C. Z. Guilmoto, 2000. *The spatial analysis of sub-populations,* technical paper prepared for the workshop on socio-cultural factors affecting demographic behaviour, Unesco and UNFPA, Paris.

Introduction

Spatial analysis aims at complementing usual data analysis by taking into account the geographical location of phenomena under study and their specific spatial features. It consists of basic operations of spatial query of varying complexity, ranging from simple description of information on maps to more refined queries involving the creation of new spatial objects. Spatial analysis is crucial to the analysis of sub-populations as it allows to relate them to their given spatial environment. Since sub-populations with specific socio-cultural characteristics are very rarely scattered randomly on a regional map, mapping is usually one of the best device to describe and identify sub-populations and their components. Thus, key geographical characteristics such as place of residence (inner city, forested areas, etc.) or place of origin (for migrants or displaced populations) are often *the* defining features of underprivileged groups as spatial segregation is one of the most common dimensions of social discrimination.

Mapping packages and more sophisticated GIS (geographic information systems) are now widely available tools used to display, explore and analyse spatial data. Combined with statistical methods, GIS operations allow specific queries that are directly applicable to the analysis of sub-populations. [1] The object of this paper is to present a selection of statistical and geographical procedures that helps to identify sub-populations through maps and other geographical analyses. We shall combine general methodological considerations, with illustrations taken from an ongoing research conducted by the author in South India.[2] Detailed descriptions of available statistical or mapping techniques are available in a variety of sources[3] and therefore, our presentation will not cover the many technicalities that these methods entail.

Here we will examine a very common situation in which the sources available relate to the whole population of entire regions or areas. This often makes the direct identification of sub-populations a difficult task. Moreover, it is often technically unfeasible to handle an exhaustive large-size dataset to extract the relevant information pertaining to specific subgroups. This paper aims at presenting various data and map processing techniques that are usually involved in these analyses.

We shall examine four types of sub-population definition and present the corresponding statistical and mapping tools that can be used to summarize and map the data in the most efficient manner. Sub-populations are defined by a set of characteristics, ranging from simple indicator to more complex indicators based on social, economic, cultural and geographical characteristics. In some cases, a single dichotomous variable (e.g. nationality) is sufficient, as will be examined in our first section. However, the large number of localities under study may result prevent a straightforward cartographic analysis as maps are not intelligible when they are based on several hundreds of geographical units. Techniques based on spatial aggregation and smoothing of large datasets are presented in Section 2. Moreover, the population concerned is often defined with a more composite indicator (e.g. "underprivileged") or with reference to its geographical location (e.g. "mountainous populations"). Sections 3 and 4 of this paper will present case studies related to these situations.

[1] A non-exhaustive list of a major packages of GIS, automated mapping and spatial analysis: PopMap, ArcView, Spatial Analyst, MapInfo, Vertical Mapper, GS+, SpaceStat. Statistical packages include: SPSS, SAS, Stata, Statistica, S+. All these packages are described in detail in their respective web sites.

[2] Maps and data used in this paper are derived from the South India Fertility Project, a programme conducted by the IRD and the French Institute. Localities refer to the 70000 villages in 1991. The area under consideration is composed of the four South Indian states of Andhra Pradesh, Tamil Nadu, Karnataka and Kerala, and the Union Territory of Pondicherry, with a combined population of 200 million.

[3] See the bibliography at the end.

1 Working with a single indicator

1.1 mapping the raw population distribution

In many cases, the definition of sub-population derives from a single indicator that describes social or cultural characteristics, or simply membership to a specific subgroup. For instance, we may want to examine some specific ethnic or socio-economic groups.

In this first example, we are showing the regional distribution of tribal population in South India. The tribal population in India displays strong spatial features, though a large number of tribals (*Adivasis*) have settled down in non-tribal areas. To prepare this map, we have used only the total tribal population per area, irrespective of the non-tribal population. Such maps are possible only when populations are concentrated in some areas. Otherwise, a map of sub-populations such as that of *Dalits* (ex-Untouchables) which are spread every where would be difficult to distinguish from a map of the total population.

