	40.576	39.346
Savona	26.973	25.710
San Remo	24.221	21.440
Propata	51	65
Fascia	48	51
Montegrosso Pian Latte	35	59
Rondanina	34	40

Conclusion

This methodology allows to predict in a very accurate way analytical data which is otherwise impossible to predict using classical multiple regression methods.

However, it is important to remember that numbers have to be handled as useful suggestions and that they should not be confused with real facts, which always require deeper assessment studies and research work.