**Supporting Text**

**Malaria Risk on the Amazon Frontier**

The purpose of this supplement is to present details about frontier malaria and our analytical strategy that should add greater clarity about this complex phenomenon than could be conveyed within the space restrictions of the paper itself. We consider four topics:

1. Spatial scaling of frontier malaria – Over the past 35 years, the state of Rondônia in Brazil has been the site of a vast diversity of settlement projects and new communities, developed under both public and private sponsorship. Although the concept of frontier malaria was formulated with a temporal pattern that was specific to a single community, we find that the same temporal pattern extends to a still coarser scale; namely, the entire state of Rondônia with its heterogeneous set of new and expanding communities. We show this evolutionary pattern in section 1.

2. Section 2 gives the details of the geo-statistical analysis used to identify distinct subareas of malaria risk as a function of time since the opening of Machadinho for settlement.

3. In section 3 we first present an intuitive introduction to Grade of Membership analysis that emphasizes interpretations relevant for malaria risk. We then show estimated pure-type frequencies for the many conditions that enter into environmental and behavioral classification of malaria risk on the Amazon frontier.

4. We conclude with an annotated photographic tour of Machadinho. In our opinion, it is impossible to fully grasp the subtlety of frontier malaria from verbal description and formal statistical analysis alone. We regard this visual supplement to our paper as an

essential component of the evidence identifying the linkages across spatial scales and over time that distinguish the notion of *frontier malaria.*

### 1. Frontier Malaria at Different Spatial Scales

The tropical rain forest assures good conditions for the spread of insects, given their high temperatures, humidity, and rainfall throughout most of the year. In the case of *Anopheles Darlingi* (the main malaria vector in the Amazon), natural breeding places are observed in the forest margins at the beginning and end of the rainy season. Inside the undisturbed forest, however, the ideal conditions for *A. Darlingi* are seldom found, since standing water is acidic and the partial shade favored by this species is absent. Manmade modifications, however, will tend to break this natural equilibrium.

In any colonization project the first environmental transformation is the process of clearing the land and preparing it for cultivation. A very common technique used in the Amazon is slash-and-burn. The beneficial aspects of burning are: the increase in the levels of pH, phosphorus, calcium and magnesium, and the decrease in the level of toxic aluminum ions. A poor burn compromises the crop yields, and contributes to the proliferation of breeding sites for mosquitoes (1). In colonization areas the quality of the burn is related to the distribution of trees on a settler's land, to individual resources (availability of chainsaw), and to credit incentives. A poor clearing and burning process can cause the obstruction of steams and leave the taller trees standing, providing ideal conditions for *A. Darlingi* breeding. Additionally, the clearing process brings up the notion of forest fringe, a frontier between the forest and the property, where mosquito density is most significant, and the risk of malaria transmission is very high.

These relationships between the environment and malaria transmission are reflected in the concept of frontier malaria proposed by Sawyer (2). This concept operates at three spatial scales, but also evolves through time in a very particular manner. Sawyer and Sawyer (3) proposed a model of malaria transition, which occurs over a period of 15 years, divided into three phases, as illustrated in Fig. 5.

The first phase, called epidemic, starts with the occupation of the area, when the natural forest gives place to men. It spreads over the first three years of the settlement process, and is characterized by a fast and dramatic rise in the Annual Parasite Index – API (number of positive blood slides per 1,000 people) up to around 3,000 per 1,000 people. During this phase, the total cleared area is still low; the quality of the house is poor; man-made transformations favor the proliferation of mosquito breeding sites; and settlers do not have the appropriate knowledge to protect themselves against malaria, among other factors. Traditional control measures are not effective, and malaria transmission reaches startling numbers.

The second phase, called transition, lasts for about five years, and is characterized by significant decreases in the API. The decline is achieved through a combination of factors, including socio-economic, environmental, and personal behavior. Cleared area increases, profits from the agricultural production allow the improvement of housing and personal care conditions, and knowledge about malaria becomes more widespread. Overall, some traditional control measures are likely to become more effective, and malaria transmission starts to decline.

The third and last phase, named endemic, starts approximately eight years after the inception of the settlement project. At that stage the area is more stable. Settlers are likely to be well established in their plots, producing different crops, living in better houses, and able to protect themselves against malaria in much better ways than at the time of the initial occupation. Local infrastructure should also have improved, facilitating the search for health care, the storage and transportation of commodities, and also the organization of community groups. Moreover, as the area develops, pollution of mosquito breeding places can contribute to a decrease in the risk of malaria transmission. Fig. 6 shows these relationships as they appear in the Machadinho project.

While Table 2 in the paper highlights that this transition is corroborated by the data at the local level, as described by the EWR observed in Machadinho, we argue that frontier malaria (and its pattern of transition) also operates at a state scale. Fig. 7 shows the relationships between deforestation, settlement patterns and malaria transmission over time. After 1970, when government colonization efforts gained importance, deforestation in the State of Rondônia increased dramatically. And so did malaria transmission. Although estimates of deforestation vary between different sources, in 1975 ≈0.3% of the State forest cover had been cleared (4). Fig. 7*A* shows that some clearing was done in the two municipalities that comprised the territory at that time, Porto Velho and Guajará Mirim. Besides that, some deforestation is already observed along BR-364, especially near Ouro Preto colonization project, which started in 1970. In that year, the Annual Parasite Index (API) in the state was 48.1 per 1,000.

As the number of colonization projects increased, deforestation became more intense, and malaria transmission assumed an epidemic character. In 1986, when 37 colonization projects had been initiated, ≈12% of the forest in the State had been cleared, and the API was 208.9 per 1,000. In 1992, with 52 colonization projects implemented, deforestation amounted to 16%, while malaria transmission starts to shift its trend, declining to 113.6 cases per 1,000 people. Finally, in 1996 the amount of forest cleared was ≈22% of the State area (5), mostly concentrated along BR-364. Considering a buffer zone of 50 km on each side of BR-364, estimates show that 33% of the forest along the highway was cleared by 1991 (6). The API, however, kept its declining path, reaching 75 cases per 1,000 people.

Fig. 8 shows the land fragmentation in Rondônia at the end of 2001: 120 colonization projects, 40 regulated protected forest areas, and 18 official indigenous reserves. The estimated cleared area in that year was ≈29%, but the API was only 32.3 cases per 1,000 people. Therefore, although new colonization areas are being opened, often registering outbreaks of malaria in the initial years of occupation, at the State level malaria rates reached a stable pattern, as described by phase 3 of the malaria transition model.

## 2. Identification of Subareas of Malaria Risk

Subareas of malaria risk in Machadinho were identified in a two-step process. First, the presence of clusters of high and low malaria rates among those plots that were occupied during the field surveys was tested using a local indicator of spatial association. Results revealed that in each of the four surveyed years (1985-7 and 1995) pockets of low and high rates were observed, which is a good indication of the existence of different subareas regarding malaria risk. However, the place where boundaries should be set between each subarea is not clear, considering that some plots were not occupied. The second step overcomes this problem. Spatial estimation of malaria rates for the whole settlement area was performed using kriging (7), a geostatistical technique for spatial prediction.

### *Clustering Pattern*

Local statistics of spatial association measure the association between each location and its neighbors based on defined distances. Consider an area divided into *n* locations, each identified with a point *i* and associated with a value $x_i$ (a realization of a random variable *X*). When focusing on location *i*, all of the others are named $x_j$. The $G_i^*(d)$ statistic is defined as (8)

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)\ x_j}{\sum_{j=1}^{n} x_j},$$

where $w_{ij}$ are the elements of a weight matrix. Standardized values of the statistic, $Z\,[G_i^*(d)]$, are frequently used. The standard null hypothesis assumes that there is no association between location *i* and its neighbors, within distance *d*. Under the null hypothesis $Z\,[G_i^*(d)]$ are asymptotically normally distributed, N(0,1), as $n \to \infty$ (8). Therefore, significant negative $Z\,[G_i^*(d)]$ reveals spatial clustering of low values of *X* within distance *d*, while significant positive $Z\,[G_i^*(d)]$ are indicative of spatial clustering of high values of *X* within distance *d*.

Assessing the significance of the $G_i^*(d)$ statistic requires careful attention. First, each location $i$ is assigned a test and the rejection or retention of the null hypothesis raises questions of multiplicity. Second, local statistics have two possible sources of spatial dependence: (i) geometric, which is caused by the fact that nearby locations share common elements in the neighborhood defined by the weight matrix, and (ii) true dependence that might exist between the values of nearby locations. The greater the number of common neighbors, the greater will be the dependence or correlation between the tests. Therefore, procedures to account for multiplicity and spatial dependency are required to properly assess the significance of local statistics. The simplest, and most widely used technique, is the Bonferroni method, which controls the probability that a true null hypothesis is incorrectly rejected. However, it is a very conservative procedure. As a result, the larger the dataset the greater the chances that truly significant differences will be missed. To overcome this problem we use a procedure introduced by Benjamini and Hochberg (9), called the False Discovery Rate (FDR), which controls the average rate that declarations of significance are truly non significant (10). An evaluation of the use of the FDR approach in appraising the occurrence of clusters detected by local indicators of spatial association showed a significant gain in identification of meaningful clusters, in comparison to more conservative approaches (11).