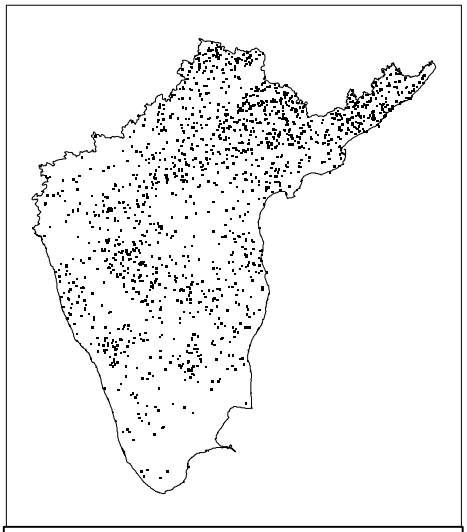

Figure 1: Tribal population in South India (one dot for 2000 people)

1.2 Ratio and percentage mapping

Some sub-populations do not display such strong geographical features at the macro scale and mapping them would often result unsatisfactory results as indicated above. In which case, their proportion in the general population must be mapped though other ratios can be computed.

New variables must computed from the original database and plotted. In the example that follows, we have plotted the districts with more than 20% tribal population. As can be observed, the geographical clustering of tribes in India is much more visible when examined in terms of percentage shares. However, our map is based on a simple, dichotomous classification. Maps of ratios are usually based on several classes of values designed to enhance insight in the data.

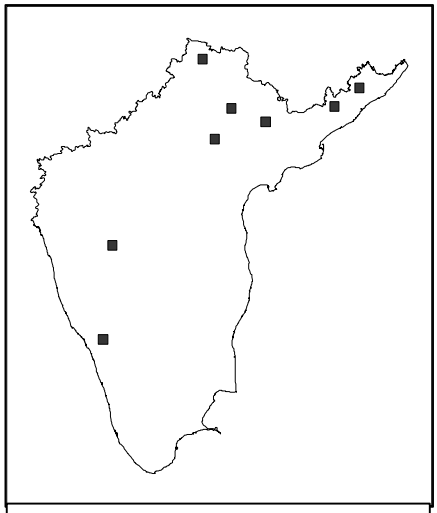

Figure 2: Districts with more than 20% of tribal population

1.3 The impact of statistical distribution

Crucial to this analysis is the examination of the statistical distribution as we are no more dealing with population figures, but with proportions. In fact, maps such as the one presented in 1.2 are possible only when there is clear-cut demarcation in the distribution of the sub-population's share between localities. For example, the previous map is based on the distribution of tribal population, which highly skewed, as most localities comprise less than 5% tribal population. However, it is worth noting that the statistical distribution is often more complex than the

normal (bell-shaped) and other unimodal distributions. In cases when the distribution has several modes, a first statistical analysis will determine how to define class intervals for mapping purposes that do not conceal the heterogeneity of the sample. Cartography manuals describe the many available solutions (quantiles, equal ranges, etc.).

1.4 Mapping indirect variables

The required variables are often not available to describe specific sub-populations and the behaviour. In such cases, map analysis may actually become spatial exploration. Plotting different variables and computed may help to identify specific phenomena that are not directly observed by surveys or censuses for a variety of reasons. In this case, variables selected for analysis will serve as indirect indicators for unobservable features.

A very simple example is taken here from our database. We have mapped the data related the sex ratio of the population (males per female) in the southwestern part of South India. Many clustered villages in Kerala State displays unusually low values, corresponding to less than 900 men per 1000 women. As it turns out, this map depicts in fact the intensity of male outmigration, a phenomenon of considerable importance to the regional economy and society. It may be noticed that no direct measurement of migration is available at this scale. Moreover, this mapping analysis demonstrates the highly concentrated character of outmigration, with visible clusters of villages with intense migration (locally known as "Gulf pockets"). This is related to the existence of strong local migration networks to Gulf countries and other destinations in India.

