# Spatial correlation and demography.
# Exploring India's demographic patterns

**S. Oliveau, UMR Géographie-cités, Paris**
**oliveau@ifpindia.org**
**CZ Guilmoto, LPED/IRD, Paris**
**guilmoto@ird.fr**

## Introduction

Demographers often use maps and regional tabulations to demonstrate the spatial dimension of population characteristics in given settings.[1] Cross-tabulations are shown to illustrate that some phenomena tend to be more prevalent in some areas than in others and maps may in addition show where high or low values of the phenomenon studied tend to concentrate. This is usually followed by the attempt to identify social, economic or cultural variables accounting for this geographic patterning or, when this fails, by the insertion of regional dummies in the model. The fact that residues are often (spatially) autocorrelated and therefore violate the basic assumption of regression analysis is rarely taken into account in the final model.

While demographers display considerable ability to test the strength of statistical links between phenomena, they prove less curious about the strength of the "geographical correlation" that may appear once demographic events are adequately mapped. The fact that populations in close-by areas tend to display rather similar demographic behaviour is not tested in the same way as the statistical correlation between demographic and other variables. Similarly, maps are shown to demonstrate the spatial clustering of various phenomena, but the cartographic analysis remains intuitive, based among others on the quality of the map and the ability of the analyst to interpret the geographic variations, often using other variables not displayed.

## Global and local tools

There are however a large gamut of tools to investigate the nature and extent of spatial correlation between demographic variables. These techniques constitute what is now called exploratory spatial

---

[1] This work is based on current research by the authors on the geostatistical analysis of spatial patterns in India and is part of the EMIS (Equipe Espace et Mesure en Inde du Sud) project supported by the CNRS (programme Société de l'Information). Data include the recent district-level data from the Census of India published in 2004 as well as original fertility and mortality estimates derived by one of the authors.

data analysis (ESDA) by reference to exploratory data analysis (EDA) popularized by Tukey in the late 1970s. Most notably, ESDA includes visual and quantitative methods to summarize the spatial properties of a variable, to describe its specific patterns in space, spot extreme values or outliers, and to identify specific geographical subsets. The availability of data in a GIS (geographic information system) format allows the systematic spatial exploration of the data in a way that was previously not feasible.

These tools examine the nature of spatial variations in the sample to distinguish the "smooth" (fitted) from the "rough" (residuals) to follow EDA's terminology. Central to this agenda is the concept of spatial autocorrelation, i.e. the correlation of a single variable between pairs of neighbouring observations. Once the concept of "neighbouring observations" is defined (using contiguity or distance matrix), the correlation between neighbours may be compared to the general variance of the sample in the same way as in ordinary correlation analysis. The resulting measure of spatial autocorrelation is a first indication of the spatialized nature of the phenomenon studied: this correlation may be non-existent, low or strong according to the variables used.

Moran introduced in 1950 the first measure of spatial autocorrelation in order to study stochastic phenomena, which are distributed in space in two or more dimensions. Moran's index has been subsequently used in almost all studies employing spatial autocorrelation. Moran's I is used to estimate the strength of this correlation between observations as a function of the distance separating them (correlograms).

It shares many similarities with Pearson's correlation coefficient: its numerator is a covariance, while its denominator is the sample variance. And like a correlation coefficient the values of Moran's I range from +1 meaning strong positive spatial autocorrelation, to 0 meaning a random pattern to -1 indicating strong negative spatial autocorrelation –although negative autocorrelation is extremely unusual in social sciences.

The precise definition of Moran's I is given below for a spatialized variable $z_i$ at location *i*.

$$I = \frac{\sum_{i,j} W_{ij}(z_i - \bar{z}).(z_j - \bar{z})}{n} / \sigma^2(z)$$

Where $\sigma^2$ is the sample variance

Usually, the proximity matrix *Wij* is everywhere 0 except for « contiguous » locations *i* and *j* where it takes the value 1. However, an extended definition of this contiguity matrix allows for the computation of Moran's I at various levels of contiguity (or distance). This provides a complete correlogram of spatial autocorrelation by distance class and the impact of distance on the strength of spatial autocorrelation for each variable can be examined. Some variables may be locally strongly

autocorrelated, but display no correlation over a slightly larger radius, while the spatial autocorrelation for other variables may be significant over a longer distance.

While the strength of Moran's I lies in its simplicity, its major limitations is that it tends to average local variations in the strength of spatial autocorrelation. This has prompted statisticians to develop local indices of spatial association. This category of tools examines the local level of spatial autocorrelation in order to identify areas where values of the variable are both *extreme* and *geographically homogeneous*. This approach is most useful when, in addition to global trends in the entire sample of observations, there exist also pockets of localities exhibiting homogeneous values that do not follow the global trend. This leads to identification of so-called hot spots -regions where the considered phenomenon is extremely pronounced across localities- as well of spatial outliers.

The index fast becoming the standard tool to examine local autocorrelation is Luc Anselin's LISA (local indicator of spatial association), which can be seen as the local equivalent of Moran's I. the sum of all local indices is proportional to the (global) value of Moran's statistic.

The local value of a LISA is computed as:

$$I_i = \frac{\sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}$$

$$with \quad \sum_i I_i = \gamma . I$$

For each location, LISA values allow for the computation of its similarity with its neighbours and also to test its significance. Five scenarios may emerge:

- Locations with high values with similar neighbours: *high-high*. Also known as « hot spots ».
- Locations with low values with similar neighbours: *low-low*. Also known as « cold spots ».
- Locations with high values with low-value neighbours: *high-low*. Potential "spatial outliers".
- Locations with low values with high-value neighbours: *low-high*. Potential "spatial outliers".
- Locations with no significant local autocorrelation.

These specific configurations can be first identified from a scatterplot showing observed values against the averaged value of their neighbours. This so-called Moran scatterplot is a useful exploratory tool. Once a significance level is set, values can also be plotted on a map to display the specific locations of hot spots and potential outliers.

**Spatial patterns of India's contemporary demography**

The second part of the paper is devoted to the application of these tools to district-level variables derived from the Indian 2001 census. The advantage in using Indian data for this investigation relates both to its size and to its heterogeneity. India is divided into 35 states that are further divided into 593 districts (mean district population = 1.7 million); the use of district data allows among others to ignore the impact of larger-scale administrative divisions such as the state units, which usually form the only basis for regional comparison within India. At the same time, India's demographic diversity is tremendous and the country includes regions at very different stages of demographic transition, with various levels of population growth, fertility and mortality.

The selected variables represent five major population indicators:

- population density (2001)
- urbanization level in % (2001)
- intercensal growth in % (1991-2001)
- fertility (TFR estimates from the 2001 census)
- child mortality (probability of survival from birth to 0-6 estimated for 2001)
- gender inequality (sex ratio below 7 years)

We have added a non-demographic variable for comparison purposes. Here we use the district-wise proportion of the Muslim population.

The basic geostatistical description of variables used here is shown in the following figure. Moran's index is computed for each variable and for different distance class, starting from 0-75 km. results are not shown at distance above 100 km when most indicators are nil or even negative.

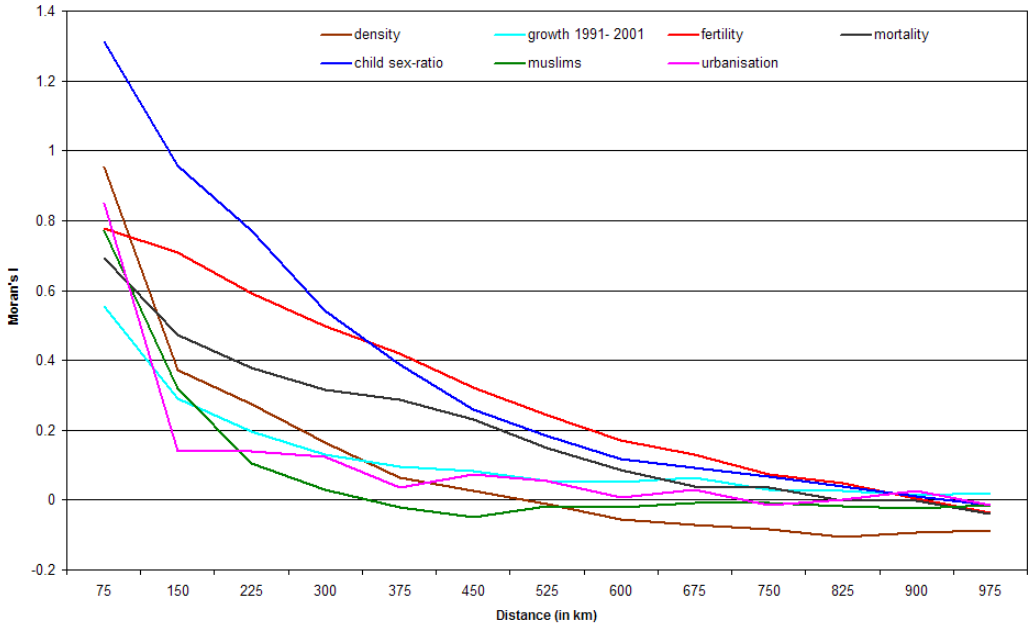



**Figure 1: Moran's index (distance correlogram)**

The main purpose of the following section is to explore India's demographic diversity through this set of indicators in order to identify the spatial patterns of these demographic features and to offer an illustration of what spatial correlation analysis can teach us about demographic trends in the country.

The analysis protocol is similar for each variable. Results are shown further below in the paper.

A. They start with descriptive statistics and a basic district map of each variable (high values are in darker colour).

B. We then compute the global spatial autocorrelation (Moran's I) according to distance lags of 75 km (distance matrix of neighbours). We have selected this distance class to offer the most detailed correlogram and preserving a sizeable sample for the smallest distance.