Assume that there are $m$ hypotheses to be tested. Among those, $R$ will be declared significant (null hypothesis rejected), $F$ are false positives (null hypothesis incorrectly rejected, or Type I error), and $S$ are the true positives (null hypothesis correctly rejected). Moreover, consider a variable $Q$ defined as the proportion of null hypotheses incorrectly rejected among all those that were rejected. This variable can be expressed as $Q = F\,(F + S)$ or $Q = F/R$, and, by definition, when $R = 0$ the variable $Q$ is set to equal zero. Benjamini and Hochberg (9) define the FDR as the expected value of variable $Q$. For independent tests it can be controlled for each test at a level $\alpha$ by the following stepwise procedure: (i) order the test statistics $p$-values ($p_i$) in ascending order ($p_1 \leq p_2 \leq \ldots \leq p_m$); (ii) starting from $p_m$ find the first $p_i$ for which $p_i \leq (i/m)\alpha$ ; and (iii) regard all tests as significant for which $p_i \leq p_{critical} = (i/m)\alpha = p_{FDR}$. Therefore, if $p_{critical}$

equals 5% it means that, on average, among the rejected null hypothesis, 5% were truly null (12). The gain in power provided by this procedure becomes larger as $m$ increases.

Choosing the right distance, and therefore the neighborhood around each plot, is a key factor. When $d$ is very small or very large (covering the total area) normality is lost. Getis and Ord (13) suggest that the maximum distance should never exceed ½ of the shorter side of the study area, while the number of neighbors should be at least 30 for large samples, and 8 for small ones. However, there is no definite rule to guide the decision. In our analysis the most appropriate neighborhood of potential exposure to malaria transmission is a function of a dynamic process between men, mosquitoes, and the local environment, which determines a lower or higher risk. Failure to account for this process will most likely result in neighborhoods that have little use, if any, for controlling the disease.

We selected the distance $d$ based on three factors associated with both the area and phenomena under study. The first factor is the flying behavior of mosquitoes, which ranges from 500 m to 3,000 m without the aid of the wind, and up to 5,000 m with the aid of the wind (14-16). The second factor is the size of the plots: an average front of 400-500 m and an average depth of 700-900 m. An extensive analysis of buffers sized between 500 to 8,000 m suggests that distances lower than 2,000 m would result in a very small number of neighbors for each plot, with a large number of them being left as "islands" with no neighbors. The third factor is related to the implementation of control measures by the local health agencies, which is done by sectors. An analysis of the distribution of sectors suggests that distances larger than 2,000 m would be more appropriate. Additionally, 3,500 m is the minimum distance necessary to guarantee that all plots have at least one neighbor in all four years.

It is important to mention that different approaches have been used to choose the most appropriate neighborhood, such as a k-nearest neighbors method in which the number of neighbors is fixed (17); using the average distance between nearest neighbors as a reference (18); comparison between the number of neighbors in a 1st order contiguity weights matrix and a matrix defined by a certain distance *d* (18); and evaluating the most effective distance to remove the spatial autocorrelation from the data through the use of a filtering procedure (18, 19). The first three approaches are not appropriate for our data since information is not available for all 1,742 plots. Some plots have no contiguous neighbor, and in those cases a fixed number of neighbors could result in a very large neighborhood with no useful meaning for purposes of malaria transmission. The last approach would imply in adopting different distances for each of the four years of data we have. However, to guarantee temporal comparability of the outcomes of the dynamic process of malaria transmission we opted to use a unique distance, which based on above described criteria was chosen to be 3,500 m.

Fig. 9 shows the clustering pattern observed in malaria rates in Machadinho using a distance of 3,500 m, and applying a FDR correction for multiple comparisons. The pattern gives an indication of the presence of distinguishable subareas of malaria risk in Machadinho. Since not all plots are occupied, and a reduced number of clusters are detected in some years, the exact borders of those subareas cannot be precisely defined. The spatial estimation of malaria rates, however, overcomes this constraint, as explained next.

***Spatial Estimation***

Consider the spatial locations $\{s_1, \ldots, s_n\}$ and the associated malaria rates $\{Z(s_1), \ldots, Z(s_n)\}$ observed at these locations. The spatial process $Z(s)$ can be decomposed into $Z(s) = \mu(s) + W(s) + \eta(s) + \varepsilon(s)$, where $\mu(s)$ is the mean structure and is called the large-scale variation; $W(s)$ is an intrinsically stationary process called the smooth small-scale variation; $\eta(s)$ is also an intrinsically stationary process, independent of $W$, called the microscale variation; and $\varepsilon(s)$ is a white-noise process, independent of $W$ and $\eta$, called the measurement error (20).

Assuming that the difference between two locations at a distance $h$ depends only on the distance $h$, $\mathrm{var}\big(Z(s+h) - Z(s)\big) = 2\gamma(h)$, where the quantity $2\gamma(.)$ is called the variogram and $\gamma$ (.) the semivariogram. If the spatial data set is intrinsically stationary, the variogram will be stable across the area studied (13). Intrinsic stationarity is defined through first differences (20), which are expressed by $\mathrm{var}\big(Z(s+h)-Z(s)\big)=2\gamma(h)$ and by $E\big(Z(s+h)-Z(s)\big)=0$.

Furthermore, when $2\gamma$ (h) can be written as $2\gamma^{0}(\|h\|)$ the variogram is isotropic. This means that $\gamma$ (*h*) is invariant under rotation. There are no directional (drift) effects (21). This is clearly seen in a contour map of directional correlations, as shown in Fig. 10. No matter the direction, the correlation is the same.

On the other hand, if the process is anisotropic, then the variogram is not only a function of the distance *h*, but some adjustment for direction has to be implemented. Fig. 11 shows an anisotropic process. Depending on the direction, the spatial correlation assumes different magnitudes.

A semivariogram is composed of three parameters: nugget effect, sill, and range, as shown in Fig. 12. The investigation of their behavior is important for guaranteeing that the variogram has the best fit. The nugget effect occurs when there is discontinuity at the origin. Ideally, $\gamma(0) = 0$, but if a discontinuity is observed, $\gamma(h) \rightarrow c_0 > 0$ as $h \rightarrow 0$, and $c_0$ is called the nugget effect. Cressie (20) decomposes the nugget effect into two components: the measurement error variance ($c_{ME}$) and the microscale variance ($c_{MS}$). So, $c_0 = c_{ME} + c_{MS}$. Estimating $c_0$ from the data are not easy, especially when the spatial lag is large. Interpolations can be made at lags close to zero, but it is not clear that this will lead to the best estimation. Many models simply assume that the measurement error is zero, but Cressie (20) proposes some alternatives to be used when it is known that the data have measurement error. For efficient estimation, the nugget effect should be well captured.

The sill can be understood as the limit of the variogram, when this limit exists. Formally, it can be expressed as $\sigma_Z^2 \equiv \lim_{\|h\|\to\infty} \gamma_Z(h)$. Considering that the measurement error has variance $c_{ME}$, and that the microscale process has a sill equal to $\sigma^{2\eta}$, the sill is the sum of three factors, $\sigma_Z^2 = partial\ \ sill\ + \sigma_\eta^2 + c_{ME}$, where the partial sill is the variance of the smooth small-scale variation.

The range is the distance beyond which Z(s) and Z(s+h) are not correlated anymore. The intuition is that, according to the First Law of Geography (Tobler, 1979) the semivariance may increase as distance increases. This behavior is observed up to a certain point (range), after which distances do not matter anymore, and the variogram will stabilize. The value of the variogram at this point is the sill, and it will eventually equal the variance.