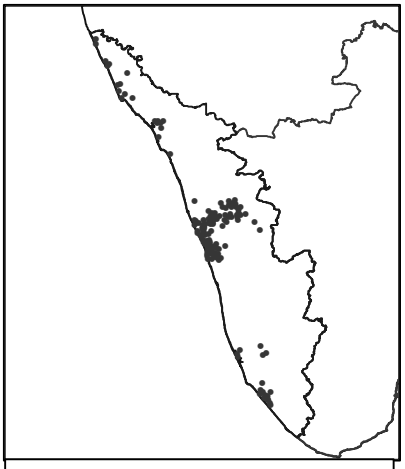

Figure 3: Migration clusters in Kerala (less than 0.9 male per female)

Exploration through maps may therefore become a very important tool to identify and map unobservable phenomena using available proxy variables.

2 Aggregation and smoothing

Illustrations given above have carefully been selected to allow an efficient mapping solution. But, this is often at the expenses of available details from the data source. For instance, the map of tribal population was based on the identification of districts with more than 20 % of tribal population. However, the number of districts in the area is limited. The same map with all individual localities (60,000) would on the contrary be illegible and time-consuming. In most of the cases, the eye can hardly distinguish trends or details of thousands of locations even when colours are used for the classification.

They are several solutions to difficulties related to the excessive number of spatial units in the analysis. Techniques based on administrative aggregation are well known, but may result in loss of information as was shown previously for districts. Other techniques of aggregation are now made possible by GIS procedures and are used intensively to summarise information. They may involve some amount of spatial interpolation and smoothing. Some methods are presented here.

2.1 administrative reaggregation

The most common way to solve these problems is by aggregating data using large administrative units. For example, data about households can be regrouped by census wards in cities, while villages will be clubbed together at the district level.

In the following map, we give an example of localities in a subregion of South India and district administrative boundaries. The resulting map of district shows localities and administrative boundaries. This straightforward technique is however often unsuitable for various reasons. First of all, the administrative units are fixed and cannot be changed. Another related problem is that administrative units are usually of uneven size, resulting in unequal levels of aggregation according areas. Districts compared on the map may comprise from 50 to 1000 villages. The definition of administrative units may vary among regions as is the case in South India. The lack of flexibility when aggregating original data in available higher-order units results in maps that lack the original richness of the raw data and introduces new biases related to the administrative zoning itself. Local patterns pertaining to specific sub-populations might completely disappear from aggregated maps. This is especially true when sub-populations are located across administrative units or represent a small minority in the new administrative aggregate.

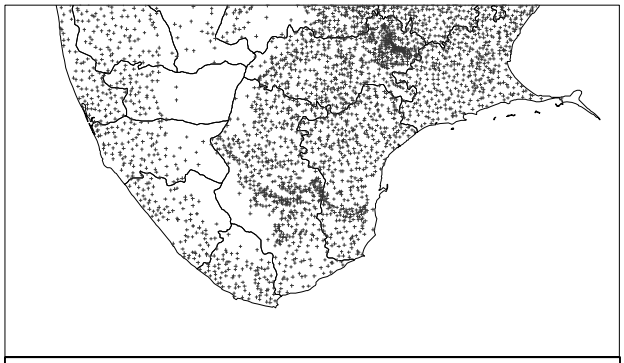

Figure 5: Localities and administrative clusters (district boundaries)

The spatial capabilities of GIS have however generated new strategies to tackle the aggregation problems. The first solution presented here refers to geographical clusters that are generated especially for the analysis, keeping in mind the characteristics of the population. A second, more powerful solution lays in spatial smoothing, a technique derived from statistical smoothing procedures and applied to two-dimensional geographical data.