C. The spatial autocorrelation is also computed for districts with common boundaries (contiguity matrix) rather than for distance. Data are first displayed in spatial scatterplots to contrast observed values with their spatial average (spatially averaged adjacent values) and detect outliers. The corresponding Moran's index for contiguity is also displayed on the scatterplot.

D. A map of local spatial correlation indices (Anselin's LISA) is finally used to display hot spots and other spatial phenomena. Hot spots are shown in red and cold spots in blue. Other values (high-low and low-high) are respectively shown in light blue and light red. Areas shown in white are devoid of any spatial autocorrelation.

   Local Moran values have been computed using first order contiguity (queen contiguity) and significance levels are based on Monte-Carlo simulations on 999 permutations.

These materials can be used in two ways. First, to describe separately the spatial properties of each variable and examine what this tells us about demographic patterns in India: are there new hot spots of female discrimination in India? To what extent do density or urbanization indicators display a lower significant level of spatial correlation and why? How far does spatial autocorrelation extend for various variables and what does it tell us about India's spatial patterning? Do we know of a demographic phenomenon devoid of spatial correlation? How do we account for spatial outliers?

Results are then used to provide a comparative analysis of spatial patterns of demographic data: why is density much less correlated than other variables? Is fertility clustering due to literacy patterns? Are hot spots of high female discrimination similar to those of mortality or fertility? As the measurement of spatial autocorrelation of demographic indices is still uncommon, comparative analysis is crucial to understand the spatial structuration of different demographic factors such as fertility, population growth, population distribution, or mortality.

**Conclusion**

The conclusion allows us to illustrate the role of spatial phenomena in the shaping of demographic change. While spatial analysis *per se* is crucial to better understand the nature of demographic differentials in a region, the presence or the absence of spatial autocorrelation also points the potential role played by diffusion mechanisms. These mechanisms tend to follow spatial and social networks and thereby to reinforce the spatial patterns of demographic disparities (spatial path dependency?), especially in countries where social and demographic changes are extremely rapid.

At the same time, we will underline some of the difficulties facing the development of spatial analysis in demography such as lack of spatialized data, impact of irregular spatial distribution and outliers on geostatistical measurements, selection of the appropriate scale for analysis, and incomplete theorization of spatial correlation mechanisms in social sciences.

**Bibliography:**


Anselin, L., 1995, « Local indicators of spatial association - LISA », *Geographical Analysis*, Vol. 27, n°2.

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

Cliff, A.D., Ord, K.J., 1981, *Spatial processes. Models and applications*, Pion, London

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

Guilmoto, C.Z., Oliveau S., Chasles V., Delage R. and S. Vella, 2004. *Mapping out social change in South India: a geographic information system and its applications*, Pondy Papers in Social Sciences 31, Institut français de Pondichéry, Pondicherry, 117 p.

Guilmoto, C. Z. and S. Irudaya Rajan, 2002. "District Level Estimates of Fertility from India's 2001 Census", *Economic and Political Weekly*, February 16, XXXVII, 7, 665-672.

Oliveau S. Ed., 2003, *Digital Atlas of South India 1991,* Cybergeo.

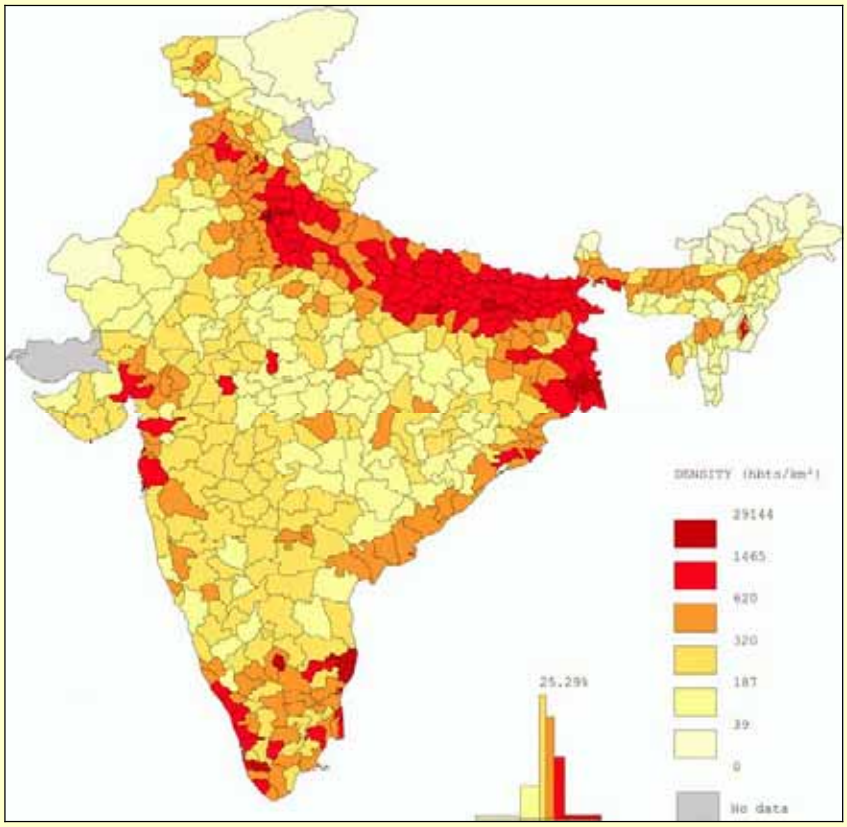
Pumain, D., et Saint-Julien, T., 1997, *L'analyse spatiale. 1. Localisations dans l'espace,* Armand Colin, Paris.

**Figure 2: State boundaries, India , 2001**

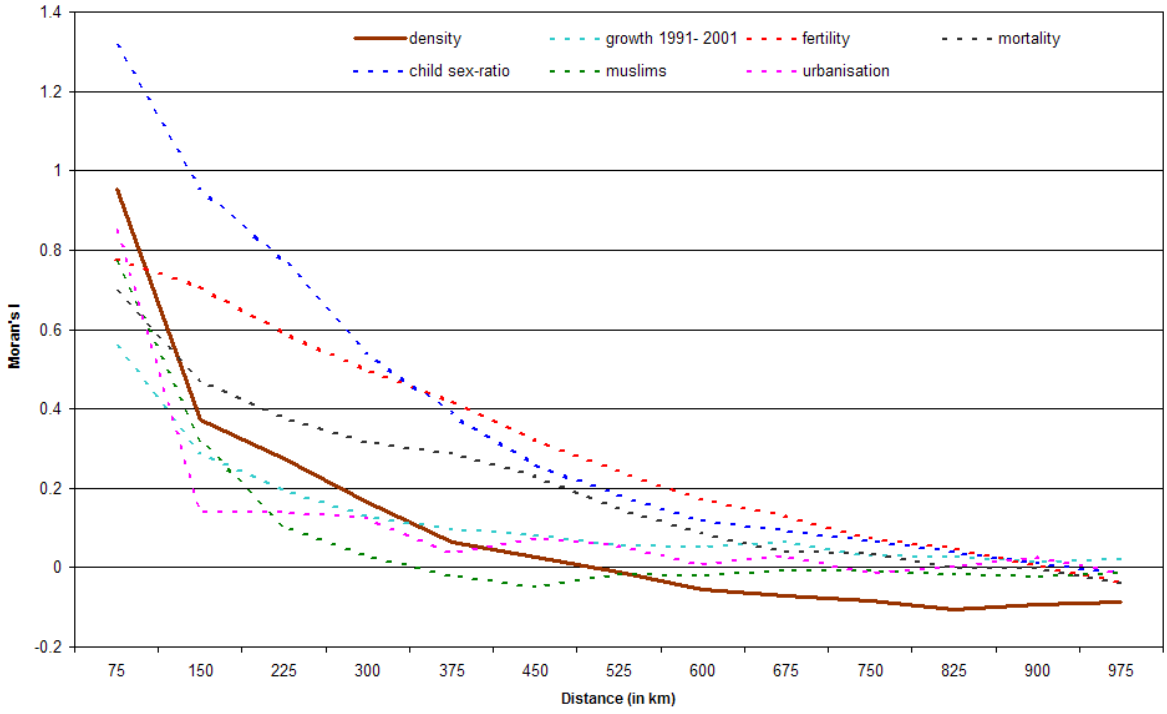

**Spatial analysis:**

- **Map of observed values**
- **Global Moran values (for various distance classes)**
- **Moran scatterplot (observed vs spatially averaged neighbouring values)**
- **Map of LISAs**

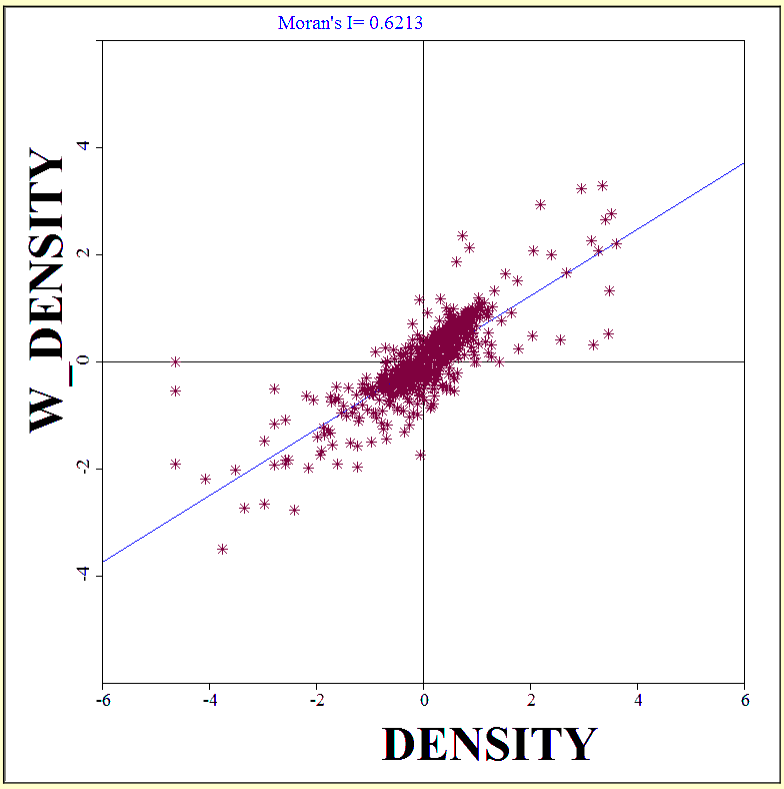
## DENSITY, 2001 (district map)