Under the assumption of intrinsic stationarity, the classical estimator of the variogram is $2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}\left[Z(s_i) - Z(s_j)\right]^2$, where $N(h) = \{(s_i, s_j){:}s_i - s_j = h;\ i, j = 1, \ldots, n\}$ (which means the number of possible pairs among the observations that are lagged by $h$) and $|N(h)|$ is the number of distinct pairs in $N(h)$. That estimator is unbiased, although not resistant, and can be affected by atypical observations. To overcome this problem, a more robust estimator is given by (20)

$$2\bar{\gamma}(h) = \left\{\frac{1}{|N(h)|}\sum_{N(h)}\left|Z(s_i) - Z(s_j)\right|^{1/2}\right\}^4 \Big/ \left(0.457 + 0.494/|N(h)|\right).$$

Once an estimate is obtained, different parametric models can be used to fit the final variogram and summarize the spatial dependence present in the data. Cressie (22) presents six models: linear, spherical, exponential, rational quadratic, wave, and power. Additional models include Gaussian, De Wijsian, and log-log (23). Each model has its own stationarity pattern, and two of the most frequently used models are the spherical and the exponential. Given a fitted variogram, and assuming it to be unbiased, optimal spatial prediction can be done using kriging techniques. Formally, kriging is a minimum-mean-squared error spatial prediction method based on the second-order structure of a spatial process $Z(s)$, where the structure is given by the variogram (20). The ordinary kriging estimator of a spatial process $Z(s)$ has two basic assumptions. The first is a model assumption of intrinsic stationarity, $Z(s) = \mu + \delta(s)$, where the mean μ is unknown and the correlated error process can be expressed as $\delta(.) \equiv W(.) + \eta(.) + \varepsilon(.)$. The second assumption is related to the predictor $p(Z; B)$, $p(Z;B) = \sum_{i=1}^{n} \lambda_i Z(s_i)$ and $\sum_{i=1}^{n} \lambda_i = 1$, where $B$ is a block with known location and geometry, over which the prediction will be done (block kriging), $\lambda_i$ are the coefficients of the linear predictor, chosen to sum up to 1 to guarantee that $p(Z; B)$ is unbiased (20). When $B = \{s_0\}$ we have a case of point (or punctual) kriging estimation, which is more appropriate for the analysis of malaria rates, since they represent point values in the area. The optimal predictor of the spatial process will minimize the mean-squared error, $\sigma_e^2 \equiv E[Z(B) - p(Z;B)]^2$. The optimal coefficients of the linear prediction can be obtained from $\boldsymbol{\lambda}_o = \Gamma_o^{-1} \boldsymbol{\gamma}_o$, where all three elements are matrices: $\lambda_o \equiv (\lambda_1, \ldots, \lambda_n, m)'$, where $m$ is a Lagrange multiplier that guarantees that $\lambda_i$ sum up to 1, $\gamma_o \equiv (\gamma(s_o - s_1), \ldots, \gamma(s_o - s_n), 1)'$, and $\Gamma_o$ is an $(n + 1)$x$(n + 1)$ matrix composed of $\gamma(s_i - s_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$; $1$ when $i = n + 1$ and $j = 1, \ldots, n$, and $0$ when $i = n + 1, j = n + 1$.

A cross-validation method can be applied to the kriging estimates as a means of checking the results. As Cressie (20) points out, this method is not a guarantee that the estimates are correct, but it is an indication that they are not incorrect. The idea behind the cross-validation technique is to delete one observation at a time, use the remaining data to predict its value, and compare the results to obtain an indication of the prediction error. Normal probability plots for the residuals of the kriging cross-validation indicate whether the estimates are biased or not.

Fig. 12 shows the estimated malaria rates for Machadinho in 1985, 1986, 1987, and 1995. For ease of interpretation, the estimates were coded as high, medium, and low risk. There is a remarkable resemblance between the overall pattern of the estimates and that related to the clustering configuration shown in Fig. 9, as expected. Malaria risk is mostly concentrated in Tract 1 and along an imaginary stripe following the course of the Machadinho River, including areas of both Tracts. It is important to emphasize the problem of difficult access in Tract 1. Some settlers living in plots with no close neighbors tend to face the effects of malaria more severely. When they get sick they have no alternative but to stay home until they are able to go to the health center or until a health agent stops by for routine surveillance.

Besides the spatial estimates presented in Fig. 13 and the patterns of local spatial association detailed in Fig. 9, two additional issues were considered for the definition of subareas for analysis of malaria risk in Machadinho. First, it is important to note that the subareas do not have to be the same for all years analyzed, especially because the spatial pattern of risk is not constant through time. That was made clear by the temporal evolution of both the clustering and kriging results. Moreover, the ultimate goal of the study is to define profiles of risk, and not to compare particular subareas over time. Second, the subareas must have enough plots occupied to allow the application of GoM analysis (detailed next). There is no particular rule regarding the minimum number of observations necessary to run those models. However, there should be enough data so that the risk profiles are stable. In this study we assumed that a minimum of 50 observations is necessary for the definition of a subarea. Given all these considerations, the definition of subareas for analysis of malaria risk in Machadinho resulted in the configuration shown in the Fig. 2 of the paper.

**3. GoM Analysis**

Characterizing *frontier malaria* risk requires an analysis of environmental and human behavioral conditions in a heterogeneous landscape, where no combination of conditions occurs at particularly high frequency. Thus, crisp classification of individuals into a discrete set of empirically defined risk categories is not feasible. However, the family of GoM models are responsive to this challenge in that they are specified by a small number of extreme categories of conditions (e.g., Low and High risk in a 2-category specification) and the assignment of grade of membership (GoM) scores to each individual, where the scores are a quantitative measure of the degree of similarity of the individual to each of the extreme profiles.

In the present setting, the plot in the Machadinho settlement project is the unit of analysis (i.e., the 'individual'). A GoM representation of plots along J dimensions identifies each plot with a point in a unit simplex. Fig. 14 shows the interpretation of the GoM scores for models with 2, 3 and 4 profiles. The vertices of the simplex correspond to extreme (or ideal) profiles of conditions. In the context of malaria risk, these will be extreme profiles of environmental conditions and behavior/personal characteristics of people on the plot. Edges of the simplex correspond to sets of conditions, some of which are from one vertex and the remainder from the second vertex at the terminal points of the edge. Thus, a plot identified with an interior point on an edge has a sufficiently heterogeneous set of environmental and behavioral conditions that components from two profiles (corresponding to the vertices) are necessary to represent the potential malaria risk features of the plot. The interiors of the faces of the simplex correspond to sets of conditions, some of which are from each of the three vertices of the triangle defining the face. More complex sets of conditions are identified with points in subsimplices defined by four and more vertices.

The relationship of this geometrical structure to plot-specific discrete-valued vectors, $\mathbf{X} = (X_1, X_2, \ldots, X_J)$ of characteristics can be seen from the following considerations. For a population of I plots, their characteristics can be summarized by counts in a J-dimensional contingency table containing $\mathbf{L_1 \times L_2 \times \ldots \times L_J}$ cells, where $\mathbf{L_j}$ = number of possible categories associated with the variable $X_j$. The typically strong dependence among the components of $\mathbf{X}$ can be represented by identifying a vector $\mathbf{g}$ such that conditional on $\mathbf{g}$, $X_1, X_2, \ldots, X_J$ are statistically independent. That such a representation is always possible for finite contingency tables was established by Woodbury, Tolley & Manton (24). They also demonstrated that $\mathbf{g} = (g_1, \ldots, g_K)$ was a set of nonnegative weights such that $g_1 + g_2 + \ldots + g_K = 1$. Furthermore, the conditional probability that each response variable $X_j$ takes on the level $l_j$ given $\mathbf{g}$ can be represented as a linear combination of the components $g_k$, of $\mathbf{g}$ multiplied by the probability that an ideal-type-k plot had response level $l_j$ on the $j^{th}$ variable. It is precisely this representation that identifies plots with points in a unit simplex. In particular, the simplex has K vertices, and $g_k$ for a plot – referred to as the plot's GoM score for the $k^{th}$ profile – indicates what fraction of the plot's characteristics are members of the $k^{th}$ ideal profile. The number of profiles, K, and the response probabilities for type-k plots are all estimated by fitting the GoM model to the J-dimensional response vectors. Technical details of the estimation process are described in Berkman, Singer, & Manton (25), Manton, Woodbury & Tolley (26), and Erosheva (27).

***Application to Machadinho Household Surveys***

A household survey was conducted on settlers at 70% of what were regarded as 'occupied plots' in 1985 and 100% of such plots in 1986, 1987, and 1995. An 'occupied plot' is one in which settlers cleared some of their land and at least lived part-time in Machadinho. Based on this definition, 13%, 20%, 33%, and 51% of the plots in Tract 1 (Fig. 1 of the paper) were occupied in 1985, 1986, 1987, and 1995, respectively. In Tract 2, these numbers were 18%, 38%, 49%, and 57% for the same period. The survey instruments included information on health, demographic, economic, social, ecological, and agricultural characteristics of the people and their immediate environment on the plot (28, 29). A core set of questions remained the same throughout all four years of data collection, thereby facilitating longitudinal data archives with the plot as the unit of analysis.

We applied GoM models with K = 2 to the household data for each subarea in each of the survey years. This led to the specification of low- and high-risk profiles that were subarea specific and to the assignment of a GoM score for each plot. Since there are only two profiles, it suffices to assign a score, g, taking values between 0 and 1 and interpreted as degree of proximity to the high-risk profile (g = 1). Table S1 shows the conditions that entered into the high- and low-risk profiles for subarea 1 in 1985.

The estimated profile probabilities must differ from the marginal frequency of the condition by either a multiplicative factor of at least 1.8 or by 10%, depending upon the value of the marginal frequency. Rationale for these criteria are presented in Berkman, Singer & Manton (25). If, for example, marginal frequencies are high e.g., 60% +, then the 10% criteria would be used for profile frequencies that are higher than that. The conditions are labeled B or E, according as they are behavioral/economic or environmental domains. Fig. 15 shows the distribution of plots in subarea 1 by GoM score.