2.2 Clusters using Thiessen polygons.

Thiessen (or Voronoi) polygons are polygonal zones generated from a set of points and derived in such a way that each polygon correspond to a "catchment area". The definition of Thiessen polygons implies that all area in the polygon is closer to the polygon's central point than to any other point.

For the map shown here, we have used the same area as for the previous map with administrative boundaries. However, the region is now divided into clusters based on Thiessen polygons. Villages were first aggregated and polygons drawn around aggregated points. Compared to the previous district map, this partitioning results in units of more similar size. Consequently, values obtained in these Thiessen polygons are more comparable than values derived from

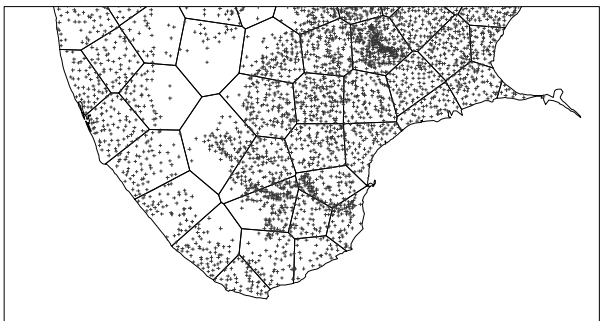

Figure 6: Localities and spatial clusters based on Thiessen polygons

district of uneven size. Thiessen polygons correspond also new geographical units of a somewhat similar shape (polygons).

Many other methods to reaggregate local data exist, such as those based on rectangular grids or on the direct aggregation of localities. The choice of techniques is large and cannot be detailed here. They all result in a significant reduction in the number of geographical units that is often required by the sheer number and complexity of the original location map. It is important to stress that "natural" aggregation (based on spatial proximity) is usually better than other principles such as geometrical aggregation (as in the case of square cells) or less flexible administrative aggregation.

2.3 Spatial interpolation and smoothing

Another way to solve the problem of excessive number of spatial units consists in interpolating and smoothing local values. These techniques are based on various mathematical formulas that average values over pre-determined areas. Usually, the map of localities is first converted into a surface map, consisting in a grid of pixels (the minimal square areas shown by the computer). Values for each pixel may be weighted average value of all surrounding localities, with weights inversely proportional to distance (the farther the locality, the less weight in the resulting smoothed value). The figure shown here corresponds to the smoothed density values in a small area of our sample; pixels are here square cells of 1 sq. km.

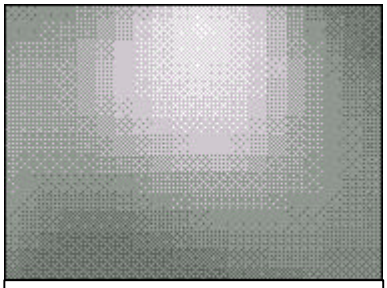

Figure 7: Pixel map of density

With proper calibration, these techniques may also reduce (or eliminate) the impact of local "outliers", i.e. isolated values that are usually irrelevant for a more global analysis. This is especially useful when variables under study may be locally of uncertain quality or very unstable because of the small size of the correspond population.

We simply cannot cover all techniques available for spatial smoothing. As indicated above, many of these techniques are based on distance weighted methods, in which closer points are assigned a greater weight in interpolating local values. Statistically speaking, the optimal solution remains however the kriging procedure, a spatial regression technique that was developed in geology. This method generates the most accurate estimates (best linear unbiased estimator). Kriging is based on the examination of spatial autocorrelation (the correlation between values as a function of distance between them). Though kriging is a computation-intensive procedure, this technique is especially appropriate when the level of local variation (between adjacent observations) is high. Various kriging procedures are available in many GIS packages, including one procedure ("universal kriging") that provides for the removal of regional trends. "Block kriging" provides the corresponding smoothing tool.

