UNIVERSIDADE FEDERAL DE JUIZ DE FORA

FACULDADE DE ECONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA APLICADA

A SPATIAL PROPENSITY SCORE MATCHING TO EVALUATE THE IMPACTS OF GROWING SUGARCANE ON SOCIAL INDICATORS IN PRODUCING REGIONS

André Luiz Squarize Chagas

Rudinei Toneto Júnior

Carlos Roberto Azzoni

TD. 021/2009

Programa de Pos-Graduação em Economia Aplicada - FE/UFJF

Juiz de Fora

2009

A SPATIAL PROPENSITY SCORE MATCHING TO EVALUATE THE IMPACTS OF GROWING SUGARCANE ON SOCIAL INDICATORS IN PRODUCING REGIONS

André Luis Squarize Chagas
IPE-USP/SP
Rudinei Toneto Jr.
FEA-USP/RP
Carlos Roberto Azzoni
FEA-USP/SP

Abstract. The expansion of sugarcane growing in Brazil, spurred particularly by increased demand for ethanol, has triggered the need to evaluate the economic, social and environmental impacts of this process, both on the country as a whole and on the growing regions. Even though the balance of costs and benefits is positive from an overall standpoint, this may not be so in specific producing regions, due to negative externalities. The objective of this paper is to estimate the effect of growing sugarcane on the human development index (HDI) and its sub-indices in cane producing regions. In the literature on matching effects this is interpreted as the effect of the treatment on the treated. Location effects are controlled by spatial econometric techniques, giving rise to the spatial propensity score matching model. We analyze 424 minimum comparable areas (MCAs) in the treatment group, compared with 907 MCAs in the control group. The results suggest that the presence of sugarcane growing in these areas is not relevant to determine their social conditions, whether for better or worse. It is thus likely that public policies, especially those focused directly on improving education, health and income generation/distribution, have much more noticeable effects on the municipal HDI.

Keywords: Spatial propensity score matching, sugarcane.

JEL Classification: C14, C21, Q18.

1. Introduction

The expanding production of sugarcane in Brazil in recent years has prompted the need to assess the economic, social and environmental impacts of this process, both for the country as a whole and for the regions where this has occurred. Doubts that can be raised, for example, concern the quality of employment, environmental impacts (soil contamination, atmospheric pollution from burning fields, water use) and dislocation of other crops to native forests, among others (Noronha et al., 2006).

Even when the balance of costs and benefits of the sector appears positive from the standpoint of the entire country (BNDES; CGEE, 2008), the benefits for cane growing regions may not be as evident. In other words, the producing regions may disproportionately bear the negative impacts of the sector's presence. Perhaps the most obvious aspect in this respect is the labor market. Many studies have analyzed the working conditions in the sector, particularly those encountered in manual harvesting (Alves, 2006, 2007; Baccarin; Alves; Gomes, 2008).

However, consideration must go to the higher value of agricultural output in cane-growing regions. Sugarcane is significantly more valuable by tilled area than many other crops, such as soybeans and corn. On the question of agricultural labor, Toneto-Jr and Liboni (2008) observe that sugarcane cultivation generates more jobs than does soybean growing, and only slightly fewer than corn cultivation. Thus, because it generates more value per hectare and more jobs as well, cane growing generates more income per area planted than other staple crops.

Given the specific aspects of sugarcane, the industrial plants (sugar mills and/or ethanol distilleries) are located near the growing fields. This tends to increase local employment even more, because of the need for industrial workers and services – transport, maintenance, etc. – increasing the sector's indirect effects on the producing region.

Despite the arguments in favor of and against the sector and its production methods, local social conditions can be more closely related to government policies (federal, state and municipal). Hence, the social indicators captured by aggregate indicators – such as life expectancy, level of schooling and per capita income – can depend more on how public funds are used locally, and only indirectly on production methods.

The objective of this chapter is to assess the effects that sugarcane growing has on the social indicators of the producing regions. We chose the municipal human development index (HDI-M) as an indicator that synthesizes the local social conditions (PNUD; IPEA; FJP, 2003). This indicator jointly measures the conditions regarding education, longevity and income, just as the HDI itself does. The HDI is composed of three dimensions: education (literacy rate and school enrollment rate), longevity (life expectancy at birth) and income (per capita GDP). To measure the regional human development index (municipal level here), we consider the same dimensions, but some indicators are different, seeking to adjust the measure to the particularities of smaller social units. To measure access to education, the HDI-M uses school attendance based on census data rather than the enrollment rate, because students can live in one municipality and attend school in another one, which distorts the regional enrollment rates. To measure income, the HDI-M calculates municipal income, again from census data, as the sum of the various per capita income sources (wages, pensions and government transfers, among others) instead of per capita GDP, since the total income generated in a local region is appropriate for its residents (PNUD; IPEA; FJP, 2003).

