MALARIA FOCI AND COLONIZATION PROCESSES ON THE AMAZON FRONTIER: NEW EVIDENCE FROM A SPATIAL ANALYSIS AND GIS APPROACH

Marcia Caldas de Castro
mcaldas@princeton.edu

Burton Singer
singer@princeton.edu

Office of Population Research
Princeton University
Wallace Hall
Princeton – NJ – 08544

Paper presented at the XXIV General Population Conference (IUSSP) in Salvador, Bahia, Brazil, August 18-24, 2001 (Session S67: Demographic dimensions of land use and land cover change).

The authors thank A. Stewart Fotheringham, Chris Brunsdon and Martin Charlton, for providing the software to run the Geographically Weighted Regression, and for the valuable support they provided in response to our many questions during the modeling process.

We are grateful to the Center for Health and Well-being and the Office of Population Research at Princeton University, the Rockefeller Foundation, the International Development Research Center (IDRC), and CAPES/Brazil for financial support of this work.

SUMMARY

INTRODUCTION

Recognition of the importance of spatial location in the study of diseases date back to at least 400 BC, with the classical work of Hippocrates – *On Airs, Waters, and Place*. In this classical work, Hippocrates attests that a properly medical investigation should consider: the seasons of the year, the winds, the waters, and the place, since the influence of cities on diseases depends on their location in the hemisphere (Stevenson, 1994-2000).

In 1883, Hirsch (Hirsch, 1883) produced a very comprehensive work, composed of three volumes (one dedicated to acute infectious diseases, including malaria, another to constitutional anomalies of general prevalence, a third one focusing on endemic morbid process) emphasizing the importance of geographical location to the study of the pathology of diseases.

The incorporation of a spatial component in epidemiological studies per se dates back to 1854, with the classical work of John Snow, who mapped the cases of cholera in Soho, London and found a pump as the putative source of the disease[1]. After Pasteur developed the germ theory of disease (late 1860s), it was widely believed that mapping diseases would be useless. However, during the 1950s and 1960s, due to the work of a physician, Jacques May, spatial issues were reintroduced to medical research (Thomas, 1990). They have persisted to the present time.

From this perspective, it is not enough to know *when* an outbreak occurred and *who* is getting sick. It is imperative to know *where* the disease is really a threat to the population. The incorporation of a spatial component allows the identification of clusters or hot spots of a disease, the optimal allocation of new facilities (hospital, clinics, health posts, etc), the identification of putative sources that increase the risk of certain diseases (nuclear installation, contaminated water reservoirs, rivers and lakes, particular industries, etc.), and the definition of high-risk populations (Jacquez, 2000).

Shifting attention to vector borne diseases of the tropics, a common denominator of all of them is that transmission is very focal and ecologically driven. Subtle variations in landscape can have profound consequences for transmission of malaria, leishmaniasis, African trypanosomiasis, onchocerciasis, and Chagas disease (Service, 1989). In this paper we will focus on malaria in the western Amazon region of Brazil, as it relates to both local ecology and larger scale human migration and land use patterns.

Previous analyses of the process of agricultural colonization in the Amazon and its connection to the incidence of malaria (Sawyer and Sawyer, 1987; Sawyer, 1992a; Sawyer, 1992b) utilize multivariate statistical methods that assume spatial stationarity. They employ global measures that describe relationships between variables and their effects on malaria transmission, considering the entire studied area as a spatially homogeneous place. That assumption is most likely to be unrealistic, and important spatial dissimilarities will be hidden by a global analysis. If spatially local measures were utilized, important findings, masked by a global approach, could represent the key factors that lead to the success of control and prevention policies. Sawyer and Sawyer (1992) called attention to this fact, highlighting the importance of introducing spatial relationships in the analysis, using maps and satellite images.

In this paper we take up this challenge. We study colonization processes and malaria transmission in the Amazon, not only from a temporal perspective, but also from a spatial one. The

[1] For a detailed description of this study, check the website *http://www.ph.ucla.edu/epi/snow.html*.

study area is a colonization project in the western part of the Brazilian Amazon named Machadinho, for which survey data is available at four different points in time, from the very beginning of the project, until ten years later. Details about the surveys are given in Sawyer and Sawyer (1987) and in Singer and Castro (2001).

Using multiple methods of local spatial analysis, we investigate the existence of significant malaria foci, their relationships with other variables, and infer plausible explanations for their locations. We clarify how spatially insensitive global measures mask important spatial differences, which are essential for understanding the determinants of malaria transmission in the Amazon.

In the next two sections we briefly review the process of frontier expansion in the Brazilian Amazon, especially in Rondônia, and its relationships with malaria incidence. In the third section we summarize the main characteristics of the Machadinho settlement project, and point out the unique features that make our survey data the best available in Brazil for following the dynamics of malaria incidence in the tropical forest. Section 4 reviews the major findings from global multivariate analyses done so far with Machadinho data. The fifth section describes the spatial statistical methods used in this paper. In Section 6 we present results of the local analysis, including maps produced with GIS tools. In the concluding section, we clarify the importance of the new evidence for the analysis of malaria risk per se and for the design of mitigation strategies, and point out future research that has to be done, refining the spatial investigation of malaria transmission.

1. COLONIZATION ON THE AMAZON FRONTIER

Large scale government intervention in the Amazon started during the first decade of the military regime (1964-1974). It was characterized by what were called “integration programs”. The large area, sparse population, and unpatrolled borders of the Amazon, shared with eight other countries - associated with the rubber boom and the mining rush – were associated with military concerns of national security (Mahar, 1979). Occupying the Amazon was also seen as a potential solution for population pressure, and a target for rural out-migration from the Northeast region of Brazil (Wood and Wilson, 1984).

The construction of highways was adopted as a means of achieving the goal of integrating the Amazon with the rest of the country. Some highway construction had a military pretext, the most notable being the Transamazon highway (Sawyer, 1984). Once roads were opened, new landless migrants came and settled either in private or government-delimited areas (colonization programs) or voluntarily in nearby locations. Gold, iron ore, and other extractable minerals were discovered and exploited. Indigenous communities were disturbed. Artificial dams were constructed, facilitating the spread of malaria, among other diseases. Small communities proliferated near the roads, with very poor infrastructure. Deforestation assumed dramatic levels.

The scenario during the 1980s was not much different. Promotion of colonization projects continued. In Rondônia, the migration process that began in the early 1970s increased significantly when large colonization projects were initiated by the Federal Government, most of them located along the BR-364 highway, which connects the capital of Mato Grosso state, Cuiabá, to Porto Velho, the capital of Rondônia (Sawyer, 1986). A publicity campaign claiming availability of fertile soils brought a dramatic influx of people to this State. While the population of the whole country was growing at an annual rate of 2.48% during this period, Rondônia registered an astonishing yearly growth rate of 16.03%, with even higher increases in the rural areas (17.69%).

In 1981, the Brazilian Government established the Northwest Brazil Integrated Development Program (POLONOROESTE), with financial cooperation of the World Bank. It was a complex, ambitious and totally new challenge both to the World Bank and to the Brazilian government. It was the Bank's first attempt to support agricultural development in the Amazon, combining many different sectors (transportation, rural development, education, health, Amerindian protection, environment and agriculture). It started in 1982, and the general goals of the program were the pavement of the existing BR-364 highway, and the promotion of a harmonious socio-economic development of the area along this road, protecting the environment and the indigenous tribes. The original plan included the creation of six colonization areas, and the settling of 15,000 farm families. Due to the marginal quality of soils only three areas were settled involving approximately 5,000 families. (World Bank, 198?).

The Machadinho settlement project was part of POLONOROESTE. It was carefully designed to be a model of ecological and social planning in the Amazon, which would overcome the problems experienced in previous initiatives. In practice, however, these expectations were not fulfilled. Deforestation was huge. According to NASA (National Aeronautics and Space Administration) reports, the damage to the forest was considered to be the largest man-made transformation to the earth's surface visible from space (Caufield, 1996). The other serious problem was malaria, which soon became an epidemic.

In summary, the problems of the frontier expansion were: uncontrolled deforestation, lack of technical support to settlers, difficulties with storage and distribution of agricultural products, emergence of disease epidemics (especially malaria), violent conflicts over land tenure, soil quality much inferior than initially publicized, and conflicts with indigenous populations (Mahar, 1979; Schmink and Wood, 1992).

2. MALARIA ON THE AMAZON

Tropical rain forest areas assure good conditions for the spread of insects, given their high temperatures, humidity, and rainfall throughout most of the year. Malaria at colonization sites in the Amazon is defined as "frontier" malaria (Sawyer, 1988), when most or all of the following conditions are present: high vector density, intense human exposure to vectors, outdoor transmission, low immunity of the exposed population, limited knowledge of the disease, high morbidity and relatively low mortality, high proportion of *P. falciparum* malaria (the disease parasite), difficulty of conventional control measures to succeed, weak institutional presence in the area, little sense of community, high population mobility, and political marginality.

In any colonization project the first environment transformation is the process of clearing the land and preparing it for cultivation, usually through the use of the slash-and-burn technique. A poor burn compromises the crop yields, and contributes to the proliferation of breeding sites for mosquitoes (Fearnside, 1986). Moreover, a poor clearing and burning process can cause the obstruction of steams and will leave the taller trees standing. This provides the necessary partial shade for *A. darlingi* – the primary Amazon vector - breeding. Furthermore, the clearing process decreases the number of wild animals that previously were sources of blood meals for mosquitoes, bringing a change in blood feeding habits (Deane, 1986).

It is worth mentioning that the relation between deforestation and malaria outbreaks has been understood for a long time. Hirsch (1883: 273) wrote: "…*the breaking up of virgin soil and other operations of that kind, the cutting down of woods, and the neglect to cultivate ground that used to be tilled, are favorable to the occurrence and prevalence of malarial disease*".

Environmental transformations redefine the notion of forest fringe, a boundary between the forest and human-occupied property, where the risk of malaria transmission in the Amazon is very high. Farmers can be more or less exposed to *A. darlingi* mosquitoes (the most common vector of the disease in Machadinho) depending upon where their houses are built and even the hours of the day when they are out of doors. Malaria is considered to be one of the major causes of dropout of settlers, and, ultimately, of the failure of colonization efforts.

During the past 50 years, malaria reached its lowest level in 1970, when only 52.5 thousand cases were registered in Brazil (Deane, 1988). After that, however, the spread of colonization programs, mining exploration, massive migration, and the construction of roads and dams brought a new social, economic, environmental, and epidemiological reality. That facilitated malaria transmission.

In 1982, the number of malaria cases rose to 200,000. They were concentrated in Amazonian settlements and mining areas. By 1988 there were half a million cases, basically concentrated in Rondônia, southern Pará, and Maranhão. In 1996, more than 99% of the national cases were in the Amazon, and the last data available show that, in 1999, the Legal Amazon registered more than 630 thousand cases, an increase of 34.21% compared to the previous year (FUNASA, 2000).

In 1970, 1980, and 1986, Rondônia represented 11%, 35% and 42%, respectively, of all cases of malaria in Brazil. In 1985, 10,492 positive cases were registered in Machadinho, representing 39% of the slides examined. In this same year, Machadinho was responsible for 7% of malaria cases in Rondônia and almost 3% in Brazil, as well as being the source of many of the cases exported to southern Brazil, through migratory movements (Sawyer and Sawyer, 1987).

Most recent data from the National Health Foundation includes Machadinho as an area of high malaria risk, with an Annual Parasite Index (API)[2] in 1999 of 252.4 per 1,000 people. Areas with an API greater than or equal to 50 are considered high risk by the government (FUNASA, 2000). However, this API is far from the dramatic numbers reached in the mid-80s, when the API in Machadinho was more than 3,000 per 1,000 people (Sawyer and Sawyer, 1987).

3. STUDY AREA

The study area is a settlement project located in the state of Rondônia, in the western part of the Brazilian Amazon, named Machadinho d'Oeste (Figure 1). The area was originally sparsely inhabited by rubber tappers, and considered not to have been a malarious location (Sawyer and Silveira, 1985).

Machadinho was officially created by the National Institute for Colonization and Agrarian Reform - INCRA[3] in 1982, as part of the POLONOROESTE Program, and occupies an area of approximately 119,700 ha divided in two tracts[4]. Tract 1 is located below the main tributary of the area (Machadinho River) and has 602 plots with 27,410 ha, 3 protected forests with 20,396 ha, and 6

[2] The API is obtained by dividing the number of positive slides by the total population.

[3] INCRA was the government institution responsible for coordinating settlement processes.

[4] The original design of the Machadinho Colonization Project included seven tracts with 5,520 plots. Later reviews of the project, based on the negative outcomes of the initial years, reduced the original 5,520 plots to only 2,934 divided into four tracts. This paper cover only Tracts 1 and 2, the first opened to the settlers. The other areas are Tract 3, with 622 plots and approximately 49,000 hectares; and Tract 6, with 570 plots and roughly 40,000 hectares.

secondary urban centers with 505 ha, comprising a total area of 48,311 ha. Tract 2, located above the Machadinho River, has 1,140 plots with 53,777 ha, 6 protected forests with 15,069 ha, an airport with 59 ha, the main urban center with 2,035 ha, and 4 secondary urban centers with 449 ha, totalizing an area of 71,389 ha (INCRA, 1985).

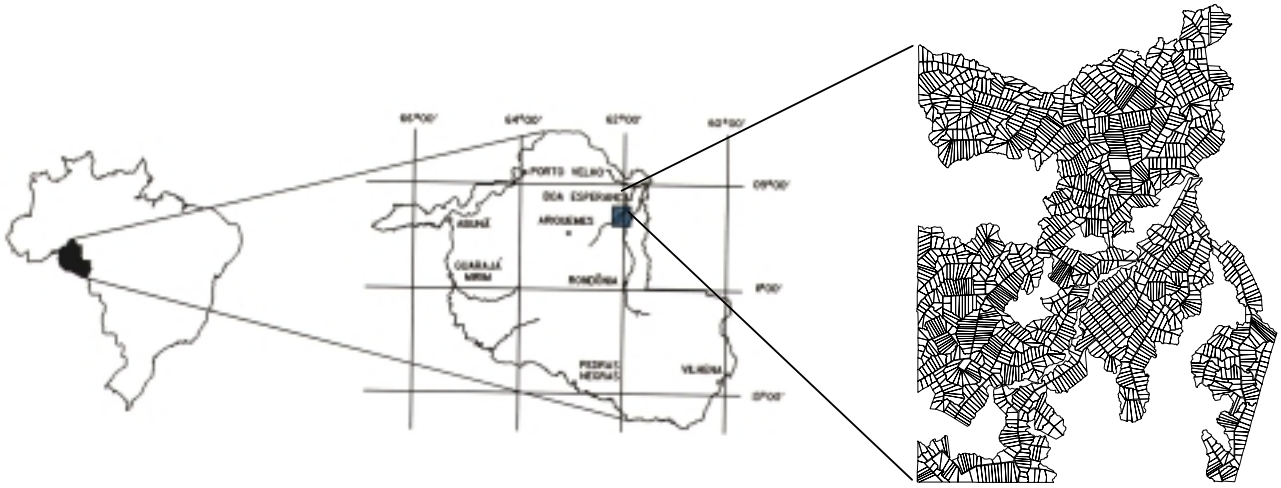


Figure 1 - Spatial location of the study area

The design of the plots was carefully planned, and was believed to reduce the potential breeding sites for mosquitoes near houses (Castilla and Sawyer, 1993). The plot layouts followed the course of the rivers and streams, so that most plots would have frontage on a road and the rear of the plot close to a source of water. Topography was also an issue in specification of plot boundaries. This arrangement of plots, distinct from the traditional fishbone pattern (Fearnside, 1986) distinguishes the Machadinho settlement in any satellite image.