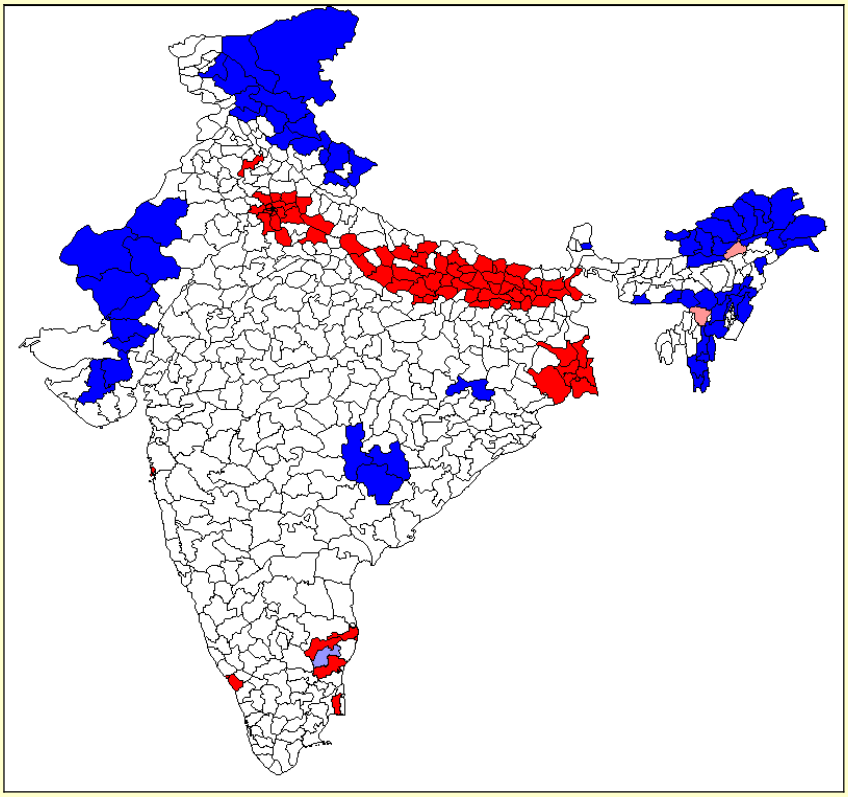

## Moran's index computed by distance class



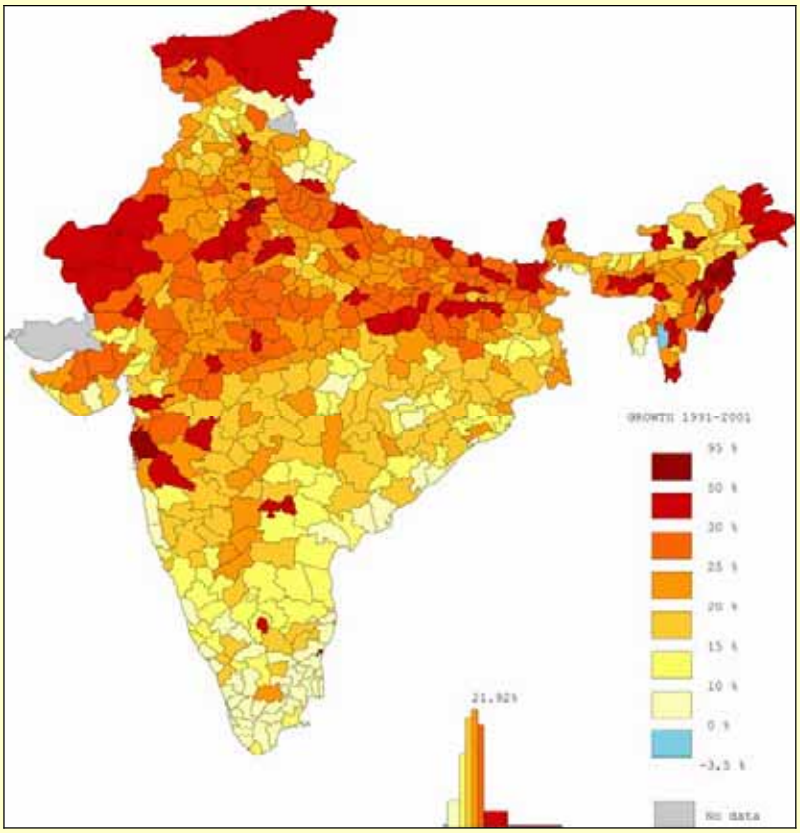
**Moran scatterplot (observed values vs. spatial average of neighbours)**



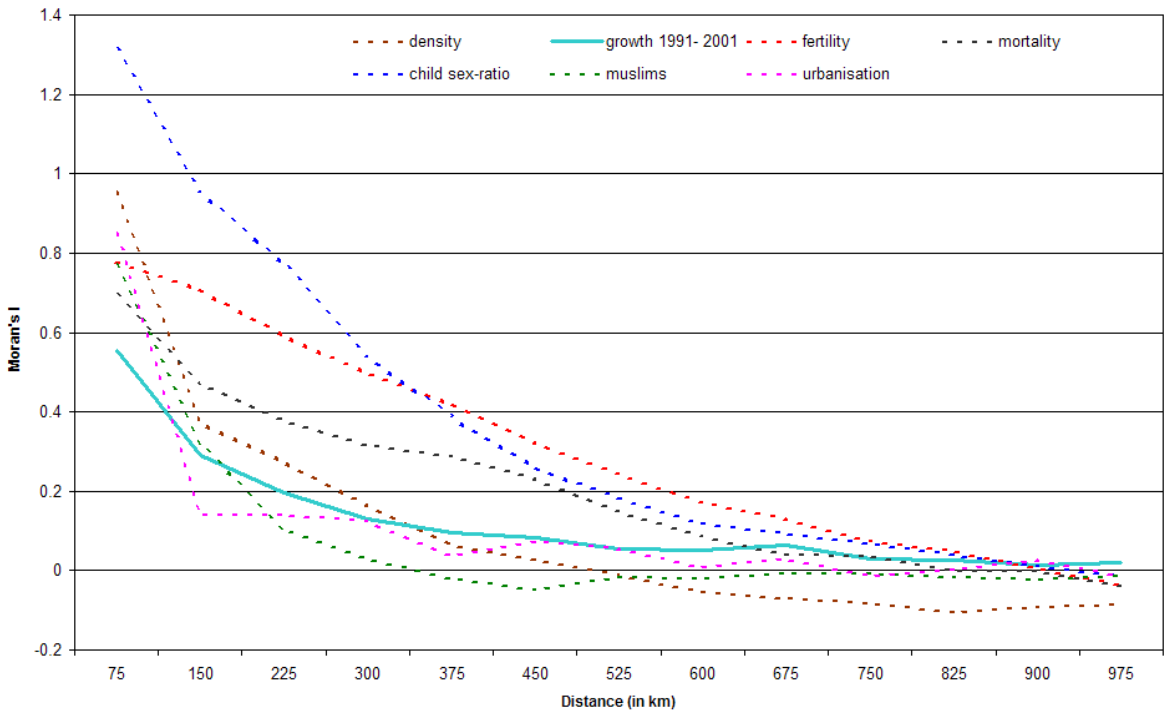

**Hot spots, cold spots and spatial outliers**



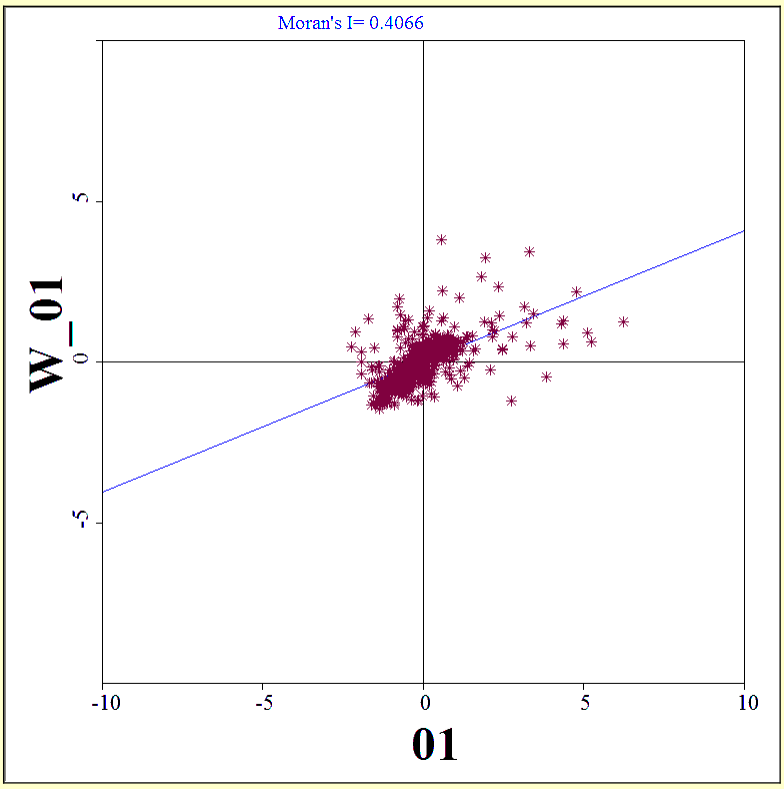
## GROWTH, 1991-2001 (district map)