This article is organized in five sections including this introduction. The next section contains a review of the literature and some considerations on the production of sugarcane and social conditions, with results for recent years. The third section presents the methodology used to identify the possible impacts of growing sugarcane on the social conditions of producing regions. The fourth section presents the results, and the fifth section contains the conclusions.

2. Production of Sugarcane and Social Conditions

The most debated social impact related of growing sugarcane is unquestionably that of the sector's working conditions (Alves, 2006, 2007; Mendonça, 2006a; 2006b; Baccarin; Alves; Gomes, 2008). Alves (2006) calls attention to the extreme physical exertion required of workers in the sector, especially those engaged in manual harvesting. Other studies, however, without analyzing the physical requirements of cane cutting, have found evidence that the wages paid are higher than in other agricultural sectors (Toneto-Jr; Liboni, 2008; Hoffman; Oliveira, 2008).

In a different approach, Piketty, Menezes and Duarte (2008), analyzing the impacts of the sector on the distribution and concentration of labor income in the period from 1992 to 2006, conclude that the sugarcane sector has not played a significant role in reducing poverty and

inequality in the country. Indeed, for the state of São Paulo (Brazil's main cane producing state), the authors conclude that the sector contributes to the concentration of income.

According to Camargo-Jr. and Toneto-Jr. (2008), there is an association between the intensity of the activities of growing sugarcane and producing sugar and alcohol and the performance on socioeconomic indicators. In general, municipalities[1] with strong involvement in the sugar-alcohol sector perform better on socioeconomic indicators, and in some cases even outperform the greater São Paulo Metropolitan Region (SPMR), the state's main region in economic terms.

Silva (2008) also observes, without considering possible cross-effects on other variables, a positive effect on social conditions in municipalities São Paulo state where sugarcane is a main crop. But in considering the fact that the sector's presence can affect local human development through its impact on other variables, he finds that the situation is reversed and the sector's presence has net negative impacts.

The problem of these recent studies is that they treat regions that are different in equal fashion. When considering data aggregated by region, one should consider that they reflect specific conditions of these locales. Thus, the effect that growing sugarcane has on these places should be compared with the situation of the same places if there were no cane growing. Obviously such a direct comparison is impossible. To overcome this drawback, matching methods can be used. These aim to estimate the impact of determined treatments on the treated subjects. In the next section we explain this methodology as applied here.

3. Methodology

3.1. Spatial Propensity Score Matching

Our objective is to estimate the effect of growing sugarcane on the HDI (total and its sub-indices) of the producing regions. In the literature on matching effects this is interpreted as the effect of the treatment on the treated subjects. We let $D_i = 1$ for the group of regions where sugarcane is grown and $D_i = 0$ for the group of regions where it is not. The regions do not have the same probability of belonging to one group or the other. Factors such as location and proximity to a mill/distillery, for example, affect this probability. We call these variables X_i.

Certainly growing cane in a determined place, from the farmer's point of view, can be interpreted as his best response in view of his available choices. And it is very likely that the fact of having other growers nearby can influence the decision process. This fact introduces a selection bias in comparing regions whose set of possibilities are different and hence whose best response (or at least whose observed response) is different. The role of the propensity score is to relax these spatial effects. In other words, the problem's spatial dimension is latent, and the introduction of spatial controls is a necessary precondition for correct identification of the effects of interest.

The propensity score method was introduced by Rosenbaum and Rubin (1983). Their method controls for the selection bias of different individuals receiving the treatment by estimating the probabilities of receiving treatment, given some observed variables. This probability, $\Pr(D_i = 1 \mid X_i)$ is called the propensity score. Individuals with similar probabilities of receiving the treatment are grouped, so that the result is conditionally independent of whether or not the individual received treatment, or

$$(Y_0, Y_1) \perp D \mid X \quad , \tag{1}$$

where Y is the result of interest, D is the treatment, and $D \in (0,1)$ and X are covariates. The aim is to estimate the average effect of the treatment on the treated, that is

$$E[(Y_0, Y_1) \mid D = 1, X] = E[(Y_1 \mid D = 1, X) - (Y_0 \mid D = 0, X)] \quad . \tag{2}$$

[1] The local governing unit in Brazil is the municipality, which is similar to a county in the United States, except it is governed by a single mayor and municipal council.

