

Cluster and outlier analysis with GIS: an exploratory study on childbearing in Italy¹

Massimo Mucciardi¹, Pietro Bertuccelli²

¹ Department D.E.S.Ma.S. “V. Pareto”, University of Messina

² Department S.E.Fi.S.A.S.T., University of Messina

e-mail: mucciard@unime.it, pbertuccelli@unime.it

Abstract: The aim of this work is to conduct a socio-economic exploratory study on childbearing in Italy by using the mapping and analytical capabilities of GIS.

Keywords: GIS, Childbearing, Global and Local Moran's I

1. Using GIS in socio-economic and demographic analysis

In the last decade the application of GIS as a tool to investigate the territory has grown exponentially. The use of GIS in the fields of geology or of earth sciences in general has become very common, if not necessary, to better understand a great number of phenomena. In particular, if we think to natural disasters such as, for example, earthquakes or volcanic eruptions, we can understand the utility of GIS in modeling these kinds of events in order to try to take prevention measures. However, the ability of GIS to map observations that at least can be referred to a specific point or portion of the territory has appealed social and economic sciences which, in turn, have seen in it a huge potential. In fact, the capability of GIS to represent, through thematic maps, the results of studies concerning a certain territory, has led social and economic scientists to approach socio-economic problems in a new way; moreover, the use of maps can help, not only common people, which usually are not familiar with these kind of arguments, to understand some aspects of society that can involve their way of life, but also policy makers to take better decisions. These reasons, in conjunction with the large variety of analytical tools existent in GIS, led scientific community to investigate social and economic research areas from the spatial perspective: we can see, to this purpose, the flourishing of studies on demography, health care, crime analysis, economy. The good results provided by GIS literature in these fields of knowledge and the potential of studying demographic and economic aspects of society in such a different way, compared to classic statistical analysis, led us to apply GIS technology to our research. Having outlined the general picture of GIS application, the aim of this paper is to spatially analyze some of the socio-economic determinants of childbearing in Italy in the last few years.

2. Childbearing: some considerations

Analyzing the determinants of childbearing from the spatial and socio-economic point of view is very important and could be useful to explain some of the society dynamics.

¹The present paper is financially supported from MURST project (2007).

The idea of having or not having children could be heavily influenced by the social and economic status of a couple. Without a doubt, several variables are very important in this respect: incomes, costs, educational level, etc. However, a fundamental role seems occupied by female participation in the labor market and, in general, by the degree of emancipation obtained by women. To this purpose, we can read the work of Di Pino and Mucciardi (2003) on the opportunity-costs of children in Italy, where the influence of women's working condition on childbearing is clearly evidenced. Another important factor to consider is the level of education. It comes out to be important for men (generally to a higher level of education corresponds high social positions and earnings), but, as assessed by Dankmeyer (1996), it influences positively women's work opportunity and expected earnings. Other determinants that seem to have a strong influence on childbearing are typically of social nature: number of brothers and sisters of the heads of family (husband and wife), parents of the heads of family living in the same house. It could depend on cultural substratum of family members; traditions and religion, for example, are often crucial elements explaining people's behaviour. So, in consideration of all these possible social and economic implications, we decided to study the childbearing in Italy from a spatial perspective. In the current state of the research we present a regional exploratory analysis of the relation of childbearing with some socio-economic variables.

3. Materials and methods

The choice about the kind of data to use to effectuate the research was quite delicate. We needed to find a source of data that could provide us a lot of economic information and a reasonable amount of demographic information. An analysis of the surveys that were conducted in Italy by the major research institutes, restricted the choice to two surveys: the "Multipurpose survey on families" by ISTAT and the "Survey of Italian Households" by Bank of Italy. The survey of ISTAT contained more demographic and social information compared to that of Bank of Italy, but it was less informative and specific with regard to the economic situation of families and of each of their members (income, insurances, costs, real estates, etc.). In the first phase we had to build the dataset to use for the research. This goal required several steps and a careful examination of the survey structure to be accomplished. Our idea was to merge several surveys from different periods (years: 2000, 2002, 2004, 2006) and to create a single database to manage the data extrapolated by the surveys, so we decided to rely on database technology. Since we didn't need any of the sophisticated characteristics of the commercial distributed databases, such as Oracle or IBM's DB2, we chose to import our data in MySQL that is free and very simple to use. The final result of all this process was a database composed of twenty tables, more than eighty thousand records in some tables and more than one thousand fields in total. In the second phase we proceeded to select the sample we need to perform the analysis. We decided to choose a subsample from the data of the surveys on the basis of the hypothesis to consider households in which the female head has an age comprised between 40 and 60 years. The reasons of this choice lie on the consideration that the majority of the women which are older than forty have completed their reproductive cycle. The next step was to use the IREG (ISTAT code of the residence region) attribute to georeference the data to a map of Italy. During this operation we observed that the sample was not so well balanced from