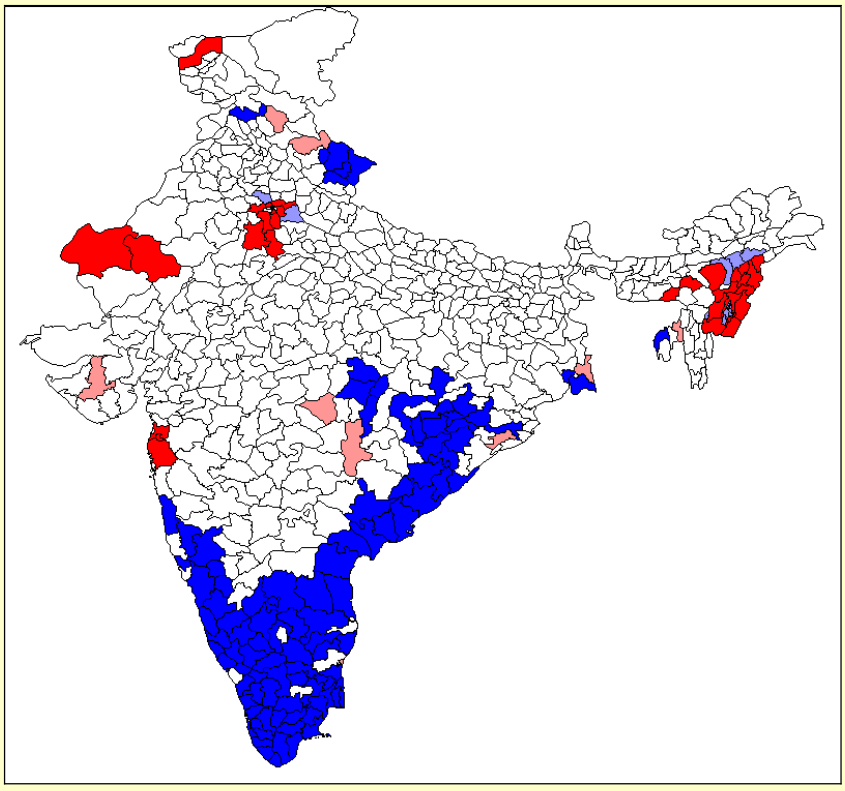
## Moran's index computed by distance class



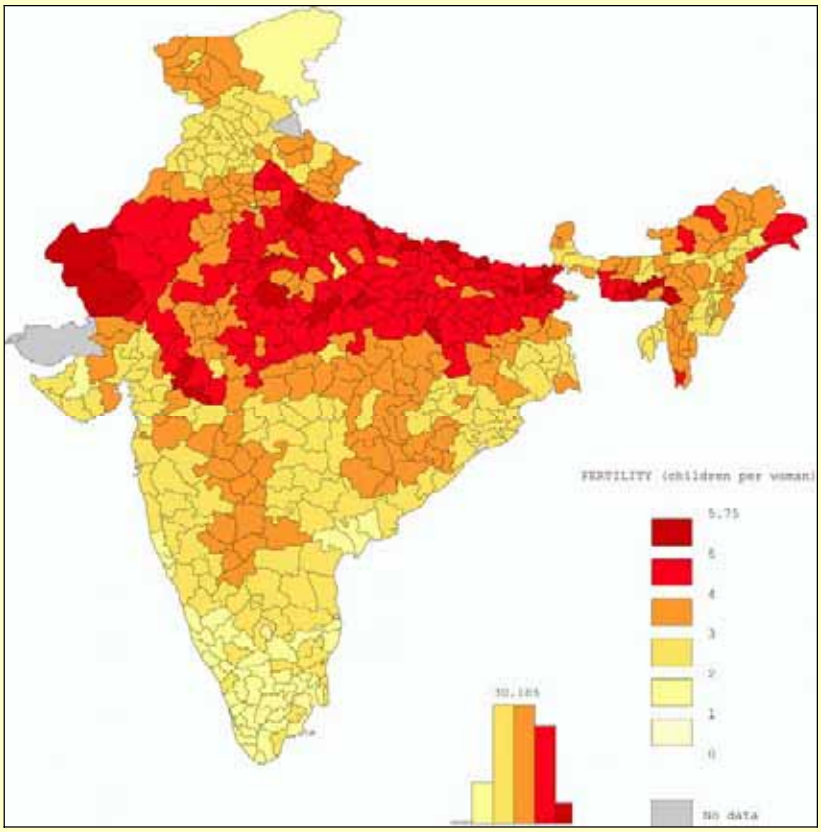
**Moran scatterplot (observed values vs. spatial average of neighbours)**



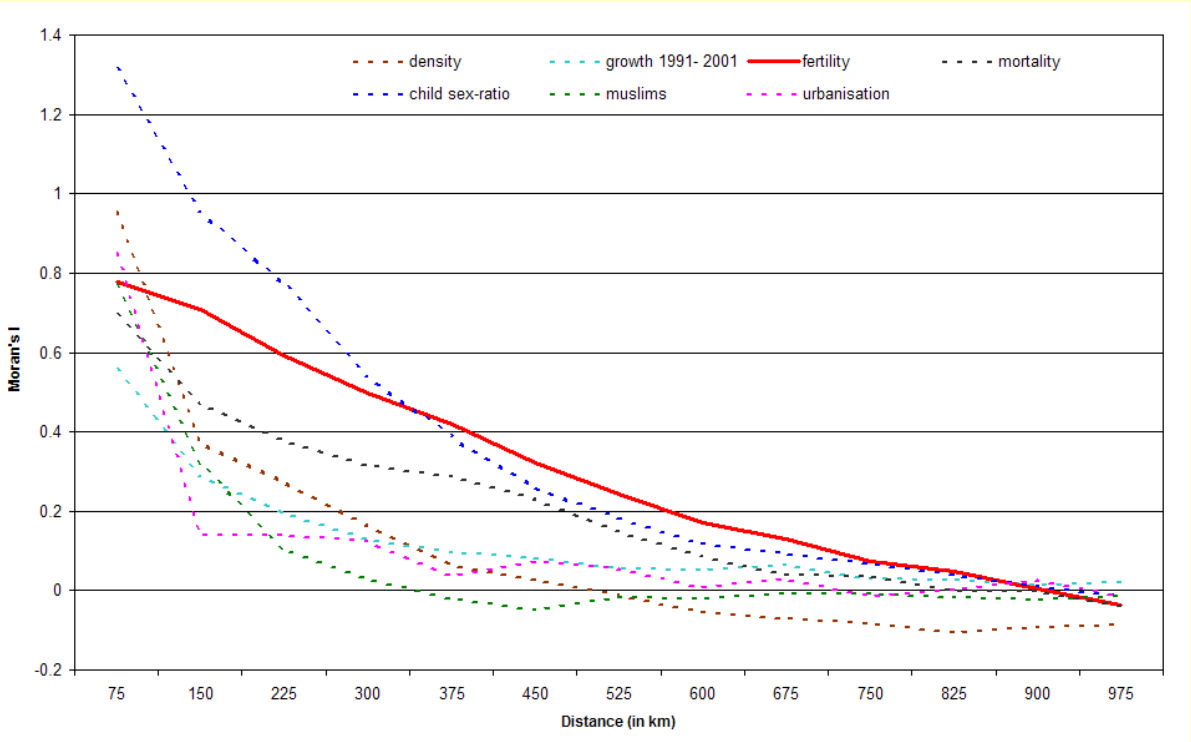

**Hot spots, cold spots and spatial outliers**



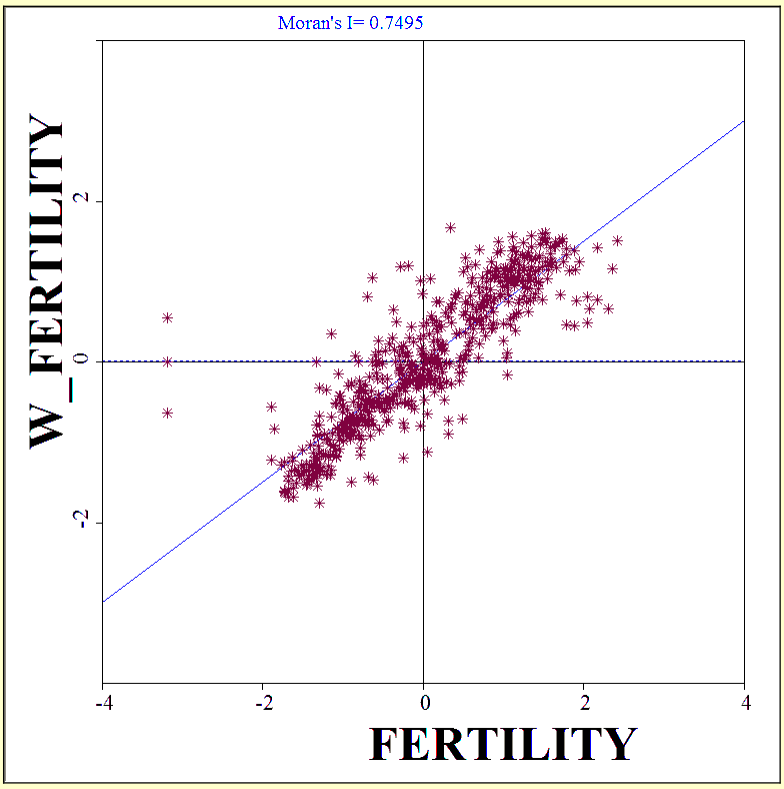
## FERTILITY (TFR), 2001 (district map)