A road network of approximately 470 km makes all plots accessible, and although not paved, they are of good quality. However, some roads require special vehicles during the rainy season (Sawyer, 1985).

The average size of the plots is 35 to 40 ha, with average frontage along the road of 400-500 meters, and average depth of 700-900 meters (Sawyer and Sawyer, 1987). The plots are the smallest spatial unit in the area, and are the spatial unit of analysis in this paper.

The settlement began in mid-1984, when roads were opened in tracts 1 and 2. Plots were distributed in July and August, too late for clearing and burning before the rainy season[5] (Sawyer and Sawyer, 1987). Many settlers abandoned their lands and later found out that they had been distributed to another applicant, increasing conflict in the area.

Altitude at the site varies between 100 and 200 meters. A dense network of streams flowing into the Machadinho River drains the area. The climate is hot with a very short dry season (during the months of June, July, and August). Because of this seasonality in the rainfall, there is great seasonal variation in the water levels of the rivers. Quiet or stagnant pools are observed at the beginning and

[5] On June the area is cleared and left to dry for 40 days. At the end of August the burning process takes place (Sydenstricker and Vosti, 1993).

end of the rainy season. Average annual temperatures are above 25°C, reaching an average monthly maximum above 32°C from July to October, and relative humidity is usually above 80% (Sawyer and Sawyer, 1987).

Field surveys were carried out in Machadinho since the beginning of the project, allowing for a dynamic analysis of how a settlement process evolves through time. To our best knowledge, this is the only available survey data in the Amazon that follows a colonization project from its onset, collecting information on health, demographic, economic, ecological, and agricultural characteristics.

Data was collected in 1985, 1986, 1987 and 1995 (Sawyer, 1985). In 1985 it is estimated that 76% of the occupied plots were covered by the survey, while in all the other years the coverage was total. The survey was designed to include all settlers, except those that did not clear any forest in their plot, and those that were not living at least part-time in the settlement project (Sawyer, 1985). The questionnaire included data on migratory history, malaria episodes, knowledge about malaria transmission, socio-demographic characteristics of the settlers, land use, ecological transformations, agricultural production, and housing conditions. The survey was organized and coordinated by the Center for Regional Development and Planning - CEDEPLAR of the Federal University of Minas Gerais – UFMG. The Remote Sensing Center - CSR of UFMG produced a digital map of tracts 1 and 2 of the Machadinho project.

4. MAJOR FINDINGS FROM A NON-SPATIAL ANALYSIS

The principal method of analysis used with Machadinho data was logistic regression with discrete-level covariates. The response variable used was the exposure weighted malaria illness rate[6], which has the number of months – out of the 12 months prior to the survey – with malaria as the numerator, and the exposure time in Machadinho as the denominator.

The age of the colonization project is an important factor. In older projects, according to Sawyer and Silveira (1985), the environment is less hospitable to *A. darlingi* and malaria transmission decreases. This phenomenon is connected to the malaria transition theory in frontier settlements proposed by Sawyer and Sawyer (1992), which assumes three stages: epidemic, transitory and endemic. The last stage starts approximately eight years after colonization is initiated, when the environment has already been drastically changed (large deforested areas, improvement in sanitation, and pollution of breeding places for mosquitos, among other factors).

Another important issue is the origin of the settler. When coming from other malarious places, they constitute a potential source of the disease, particularly if no effective anti-malarials are taken. On the other hand, if they were previously living in places with no malaria, they have no immunity against the disease, and probably no knowledge about the transmission process and preventive measures (Sawyer and Silveira, 1985). This places new settlers at increased risk of debilitating episodes of malaria.

Sawyer (1988) analyzed the 1985 and 1987 Machadinho data, identifying important relationships. A lack of knowledge about the transmission process and the vector of the disease was an important variable contributing to high malaria incidence in 1985. However, individual knowledge about malaria was less important in 1987. The same phenomenon was observed for the level of education of the head of the household's wife. People who worked on agriculture or that had other

[6] From this point on, we'll refer to this measure just as malaria rates.

activities near the forest, river and streams were exposed to a higher risk of malaria transmission. Settlers living in bigger, better and DDT sprayed houses, far from the forest, in plots with a large cleared area, exhibited lower malaria rates. It is interesting to note, however, that the amount of cleared area became an important variable for reducing malaria rates only at later stages of the settlement. During the onset of colonization, despite the cleared area, there was a still significant contact with the forest. This fact conforms to the statement of Sawyer (1992a: 12) that "*deforestation initially provokes epidemics*". The author shows that at later stages the increase in deforestation implies less malaria, unless the crops reproduce the forest environment.

The income of the household had a weak differential in malaria rates (Sawyer, 1992a). The distance from the house to a secondary growth was significant, with small distances associated to more malaria. Plots with pasture or large areas of annual or permanent crops had less malaria (Sawyer, 1992b).

In terms of odds ratios, Sawyer and Sawyer (1987) showed that those who live in a house at a distance equal to or higher than 51 meters from the forest, in 1985, have an odds ratio of 0.4043, a decrease of almost 60% in malaria rates when compared to those living at a distance lower than 51 meters. Considering the number of goods owned by the household (including radio, clock, hunting gun, bicycle, sewing machine, gas lamp, pressure pan, water filter, gas stove, refrigerator, chainsaw, planter, horse, wagon, car, stereo, TV, satellite dish, and VCR), those with five goods or more have an odds ratio of 0.7731, which represents a decrease of 23% in malaria rates compared to those owning less than five goods.

The only attempt to analyze spatial differences in malaria rates was done by Sawyer and Monte-Mór (1992). The authors used an aggregation of plots by sectors (three in Tract 1, and four in Tract 2) as a proxy of the macro-environment of the area, and included this variable as one of the covariates in a logistic model. Coefficients were obtained for each sector, and for interaction of them with other variables. Sectors of Tract 1 showed a higher risk of malaria, while in Tract 2 only the sector with most precarious access presented higher risk.

Although extremely informative and able to reveal spatial differences in the incidence of malaria in Machadinho, this approach does not take into account the spatial association of the data. The definition of sectors is arbitrary, based only on location, and not on the spatial dependency of malaria rates. Table 1 summarizes the major findings from the global multivariate analysis.

Table 1 – Expected effects of selected variables on malaria rates

Variable	Expected effect
1. Large amount of annual or permanent crops	-
2. Large pasture area	-
3. Small distance from the house to secondary growth	+
4. No knowledge about malaria transmission	+ (at early stages)
	None (at later stages)
5. Level of education of the head of the household's wife	+ (at early stages)
	None (at later stages)
6. Working in agriculture	+
7. Living in sprayed houses	-
8. House located far from the forest	-
9. Cleared area	+ (at early stages)
	- (at later stages)
10. Large number of goods	-
11. Living in Tract 1	+

Finally, it is important to note that some of the relationships described here can change in magnitude, and even direction, when interactions between covariates are introduced into the model. However, our intention is just to summarize the global associations more nuanced.

5. SPATIAL ANALYSIS

Any study of spatial analysis has to consider two important issues: the modifiable areal unit problem and the spatial autocorrelation problem (Arbia, 1989). The first is related to the arbitrary way in which regions (like counties, for example) are divided. The second issue is based on Tobler's First Law of Geography that "*everything is related to everything else, but near things are more related than distant things*" (Tobler, 1979).

The unit considered in this analysis is the smallest spatial unit available in Machadinho, the plot. Different aggregations are available, like the concept of sectors used by Sawyer and Monte-Mór (1992), the areas used as reference by the National Health Foundation, or, ultimately, the notion of tracts. Considering the plot as the spatial unit, and taking into account the spatial autocorrelation of the data, reduces both the problems of atomistic and ecological fallacies.

Spatial association can be measured with different statistics. In general matrix notation, spatial association can be expressed as

$$\Gamma = \sum_{i}^{n} \sum_{j}^{n} w_{ij} y_{ij} \,, \tag{1}$$

where Γ is a particular measure of spatial association, w_{ij} are the elements of a matrix **W** (weight matrix) representing the relationship of each location to the others, and y_{ij} are the elements of a matrix **Y** representing the interaction between all combinations of locations (Getis, 2000). The weight matrix **W** is derived from knowledge about the spatial relationships in the area of study. If one assumes that closer geographical areas are similar, then near-neighbor locations should have a higher weight than distant ones. The interaction represented in the **Y** matrix can be expressed as an addition, subtraction, multiplication, division, covariance or combinations of the first four types. Each one will generate a different kind of statistic.

These statistics can be global or local. The first considers many locations simultaneously. An analogy can be made between these measures and global effects of a regression model, the sense that both mask important localized details. The higher the number of locations, the less information we retain about local nuances.

Local statistics measure the association between each location and its neighbors, based on defined distances. They were named by Anselin (1995) as LISA (Local Indicators of Spatial Association) statistics. Its matrix formulation is similar to (1), but specifically for each location *i*

$$\Gamma_i = \sum_{j}^{n} w_{ij} y_{ij} \,. \tag{2}$$

Both matrices **W** and **Y** have the same interpretation as in (1), but in (2) all the associations are between location *i* and the other locations *j*.

Getis and Ord (1996) highlight six LISA statistics to measure local association: Moran's I_i - based on covariance, Geary's c_i, K_{1i} and K_{2i} - based on differences, $G_i(d)$ and $G_i^*(d)$ - based on

additive interactions. Additive measures are particularly useful for the identification of hot spots of diseases.

5.1. Identifying clusters

As Marshall (1991: 423) points out, testing for clustering has the objective of answering two questions: "*First, is there a general tendency for clustering to occur and, if so, where? Second, do clusters occur in specific areas, e.g. near suspected environmental hazards?*" Identifying, testing the significance, and searching for the possible explanations for the occurrence of clusters are important steps in understanding the aetiology of a disease. The first two steps can be achieved by the use of spatial statistical analysis.

Four types of tests can be applied in order to identify clusters: a focused test, a global test, a local test and a test for evaluation of cluster alarms (Kulldorff, 1998). The first type is applied to a smaller areas located near a potential high risk factor (putative source), like a nuclear installation or a dam (according to Besag and Newell (1991) the test is useless when a clear and precise scientific explanation is available for the observed high incidence near the putative site, in which case no testing is actually necessary). The global test is applied to a large area, with no specific concern for where the clusters are located. Local tests focus on the location of clusters and their significance, and are the most appropriate tests to achieve the goals of this paper. Finally, the evaluation of cluster alarms focus on particular areas where the high/low rates do not seem to be generated by a putative source, but by the data itself.

The $G_i(d)$ and $G_i^*(d)$ statistics allow the identification of clusters of high or low values surrounding a particular location i within a radius of distance d from i. The first statistic doesn't consider the value of location i itself, and is better used for spread or diffusion studies. The second statistic takes the value of location i into account and is the most appropriate for the identification of clusters (Aldstadt, Chen and Getis, 1998).

Consider an area divided into n locations, each identified with a point i and associated to a value x_i (a realization of the random variable X). When focusing on location x_i, all the others are named x_j. The $G_i^*(d)$ statistic is defined as (Getis and Ord, 1996)

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)\ x_j - \bar{x}\sum_{j=1}^{n} w_{ij}(d)}{s\left\{\left[n\sum_{j=1}^{n} w_{ij}^{\ 2}(d) - \left(\sum_{j=1}^{n} w_{ij}(d)\right)^2\right] \Big/ (n-1)\right\}^{1/2}}, \tag{3}$$

where w_{ij} are the elements of the weight matrix and

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad \text{and} \qquad s = \sqrt{\frac{\sum_{j=1}^{n} x_j^{\ 2}}{n} - (\bar{x})^2}\ . \tag{4}$$

The null hypothesis assumes that the results for a particular location i, given a distance d, are the same as those obtained considering all the locations. In other words, there is no association between location i and its neighbors, within distance d. Under the null hypothesis, the $G_i^*(d)$ statistic is asymptotically normally distributed, N(0,1), as $n \to \infty$ (Getis and Ord, 1992). When d is very small

or very large (covering the total area) normality is lost, which emphasizes the importance of the selection of an appropriate distance. As a rule of thumb, the maximum distance *d* considered should never exceed ½ of the shorter side of the study area.

Given a significance level, the results can be compared to standard values of the normal distribution. Significant negative $G_i^*(d)$ reveals spatial clustering of low values, while significant positive $G_i^*(d)$ are indicative of spatial clustering of high values. Special care, however, has to be taken. The $G_i^*(d)$ values for different locations tend to be correlated, since neighboring points are shared by multiple locations, given a distance *d* (Anselin, 1995). Given this multiple testing scenario, Ord and Getis (1995) propose a Bonferroni correction to control for the overall probability of Type I error[7]. Using this correction, the level of significance to be chosen is given by $1-(1-\alpha)^{1/n}$, where α is probability of Type I error, and *n* is the number of observations, which is assumed to be equal to the number of multiple comparisons. Standard measures for different percentiles of the normal distribution, and different *n*, are presented in Ord and Getis (1995: 297).

As Anselin (1995) points out, the Bonferroni correction can become too conservative as *n* increases. To overcome this problem, Getis is developing a new correction that takes into account the proportion of overlap between nearby points (Getis, personal communication).

No matter what metric is selected as appropriate, it is important to keep in mind that boundary problems are very likely to occur. They are a common feature of geographical studies. In other words, given a particular location *i* and a buffer with a radius *d*, the buffer area can include only other plots (the ideal case), non-inhabited areas (like rivers and protected forests, in the case of Machadinho), or areas for which there is no available data (in the case of edge locations). We will discuss this issue in Section 6.1.

5.2. Measuring spatial non-stationarity in a multivariate context

The multivariate analytical tools applied thus far to the Machadinho data assume that the results are stationary over space. The major findings from regression analyses, highlighted previously in Section 4, are supposed to apply equally to all areas of Tracts 1 and 2. However, as Sawyer and Monte-Mór (1992) showed, malaria risk is significantly different between sectors. This suggests that the distribution of the disease in the area is a spatial non-stationary process. In this case, the coefficients of a regression analysis are global measures that can hide spatial variations in the associations between variables. The most appropriate methods, therefore, are those with a local focus that highlight differences across space and are easy to map.