2.4 Contours and spatial trends

It should be emphasized here is that spatial interpolation and other smoothing procedures generates an entirely new type of map. Original maps were based on discrete spatial units: points or regions corresponding to a limited set of spatial units such as villages or administrative units. However, interpolation allows the estimation of values for all the pixels,

including in intervening areas where no observation is available. Consequently, smoothed maps are much more detailed and the corresponding files may include millions of pixels. The entire map of South India based on the grid shown above comprises more than 1.2 million pixels.

Contouring is a basic tool of sophisticated GIS packages that transforms complex pixel ("raster") maps in different zones by drawing contour lines, also known technically as isolines. Contour lines are paths along which interpolated values are constant. Topographic maps are well known contour maps, but similar techniques can be employed for any other information shown on maps.

The contouring of smoothed data converts pixel data into regions and help to discern "spatial trends". This operation significantly improves the legibility of maps as the map of sex ratio among children may illustrate. Each line corresponds to a specific value, starting here from 1100 boys per girls to the darkest areas with more 1500. It clearly identifies the pockets affected by a deficit of girls. Incidentally, this map is also an indirect indicator, in this case for the intensity of female discrimination in Tamil Nadu State. Actually, high sex ratio values correspond closely to the prevalence of both female foeticide and infanticide in this region of India. The map shows this phenomenon to very concentrated in some small areas where the sex ratio among children is disproportionate. At the same time, excess female mortality is almost completely absent from the rest of the State.

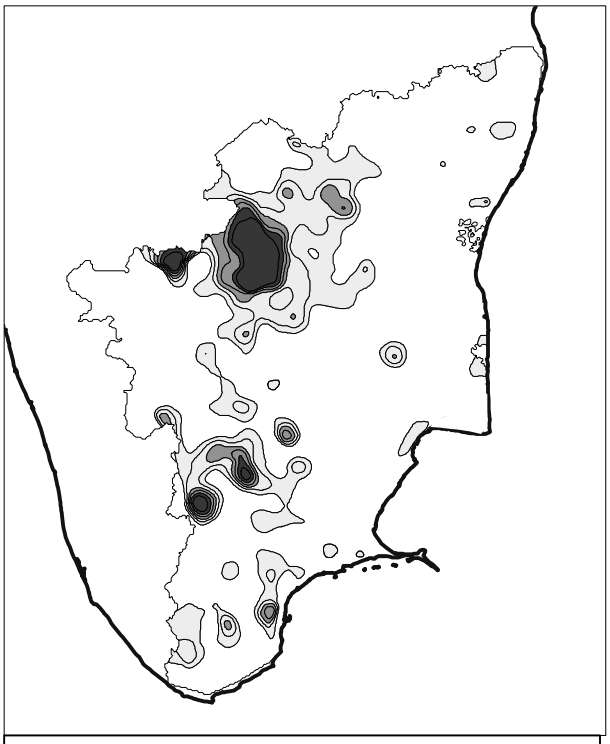

Figure 8: Female infanticide and foeticide (low proportions of girls among children in grey)

Contouring of pixel data summarises existing spatial trends and helps to create a new type of zoning. This is a powerful device to visualise data and to focus intervention on specific target groups. In the case given above, the concentration of infanticide in some areas is a crucial information to address the problem and its social causes.

3 Statistical analysis

Another common problem faced with the identification of sub-populations relates to the absence of a single direct or indirect indicator. In some cases, it is only through use of several variables that a new defining characteristic can emerge. In fact, many definitions of social groups combine various educational, economic and demographic indicators. In other cases, the continuous distribution of a single variable makes it impossible to classify individuals or localities.

In such situations, it is more appropriate to apply first special statistical methods to reclassify observations or attributes. We will present two among the best-known techniques and see how they can fruitfully applied to a data set to identify important subgroups and map them. These

techniques relates to a standard methodology called *factorial ecology* widely used in in social geography.