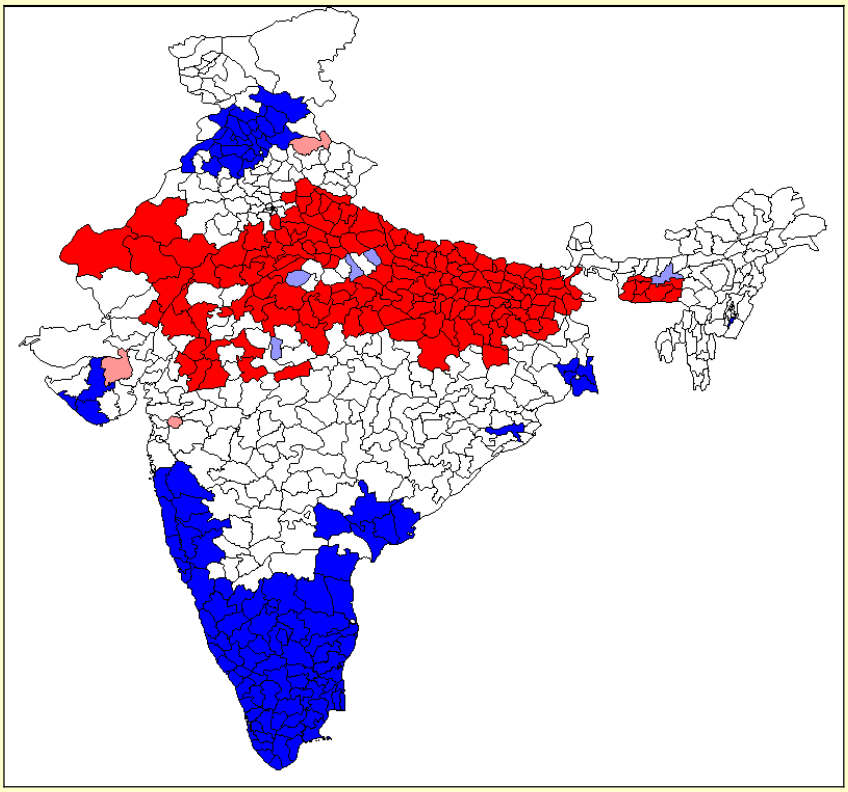

## Moran's index computed by distance class



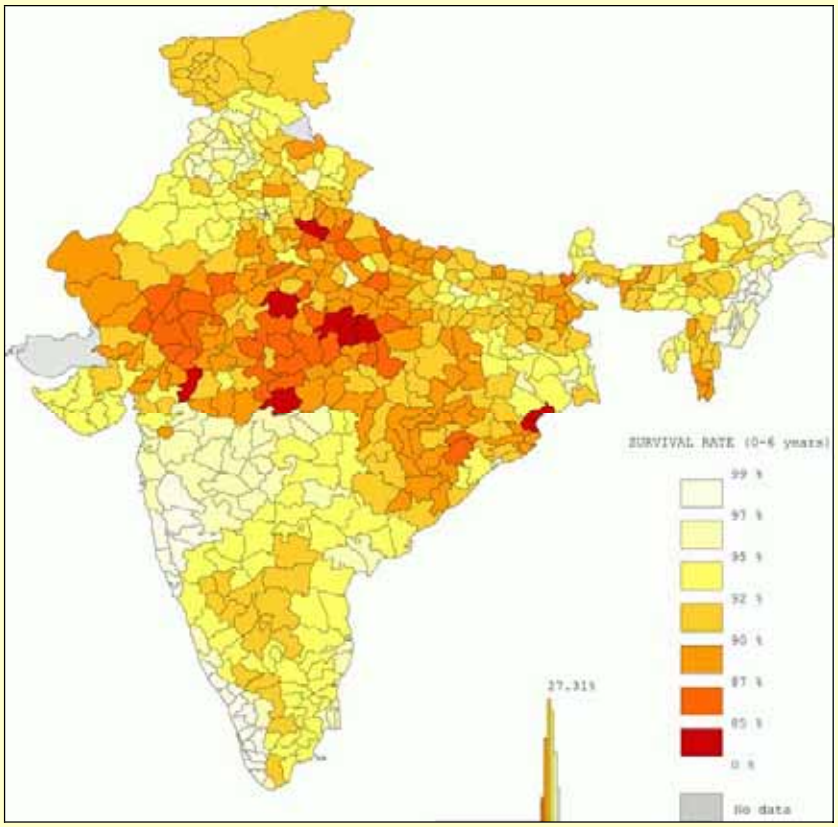
**Moran scatterplot (observed values vs. spatial average of neighbours)**



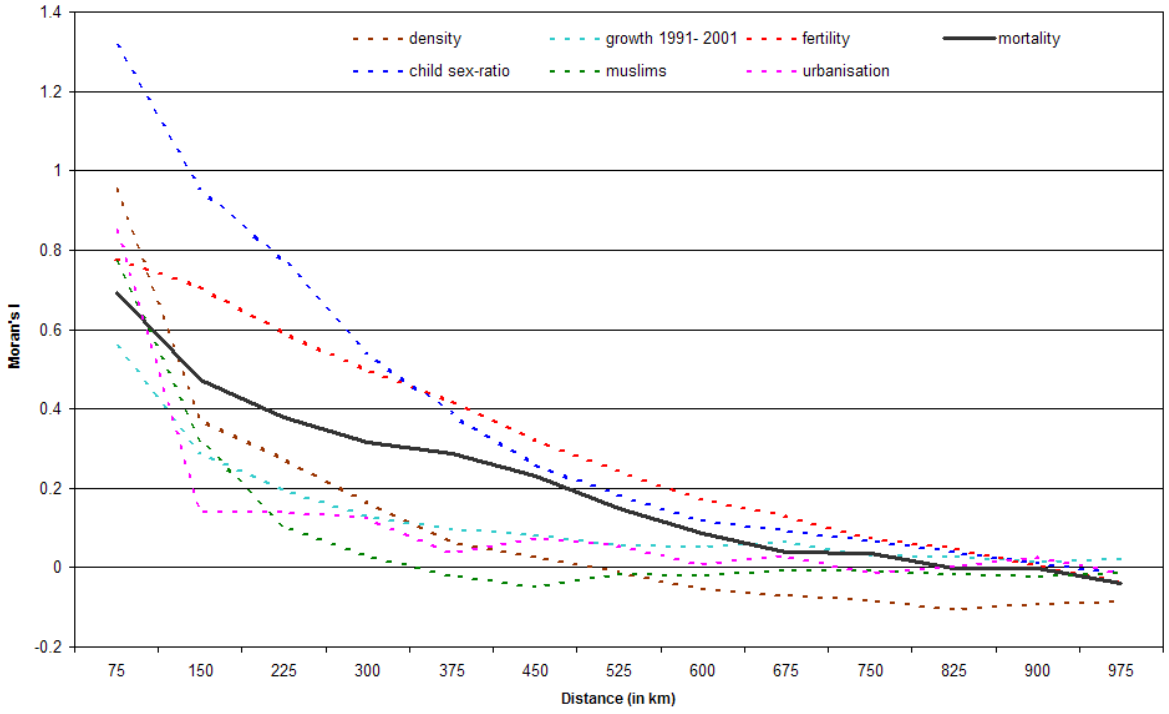

**Hot spots, cold spots and spatial outliers**



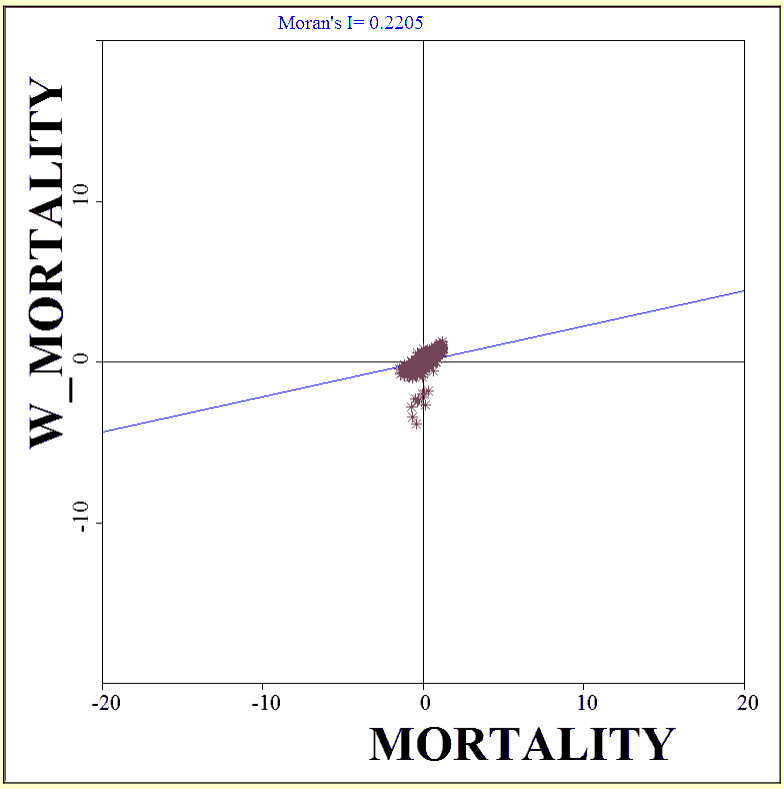
## SURVIVORSHIP, 2001 (district map)