The value of the counterfactual effect of no treatment on the treated, $E(Y_0 \mid D=1, X)$, is approximated by the average result of the self-selected group of untreated individuals $E(Y_0 \mid D=0, X)$ (Heckman; Ichimura; Todd, 1998). Instead of using various conditional covariates, we use the propensity score P(X) = Pr(D = 1 | X), that is, the probability of belonging to the group of cane growing regions, given determined observed characteristics.

The probability of belonging or not to the group of producing regions is not a random variable. Spatial factors interfere in this choice, such as climate, quality and land availability, among others. These locational factors can be controlled by the proximity to other producing regions.

Moreover, according to Heckman, Lalonde and Smith (1999), an additional condition for the use of the propensity scoring is the existence of a common support, i.e., that there exist units in both the treatment and control groups for each characteristic X for which comparison is desired. The condition that $0 < P(X) < 1$ assures that for each treated individual there is another matched untreated one, with similar values of X.

The estimation of P(X) = Pr(D = 1 | X) is done by means of a probit or logit model. However, when there are lagged or spatial effects, conventional models, calculated by maximum likelihood, are not adequate. By construction, the errors of a spatial logit model are heteroskedastic, and estimates based on the hypothesis of homoskedasticity in the presence of heteroskedastic errors are inconsistent (Greene, 2000; Wooldridge, 2001).

The general model, considering spatial lags in the dependent variable and the residuals, called the spatial autocorrelation (SAC) model (Lesage, 1999; Chagas, 2004), can be described in the following form:

$$
\begin{aligned}
&y = \rho \mathbf{W}_1 y + \mathbf{X}'\beta + u \\
&u = \lambda \mathbf{W}_2 u + e \\
&e \sim N(0, \sigma^2 \mathbf{V}) \\
&V = \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & v_n \end{bmatrix}
\end{aligned}
\qquad (3)
$$

where y is a dummy variable that assumes values of 1 and 0, $\mathbf{X}$ are covariates, $\mathbf{W}_1$ and $\mathbf{W}_2$ are variance matrices that control for the effects of the spatial lag; and v_i, i = 1, ..., n are parameters (associated with the heteroskedasticity[2]) to be estimated, which capture the model's heteroskedasticity. The parameters ρ and λ are, respectively, the effects of the spatial autocorrelation and of the spatial correlation of the residuals. If $\mathbf{W}_1$ and $\mathbf{W}_2$ are the same, it is possible to estimate this general model, but its identification is problematic (LeSage, 1999).

Alternatively, a less general model can be estimated, considering only the spatial autocorrelation, called the spatial autoregression (SAR) model:

$$y = \rho \mathbf{W} y + \mathbf{X}'\beta + e. \qquad (4)$$

Another possibility is the spatial error model (SEM), which considers the spatial effect only in the residuals:

$$
\begin{aligned}
&y = \mathbf{X}'\beta + u \\
&u = \lambda \mathbf{W} u + e.
\end{aligned}
\qquad (5)
$$

[2] Since the dependent variable of a probit model (y) assumes the values 0 or 1, the errors of a spatial autocorrelation model, for example, take on values $-\rho Wy - X\beta$ when y = 0, and $\iota - \rho Wy - X\beta$, when y = 1. The error term depends on a parameters vector (β) and a constant (ρ), which induces heteroskedasticity (Wooldridge, 2002, p. 470).

A strategy to choose among these models is first to estimate the most general one (SAC). If the coefficients of the two spatial effects are accepted, this is the best model among the three. If not, the model is estimated associated with the significant knowledge from the previous step.

In the form specified, the models have many more parameters to be estimated than degrees of freedom, preventing the use of the usual techniques. LeSage (1999, 2000) introduced Bayesian estimates, employing techniques based on Monte Carlo Markov chains (MCMC) by means of Gibbs and Metropolis-Hastings sampling.

The basic idea of the Monte Carlo method is to characterize the joint (posterior) distribution of the quantities of interest (parameters), and using modern computational techniques, simply to generate a sample of the distribution (taking selections randomly) and calculate the statistics from this sample. With a sufficiently large number of draws, the statistics can approximate the population parameters. Since the initial draws are performed based on an initial (prior) estimate, Franzese Jr. and Hays (2007) suggest that 5,000 to 10,000 draws be taken, and to discard the first 1,000 (called burn-in)[3].

Another model selection criterion arises from this procedure. At each step of the simulation, the cases are recorded when ρ and λ lie within the acceptance interval (-1 to 1). If this rate is very low, the model might be misspecified.