Analogous tables and GoM score distributions for each subarea in each survey year were obtained, and this information was the basis for the summary Table 1 in the main text.

## 4. Photographic Tour

Next we show a series of photos that illustrate characteristics of Machadinho, which facilitate the understanding of the relationships between environmental change and malaria transmission.

### *Project design*

Instead of adopting the traditional fishbone pattern frequently used in settlement projects in the Amazon, Machadinho had an original and carefully planned plot design. The shape of the plots was irregular, following the course of rivers and streams, and was influenced by the topography. This design can only be implemented in areas of jungle where there is no previous construction or agricultural practice in place. The information available for the specification of the design was a set of aerial photos taken in 1979 during the dry months of June and July. The photos were shot at a distance of 8,600 m. Each picture covered an area of ≈300 km$^2$. A team of technicians from the National Agency for Agrarian Reform (INCRA) developed a photo interpretation technique to analyze the aerial photos. The team selected areas with very irregular elevation for preservation as protected forest, despite the fact that these areas were already sparsely occupied by rubber tappers. They also designed the ideal location of roads in such a way that during the rainy season the increased water volume of rivers and streams would not leave them impassable. Fig. 16 shows an example of the interpretation done by the team with the aerial photographs. Blue lines were used to highlight the streams in the area. Thin red lines buffered around the blue lines define the possible area that can be affected during the rainy season. Green dotted lines suggest optimal location of roads.

***Road construction***

Roads were opened in Machadinho by INCRA, and represented the largest costs in the budget planned for the project. Three types of roads were constructed guaranteeing access to all plots in the area: collector, access and penetration roads. The differences between the types related to the investment during construction, and consequently on the final quality. Collector roads presented the best quality. None of them were paved, but the material used guaranteed that they would be passable even during the rainy season. Collector and access roads were gravel-surfaced (as shown in Fig. 17), widening 6 and 4 m, respectively. Penetration roads did not have a gravel surface and were ≈4 m wide. During the rainy season these were the only roads that required vehicles with special traction. Regarding this issue, culverts were built on both sides of collector roads, as shown in Fig. 18, to allow the flow of water. A total of 468 km of roads were constructed in Tracts 1 and 2 before the plots were distributed. Machadinho was the first settlement project in the Amazon that had a complete road network built before its occupation.

***Housing***

Assistance with housing construction was limited due to scarce resources. A total of 1,094m$^3$ of wood were distributed by INCRA in 1984 to allow the construction of 20m$^2$ houses. It benefited 547 families on a first-come, first-serve basis (30). Fig. 19 shows the typical house that was built using that support. Although this particular house was photographed in 2001, when it was abandoned, without windows and without the original door, the overall sealing of the walls gives much better protection against malaria than most of the houses built during the early stages of occupation. Houses were often made with plastic, cardboard, sawmills leftover, and thatch, as shown in Figs. 20, 21, and 22. As a result, the sealing of the houses was extremely poor, offering limited protection against mosquitoes, if any at all. Yet, the National Health Foundation was spraying the walls of these houses in the early years of occupation. As the colonization project consolidates, those settlers who have been living in the area for longer periods of time are likely to be those who generate some income from their agriculture production. Consequently, they are able to improve their housing conditions. In contrast, settlers who recently moved to the area are expected to face similar challenges as those observed in the initial occupation. Fig. 23 shows the contrast between new and old settlers, their house conditions, crop production, and the outcome regarding malaria transmission. Panel (A) shows the house of a settler living in the border of a forest reserve. Due to the narrow and elongated shape of the plot, the settler lives in close proximity to a stream (which is the main source of water for the family, and the place where clothes are often washed) and to the forest fringe. The quality of the soil is poor, and the settler cannot afford the use of fertilizers. The main crop in the plot is coffee, but it is small and of poor quality. Malaria transmission has been consistently high in the area, and the plot had multiple owners. Panel (B), in contrast, shows the house of a settler living in a fertile area in Machadinho. This area was the first one to produce coffee, and was significant for clusters of low malaria rates since 1986. The house, photographed in 2001, is a high quality construction located far from the forest fringe or streams, which offers excellent protection against mosquitoes. Coffee production in this plot is large and profitable.

***Land Use***

Based on studies of soil reconnaissance of Machadinho (31), the government proposed two models of agriculture production (32). The first, recommended for fertile soils, consists of the cultivation of cocoa. The second, appropriate for soils with low fertility, suggests the production of rubber trees. Official estimates were that 95% of the plots would produce rubber trees, and only 5% would produce cocoa (32). Field survey data show that in practice the agriculture production in Tracts 1 and 2 do not conform to these proposed models and expectations. Among plots surveyed in 1995, 23% had planted rubber trees, and 24% were cultivating cocoa. However, 86% produced coffee, which became the most important crop in the area. The productivity obtained with coffee is below the national average. Nevertheless, coffee cultivation progressively gained in importance in the area, as a consequence of a municipal government incentive that produced and distributed seedlings to settlers. Most often a combination of several types of crops is observed (Fig. 24), including subsistence and commercial types. Also, pasture gains importance over time due to a combination of factors (33-36). While in 1985 22% of the surveyed plots in Machadinho had a pasture area of any size (Fig. 26), in 1995 this number raises to 84%. Due to the overall low quality of the soil, some areas are abandoned after one or more attempts to develop agriculture (Fig. 25).

***Urban Area***

The main urban area of Machadinho is situated 347 km from Porto Velho, Rondônia's capital. It was planned to be an administrative and technical nucleus. Different government agencies were installed to facilitate the coordination of the settlement project. The area grew up very fast, with the installation of commerce (pharmacies, bars, food and agricultural implement shops, etc) and industries (mainly related to sawmill). In the early years it faced problems of infrastructure, due to the dramatic migratory movement. The first plots in Machadinho were distributed to settlers in July and August, too late for starting the clearing and burning process before the rainy season. As a consequence, few settlers effectively occupied them, and many stayed in the urban area. Machadinho quickly became a boom city and is currently the central area of the municipality of Machadinho D'Oeste, created in 1988. Fig. 27 shows the spectacular environmental transformation that took place in the area. Panel (A) shows the urban area in its early stages (in 1985), with temporary houses. On the background of the incipient city one can see the haze, a result of the slash-and-burn agriculture practice that was being practice by every new settler who moved into the area. Panel (B) shows an aerial view of the same area in 2001. In 17 years the forest was replaced by an urban area with almost 24,000 people. Fig. 28 shows the main street in the urban area of Machadinho, where several government buildings are located, such as the City Hall (Fig. 29).

### *Settlers*

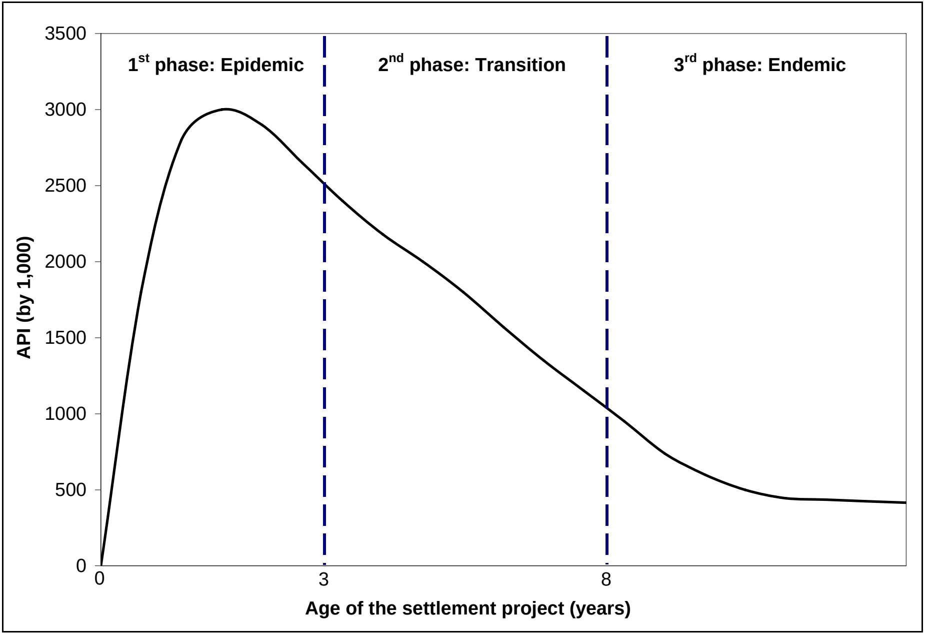
On average, settlers can be characterized in the following manner. They were migrants, some with previous agricultural experience, but most of them with no knowledge of agricultural potential or the techniques necessary for farming in a tropical rain forest area. The migrants had no previous exposure to malaria, and consequently no immunity against it. They had very little knowledge of how malaria was transmitted and treated. They were poor people, attracted by cheap land and promised government support. The characteristics of settlers reflected the selection process imposed by INCRA. Candidates for settlement filled out a form that asked many demographic questions and information about previous work experience. Demographic information included sex, age, education, marriage status, and place of birth. This information was completed for the candidate and for each member of the family, including the type of relationship with the head of the household. Regarding work experience, candidates for land listed their previous agricultural and livestock experience, including type of crops, type of management adopted in the production (manual, mechanized, etc), type of animals, and the time spent in that activity. Weights were assigned to the answers according to the definition of who is a potential settler in colonization areas. Those definitions were established in 1964 by a federal law. As a result, higher weights were given to those candidates that had larger families, lower income, ages between 36 and 45, and more than five years working in agricultural activities. Fig. 30 portrays these migrants. Their clothing offer no protection against mosquito bites (partly because of the heat, partly because of the lack of appropriate knowledge), and it is common for children to be shirtless, as seen on Panels (A) and (C). Their meals often are made and eaten in a partly open area, as shown in Panel (B).