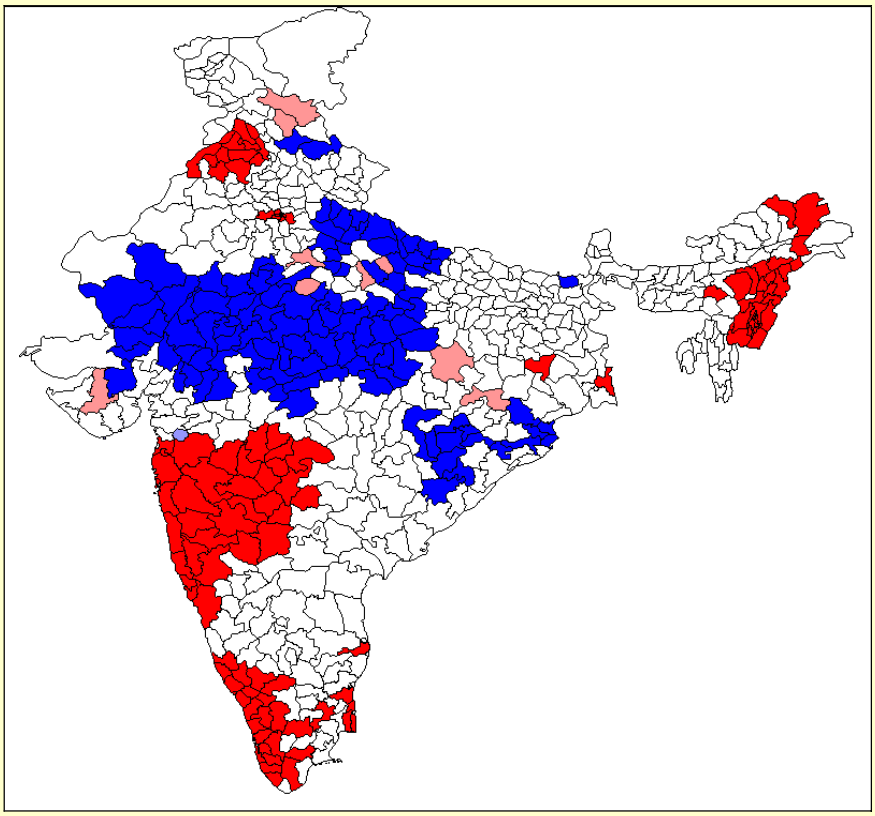

## Moran's index computed by distance class



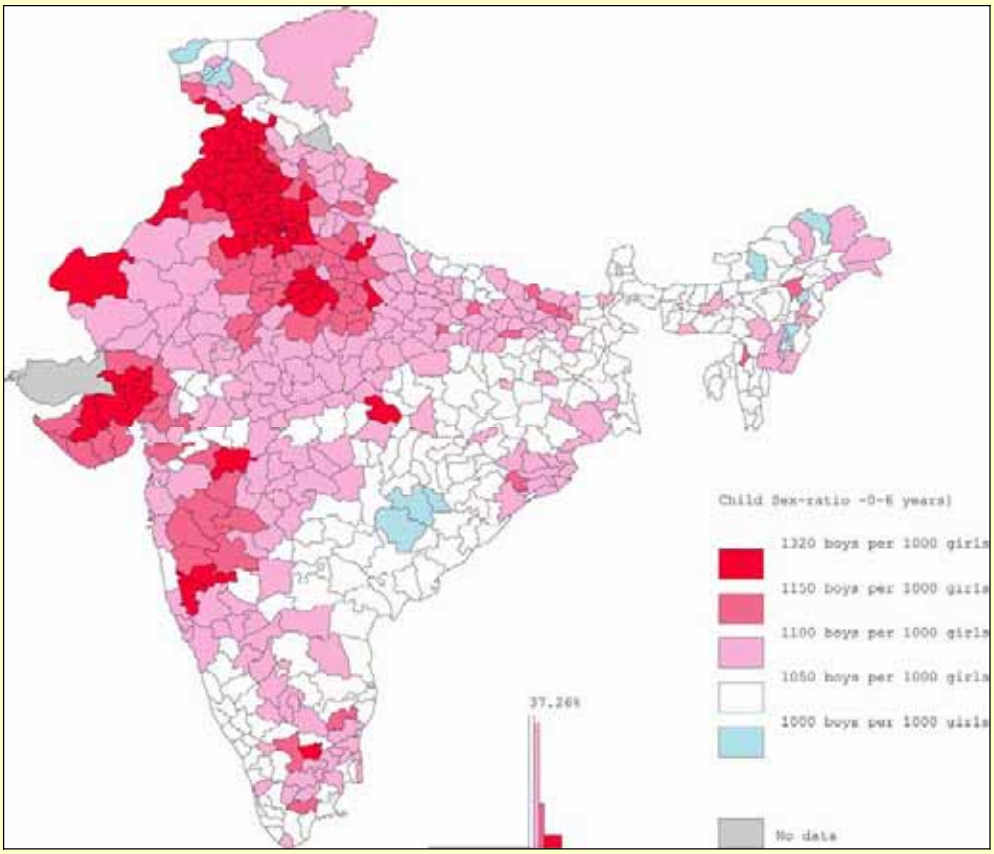
**Moran scatterplot (observed values vs. spatial average of neighbours)**



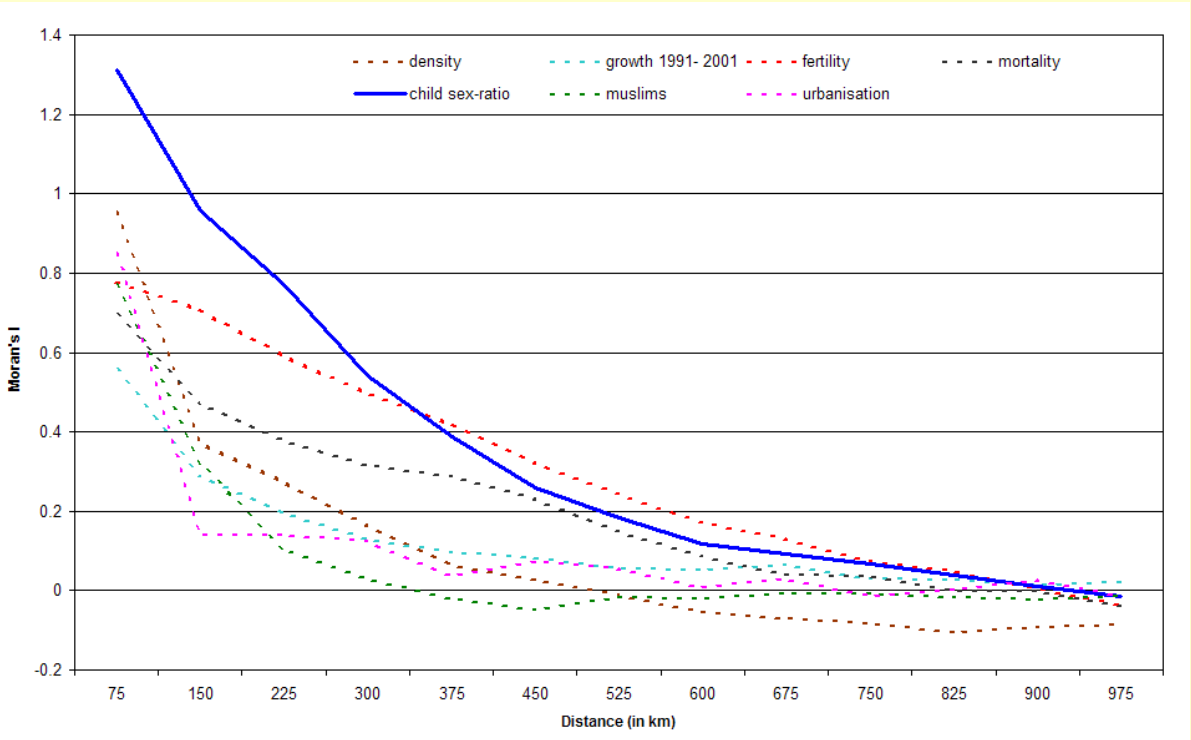

**Hot spots, cold spots and spatial outliers**



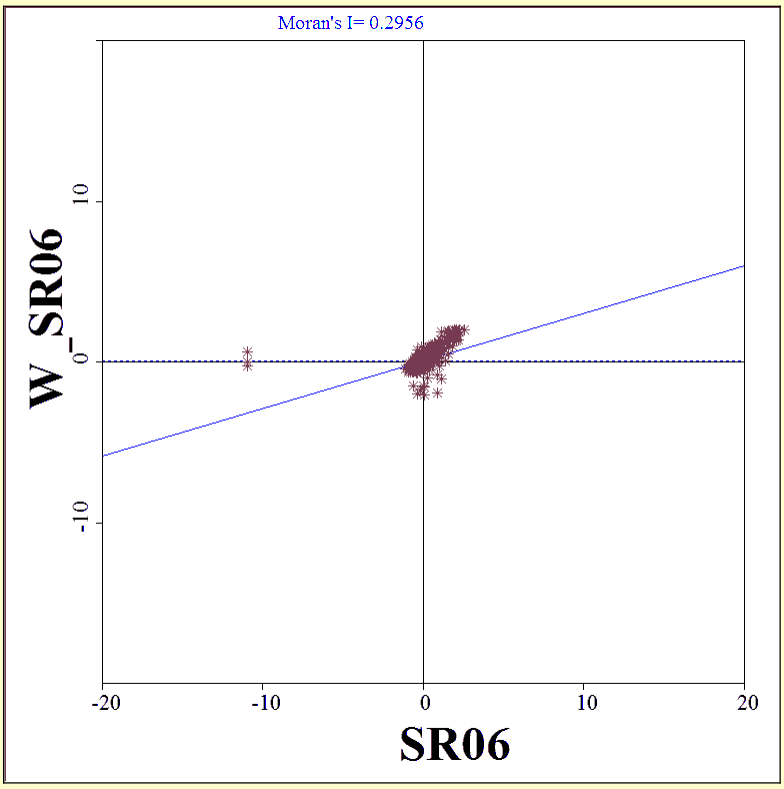
## CHILD SEX-RATIO, 2001 (district map)