Since the early 70's, some methods have been proposed to measure spatial non-stationarity in multivariate data. Fotheringham, Charlton and Brunsdon (1997) mention four of these methods, and propose a new one that solves the problems related to the first four. In summary they are:

1. Expansion Method (Casetti, 1972; Jones and Casetti, 1992) – Measures spatial non-stationarity (named spatial drift by the authors) through the incorporation of the spatial coordinates of each location into the model. The results are trends over space that depend on the expansion equations used, and can hide some local variations (Fotheringham, Charlton and Brunsdon, 1997).

[7] Two different types of error can be made in hypothesis testing. Type I error is when a true null hypothesis is incorrectly rejected. Type II error is when a false null hypothesis fails to be rejected (Hoaglin, Mosteller and Tukey, 1991).

2. Spatial Adaptative Filtering – SAF (Foster and Gorr, 1986; Gorr and Olligschlaeger, 1994) – The main problem is that parameter estimates cannot be tested statistically.
3. Random Coefficients Model (Aitkin, 1996) – The parameter estimates are assumed to be random variables, modeled as mixture distributions. One of the problems is that no assumption is made about the spatial dependency on the data. Moreover, the parameter estimates are sensitive to the distributions chosen.
4. Multilevel Modeling (Goldstein, 1987) – Takes explicit account of different levels in the fitting process (e.g., individual, family and community, or neighborhood, city, county and state). Two problems can be identified in the use of this model for measuring spatial non-stationarity. The first is that the definition of the spatial aggregations that represent the levels is done a priori by the researcher, and not through an analysis of the spatial associations observed in the data. The second problem, a consequence of the first, is that the spatial process being modeled is discontinuous. It brusquely changes when the border between different levels are crossed. Most spatial process, however, are continuous (Fotheringham, Brunsdon and Charlton, 2000).
5. Geographically Weighted Regression – GWR (Brunsdon, Fotheringham and Charlton, 1996) – This method overcomes the problems of the previous ones, and is an extension of classical regression analysis, generating local, instead of global, coefficients. Therefore, for each covariate, the model estimates as many coefficients as locations in the data.

In this paper we apply the GWR method. Details of the specification are presented in the next section.

5.2.1. Geographically Weighted Regression

Consider a data set with $i = 1, \ldots n$ observations and $j = 1, \ldots k$ covariates, available for a set of spatial coordinates (u_i, v_i). The GWR model can be written as

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i, \qquad \varepsilon \sim N(0, \sigma^2), \tag{5}$$

where y is the dependent variable, x_{ij} are the covariates, β's are estimated coefficients for each location (u, v), and ε's are random errors, assumed to be independently and normally distributed (Brunsdon, Fotheringham and Charlton, 2000). From (5), it is clear that the global regression model defined as $y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$ is simply a particular case of GWR, in which the location is assumed to be constant and equals the whole studied area.

To estimate the GWR model, a calibration needs to be performed, since there are more unknowns than observed variables. The calibration process is based on Tobler's First Law of Geography (Tobler, 1979). Thus, data near a certain point (u_i, v_i) have more influence on the estimated parameters for (u_i, v_i) than a point far away. In other words, GWR can be compared to a weighted least squares model, where each point is weighted according to a function that describes the proximity between locations. In matrix notation, the GWR estimator is given by

$$\beta(u_i, v_i) = \left(\mathrm{X}^T \mathrm{W}(u_i, v_i) \mathrm{X}\right)^{-1} \mathrm{X}^T \mathrm{W}(u_i, v_i) \mathrm{y} \tag{6}$$

where $\mathbf{W}(u_i, v_i)$ is a diagonal matrix of weights w_{ij} for each observation (Brunsdon, Fotheringham and Brunsdon, 2000). Except for the weight matrix and the fact that an estimate is obtained for each coordinate location, the GWR estimator resembles the OLS estimator: $\beta = \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{y}$.

The idea behind the weighting process is close to that of kernel regression, which weights data according to the absolute differences among observations (Hastie and Tibshirani, 1990). In GWR, kernels are also obtained, but with weights based on the closeness of data in space. Depending on the weighting function used, the kernels can have fixed or variable bandwidths.

Different weighting functions can be used. However, the ideal one has to avoid the problem of discontinuity mentioned in the multilevel models. Functions that assume that data at a certain distance from location (u_i, v_i) should be weighted with a constant value (one, for example), and further from that distance should be zero, must be avoided. The ideal scenario is to define the weights w_{ij} as a continuous function of a certain distance (Fotheringham, Charlton and Brunsdon, 1997).

Kernels with fixed bandwidths are obtained by the use of a Gaussian weighting function, with weights given by

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{h^2}\right), \tag{7}$$

where h is the kernel bandwidth, d_{ij} is some distance between locations i and j, and the weights decrease as d_{ij} increases (Fotheringham, Charlton and Brunsdon, 1998). Observe that we have now relabeled locations as i and j, rather than (u_i, v_i).

On the other hand, kernels with variable bandwidths are obtained when a bi-square function is applied. This function is based on the notion of nearest neighbors, and is defined as

$$w_{ij} = \begin{cases} \left\{1-\left(d_{ij}/h_i\right)^2\right\}^2 & \text{if } d_{ij} < h, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In this case, h_i is the bandwidth, which varies according to the number of neighbors that each location must have. In other words, observations at a distance greater than the bandwidth receive weight zero. They are not taken into account in the estimation process. These types of kernels are called adaptive spatial kernels. They are characterized by small bandwidths in areas where data is densely distributed, and by large bandwidths where data is sparsely distributed (Fotheringham, Brunsdon and Charlton, 2000).

Given a weighting function, the problem is to get an appropriate bandwidth. It shouldn't be so small, that it produces results extremely dependent on close observations, and not so large, that can end up generating estimates that are similar to the OLS model.

To solve this problem, Fotheringham, Charlton and Brunsdon (1997) propose the use of the cross-validation (CV) technique, a method of assessing estimation and prediction through the exclusion of one observation at a time, predicting its value with the remaining data, and comparing the results with real values (Mosteller and Tukey, 1968; Stone, 1974; Geisser, 1975; Cressie, 1993). The authors use the formulation introduced by Cleveland (1979), and propose a CV score z given by

$$z = \sum_{i=1}^{n}\left[y_i - y_{\neq i}(h)\right]^2, \tag{9}$$

where $y_{\neq i}$ is the fitted value of y_i when observation i is omitted. The ideal h will be the one that minimizes z, and can be obtained by an optimization algorithm. Brunsdon, Fotheringham and Charlton (1996) chose the Golden Section search[8] for optimization purposes (Greig, 1980).

The importance of choosing an appropriate h can be better understood by a brief discussion of the bias-variance trade off. Considering that the GWR estimates are a function of the location of observations in space, a totally unbiased estimate is impossible, although results with a small bias can be achieved. Briefly, if the bandwidth is too small, a reduced number of observations are considered, which reduces the bias, but the variance increases due to the small sample size used to obtain the estimate. On the other hand, if the bandwidth is too large, many more points are considered, reducing the variance, but increasing the bias (Fotheringham, Charlton and Brunsdon, 1998).

Once h is chosen, local estimates of the regression coefficients, standard errors, t-values, and R^2 can be obtained and mapped. The resulting analysis reveals local variations present in the effects of each covariate considered in the model.

6. SOME FINDINGS FROM A SPATIAL ANALYSIS

The first attempts to map the malaria risk in Machadinho started 5 years ago. Marchesini, Spencer and Lima (1996) used as spatial unit 133 localities defined by the National Health Foundation, each containing a variable number of plots. The Annual Parasite Index (API) was mapped based on 1994 data, and the major visual pattern of the map showed that the largest API's were located near the Machadinho River and the protected forest reserves. Moreover, the spatial distribution of the API did not seem to be random, and similar values were closer in space than distinct ones.

Reis and Assunção (1996) mapped the malaria rates for 1986 in Machadinho using a Bayes approach, with the plot as the spatial unit of analysis. A global and a local[9] empirical Bayes estimator were applied to the 1986 and 1987 data. The results obtained for both years were very similar and no defined spatial pattern of risk was observed. Moreover, a comparison between the raw rates and those estimated by the Bayesian method also revealed similar results, with significant differences occurring only for the cases of rates equal to zero or one (the extremes of the distribution). The question, however, is how relevant for the analysis is the smoothing of extreme values, since these can actually be important factors in the determination of risk patterns. Probably the best approach in this case is to start the investigation with an exploratory data analysis routine, using transformation functions, and identifying possible outliers.

In 1999, another application of the Bayes estimator was done (Assunção, Reis et al., 1999). This time the estimation considered also 1995 data. The major results are a decline in malaria rates during the period of analysis, and a smoothed pattern of rates that reveals higher rates near the forest reserves and the Machadinho River.

---

[8] Given three points of a function, (a,b,c), which contain a minimum inside, the golden section search chooses some point x, defining the new triplets (a,x,b) and (b,x,c) in a way that at each new interaction the interval is *0.61803* times the size of the preceding.

[9] The local estimator follows the same definition used for the global estimator, but the results for each plot are obtained considering a neighborhood area within a certain distance from the plot. The definition of the distance depends on the knowledge of the risk patterns of malaria in the region. In the study of Reis and Assunção (1996) a circle with a 5 km radius was defined as the neighborhood area for each plot.

Castro and Singer (2000) mapped the malaria rates for 1985, 1986 and 1987, highlighting their association with soil quality and land use. The authors propose that areas close to rivers and to protected forest reserves would be exposed to a higher risk of malaria transmission during the early years of the colonization project, although as the project ages goes by this effect would be explained by a group of variables that, ultimately, characterize the profile of high-risk malaria transmission in the area.

Finally, Singer and Castro (2001) applied spatial statistical analysis tools to Machadinho data in 1985, 1986, 1987 and 1995. Variogram analyses were used to model the spatial association of malaria rates, and were the starting point for kriging estimation. Estimates of malaria rates were obtained for all the plots, including those not yet occupied, based only on the spatial association of the rates. The results show the spatial evolution of risk through time, and emphasize the strong contribution of the local component. They also provide a clear picture of which areas might be productively occupied and those which should not be occupied at all. It also provides an indication of what would be the observed pattern of malaria transmission in demographically and ecologically similar areas in the Amazon.

6.1. Temporal and spatial evolution of risk in Machadinho

Before we apply the spatial statistical modeling to the Machadinho data, it is important to have a sense of how the settlement project developed through time, including the process of plot turnover and the associated variation in malaria rates. Figures 2, 3 and 4 show the variation in malaria rates and spatial distribution of plots occupied for the time periods 1985/86, 1986/87, and 1987/95, respectively.

In the first two periods (Figures 2 and 3), some of the plots that had no malaria in the previous year are those that were owned by a settler that spend most of his time outside Machadinho. So the absence of malaria cannot be translated as no risk, but as no exposure to the risk of transmission. This situation was common in the early stages of the settlement. Malaria exposure was influenced by the times in year that the settler had access to the land. If that occurs at a late stage in the agriculture schedule, settlers would not have enough time to prepare the land for cultivation before the rainy season. Although some clearing would be carried out, he would most likely live in the urban area until the season is over. As the settlement process consolidates, this situation becomes less frequent. In numbers, the cases of no exposure represented 15% of the malaria rates equal to zero in 1985. In 1986, this number jumped to 76%, going down to 14% in 1987, and only 1% in 1995.

A possible explanation for the occurrence of zero malaria rates when effective exposure was observed can be based on the effect of deforestation in different stages of the settlement process (Sawyer, 1992a). We argue that these cases are most likely to be associated with a low level of deforestation during the first three years of the project. In 1995, however, the explanation is more complex, and the effects of different socio-economic variables, as well as their interactions, play a role.

Still considering the first two years, it is clear that the occupation process was initially much more concentrated in plots of Tract 2, probably because of its easier access and proximity to the urban area.

With respect to the abandoned plots, 20% of the plots abandoned in 1985 had high malaria rates during that year, and in 1986 this number jumped to 44%. In 1987, the analysis is not so straightforward, since during the lag period of eight years between 1987 and 1995 the detailed surveys of occupation process were not carried out.

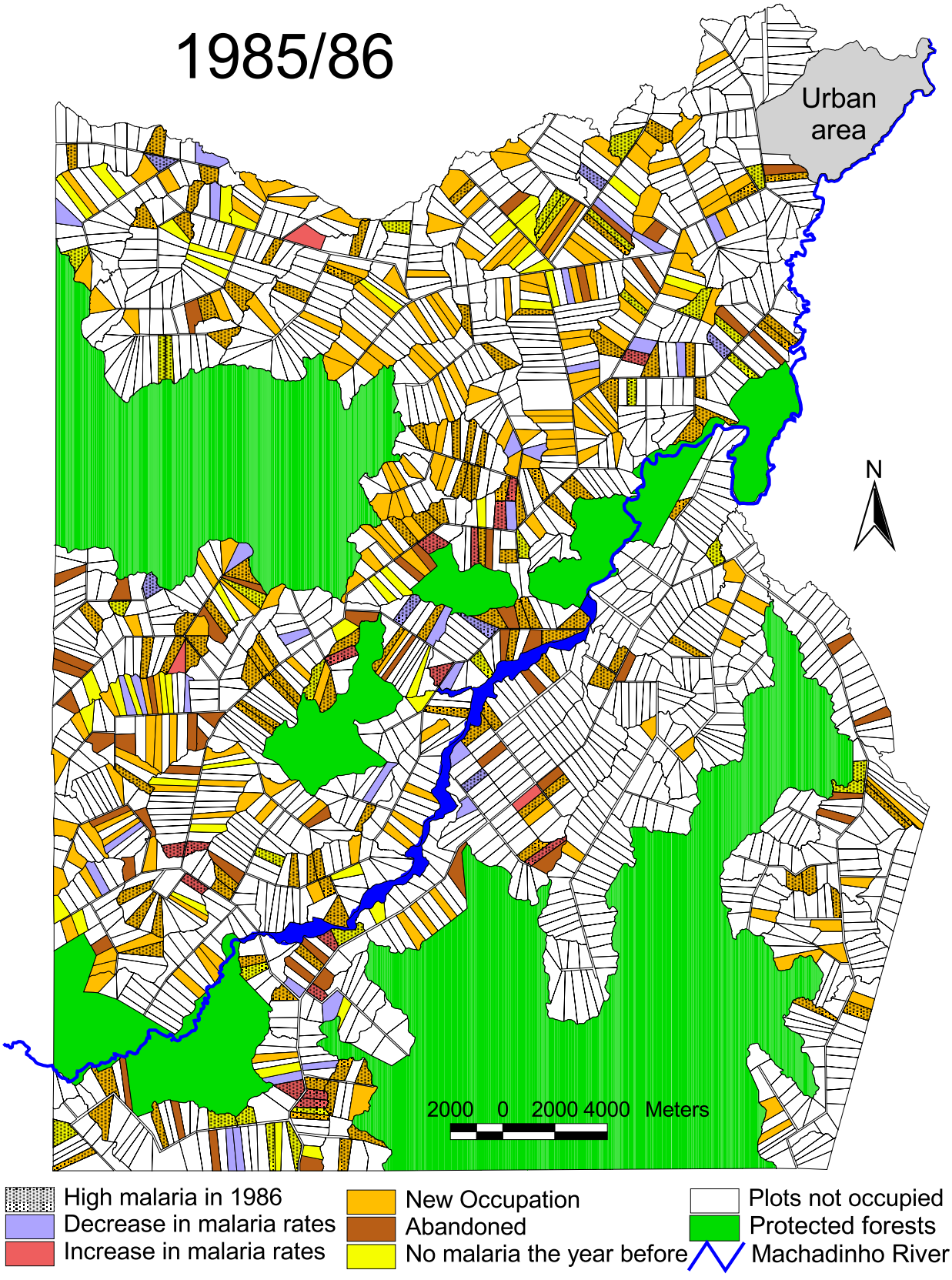


Figure 2 – Variation in malaria rates and in occupation of plots – Machadinho – 1985/86