### *Rubber Tappers*

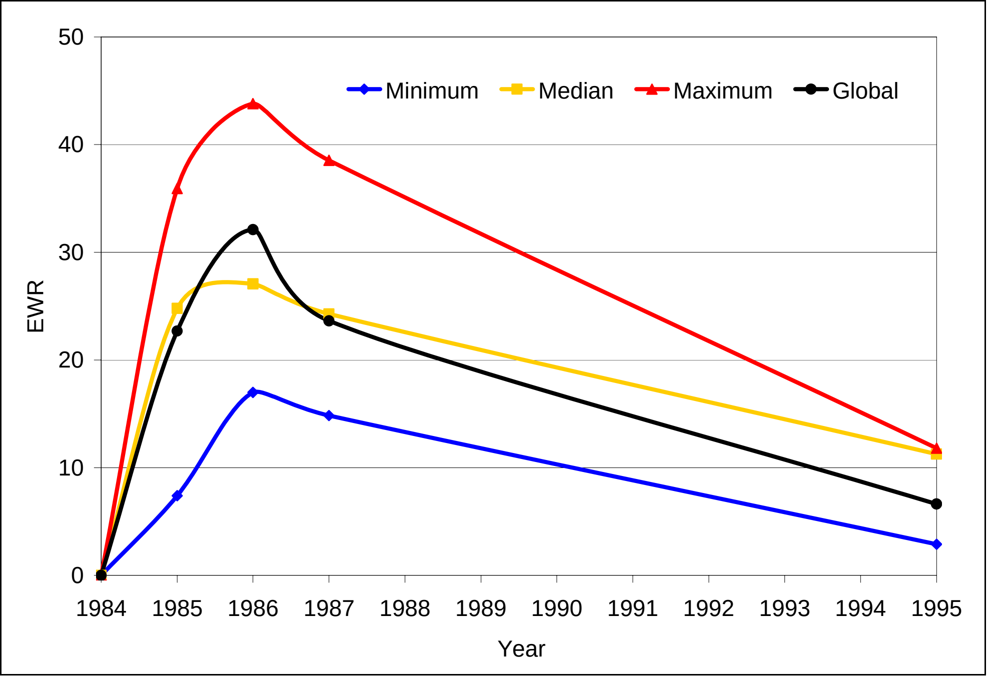
Protected forest reserves in Machadinho should not foster any settlers, and therefore deforestation should not take place. However, satellite images highlight human activity, mainly from rubber tappers (Fig. 31), who sparsely populated Machadinho before the area was opened for occupation. All forest reserves register rubber extraction for 50 years or more. The most populated is the Aquariquara forest reserve: in 1993 there were 51 different areas of rubber extraction in the reserve, with a population of 181 people. Moreover, agriculture was practiced in the area not only for subsistence, but also as a means of complementing family income (37). It is expected that rubber tappers have acquired immunity because of continued exposure to the malaria parasite. Therefore, they are carriers of the parasite. Considering that some of these rubber tappers received plots on the border of the forest where they used to work, they probably contributed to the fast spread of malaria in those locations. Fig. 32 shows a typical rubber tapper's house, located at the border of the forest reserve (at the forest fringe), and near a stream.

### *Forest Fringe*

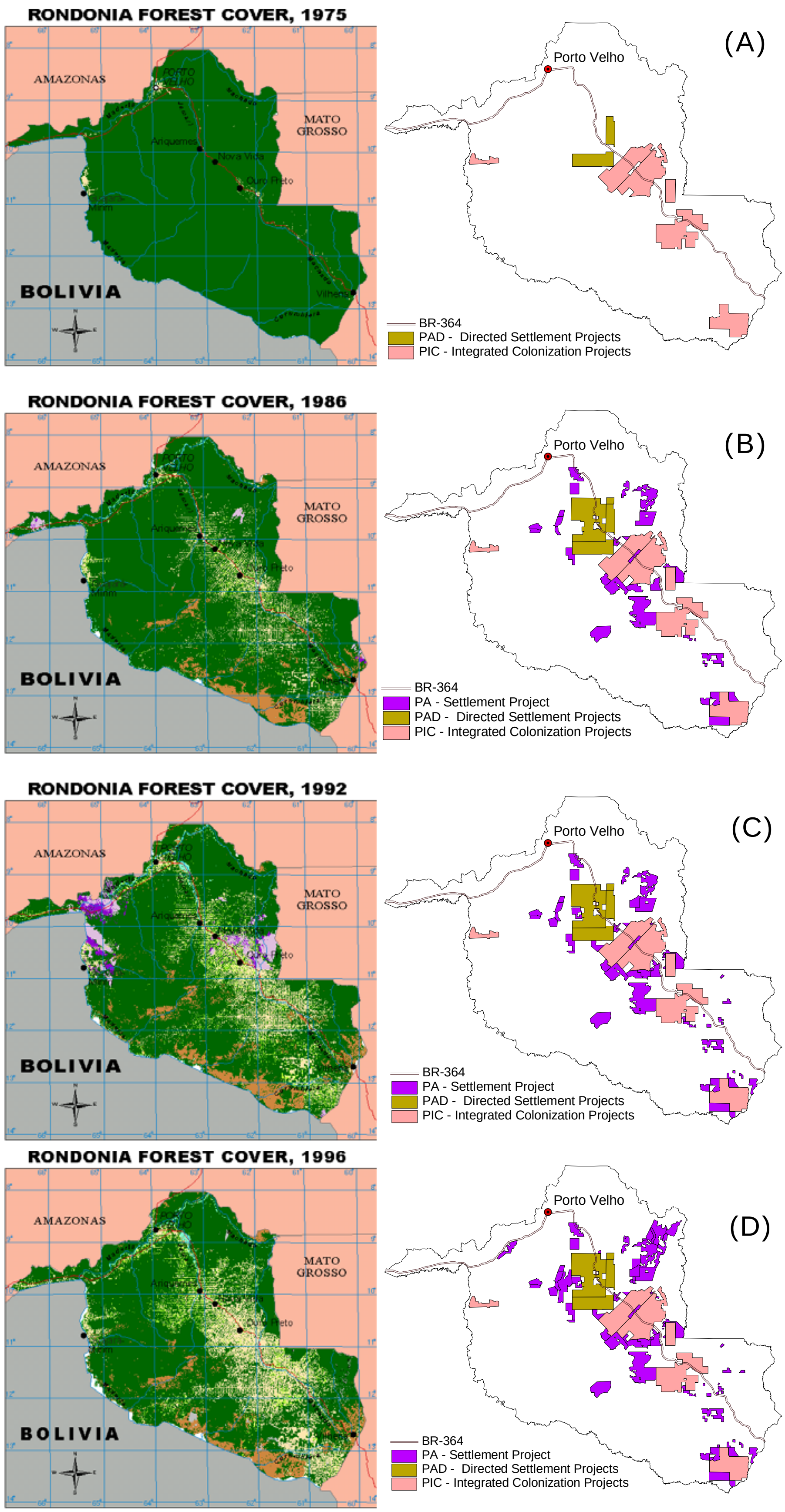
An important issue for the understanding of frontier malaria is the notion of forest fringe: a frontier between the forest and the property, where the risk of malaria transmission is very high. The forest fringe is generated by dramatic ecological changes induced during colonization projects, especially during the early years of occupation. Farmers can be more or less exposed to *A. Darlingi* mosquitoes depending upon how far from the forest fringe their houses are built and even the hours of the day when they are out of doors. Fig. 33 shows the early stages of occupation of a plot in 1986. The settler (wearing shorts and short sleeves), built a temporary house with precarious conditions, and is on the process of clearing the land around the house, so that agriculture can be initiated. At that stage, the forest fringe is in close proximity to the house and to his working place, maximizing his exposure to the risk of being bitten by a mosquito.