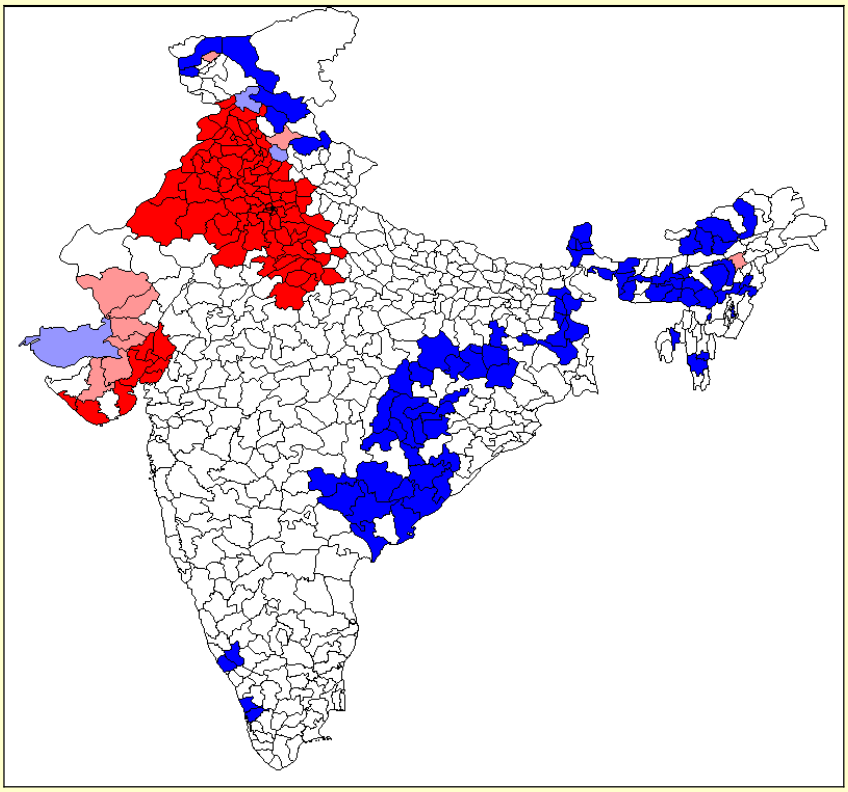

## Moran's index computed by distance class



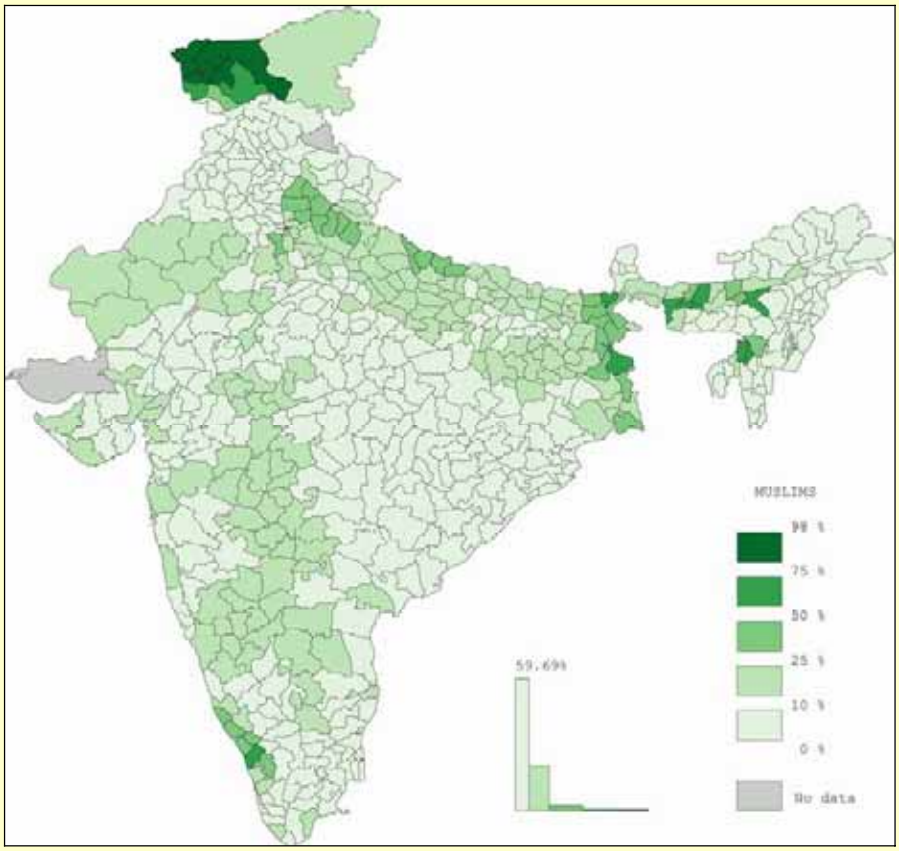
**Moran scatterplot (observed values vs. spatial average of neighbours)**



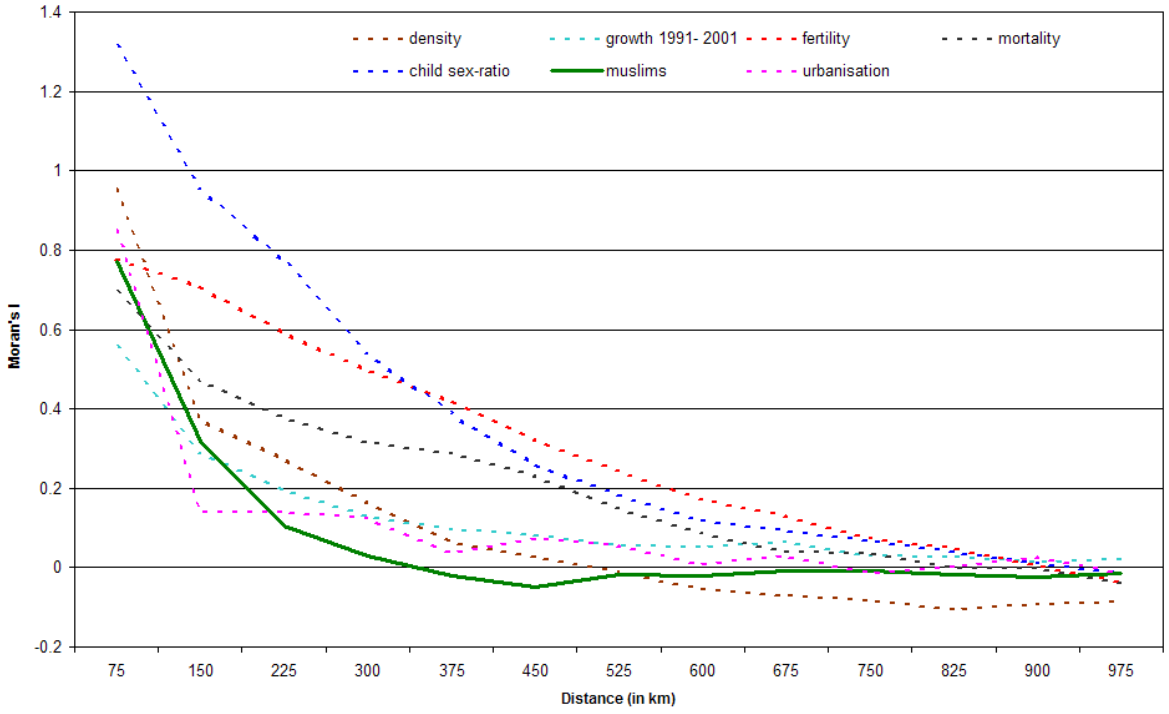

**Hot spots, cold spots and spatial outliers**



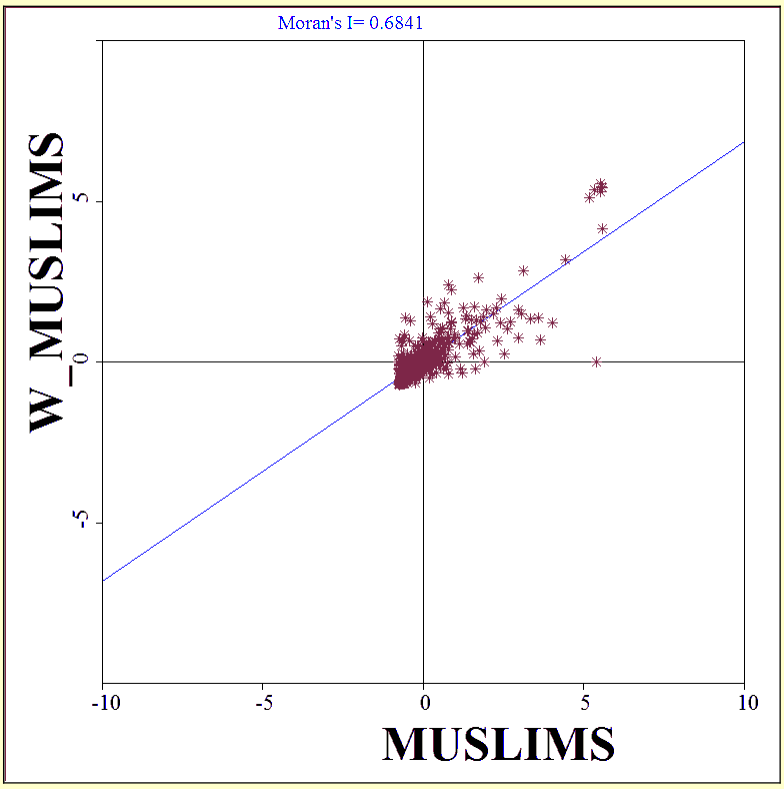
## MUSLIMS, 2001 (district map)