3.1 Factor analysis

The main applications of factor analytic techniques are: to reduce the number of variables and to detect structure in the relationships between variables. The principal components method is the most common technique applied to reduce the number of characteristics for the analysis. Starting from a set of variables, the technique will extract the main *factors* (linear combinations of variables). Usually, the first, most significant factor is a very useful proxy for a general character that is not directly observable.

In the example given here, we have analysed all localities in our sample to examine their infrastructure and amenities. These infrastructures were described in our original database by more than 30 odd variables, ranging from primary school, to telephone facilities or bus stop. The plotting of individual maps for each infrastructure is not a feasible solution. Instead, we have used the principal components method to define a combined infrastructural factor. After correction for the size of the localities and spatial smoothing, this factor is then plotted on the map shown here. Only areas with deficient infrastructures are shown (in dark grey). This map is a very useful summary of a complex situation. Incidentally, it may be noticed that it bears some resemblance with the distribution of tribal population.

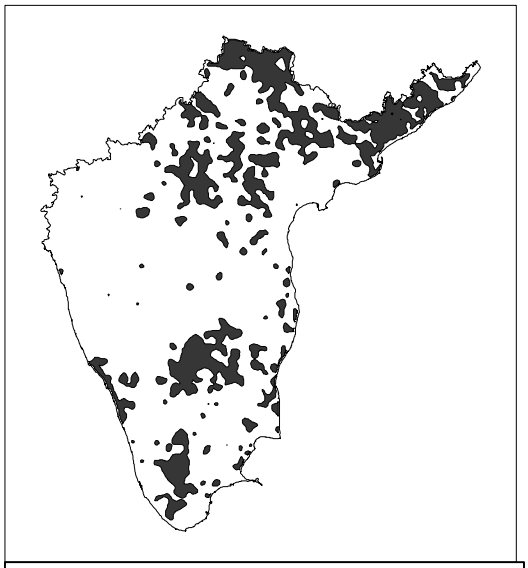

Figure 9: Underdeveloped villages (result from a factor analysis on local infrastructures)

3.2 Cluster analysis

It is important to stress at the outset that cluster analysis as presented in this section is very distinct from spatial clustering presented above. Cluster analysis is the most famous among various reclassification methods. Instead of classifying variables as in the factor analysis, cluster analysis reclassifies observations into distinct groups. Its first benefit is again to simplify information. A cluster based on infrastructure data would result in different categories of localities. The cluster of the most under-equipped villages would closely resemble the map obtained from factor analysis above.

Analysts might want to include location variables as variables used in the cluster analysis. Typically, location characteristics will be given as two variables (such as the geographical co-ordinates of longitude and latitude), though other solutions exist (such as the distance to a fixed point, etc.). Cluster analysis will then treat geographical co-ordinates as two further variables and the role is the results is likely to be very significant if the phenomenon studied displays real geographical concentration. However, map analysis of non-spatial data is usually a much safer procedure.

4. Spatial queries

Spatial queries require the processing of geographical information. Sub-populations will be identified through some of the specific spatial features. While the most common spatial query relate to the (absolute) location of populations, another important type of analysis that GIS make possible corresponds to distance, i.e. the position of localities relative to other geographical entities.

4.1 Spatial subsets

The more common analysis relates to overlay analysis that manipulates spatial data organised in different layers of information (elevation, density, etc.). Resulting spatial features depend on specific selection criteria. Intersection and union are the basic operators of this algebra. However, detailed raster data (available for all map pixels) may lead to more sophisticated queries based on a combination of factors. However, pixels correspond to spatial co-ordinates and not to social entities such as households or villages. Consequently, this type of spatial analysis is of greater use in natural sciences than in the analysis of social phenomena.