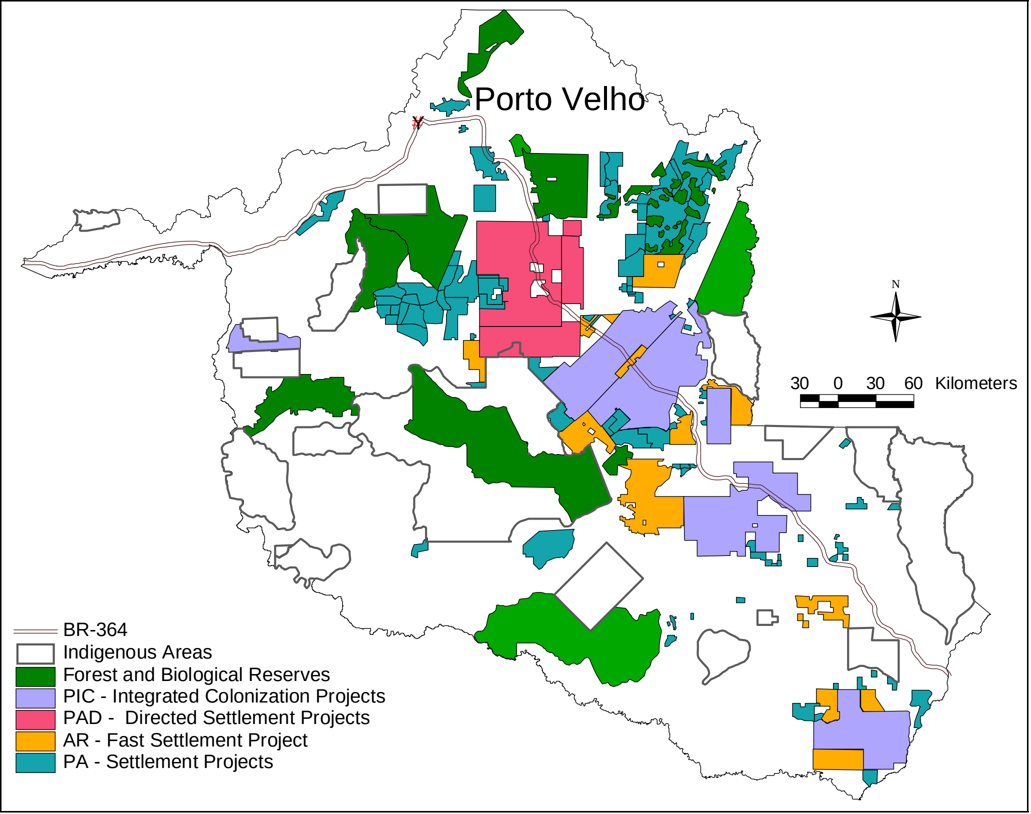


1st phase: Epidemic
2nd phase: Transition
3rd phase: Endemic
3500
3000
2500
2000
1500
1000
500
0
API (by 1,000)
0
3
8
Age of the settlement project (years)

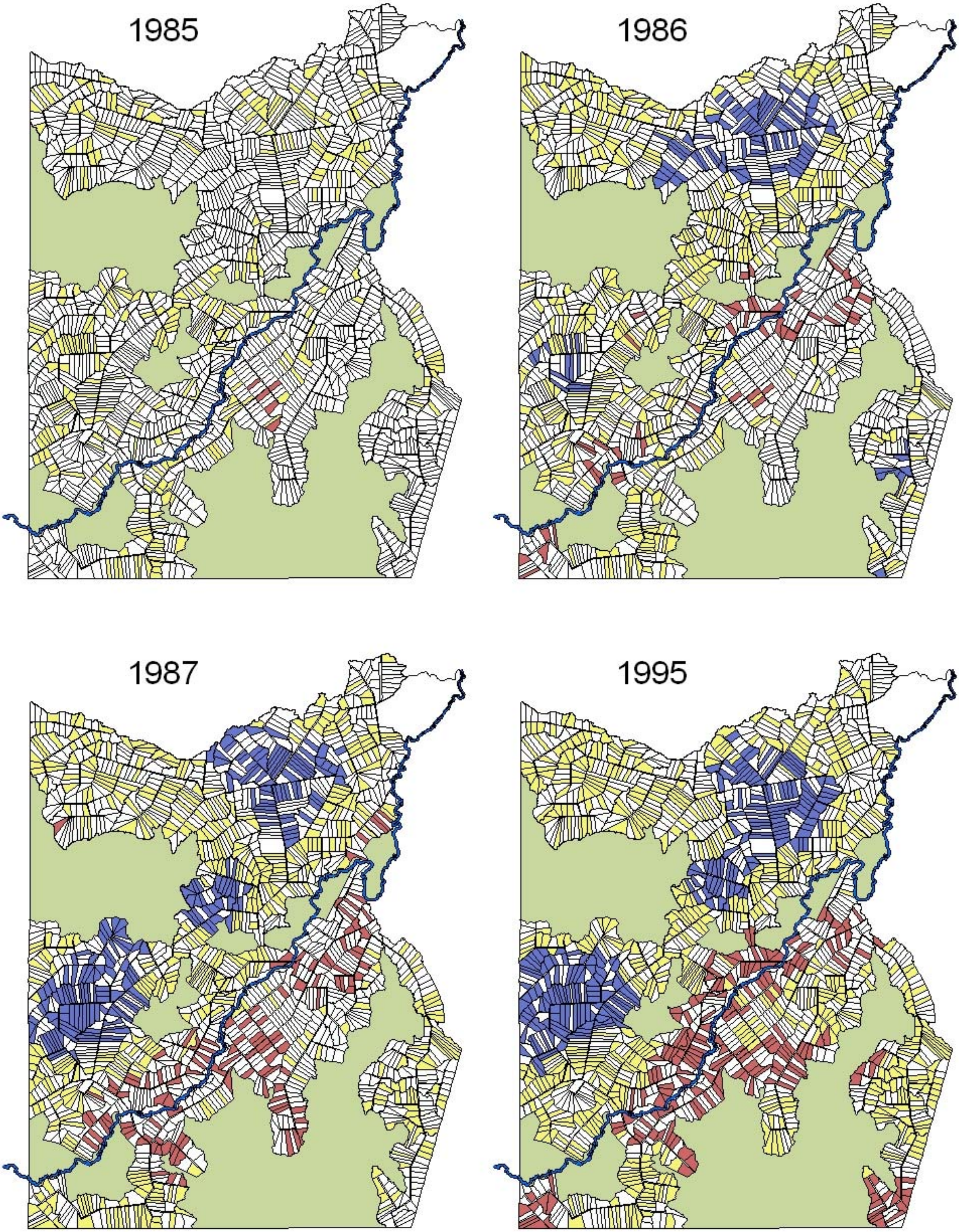


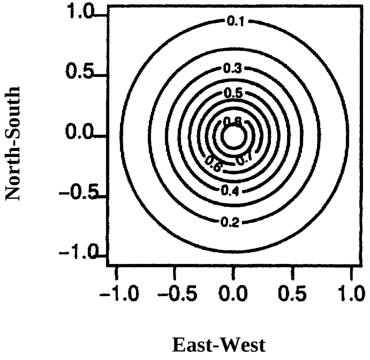
50
40
30
20
10
0
EWR
Minimum
Median
Maximum
Global
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
Year

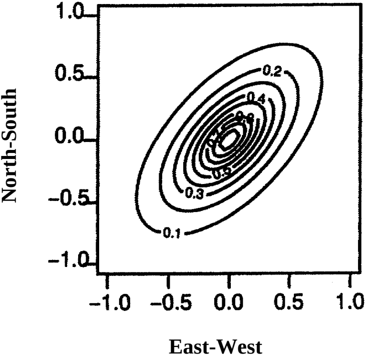

RONDONIA FOREST COVER, 1975
AMAZONAS
MATO GROSSO
BOLIVIA
(A)
Porto Velho
BR-364
PAD - Directed Settlement Projects
PIC - Integrated Colonization Projects
RONDONIA FOREST COVER, 1986
AMAZONAS
MATO GROSSO
BOLIVIA
(B)
Porto Velho
BR-364
PA - Settlement Project
PAD - Directed Settlement Projects
PIC - Integrated Colonization Projects
RONDONIA FOREST COVER, 1992
AMAZONAS
MATO GROSSO
BOLIVIA
(C)
Porto Velho
BR-364
PA - Settlement Project
PAD - Directed Settlement Projects
PIC - Integrated Colonization Projects
RONDONIA FOREST COVER, 1996
AMAZONAS
MATO GROSSO
BOLIVIA
(D)
Porto Velho
BR-364
PA - Settlement Project
PAD - Directed Settlement Projects
PIC - Integrated Colonization Projects