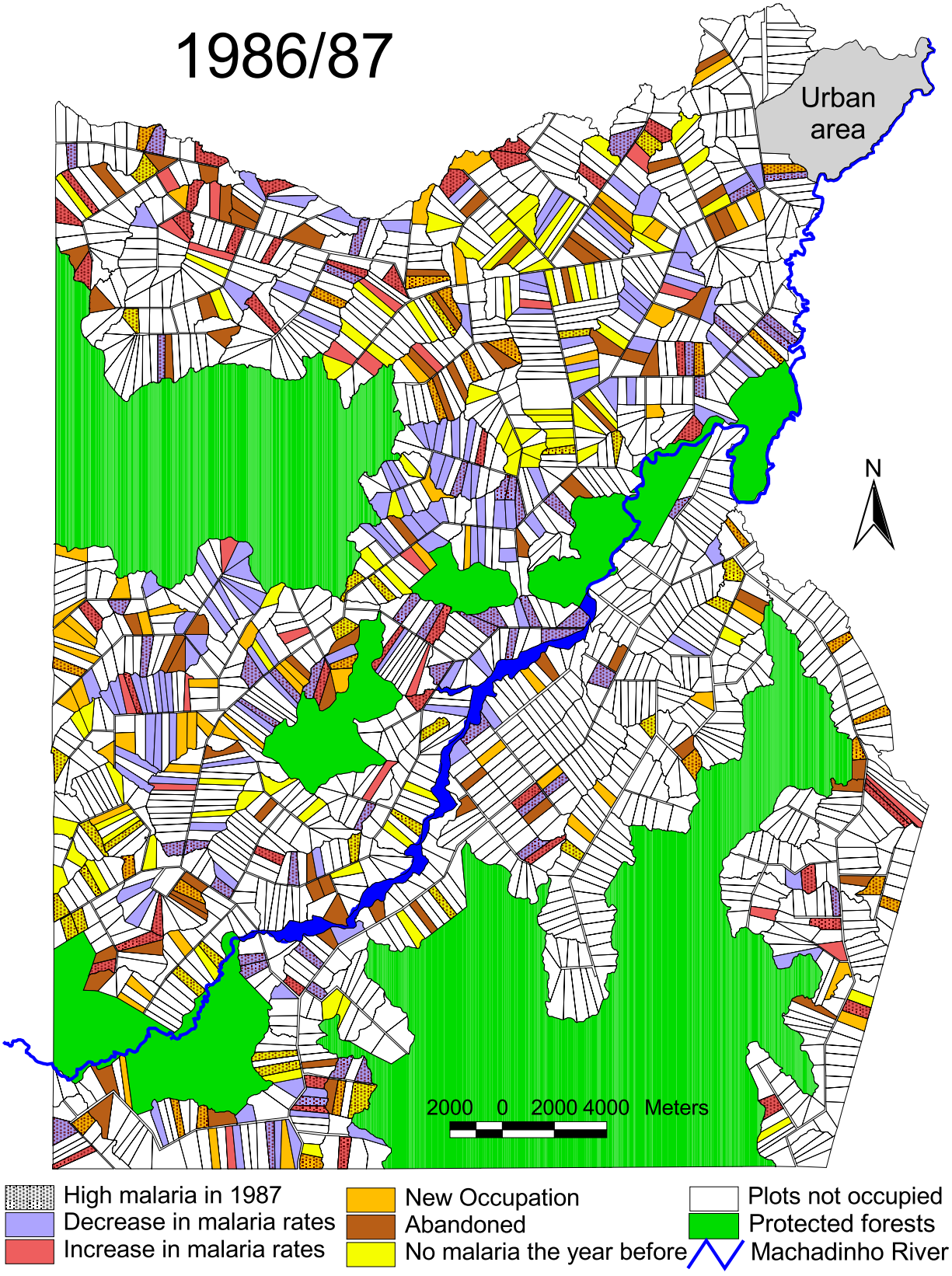


Figure 3 – Variation in malaria rates and in occupation of plots – Machadinho – 1986/87

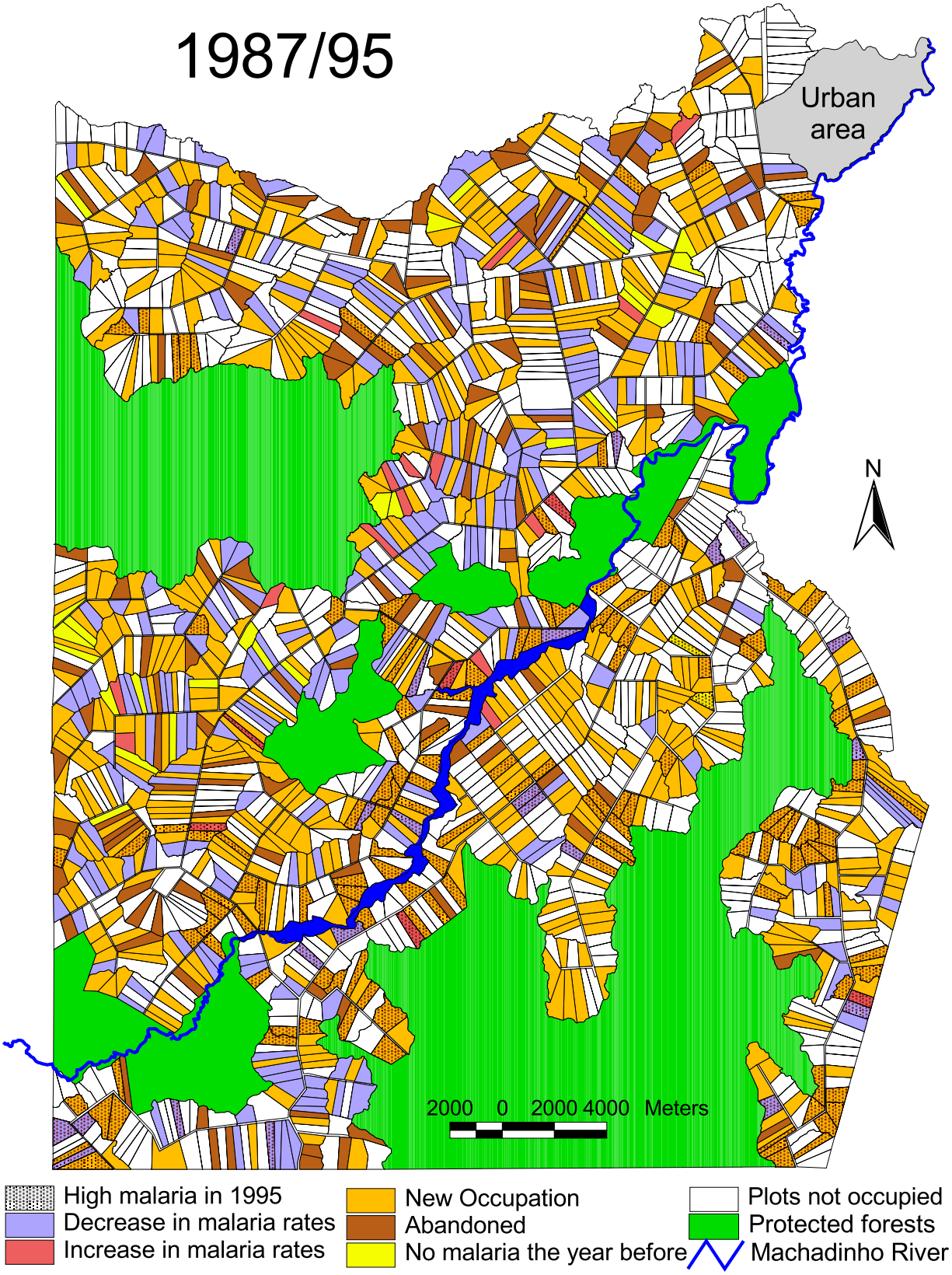


Figure 4 – Variation in malaria rates and in occupation of plots – Machadinho – 1987/95

Between 1985 and 1986 (Figure 2), a few plots were able to reduce their malaria rates, although most of the action during this period was related to the occupation process itself. It is remarkable to note that out of 294 plots newly occupied in 1986, 40% registered high malaria rates.

During the next period, however (Figure 3), major decreases in malaria rates were observed, especially in Tract 2, although in some areas near the protected forest reserves and the Machadinho River, reductions did not occur.

For the eight-year period between 1987 and 1995 (Figure 4), a general decrease in malaria rates is still observed for plots located in Tract 2. Most of the plots with high malaria rates in 1995 are those in Tract 2 that are near the protected forest and the Machadinho River, as well as those in Tract 1.

In summary, the results for the three periods suggest that some areas are candidates for pockets of high malaria in the area, and that although some settlers are able to reduce the incidence of malaria, this does not imply that the plot will register low rates from then on. The spatial location of the plot seems to play an important role in addition to variables that capture the social, ecological, economic and environmental characteristics of the plot itself. Firm conclusions cannot be drawn from the variation maps alone, since there is no statistical significance associated with the results. They are suggestive visual patterns that require further testing. Such steps are needed if we are to avoid the major mistake that Jacquez (1998) calls as the *gee-whiz* effect.

6.2. Malaria foci in Machadinho

The first question to be addressed is whether or not malaria is clustered in Machadinho. If so, we need to investigate where clusters are located, as well as how they evolved through time and space. To approach this issue we apply the local indicator of spatial association described in Section 5.1., the $G_i^*(d)$ statistic, computed using the software Point Pattern Analysis – PPA (Aldstadt, Chen and Getis, 1998), developed at San Diego State University.

Before presenting the results, two methodological issues have to be discussed. The first one is the appropriate distance d to use with Machadinho data. The second is related to the level of significance to apply (whether to use the Bonferroni correction or not). As emphasized in Section 5.1, the distance d must have an interpretation relative to the phenomenon under study. For Machadinho, the appropriate distance should be related to the flying behavior of the mosquitoes, and the average size of the plots.

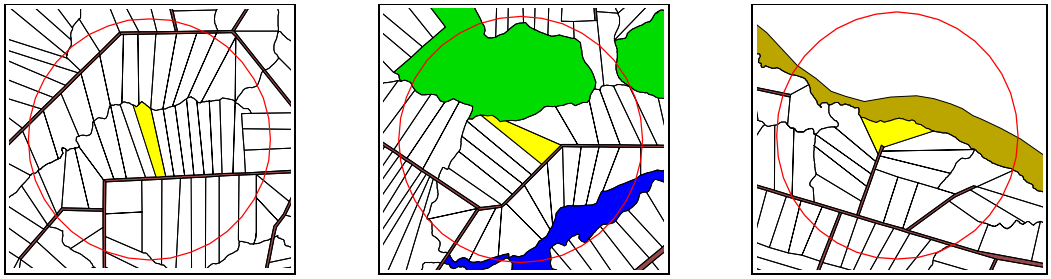
The flying behavior of mosquitoes varies among species, and the same species can have different behavior depending on its physiologic state and on the habitat where they live. Studies show that the lifetime distance traveled can vary from hundreds of meters to several kilometers. Gillies (1988) argues that instead of considering the flight range one should analyze the ambit, which is the "*the distance covered in the search for sources of food and for breeding sites*". In this sense, the main determinant of the distance is the habitat, and not the species. However, no precise information is available concerning the ambit for *Anopheles darlingi* in Machadinho per se. However, guided by entomological studies in other parts of the Amazon, a reasonable distance would be between 2,000 and 3,000 meters.

Considering the size of the plots, their irregular shape and the existence of non-inhabited areas in some places, makes the choice of the appropriate distance a delicate issue. The analysis of buffers sized between 500m to 8,000m, drawn around plots located in different places, suggests that distances lower than 2,000m would be too small, and would result in a very small number of neighbors. This

could seriously compromise the normality of the statistic, $G_i^*(d)$ (Getis and Ord, 1996). On the other hand, distances larger than 6,500 have no practical meaning. As they keep increasing, they cover such a large area that they can no longer be considered to be a local measure.

Considering all of these factors, we utilized distances of 2,500m and 6,250m. Some additional points must also be considered. Three different situations can occur when a buffer is drawn around a certain plot, as shown in Figure 5.

In Figure 5(a), the buffer covers an area that includes only other plots, and this reflects the ideal situation. In Figure 5(b), the buffer covers an area that includes places like rivers and protected forests, which are not inhabited and, consequently, do not contribute any significant information for the determination of clusters. Finally, in Figure 5(c), the buffer is derived from a plot on the edge of the area. Almost half of it includes both Igarapé Preto (a small and narrow river, with almost stagnant waters, which only allows the sailing of small boats or canoes) and other populated places for which information is not available. This is called the boundary problem, which is peculiar to most geographical studies. The analysis is limited to the area for which there is available information, or to a specific targeted area. One of the possible problems with the cases represented by Figures 5(b) and (c) is that the number of neighbors can end up being very small, and normality of summary statistics is approached more slowly (Ord and Getis, 1995).