the geographical point of view, so we decided to merge the data referring to Valle d'Aosta (which counts too few cases) and Piemonte and the polygonal spatial units of the map. The final result was a map composed of 19 polygons (the eighteen regular regions and the Piemonte-Valle d'Aosta) and an attribute table containing 18 factors: we began to lead some spatial statistical analysis, although the spatial numerosness was not particularly large for this kind of study (unfortunately the data lacked provincial information).

4. Results and conclusion

We will highlight now the most important results we obtained by considering spatial cluster and outliers analysis. They are based on the Global and Local Moran indices calculated and mapped by GIS (ARCGIS 9.1). The weight matrix we used was generated through the application of the "maxmin" algorithm (threshold distance of 379 Km) implemented on the stand-alone package S-Joint (Mucciardi-Bertuccelli, 2007). Taking into account the entire sample, we can notice a significant spatial autocorrelation of the CQ (childbearing quote) (79.5 %) with regard to national territory (table 1). The variable which appears to mostly clusterize the Italian territory is surely the number of husband's brothers in the "0" modality. In fact, we observe in this modality a CQ of 77%, with a Moran Index of 0.36 and two quite evident spatial clusters in north and south Italy (see, type of local relation in table 1). The same analysis can be extended, even if in a less evident way, to the number of wife's brothers in the "0" modality. In the other two modalities (1 and 2+) significant spatial autocorrelation indices are not found. For what concerns the educational level, both man and woman, the "middle" qualification level generates the highest spatial divergence of the CQ between northern and southern areas. This evidences that for low and high qualification levels, spatial variability of the CQ is fairly high. Another important factor in the CQ analysis is the presence, at the time of the interview, of fatherly and motherly grandparents. It appears that the presence of grandfathers positively influences the CQ even if at spatial analysis level different situations are found. Specifically, the presence of husband's mother generates, at national level, a CQ of 85% along with a cluster including the north-western regions of Italy (Pie-Val, Lig, Emi and Tos), whereas the absence of wife's father, which produces a CQ of 77% at national level, let emerge a similar cluster composed of the north-western regions (Pie-Val, Lig and Emi) which has very low values and an hotspot (Campania). Lastly, let us analyze the effect of income on the CQ. For what concerns women, we can notice a "U" classical effect (Becker and Lewis, 1973). Spatially, a great gap between northern and southern regions emerges. Particularly, we can notice that an income increase (15000 € class) produces in the southern regions (Bas, Cal and Sic) a significant increase of the CQ, whereas in the northern regions this increase is less evident. On the contrary, we observe an increase of the CQ concerning men's income when we consider higher income classes, as it is well evidenced in the literature. With respect to territorial differences we can remark that in the highest class (>25000 €) a significant value of the Moran index (0.31) is observed along with clusters concerning the northern and the southern regions. As for women, an income increase produces greater increases for the CQ in the South rather than in the North. Other variables which registered a certain influence on the CQ at territorial level (here we don't show them) appeared to be the property of the main real estate and the

not working status regarding the woman. Even in this case there are notably significant territorial differences (North - South). It can be observed that the analysis of this dataset through the use of other standard spatial clustering methods, such as G^{*}i statistic (Ord and Getis, 1995), confirms, even with some differences, the previously obtained results (results are not shown). In conclusion, it seems interesting to highlight how the analysis of spatial variation of the CQ can be a tool of specific interest. Local index of spatial correlation allows us to detect a region with significant deviation from the national average of CQ. In fact, the paper shows that, as a function of the socio-economic characteristics of the family, CQ markedly changes between different areas.