4.2 Buffers

Buffers are areas constructed by extending outward from given geographical objects. In a buffer operation, the buffer zones are created from a spatial object and a given distance such as areas within 5 km from a city or villages lying more than 20 km from the road. Buffering is one of the most commonly used techniques of spatial analysis and helps to easily determine localities with respect to their distance to some other geographical unit. It is especially powerful to identify the potential impact of local features on surrounding population. Examples can include populations vulnerable to flood or cyclone (proximity to rivers or to the coastline), populations served by various infrastructure (transportation network, post office, health centre, etc.), rural populations living close to urban agglomeration, etc.

The sample map given here relates to the mountainous Kodagu district region of South India. We have tried to identify remote localities for further analysis of their socio-economic characteristics by creating a buffer along major roads. Only villages that lie at least 2 km away from main roads are shown on the maps. The 2-km buffer is also displayed in grey on the map, along with the road network and local towns. As may seem obvious enough, all these localities are precisely defined by their spatial location and this constitutes a classical example of the new possibilities offered by spatial analysis.

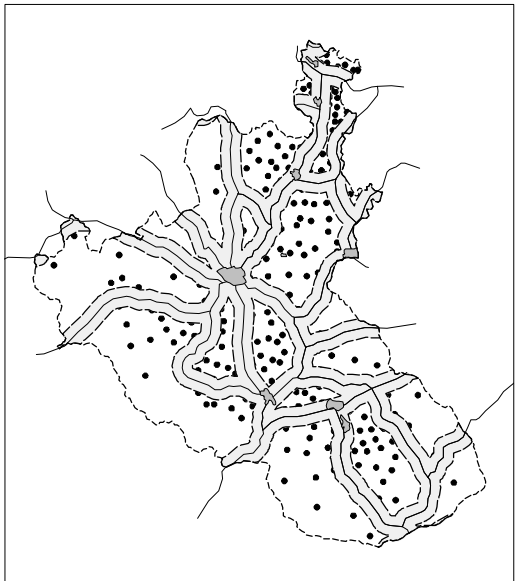

Figure 10: 2-km road buffer and localities in Kodagu district

4.3 Spatial correlation analysis

The primary objective of correlation analysis is to reveal relationships between different layers of information. Correlation analysis determines

whether the distribution of one type of feature is related to the distribution of another type. However, it may not be relevant to elaborate more on these techniques in our paper, as they it proceeds directly from ordinary statistical analysis. Moreover, the fact that data are spatialized hardly changes the interpretation of results.

However, spatial correlation between different observations for the same variable (*spatial autocorrelation*) is a more powerful analysis tool. It aims at assessing the real intensity of geographical clustering that cartography suggests. It measures the extent to which the occurrence of one characteristic is influenced by the distribution of similar characteristics in the adjacent areas. Positive spatial autocorrelation is the most common situation, corresponding to a pronounced geographical clustering of values: the closer the locations, the more similar are observed values.

The most commonly used statistic for autocorrelation is Moran's coefficient, with values ranging from 1 (clustering) to 0 (no random pattern) to –1 (scattered pattern). It helps to assess the range of geographical influence and compare the degree of spatial autocorrelation among various attributes.

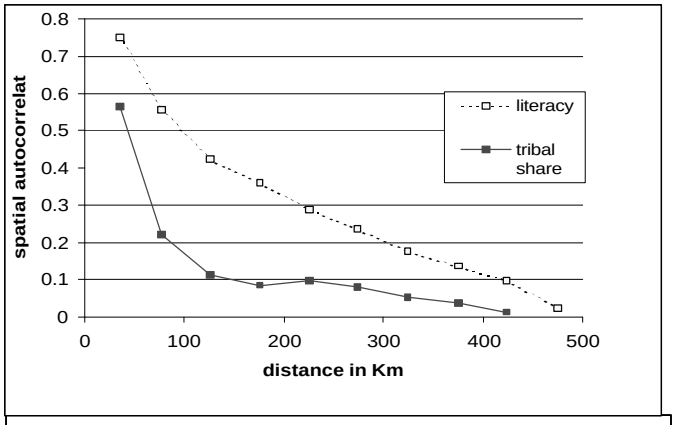

Figure 11: Spatial autocorrelation for literacy and tribes (up to 500 km)