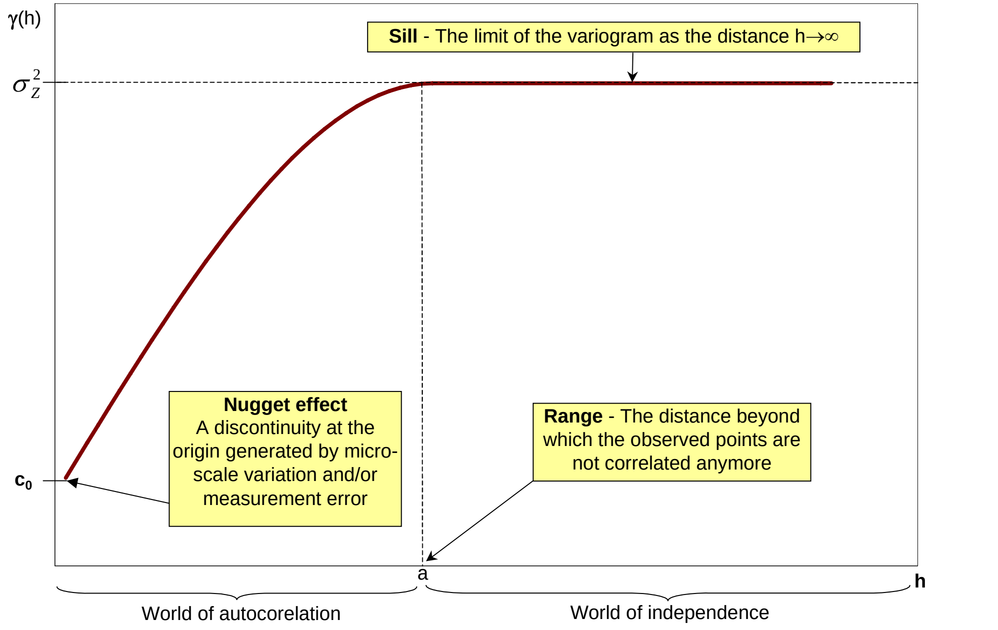
Porto Velho
N
30 0 30 60 Kilometers
BR-364
Indigenous Areas
Forest and Biological Reserves
PIC - Integrated Colonization Projects
PAD - Directed Settlement Projects
AR - Fast Settlement Project
PA - Settlement Projects


1985
1986
1987
1995

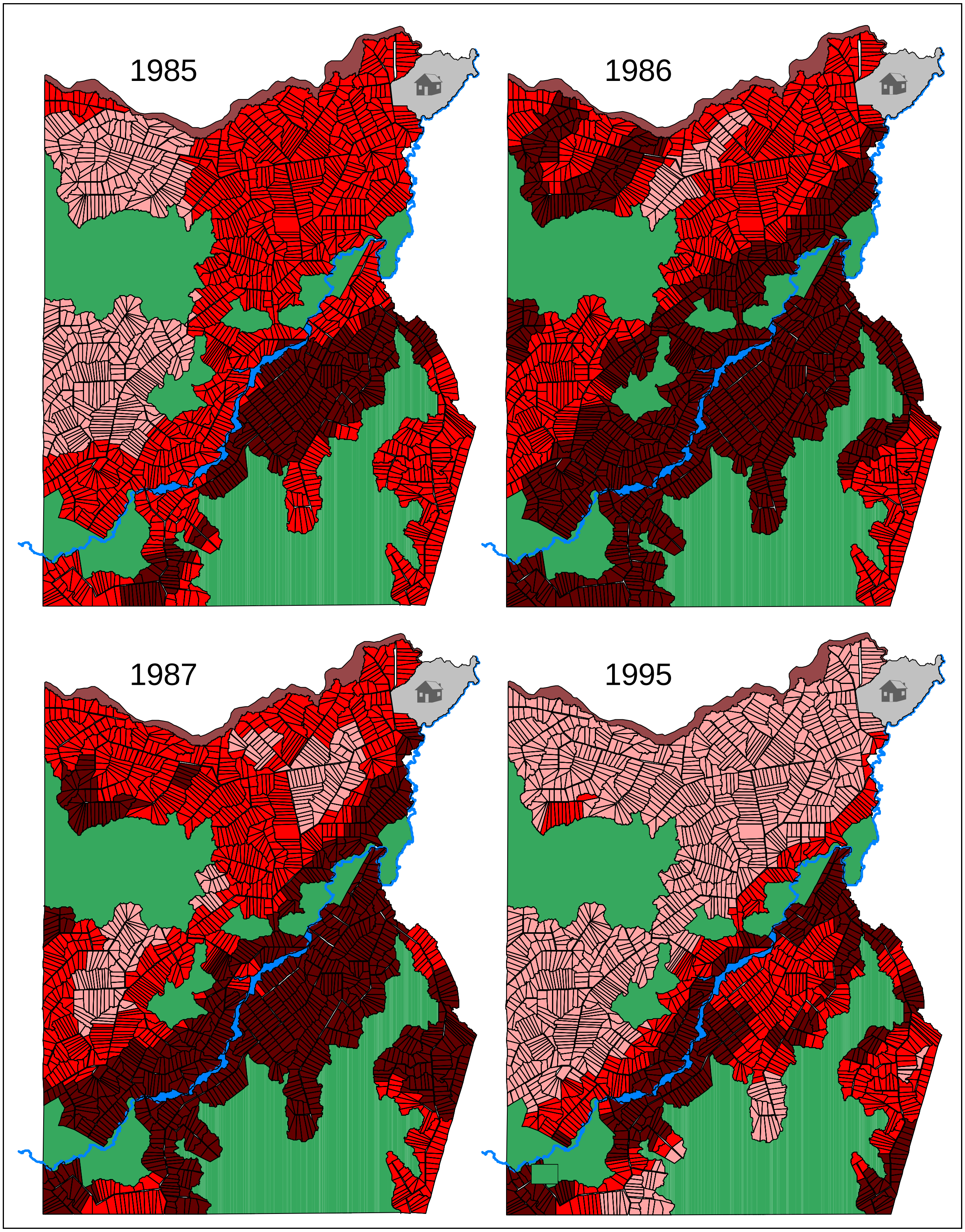

1.0
0.5
0.0
−0.5
−1.0
North-South
0.1
0.3
0.5
0.8
0.7
0.6
0.4
0.2
−1.0
−0.5
0.0
0.5
1.0
East-West