(a) (b) (c)

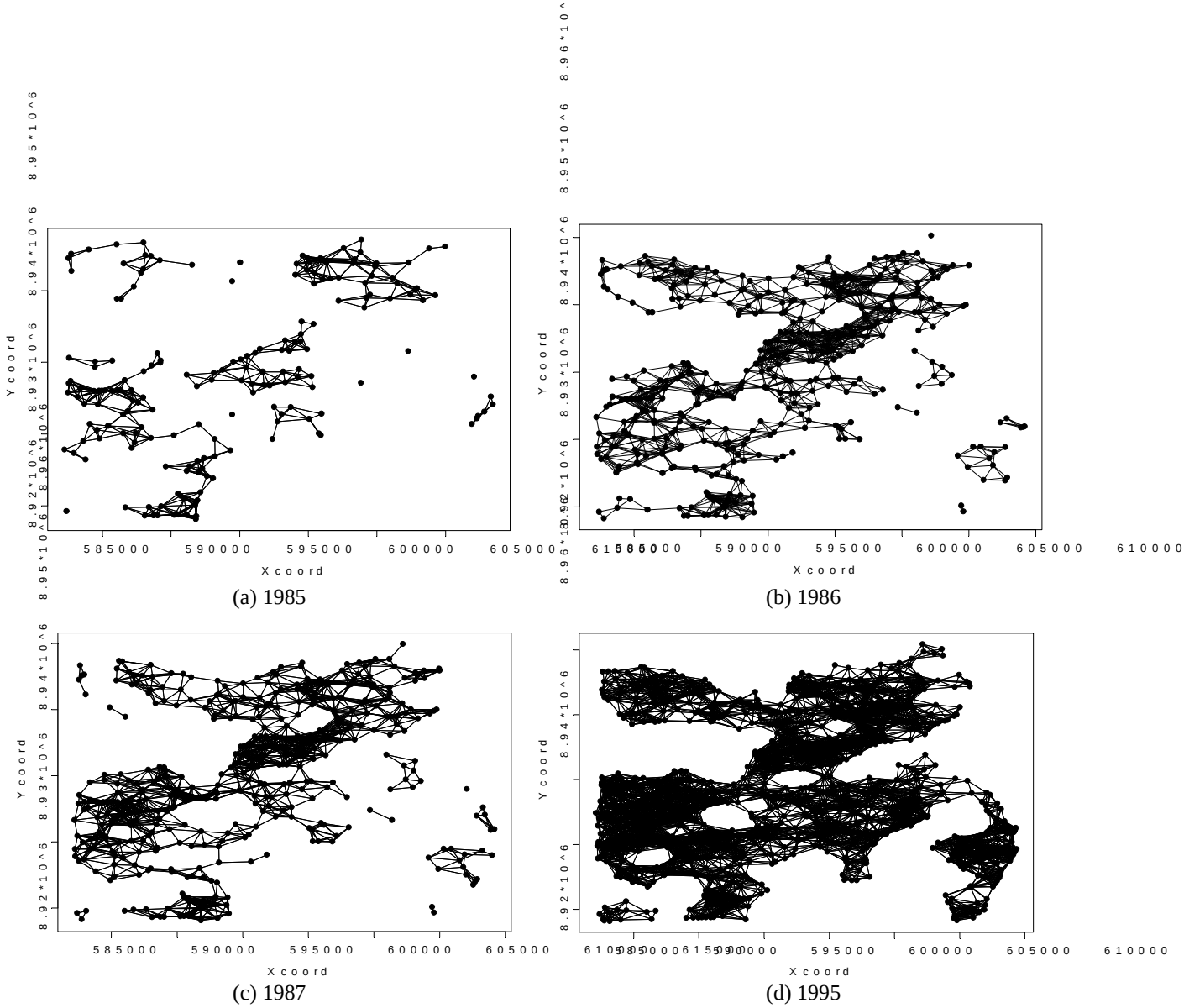
Figure 5 - Buffers of 2,500 meters radius around selected plots

Using *d*=2,500m, the derived lattice for Machadinho data, in each of the four years, is shown in Figure 6. The Xcoord represents the coordinates in the West-East direction, while the Ycoord those in the North-South direction. In other words, the upper left corner of the graph shows the plots located in the most North and West section of the Machadinho project.

Similar graphs for *d=6,250* are not worth for presenting, since the number of neighbors increases so much that the lattice is just a dark surface composed of very dense connections between the plots. Table 2 highlights this fact, showing the maximum, minimum and average number of neighbors at each distance.

Table 2 – Summary of neighborhood configuration at different distances

Year	d=2,500			d=6,250		
	Minimum	Average	Maximum	Minimum	Average	Maximum
1985	0	6.07	15	3	25.16	42
1986	0	5.35	17	5	53.07	97
1987	0	10.67	25	6	53.23	91
1995	3	20.71	39	14	100.40	174



(a) 1985 (b) 1986

(c) 1987 (d) 1995

Figure 6 – Lattice for Machadinho data, considering a neighborhood distance of 2,500m

Although the number of neighbors is small for *d=2,500*, considering that the distribution of malaria rates is not particularly skewed, the results are not likely to be biased, and we would prefer to work with that distance. A simple test can be done by checking if the 2.5% highest and the 2.5% lowest $G_i^*(d)$ values are themselves included in spatial clusters. If they are, one can conclude that they are statistically significant at the 5% level (Getis, personal communication). We applied this test to our results, and the 2,500m-distance seems to be very appropriate.

Moving to the issue of level of significance, we use the tabulated values presented by Ord and Getis (1995: 297) to obtain the critical values to be used in each year, given the adoption of a Bonferroni correction. The results for a 95% level of confidence are presented in Table 3. They are very conservative results, since the correction proposed assumes that the number of multiple comparisons is equal to the number of observations.

Table 3 – Critical values for the $G_i^*(d)$ statistic, assuming a Bonferroni correction and 95% of confidence

	1985	**1986**	**1987**	**1995**
n	186	425	431	927
Critical value	3.38017	3.63381	3.64017	3.87036

In the case of Machadinho, as pointed out in Table 2, the small number of neighbors for a 2,500m distance does not justify the use of the Bonferroni correction. The simple test proposed above (Getis, personal communication) was used to validate the results.

Before showing the maps with the significant clusters of malaria in Machadinho, we graphed the results of the $G_i^*(d)$ statistic for plots with very high and very low results. The graphs are shown in Figure 7, and allow one to understand both the features of the statistic and the conservative characteristics of the Bonferroni correction.

The graphs show two groups of distinct trajectories, both for high and low values, especially from 1986 on, when the level of occupation is higher. In the first group, taking low values as an example, the $G_i^*(d)$ statistic becomes more negative as the distance increases, which means that the extra neighbors added at every additional 500m are contributing to the significance of the cluster. In the second group, the plots' trajectories reach a minimum $G_i^*(d)$ at which there are significant clusters of low levels of malaria. However, after the distance where this minimum was observed, the new neighbors included do not contribute to the clustering. This is the case for the plots that are close to clusters of high values, or of those that are on the edge, and that naturally have a reduced number of neighbors. In 1995, when low rates are more concentrated in Tract 1, while high rates are observed in Tract 2, the graph of $G_i^*(d)$ for low values is a very dense cloud of negatively increasing lines, except for the plots that are on the border of the areas with high malaria rates.

Comparing the critical values of the normal distribution for a level of significance of 95% and those given by the Bonferroni correction proposed by Ord and Getis (1995), one can see how conservative the correction is for smaller distances, which have a reduced number of neighbors. It is very interesting to note, however, that if the Bonferroni correction is applied to the results of the $G_i^*(d)$ at a 6,250-distance, the significant clusters bear a resemblance with those identified at 2,500m. In summary, based on all of the above discussion, we decided to consider a distance equal to 2,500m, and to use the critical value of the normal distribution, instead of the Bonferroni correction, as the criterion for significance testing.

Having defined the distance and the level of significance to work with, maps of the $G_i^*(d)$ statistic allow the visualization of the spatial location of malaria foci in Machadinho. As a convention, clusters of high malaria rates are colored in red, and circled by a radius based on the plots inside the clusters and the 2,500m distance. Clusters of low malaria rates are colored in light blue, also circled by a buffer. Light yellow colored plots are those that had a $G_i^*(d)$ value not significant for clustering. The maps are shown in Figure 8.

Three clusters are identified in 1985 – two of low malaria rates and a big one of high values. The latter, which we'll call cluster H, is located near the Machadinho River and the protected forest areas. A possible explanation for this fact is related to the natural process of mosquito breeding. First, the forest is a natural reservoir of mosquitoes, and plots close to it, up to a certain distance, are more likely to be in the danger of mosquito biting activity, all other factors assumed to be the same. Second, it is well known that mosquito-breeding sites are very frequent at the margins of rivers, especially at the beginning and at the end of the rainy season, given that the water is clear. Therefore, proximity to the Machadinho River may mean proximity to mosquito-breeding sites and, consequently, to higher malaria risks[10].

[10] Although it seems that these particular locations should be automatically associated with clusters of high malaria rates, even dispensing with a statistical test, it must be emphasized that examples of environmental management and vector control in the past have proven to be successful in drastically reducing the risk that these types of locations pose.

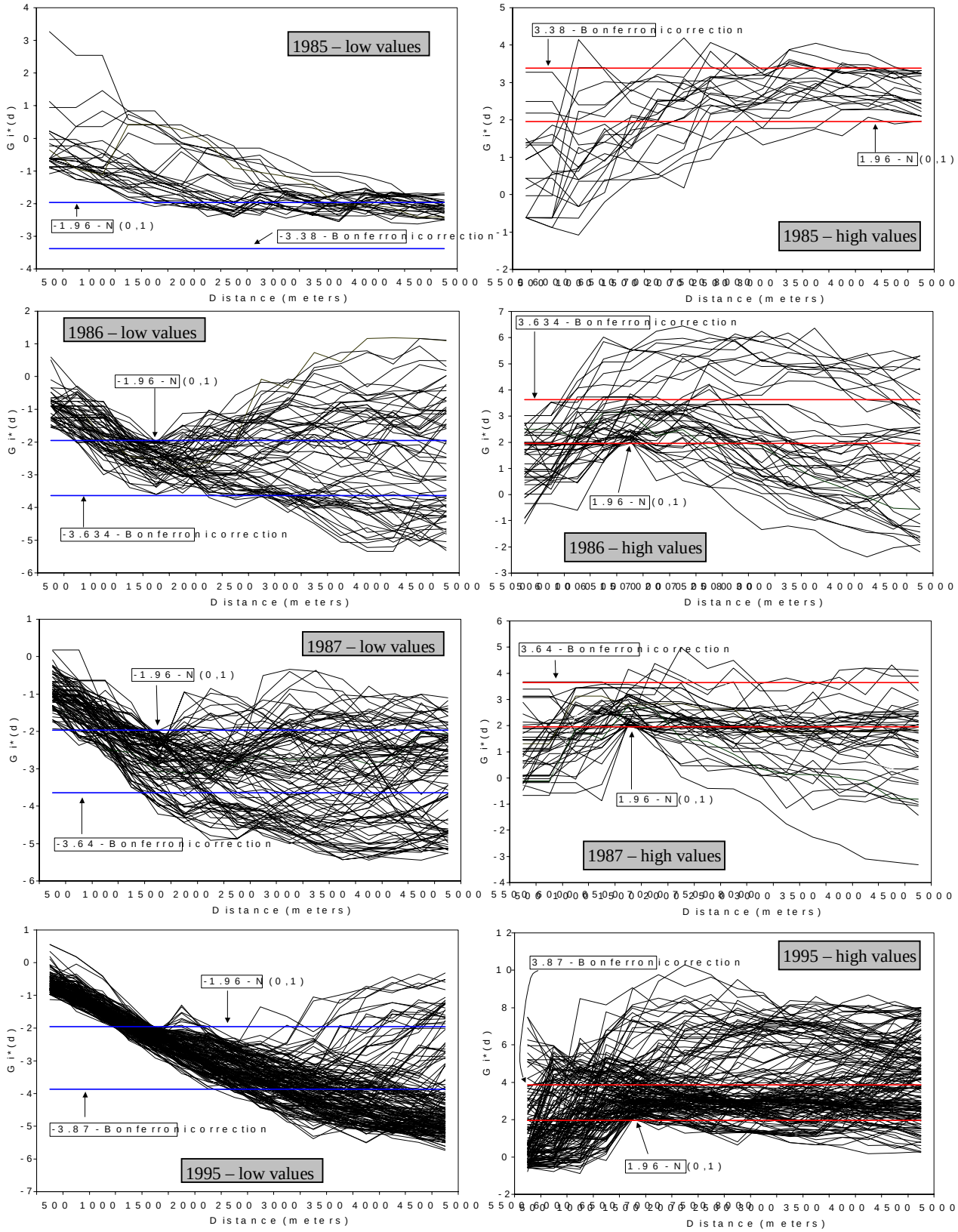


Figure 7 – Gi*(d) statistics for plots with significant clustering of malaria rates

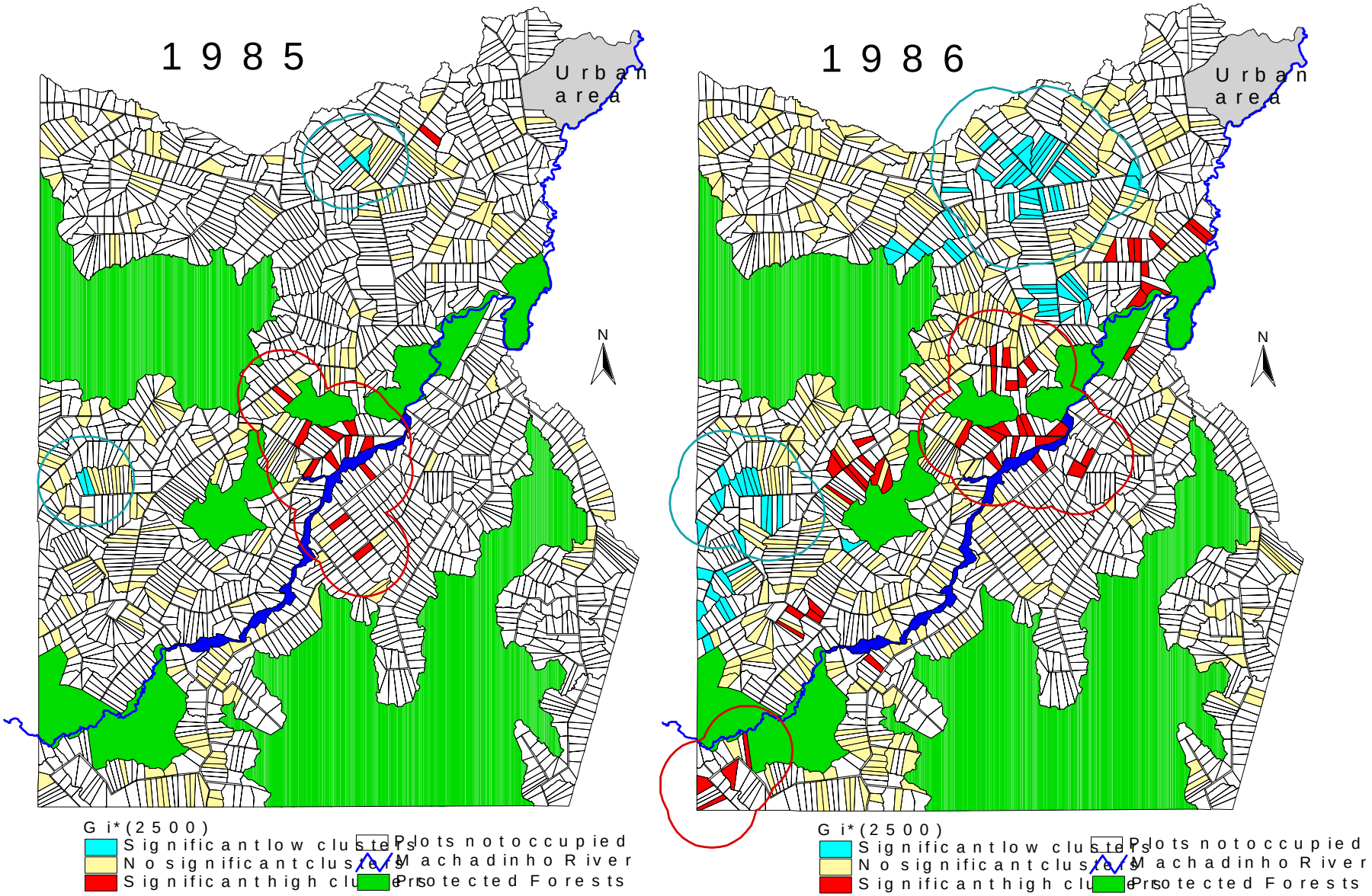


Figure 8 – Results of the $G_i^*(d)$, for d=2,500 – Machadinho – 1985/95 (cont'd)