Table 1: *Principal results for some spatial indices regarding CQ by factor.*

Factor (determinat)		CQ (%)	Moran I ^a	Sign .	Clusters*/Outliers (Local spatial analysis)	Type of local relations
"Full sample"		79.51%	0.22	∞	Pie_Val*, Lig*, Emi*, Cam	LL, LL, LL, HH
Real estate property	Yes (No)	79.00%	0.18	°	Pie_Val*, Lig*, Cam	LL, LL, HH
Education woman	Medium (Low-High)	84.00%	0.23	∞	Pie_Val*, Lig*, Emi*, Bas*, Cal*, Sic*, Ven	LL, LL, LL, HH, HH, HH, HL
Education man	Medium (Low-High)	84.00%	0.46	∞	Pie_Val*, Lig*, Emi*, Lom*, Tre*, Bas*, Pug*, Mol*	LL, LL, LL, LL, LL, HH, HH, HH
Husband's father alive	No (Yes)	77.00%	0.15	°	Pie_Val*, Emi*	LL, LL
Husband's mother alive	Yes	85.00%	0.22	∞	Pie_Val*, Lig*, Emi*, Tos*	LL, LL, LL, LL
	No	74.00%	0.18	°	Pie_Val*, Lig*, Cam	LL, LL, HH
Wife's father alive	No (Yes)	77.00%	0.27	∞	Pie_Val*, Lig*, Emi*, Cam	LL, LL, LL, HH
Wife's mother alive	Yes (No)	85.00%	0.2	°	Pie_Val*, Lig*, Emi*, Cam	LL, LL, LL, HH
	0 €	80.00%	0.17	°	Pie_Val*, Lig*, Emi*	LL, LL, LL
Woman income (classes)	(0-7000 €)	77.00%	-0.03		Tre	HL
	(7000 – 15000 €)	79.00%	0.01			
	>15000 €	81.00%	0.24	∞	Bas*, Cal*, Sic*, Pie-Val	HH, HH, HH, LL
	(0 – 18000 €)	78.00%	0.09			
Man income (classes)	18001 – 25000 €	80.00%	0.21	∞	Emi, Cam*, Cal*, Sic*	LL, HH, HH, HH
	>25000 €	82.00%	0.31	∞	Pie_Val*, Emi*, Lom*, Tre, Bas*, Cam*	LL, LL, LL, HL, HH, HH
Woman employment	Not working (working)	77.00%	0.32	∞	Pie_Val*, Lig*, Emi*, Bas*, Cam*	LL, LL, LL, HH, HH
Woman civil status	Married (single, divorced)	80.00%	0.19	°	Pie_Val*, Lig*, Emi*, Cam	LL, LL, LL, HH
Husband's brothers number	0 (1,2+)	77.00%	0.36	∞	Pie_Val*, Lig*, Emi*, Lom*, Tos*, Cam*, Bas*, Pug*, Cal*, Sic*	LL, LL, LL, LL, LL, HH, HH, HH, HH, HH
Husband's sisters number	1 (0)	80.00%	0.18	°	Emi	LL
	2+ (0)	81.00%	0.16	°	Pie_Val*, Lig*, Emi*, Bas	LL, LL, LL, LH
Wife's brothers number	0 (1,2+)	78.00%	0.30	∞	Pie_Val*, Lig*, Emi*, Cal	LL, LL, LL, HH
Wife's sisters number	2+ (0,1)	82.00%	0.26	∞	Pie_Val*, Lig*, Emi*, Umb, Cam	LL, LL, LL, HL, HH
Between brackets: modalities of factors presenting not significant Moran index					--	^a threshold distance = 379 Km
°p<0.05 ∞°p<0.01 -- LL=low-low; HH=high-high; LH=low-high; HL=high-low					--	2+ = greater than or equal to 2

References

- Becker G., Lewis H. G. (1973) On the interaction between quantity and quality of children, *Journal of Political Economy*, vol. 82, pp. 279-288.
- Dankmeyer, B. (1996) Long run opportunity-costs of children according to education of the mother in the Netherlands, *Journal of Population Economics*, n.9, pp. 349-361.
- Di Pino A., Mucciardi M. (2003) An endogenous dummy variable estimator of women's opportunity costs of children in Italy in: *Household Behaviour, Equivalence Scales, Welfare and Poverty*, Dagum-Ferrari (Eds), Physica, pp. 55-74.
- Mucciardi M., Bertuccelli P. (2007) S-Joint: a new software for the analysis of spatial data in: *Proceeding of Italian Statistical Society*.
- Ord J.K., Getis A. (1995) Local spatial autocorrelation statistics: distributional issue and an application, *Geogr. Anal.* n.27, pp. 286-306.