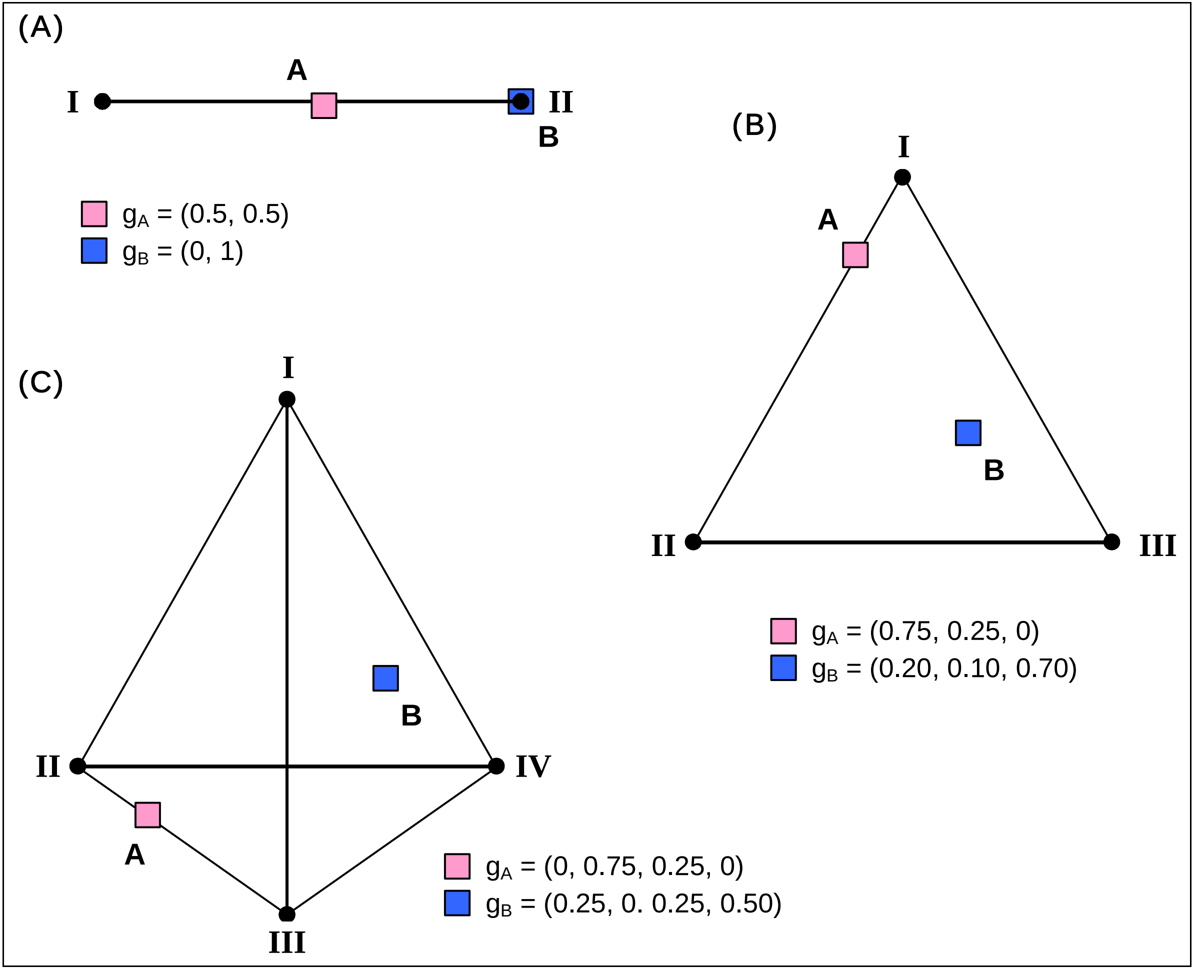


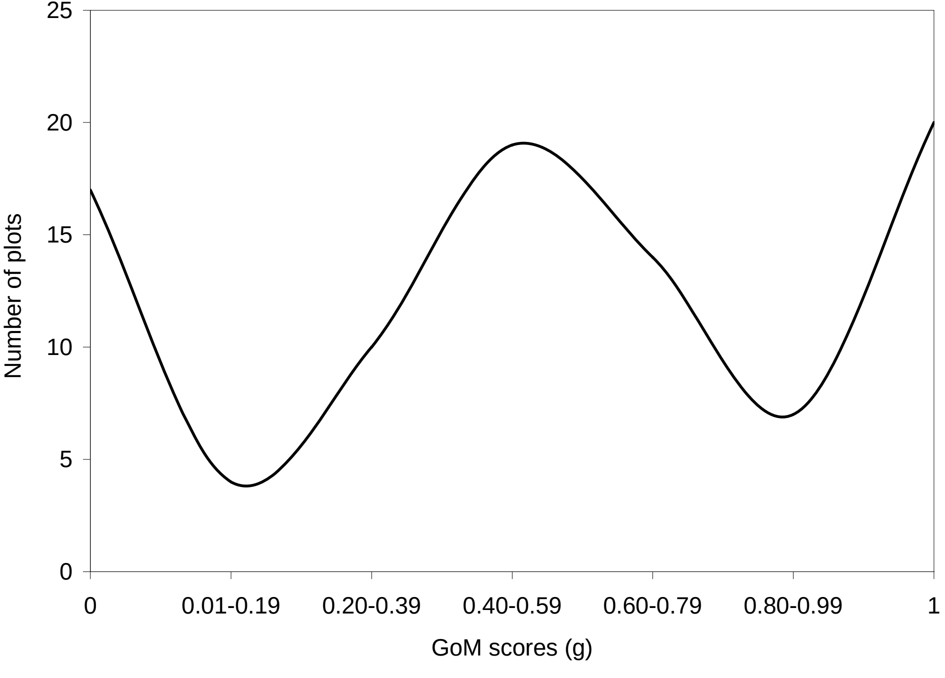
1.0
0.5
0.0
−0.5
−1.0
North-South
0.2
0.4
0.3
0.3
0.1
−1.0
−0.5
0.0
0.5
1.0
East-West

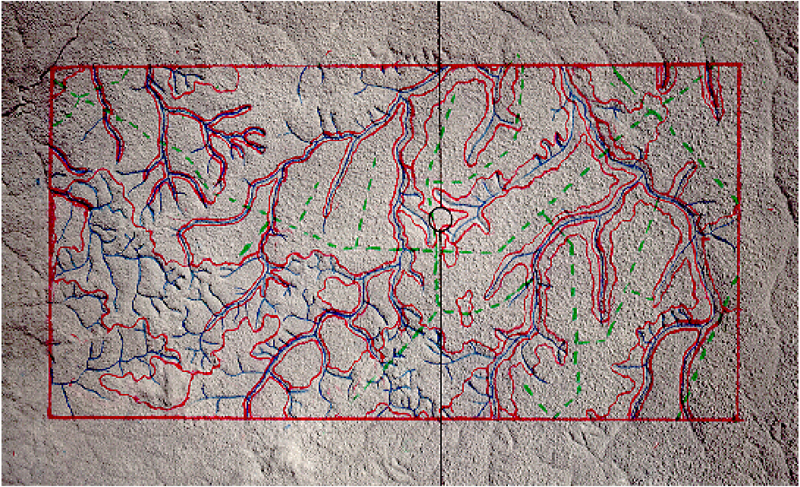


γ(h)
Sill - The limit of the variogram as the distance h→∞
$\sigma_Z^2$
Nugget effect
A discontinuity at the origin generated by micro-scale variation and/or measurement error
$c_0$
Range - The distance beyond which the observed points are not correlated anymore
a
h
World of autocorelation
World of independence

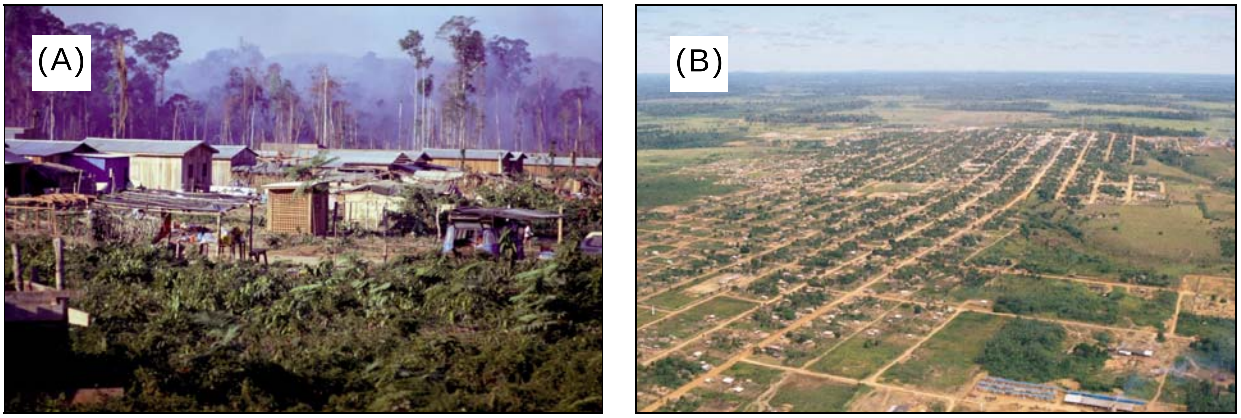

1985
1986
1987
1995


(A)
A
I
II
B
gA = (0.5, 0.5)
gB = (0, 1)
(B)
I
A
B
II
III
gA = (0.75, 0.25, 0)
gB = (0.20, 0.10, 0.70)
(C)
I
B
II
IV
A
III
gA = (0, 0.75, 0.25, 0)
gB = (0.25, 0. 0.25, 0.50)

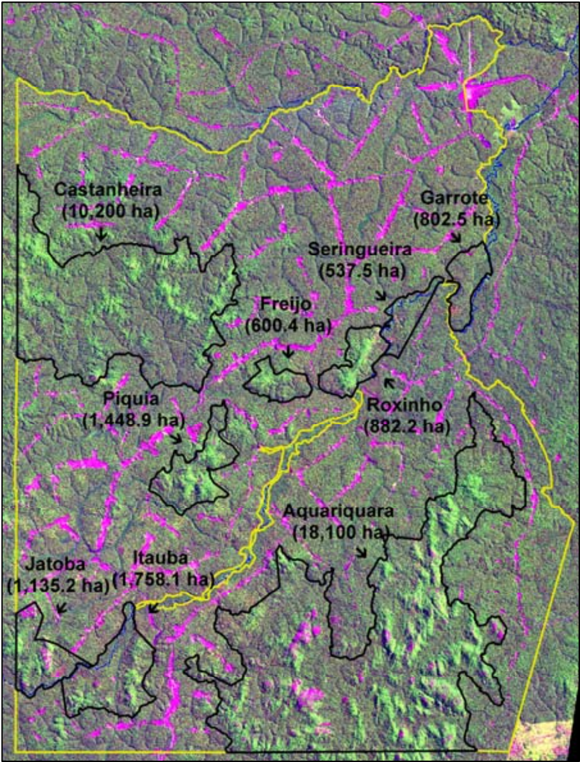

Number of plots
25
20
15
10
5
0
0
0.01-0.19
0.20-0.39
0.40-0.59
0.60-0.79
0.80-0.99
1
GoM scores (g)




(A)
(B)


(A)
(B)


(A)
(B)
(C)



Castanheira
(10,200 ha)
Garrote
(802.5 ha)
Seringueira
(537.5 ha)
Freijo
(600.4 ha)
Piquia
(1,448.9 ha)
Roxinho
(882.2 ha)
Aquariquara
(18,100 ha)
Jatoba
(1,135.2 ha)
Itauba
(1,758.1 ha)