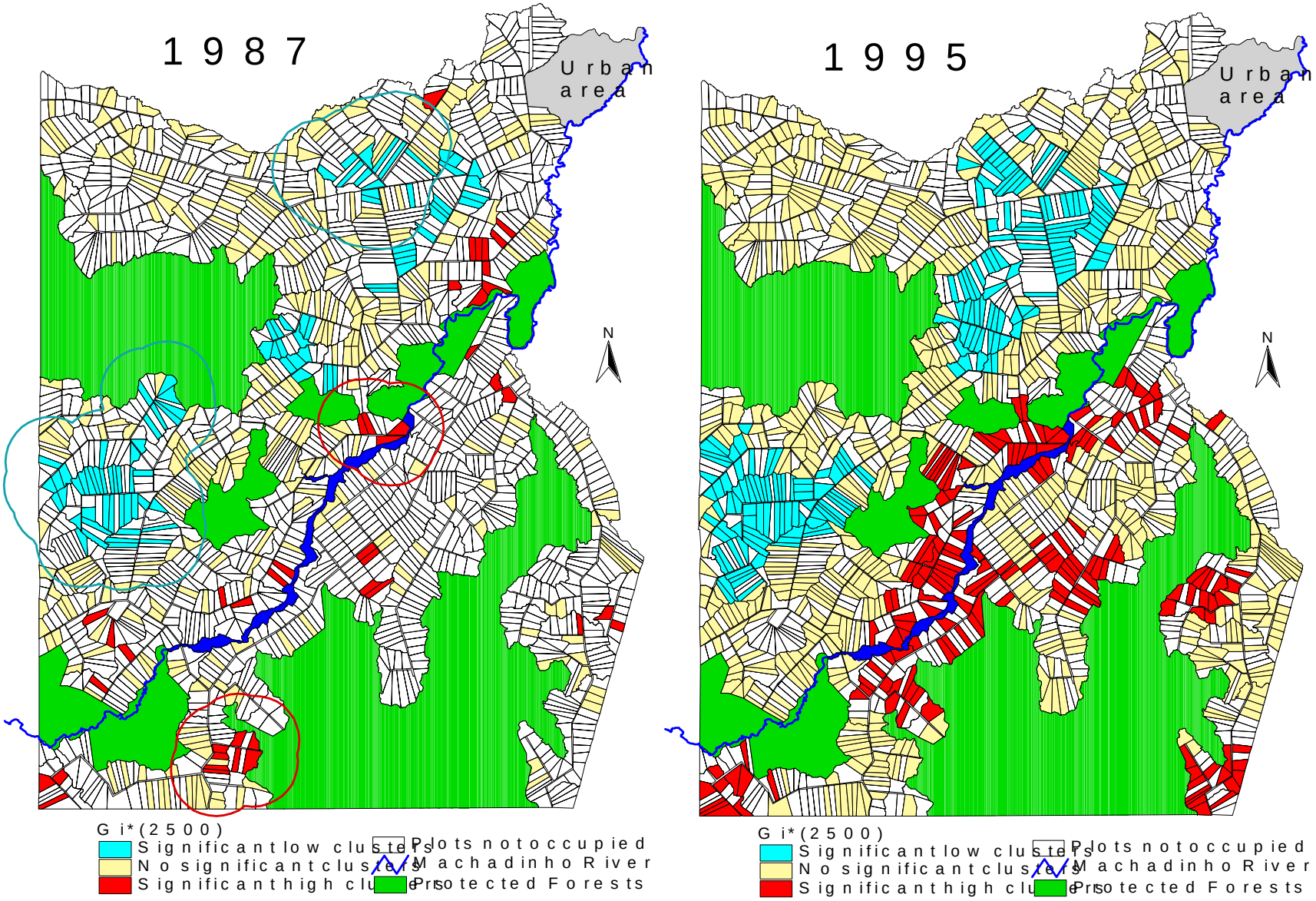


Figure 8 – Results of the $G_i^*(d)$, for d=2,500 – Machadinho – 1985/95 (conclusion)

One of the clusters of low malaria rates, at the west side of the high malaria clusters, is located in a very particular area in Machadinho, which has been successful in agricultural production since the onset of the settlement. It was the first place where coffee grew in the area (Monte-Mór, personal communication). We'll call this area Cluster L.

In 1986, the clusters identified the year before are maintained and even expanded. Additional clusters of high values are observed, which conforms to the fact that in 1986 malaria was a serious problem in Machadinho. In 1987, the clusters of low malaria rates are maintained, and it is clear that Cluster L expands even more. The hot spots seem to clear a little, being more sparsely located in areas of Tract 1.

Finally, in 1995, there is no need to draw a buffer for the clustered values, since the spatial division is perfectly clear. Clusters of low malaria are concentrated in Tract 2, while clusters of high values are concentrated in Tract 1 and in areas of Tract 2 near the Machadinho River and the protected forest reserves.

The spatial evolution of clusters over the years suggests that different variables play distinct roles at certain times. Not all plots sharing borders with the forest are clusters of high values, neither are all those near the Machadinho River. In the same way, not all plots located in the interior part, away from forest and rivers, are significant clusters of low rates. There is much more to be explained, than just factors based on the physical environment or the level of education of the settler, for example.

The next issue to be investigated is whether some of the variables previously identified to have a global effect on malaria incidence are also clustered themselves. Moreover, it is important to find out if their cluster pattern bears a resemblance to the pattern of malaria rate clustering. To address this question, the $G_i^*(d)$ statistic was calculated for other variables in all four years. Similar maps were made and the location of clusters compared to those shown in Figure 8.

In 1985, considering the *distance from the house to the river*, all significant clusters of high distances are located in Tract 2, while those of low distances are basically in Tract 1 and in some areas of Tract 2 near the protected forest and the Machadinho River. The buffers shown in Figure 8 match the cluster of distance to the river in the expected direction: the buffer associated with high malaria rates contains clusters of low distances from the house to the river, while the opposite happens for the buffers representing clusters of low malaria rates. The exact same pattern is observed for the variable *education of the head of the household*. On the other hand, the variable *number of goods* had only one significant cluster of low values, but it was located at the place of Cluster H.

Analyzing the same variables for 1986, only *education of the head of the household* maintains the same pattern. The other two variables are not as clustered as before, and their relation with malaria clusters is no so clear anymore. This latter pattern is also observed in 1987, for the three variables.

In 1995, although the clustering pattern observed for *education of the head of the household* and for *number of goods* follows the major characteristic identified for clusters of malaria rates, they are not so large as in the case of malaria.

In summary, the results show that malaria is, indeed, spatially clustered in Machadinho. In the early years of the colonization project, clusters of high rates were located in areas that suggest a strong influence of environmental factors. On the other hand, ten years after the onset of the settlement, malaria foci appear to be strongly driven by a combination of many different factors, which leave high rates basically only at Tract 1. The comparison of clusters of malaria with clusters of other variables, evoke the possibility that their effects vary over time and space. This suggests that a local multivariate analysis could shed some light on the determinants of malaria transmission in Machadinho.

6.3. Spatial multivariate analysis in Machadinho

A local multivariate analysis was carried out utilizing geographically weighted regression. The model was fitted with Gaussian weighting functions – Equation (7), Section 5.2.1. - (and, consequently, kernels with fixed bandwidths are applied to each point). The cross-validation process to obtain the bandwidth always choose as initial estimates the length of the smallest side of the area for which there is available data (this number varies from year to year due to the pattern of plot occupation). Figure 9 shows the final bandwidths for each year, obtained after the optimization process converged.

Models with different sets of variables were tried for each year, having as response variable the malaria rates. The most appropriate models, for each year, given the overall significance of the covariates, include the following variables:

- 1985 – Minimum distance from the house to the forest[11], number of goods, and level of education of the head of the household.
- 1986 – Minimum distance from the house to the river, number of goods, level of education of the head of the household, and level of education of the head of the household's wife.
- 1987 – Cleared area, minimum distance from the house to the forest, and level of education of the head of the household's wife.
- 1995 – Cleared area, area cultivated with rice, minimum distance from the house to the forest, minimum distance from the house to the river, number of goods, income generated from the activities in the plot, level of education of the head of the household, and level of education of the head of the household's wife.

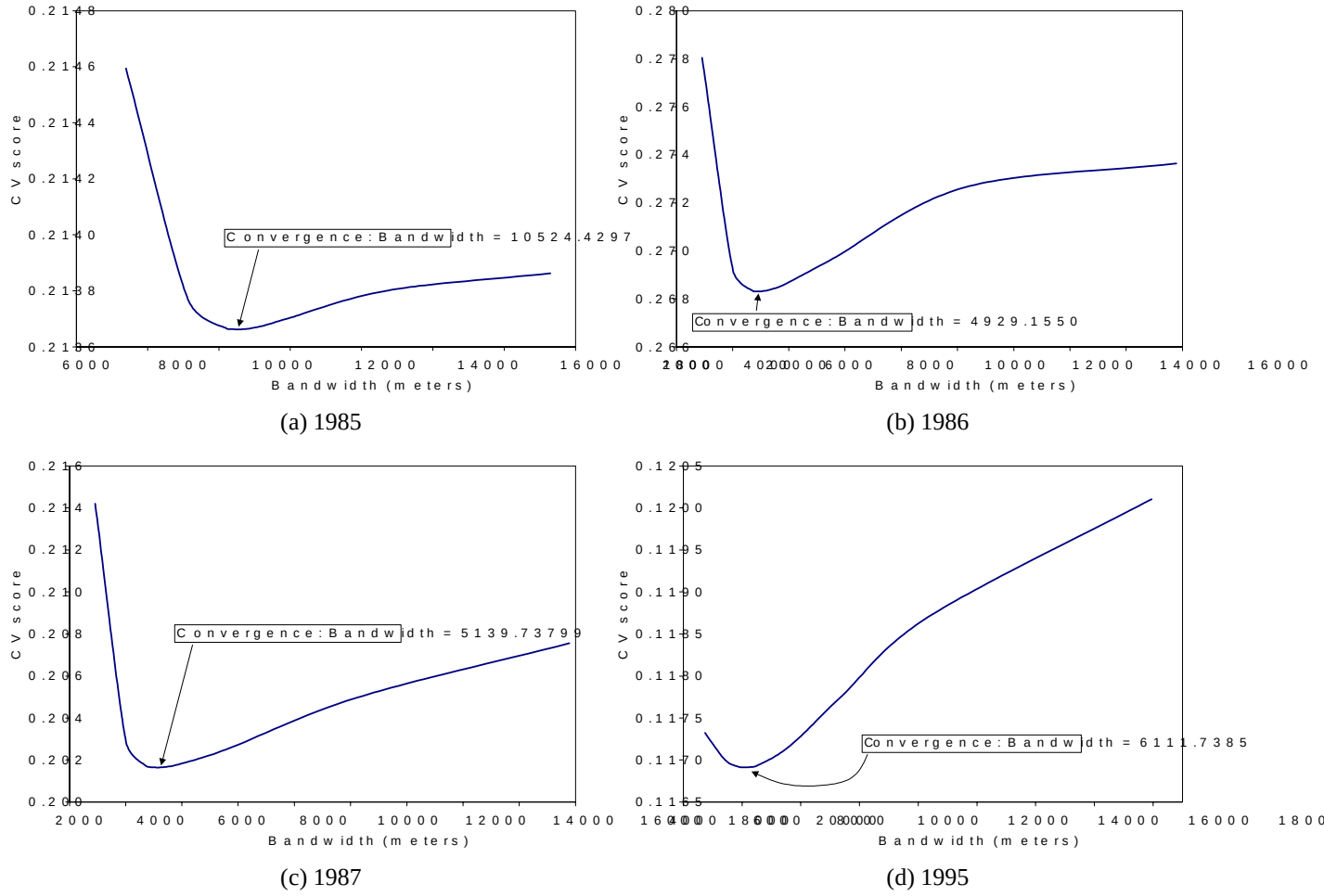
The coefficients obtained from the global model show the expected direction in effects. As the distances from the house to the forest and to the river increase, the lower is the malaria rate. The greater the number of goods, the lower the rate. Increases in the level of education of the head of the household's wife imply lower malaria rates. Increases in cleared area and in income are also associated with reduced malaria rates. However, although significant, the level of education of the head of the household has the expected effect only in 1985.

The local regression produces coefficients for every occupied plot, revealing features hidden in the global coefficients. Although estimates can be generated for all plots in the area, we chose to map only the results for the plots occupied in each year, in order to make some comparisons with the cluster analysis previously discussed.

Figure 10 shows the mapped intercepts modeled for each year by the geographically weighted regression method. The maps show the pattern of malaria rates after the spatial variation of the covariates have been taken into account. Two points are clearly evident. First, the covariates included in the model do not capture all the spatial variation. Much is still left to be explained. Second, the maps bear a resemblance to the cluster maps shown in Figure 8, in the sense that the larger values of the intercept are mostly located in areas previously identified as clusters of high malaria rates, and

[11] Since our unit of analysis is the plot, those that have more than one household will have different outcomes for certain variables. In the case of distance from the house to the forest or to the river, we consider the minimum distance, since this represents the worst possible scenario for malaria transmission. In the case of number of goods, we work with a weighted average, with the number of people in each household as the weights. For level of education of the head of the household and of his wife, we computed an arithmetic mean.

vice-versa. In 1995, despite the fact that eight explanatory variables were included in the model, the intercept still reveals the same pattern of high values in Tract 1 and in plots of Tract 2 near protected forests and the Machadinho River.