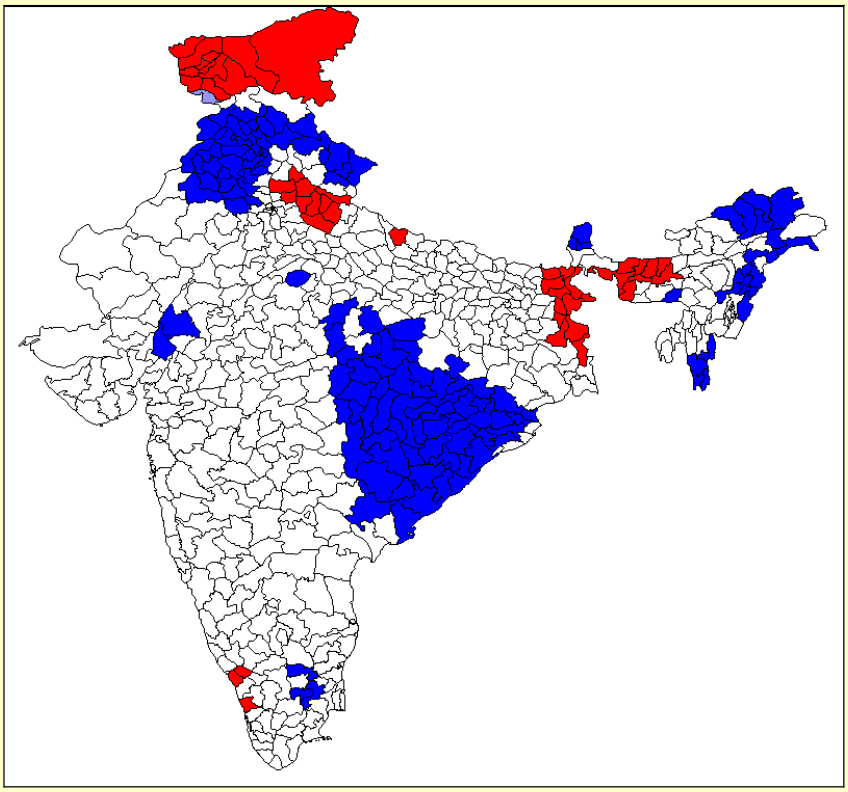

## Moran's index computed by distance class



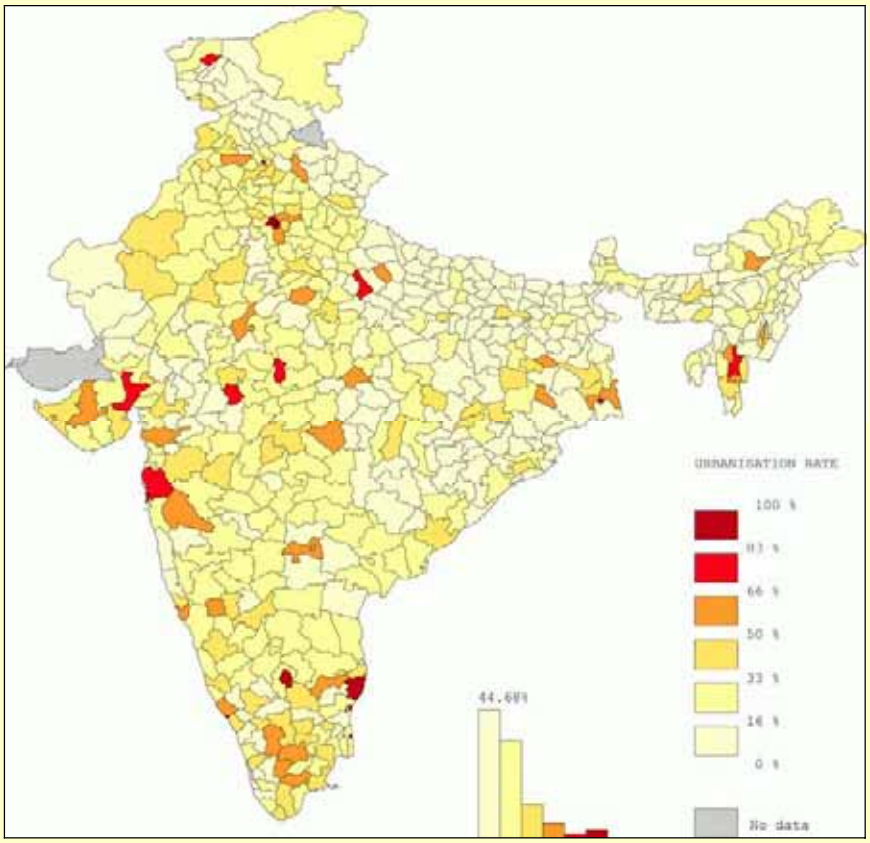
**Moran scatterplot (observed values vs. spatial average of neighbours)**



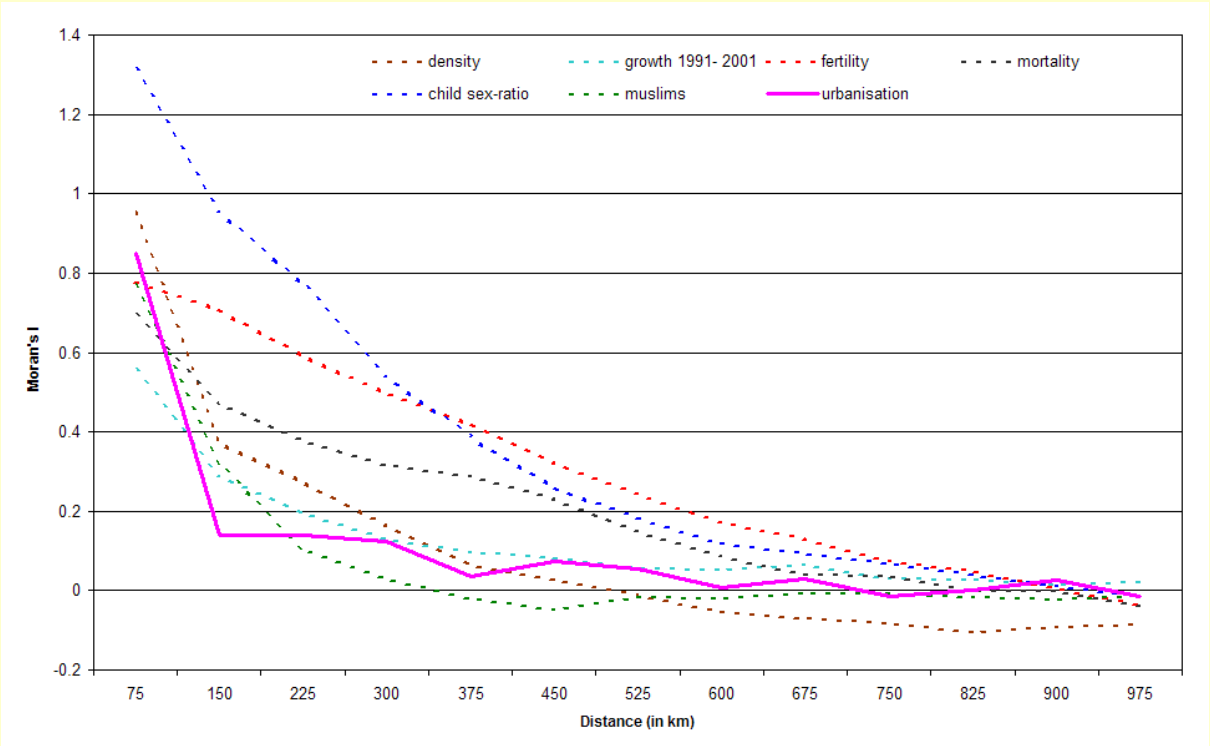

**Hot spots, cold spots and spatial outliers**



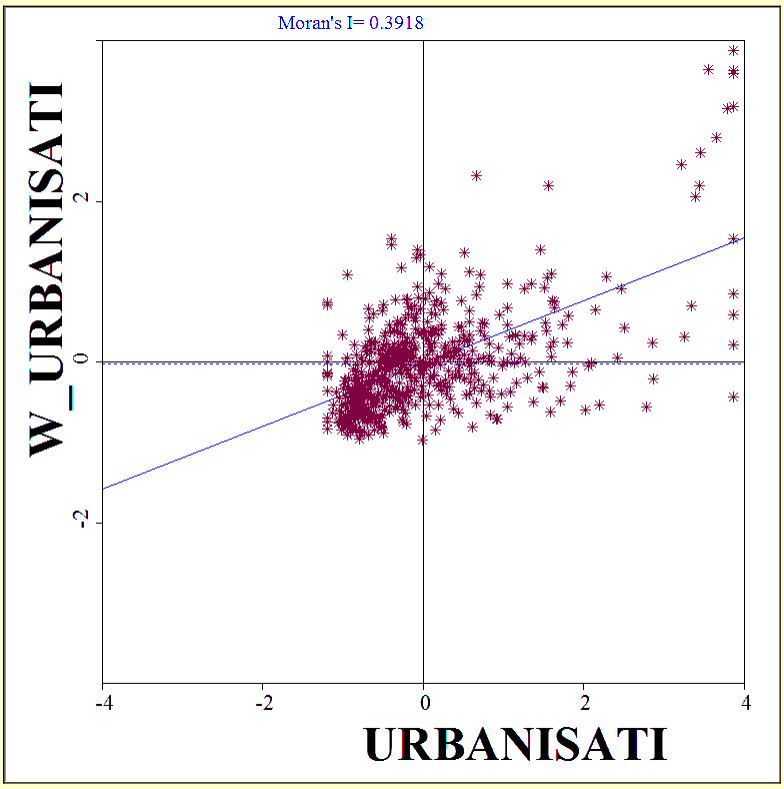
## URBANISATION, 2001 (district map)




## Moran's index computed by distance class



**Moran scatterplot (observed values vs. spatial average of neighbours)**




**Hot spots, cold spots and spatial outliers**