The figure shown here is a correlogram, plotting Moran's index for two variables in our South Indian sample. The index decreases regularly as distance between observations increases and disappears almost completely after 400 km. Literacy displays a very strong spatial concentration, with the closest localities exhibiting almost similar literacy rates. The trend is still manifest when localities are more than 200 km apart. On the contrary, tribal settlements are clustered at a very close distance. This clustering is almost invisible for distance greater than 100 km.

Apart from correlogram, the analysis of spatial auto-correlation may also be based on the semi-variogram (used for kriging interpolation) and Geary's coefficient.

Summary

This paper introduced general principles for exploring the characteristics of sub-populations using mapping and GIS tools. The benefits and limitations of these methods were briefly summarized and illustrated through maps and statistics related to sociodemographic data. These concrete illustrations may demonstrate the effectiveness of these techniques in processing and summing up available empirical information from a macro database and turning it through maps into powerful visualisation and analysis devices at the micro level.

However, unusual constraints often exist that require specially designed methods or deep transformations of the existing database. Moreover, the implementation of many a procedure described in this paper often entails serious technical difficulties that are out of the scope of

this paper. More methods are described in recent books and manuals related to cartography, data exploration, geostatistics, quantitative geography or demography (see bibliography below).

From an empirical viewpoint, the mapping approach is especially suitable for projects requiring direct interventions on concerned sub-populations. It will therefore help to target interventions by identifying the sub-populations, visualising and quantifying relevant characteristics in its geographical setting, and finally reaching the most efficient solutions. From a more theoretical viewpoint, spatial analysis is the basis for understanding how sub-populations are social groups inseparable from their spatial context. Spatial analysis as a specific tool to explore social data seems destined to flourish, in response to both technological improvements and the growing awareness of the geographical embeddedness of social and cultural phenomena.

A selection of books related to spatial analysis (in English only)

Airlinghaus, S. L., eds., 1996. *Practical Handbook of Spatial Statistics*, CRC Press, Boca Raton and New York.

Bailey, T. C., and A. C., Gatrell, 1995. *Interactive Spatial Data Analysis*, Longman, Harlow.

Bocquet-Appel, J.P., Courgeau, D., and Pumain, D., eds., 1996. *Spatial Analysis of Biodemographic Data*, John Libeey and INED, Paris.

Chou, Yue-Hong, 1997. *Exploring Spatial Analysis in GIS*, Onward, Santa Fe (USA).

Clarke, K. C., 1995. *Analytical and Computer Cartography*, Prentice Hall, Englewood Cliffs.

Fotheringham, S. and Rogerson, P., eds., 1994. *Spatial Analysis and GIS*, Taylor and Francis, London, 1994.

Fotheringham, S., Brundson, C, and Charlton, M., 2000. *Quantitative Geography. Perspectives on Spatial Data Analysis*, Sage, London.

Haining, R., 1990. *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press, Cambridge.

Houlding, S. W., 2000. *Practical Geostatistics. Modeling and Spatial Analysis*, Springer Verlag, Berlin-Heidelberg.

Isaaks, E., and Srivastava, R. M., 1989. *An Introduction to Applied Statistics*, Oxford University Press, Oxford.

Kraak, M. J., and Ormeling, F.J., 1998, *Cartography. Visualisation of Spatial Data*, Longman, Harley.

Longley, P., and Batty, M., eds., 1996. *Spatial Analysis. Modelling in a GIS Environment*, GeoInformation International,, Cambridge.

Plane, D. A., and Rogerson, P. A., 1994. *The Geographical Analysis of Population, with Applications to Planning and Business*, John Wiley and Sons, New York.