(a) 1985 (b) 1986

(c) 1987 (d) 1995

Figure 9 – Cross-validation scores and optimal bandwidths

To clarify what is revealed by geographically weighted regression we highlight the major findings and illustrate the application with two maps: minimum distance from the house to the forest in 1985 (Figure 11) and number of goods in 1986 (Figure 12)

The variable *minimum distance from the house to the forest* was included in three years: 1985, 1987 and 1995. In the former two years the parameter estimates are all negative, as one would expect. However, in 1985 (Figure 11) they are more negative in areas close to protected forest reserves, as well as in the location identified before as Cluster H. In 1995, although the most negative values are all observed in Tract 1 and in those plots of Tract 2 near the protected forests and the Machadinho River, two particular areas show positive coefficients, an unexpected relationship. One of the areas is the one previously identified as Cluster L. The fact that positive values occur only in 1995, and in areas of low malaria rates, can be an indication that at later stages of the settlement process, and in areas where the colonization is more consolidated, this variable is not a key factors as it was in the earlier years.

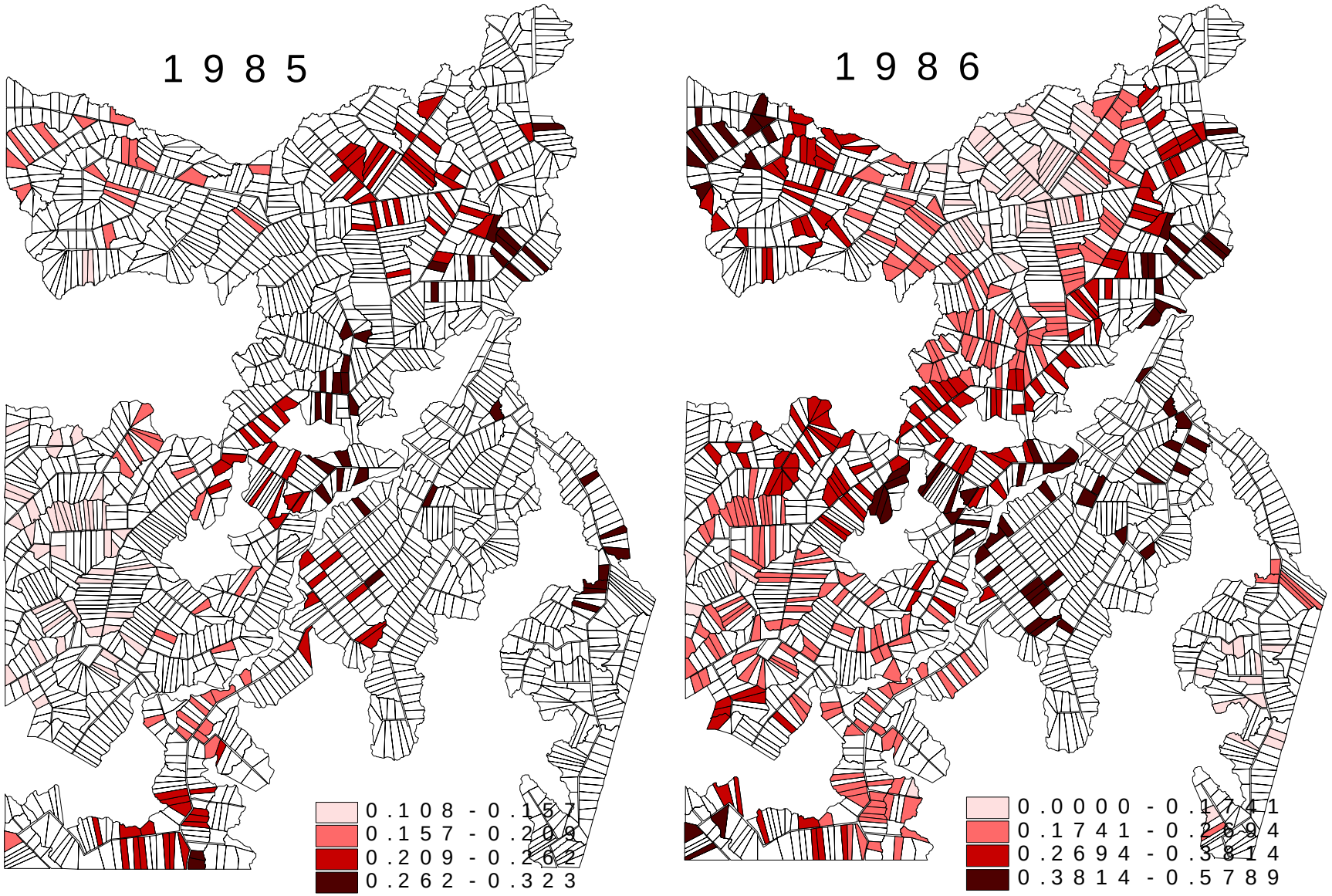


Figure 10 – Spatial distribution of the intercept of the geographically weighted regression model – Machadinho – 1985/95 (cont'd)

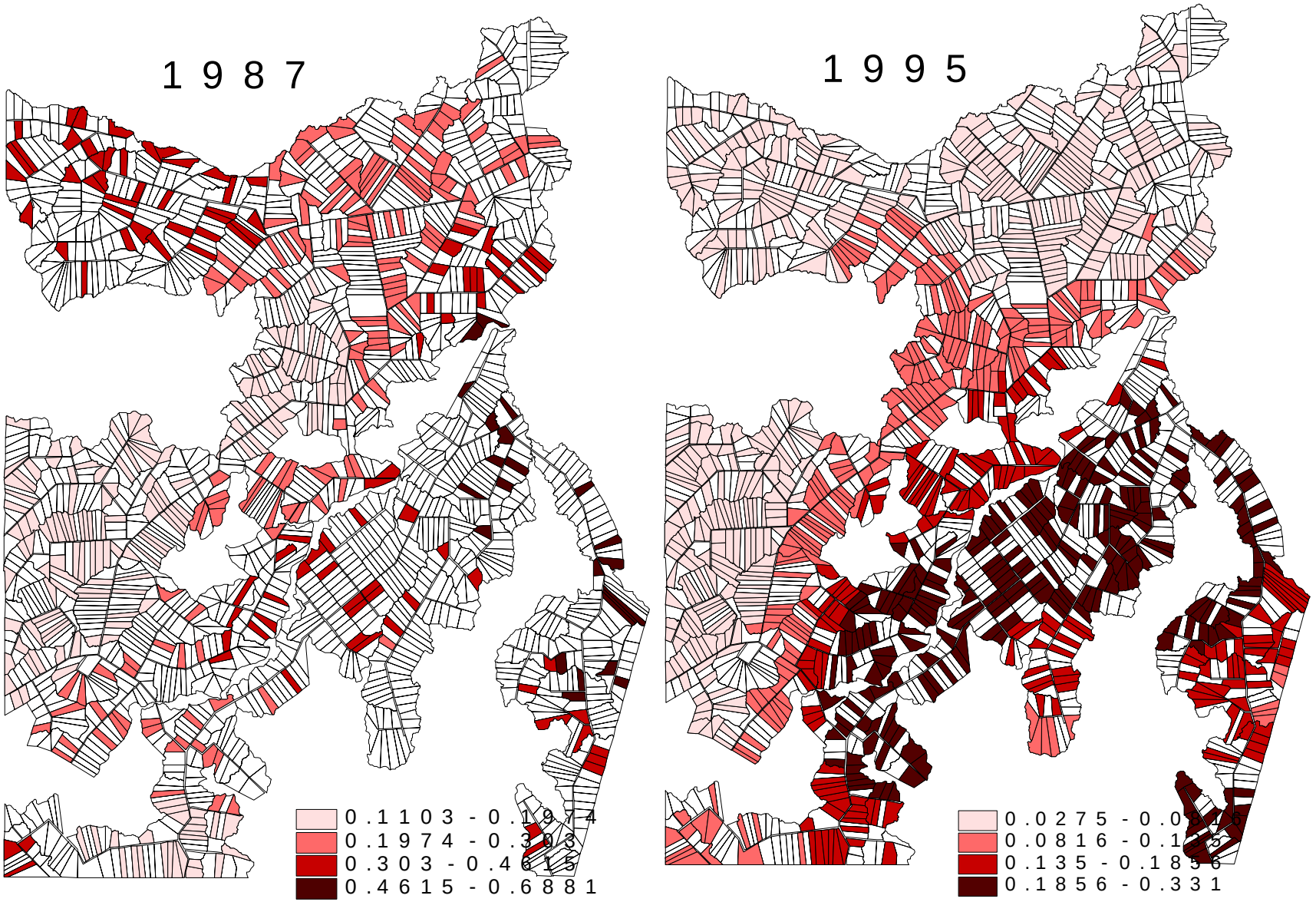


Figure 10 – Spatial distribution of the intercept of the geographically weighted regression model – Machadinho – 1985/95 (conclusion)

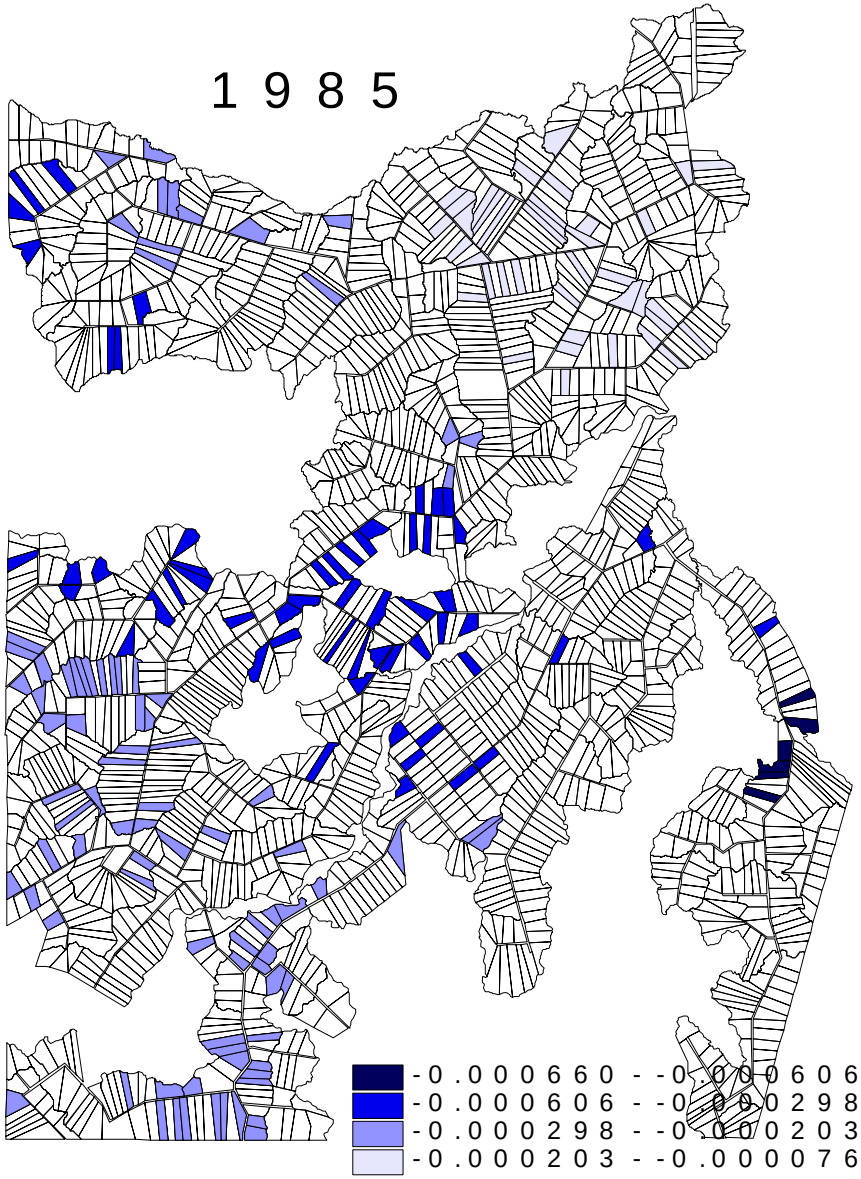


Figure 11 – Spatial distribution of the coefficient for minimum distance from the house to the forest – Machadinho – 1985

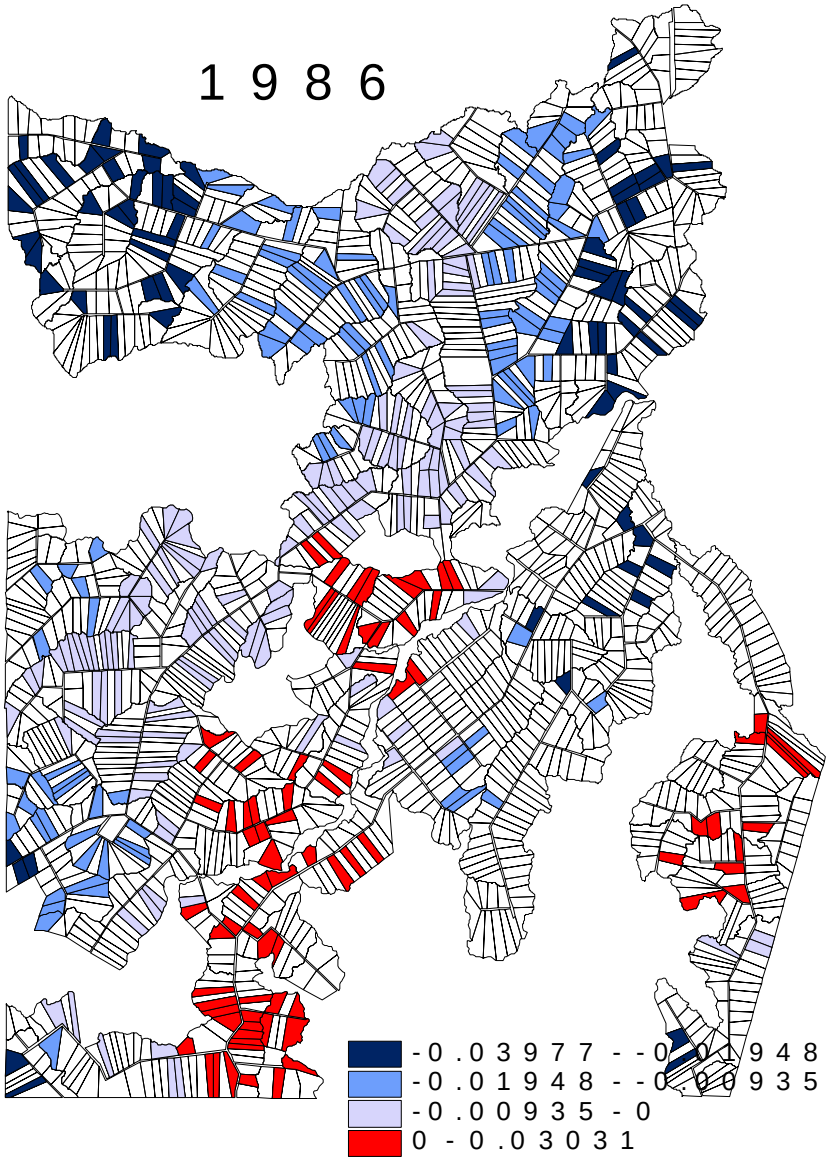


Figure 12 – Spatial distribution of the coefficient for number of goods Machadinho – 1986

Concerning the *minimum distance from the house to the river*, although almost every plot has a source of water at the back (due to the regional design that follows the network of streams), the greatest negative effect of this variable on malaria rates is located near the Machadinho River and the protected forest reserves. These are areas where malaria is usually high.

When comparing 1986 and 1995, the coefficients of this variable are much more negative in the latter than in the former year. Although the positive coefficients in 1995 are of a small magnitude, it is an interesting pattern that is most likely to be explained only by a combination of factors.

The variable *cleared area* is a good example of interesting spatial variations hidden by a global coefficient. As we mentioned before, the global regression estimated negative coefficients for this covariate. However, in 1987, almost half of the plots had positive parameters, which must be intrinsically related to the fact that in the earlier stages of the settlement, deforestation contributes to the increase in malaria (Sawyer, 1992a). In 1995, the negatives values are much more frequent, being more negative in the west part of Tract 1, and in a small area in the northwest part of Tract 2. None of these details could be captured by a global measure.

The *number of goods* variable also reveals an interesting spatial pattern (Figure 12). In 1986, although most of the plots have negative coefficients, with the greatest effect in the north section of Tract 2, positive coefficients are observed exactly in the areas that have a high risk of malaria transmission (Cluster H identified before, and other parts of Tract 1). We can argue that in the early stages of occupation, when agricultural production is not consolidated and the settler does not have enough income generated from his own production, an increase in the number of goods can have the effect of reducing the available financial resources. Such resources are necessary for malaria treatment, especially in areas of high-risk. This hypothesis is supported by the spatial distribution of the coefficients in 1995. Only 2% of the plots show a positive coefficient, all located in the southern part of Tract 1. It is important to note that the number of goods is just a proxy for economic condition. Having one or more of the goods considered here won't reduce the chances of the settler having malaria. However, the variable is certainly an indication of wealth of the household. It is possible that there is a threshold in the number of goods during the early stages of settlement, which determines the direction of the effect of the variable.

As one would expect, the spatial distribution of the coefficients of the variable *income generated from the activities in the plot* is analogous to the variable *number of goods* in 1995, since both capture the economic effect on malaria. The most negative values are all located in Tract 1, where the risk of malaria transmission is high.

The two variables related to the level of education show very curious spatial distributions. The analysis described in Section 4 showed that these variables had a strong positive effect at early stages, losing their potential effect later on. The spatial distribution of the coefficients given by the local model changes through time. Education of the head of the household and education of his wife evolve in opposite directions. For the head of the household an increase in education is associated with increased malaria through time. The parameter for education of the head of the household's wife becomes more negative as the settlement ages. One hypothesis is that the increase in education acts in different ways for the head of the household and for his wife. For the former, an increase in education can be translated into more access to agricultural practices, forms of land use and improvement of the soil, which ultimately can expose him to a higher risk of malaria transmission. On the other hand, for the wife, an increase in education can imply adoption of preventive measures against malaria, and more regular visits to the health centers. These covariates can have very strong interactions with other variables, and their effect alone does not reveal a clear relationship.

In summary, the results show that many of the variables previously identified as having a strong global effect on malaria rates, present important spatial variations not always easy to explain or understand. In many cases, although the global effect has the expected direction, the local coefficients reveal interesting relationships that, ultimately, can contribute to the success or failure of local malaria control policies. These features make clear that ignoring the spatial variation in the discussion of the determinants of malaria risk, and in the design of surveillance and mitigation strategies, is a serious mistake.

CONCLUSION

Global models that assume spatial stationarity have been widely used in social science. The study of malaria in the Brazilian Amazon is just one example of its main applications. We argue in this paper that this assumption can be too strong, hiding important features of the case under study. Taking malaria as an example, we review the main conclusions of the global models, and apply local models in order to see if the malaria risk does have a significant spatial variation.

Using a spatial analysis technique to identify clusters, we showed that malaria in Machadinho is indeed clustered, as well as many other variables assumed to have an effect on the risk of transmission. Hot spots of malaria are observed near the forest reserves and the Machadinho River, especially in the early years of the settlement process. Later on, they are mostly concentrated in Tract 1, the area of most difficult access in the project. Clusters of low malaria rates were also observed. Interestingly enough, they are present in the area since the onset of the project, revealing that some groups of settlers could, indeed, manage to have low malaria in an epidemic area. The identification of spatial clusters is a powerful technique, especially for a better understanding of the disease mechanisms and for surveillance.

Moving to a multivariate analysis, the spatial factor was incorporated through the use of geographically weighted regression, and the results revealed hidden spatial patterns, interesting effects, and surprising relationships. Variables that had large global effects can turn out to have a strong influence only at some particular locations.

These results, in summary, answer many questions, uncover a lot of hidden spatial patterns, but also bring new questions to the analysis. The strong spatial component left in the intercept of the regression results raises the issue of how far these models can go in trying to characterize the determinants of malaria transmission. Can a regression model fully account and explain the variation in malaria rates? Is this the best approach to define low-risk and high-risk malaria?

One possible alternative is the use of fuzzy-set analysis (Ragin, 2000). In this approach, the focus is not on homogeneity, but on understanding diversity. If we consider that different areas are more likely to have distinct profiles of risk, with a set of specific variables acting as the major determinants of malaria transmission, and that these profiles are not the same through time, the fuzzy-set analysis is very appealing.

To conclude, the evidence presented in this paper makes it clear that for in-depth understanding of malaria foci in dynamic areas such as colonization projects, the data cannot be modeled as spatially stationary. The use of GIS and spatial analysis methods help fill in some gaps in knowledge never investigated or even known. Ultimately, we want to be able to predict malaria rates in all areas, even those not yet occupied, based on the relationships of malaria rates and other variables, as well as their spatial variation.

REFERENCES

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**: 251-262.

Aldstadt, J., Chen, D., and Getis, A. (1998). *Point Pattern Analysis*. website: http://xerxes.sph.umich.edu:2000/ppa/doc/html/ppa.html

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis* **27**(2): 93-115.

Arbia, G. (1989). *Spatial data configuration in statistical analysis of regional economic and related problems*. Dordrecht, Kluwer Academic.

Assunção, R. M., Reis, E. A., et al. (1999). Mapas de malária em Rondônia usando o estimador Bayesiano empírico para dados binários. *Revista Brasileira de Estatística* **60**(213): 69-94.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A* **154**: 143-155.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* **28**(4): 281-298.

Brunsdon, C., Fotheringham, S., and Charlton, M. (2000). *Geographically weighted regression as a statistical model*. Newcastle-Upon-Tyne, Spatial Analysis Research Group, Department of Geography, University of Newcastle-Upon-Tyne.

Casetti, E. (1972). Generating models by the expansion method: applications to geographic research. *Geographical Analysis* **4**: 81-91.

Castilla, R. E. and Sawyer, D. O. (1993). Malaria rates and fate: a socioeconomic study of malaria in Brazil. *Social Science & Medicine* **37**(9): 1137-1145.

Castro, M. C. de and Singer, B. (2000). *Agricultural colonization, environmental changes, and patterns of malaria transmission in the tropical rain forest: the case of Machadinho d'Oeste, Rondônia, Brazil*. Annual Meeting of the Population Association of America, Los Angeles, California.

Caufield, C. (1996). *Masters of illusion: the World Bank and the poverty of nations*. New York, Henry Holt.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368): 829-836.

Cressie, N. (1993). *Statistics for spatial data*. New York, J. Wiley.

Deane, L. M. (1986). Malaria vectors in Brazil. *Memórias do Instituto Oswaldo Cruz* **81**(suppl.II): 5-14.

Deane, L. M. (1988). Malaria studies and control in Brazil. *The American Journal of Tropical Medicine and Hygiene* **38**(2): 223-230.

Fearnside, P. M. (1986). *Human carrying capacity of the Brazilian rainforest*. New York, Columbia University Press.

Foster, S. A. and Gorr, W. L. (1986). An adaptative filter for estimating spatially varying parameters: application to modeling police hours spent in response to calls for service. *Management Science* **32**: 878-89.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000). *Quantitative geography: perspectives on spatial data analysis*. London, Sage Publications.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1997). Measuring spatial variations in relationships with geographically weighted regression. In: M. M. Fischer and A. Getis (eds.). *Recent developments in spatial analysis: spatial statistics, behavioural modelling, and computational intelligence*. New York, Springer. p.60-82.

Fotheringham, A. S., Charlton, M., and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment & Planning A* **30**: 1905-1927.

FUNASA (2000). *Plano de Intensificação das Ações de Controle da Malária na Amazônia Legal*. website: http://www.funasa.gov.br/epi/malaria/pdfs/resumo_exec_pcim.PDF

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**(350): 320-328.

Getis, A. (2000). *Measures of spatial association*. Manuscript

Getis, A. and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**(3): 189-206.

Getis, A. and Ord, K. J. (1996). Local spatial statistics: an overview. In: P. Longley and M. Batty (eds.). *Spatial analysis: modelling in a GIS environment*. Cambridge. p.261-277.

Gillies, M. T. (1988) Anopheline mosquitos: vector behaviour and bionomics. In: Wernsdorfer, W. H. and McGregor, I. (eds.) *Malaria: Principles and Practice of Malariology*. Edinburgh, Churchil Livingstone. p.453-485.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London, Oxford University Press.

Gorr, W. L. and Olligschlaeger, A. M. (1994). Weighted spatial adaptative filtering: monte carlo studies and application to illicit drug market modeling. *Geographical Analysis* **26**: 67-87.

Greig, D. M. (1980). *Optimisation*. London, Longman.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. London, Chapman and Hall.

Hirsch, A. (1883). *Geographical and historical pathology*. London, The New Stdenham Society.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1991). *Fundamentals of exploratory analysis of variance*. New York, Wiley.

INCRA (1985). *Parcelamento geral do P. A. Machadinho*. Rondônia, INCRA.

Jacquez, G. M. (1998). GIS as an enabling technology. In: A. C. Gatrell and M. Löytönen (eds.). *GIS and health*. London, Taylor & Francis. p.17-28.

Jacquez, G. M. (2000). Spatial analysis in epidemiology: nascent science or a failure of GIS. *Journal of Geographical Systems*(2): 91-97.

Jones, J. P. and Casetti, E. (1992). *Applications of the expansion method*. London, Routledge.

Kulldorff, M. (1998). Statistical methods for spatial epidemiology: tests for randomness. In: A. C. Gatrell and M. Löytönen (eds.). *GIS and health*. London, Taylor & Francis. p.49-62.

Mahar, D. J. (1979). *Frontier development policy in Brazil: a study of Amazonia*. New York, Praeger.

Marchesini, P. B., Spencer, B., and Lima, M. de C. (1996). *Distribuição especial da malária no município de Machadinho/RO, 1994*. Anais do X Encontro Nacional de Estudos Populacionais, Caxambú, ABEP.

Marshall, R. J. (1991). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society A* **154**: 421-441.

Mosteller, F. and Tukey, J. W. (1968). Data analysis including statistics. In: G. Lindzey and E. Aronson (eds.). *The handbook of social psychology*. Reading, Mass., Addison-Wesley Pub. Co.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* **27**(4): 286-306.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago, University of Chicago Press.

Reis, E. A. and Assunção, R. M. (1996). *Mapeamento da malária em Rondônia usando o estimador empírico de Bayes*. Anais do X Encontro Nacional de Estudos Populacionais, Caxambú, ABEP.

Sawyer, D. O. (1988). *Notas sobre diferenciais comparativos da prevalência de malária em Machadinho-Ariquemes nos anos de 1985 e 1987*. Belo Horizonte, CEDEPLAR.

Sawyer, D. O. and Monte-Mór, R. L. (1992). *Malaria risk factors assessment.* Interregional Meeting on Malaria, Brasília.

Sawyer, D. O. and Silveira, C. M. (1985). *Malaria in the Machadinho settlement project.* Belo Horizonte, CEDEPLAR.

Sawyer, D. R. (1984). Frontier expansion and retraction in Brazil. In: M. Schmink and C. H. Wood (eds.). *Frontier expansion in Amazonia*. Gainesville, University of Florida Press. p.180-203.

Sawyer, D. R. (1985). *Research design and feasibility in the Machadinho settlement project*. Belo Horizonte, CEDEPLAR.

Sawyer, D. R. (1986). Malaria on the Amazon frontier: economic and social aspects of transmission and control. *Southeast Asian Journal of Tropical Medicine and Public Health* **17**(3): 342-345.

Sawyer, D. R. (1988). *Frontier malaria in the Amazon region of Brazil: types of malaria situations and some implications for control*. Brasilia, PAHO/WHO/TDR.

Sawyer, D. R. (1992a). *Deforestation and malaria on the Amazon frontier*. Seminar on Population and Deforestation in the Humid Tropics, International Union for the Scientific Study of Population, Campinas, Brazil.

Sawyer, D. R. (1992b). *Malaria and the environment*. Brasília, Instituto SPN.

Sawyer, D. R. and Sawyer, D. O. (1987). *Malaria on the Amazon frontier: economic and social aspects of transmission and control*. Belo Horizonte, CEDEPLAR.

Sawyer, D. R. and Sawyer, D. O. (1992). The malaria transition and the role of social science research. In: Chen, L. C., et al (eds.). *Advancing the health in developing countries: the role of social research.* Westport, Auburn House. p.105-122.

Schmink, M. and Wood, C. H. (1992). *Contested frontiers in Amazonia*. New York, Columbia University Press.

Service, M. W. (ed.) (1989). *Demography and Vector-Borne Diseases.* Boca Raton, FLA, CRC Press.

Singer, B. and Castro, M. C. de. (2001). Agricultural colonization and malaria on the Amazon frontier. In: A. Hermalin, M. Stoto, M. Weinstein (eds.). *Population health and aging: strengthening the dialogue between epidemiology and demography*. Forthcoming.

Stevenson, D. C. (1994-2000). *The Internet Classics Archive, Web Atomics*. website: http://classics.mit.edu//Hippocrates/airwatpl.html

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* **36**(2): 111-147.

Sydenstricker, J. M. and Vosti, S. A. (1993). *Household size, sex composition, and land use in tropical moist forests: evidence from the Machadinho colonization project, Rondônia, Brazil*. Meeting of the Population Association of America, Cincinnati, Ohio.

Thomas, R. W. (1990). Introduction: issues in spatial epidemiology. In: R. W. Thomas (ed.). *Spatial epidemiology*. London, Pion. p.1-14.

Tobler, W. R. (1979). Cellular geography. In: S. Gale and G. Olsson (eds.). *Philosophy in geography*. Dordrecht, D. Reidel Pub. Co. p. 379-386.

Wood, C. H. and Wilson, J. (1984). The magnitude of migration to the Brazilian frontier. In: M. Schmink and C. H. Wood (eds.). *Frontier expansion in Amazonia*. Gainesville, University of Florida Press. p.42-152.

World Bank. (198?). *World Bank approaches to the environment in Brazil: a review of selected projects - The POLONOROESTE Program*. Washington, World Bank. Internal document.