3.2. Kernel Matching

The effect of the treatment on the treated is calculated by comparing the performance of the treated group (denoted by Y_1, indexed by I_1) with that of the untreated group (denoted by Y_0, indexed by I_0), through the following equation (Heckman; Ichimura; Todd, 1998):

$$E[(Y_1 - Y_0) \mid D = 1, P(X)] = \frac{1}{N_1} \sum_{i \in I_1} [Y_{1i} - \sum_{i \in I_0} W_{N_0 N_1}(i, j) Y_{0j}] \quad (6)$$

where $W_{N_0 N_1}(i, j)$ is usually a matrix of positive weights, defined so that each $i \in I_1$, $\sum_{i \in I_0} W_{N_0 N_1}(i, j) = 1$, and N_0 and N_1 are the numbers of observations in I_0 and I_1, respectively.

A kernel estimator is used to choose the weights so that the observations that are nearer in terms of their distances measured by $|P(X_i) - P(X_j)|$ receive greater weight. This weighting is given by a kernel function. This function must integrate to one and be continuous and symmetric about the origin (Härdle; Linton, 1994).

$$K(u) = K(-u) = \int_{-1}^{1} K(u) du = 1 \quad (7)$$

A frequently used functional form is the "biweight" (or quartic), expressed by

$$K(u) = \frac{15}{16}(u^2 - 1)^2 \text{ for} \mid u \mid < 1$$
$$= 0 \text{ other case} \quad (8)$$

where $u = \frac{P(X_i) - P(X_j)}{h}$.

Implementing the estimation via a kernel function requires choosing a suitable bandwidth (h). The smaller h is, the less weight is given to larger distances and the greater the weight given to more proximate observations. The consistency of nonparametric estimators requires the bandwidth to approach zero as the sample size increases, but not necessarily at the same speed (Todd, 1999).

The approximation of the score distribution, by means of the kernel function, is

[3] In our spatial propensity score estimates, we used 10,000 drawings and discarded the first 1,000.

$$\hat{f}_h(P(X)) = \frac{1}{n}\sum_{i=1}^{n} K_h(P(X) - P(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_h\left(\frac{P(X) - P(X_i)}{h}\right) \quad (9)$$

Sensitivity tests are performed to check the sensitivity of the results to the choice of the bandwidth *h*. Following Härdle and Linton (1994), here we consider a bandwidth of 0.2 as a base.

3.3. Considerations on Bias

Let *V* be the bias in the estimate of the treatment effect on the treated, defined as

$$V = E[(Y_0 \mid D = 1, P(X)) - (Y_0 \mid D = 0, P(X))], \quad (10)$$

that is, the bias that comes from utilizing the average results of the comparison group as a proxy for the average results of the participants in the program if they had not participated. According to Heckman, Ichimura and Todd (1997), the bias can be divided into three basic components: the first component arises from the lack of a common support; the second comes from unobservable errors; and the third is due to differences in the results that remain even after taking into consideration the observable characteristics and performing comparisons in a common support region. This last component is due to the differences in the unobservables, known as selection bias. This bias arises when for given values of *X* there is a systematic relation between participation in the program and the results, i.e., there are unobserved variables that jointly influence the results and participation in the program, conditional on the observable variables. To deal with this bias, the best way would be to allocate subjects to the program at random, because in this way one can guarantee that participants and nonparticipants would have the same expected outcome without the program.

We should remark that matching methods (as is the case of the propensity score) only eliminate two of the three sources of bias. The first type is eliminated by the matching within a region with common support. The careful matching of the comparison group, based on observable characteristics, eliminates the second bias component. However, matching methods only deal with observable characteristics, leaving the problem of latent heterogeneity, which causes a possible bias in estimating the program's impact.

So, the propensity score method permits reducing, but not eliminating, the bias caused by unobservable factors. How much the bias is reduced crucially depends on the richness and quality of the control variables used to compute the propensity score and carry out the matching (Becker; Ichino, 2002).

3.4. Sensitivity Analysis Utilizing the Rosenbaum Bounds Method

As previously mentioned, unobserved factors can bias the estimates of the treatment effect on the treated when using the propensity score method. Since it is impossible to estimate the magnitude of the selection bias in non-experimental data, an important tool can be employed to evaluate it in estimating the causal effects. This method is called bound analysis. It evaluates the potential impact of the selection bias that arises due to unobserved variables. Here we use the method known as Rosenbaum bounds (Rosenbaum, 2002; Diprete; Gangl, 2004), whose idea is to estimate the influence of a possible omitted variable on the selection bias existing in the probability of participation in the treatment, which can possibly impair the conclusions on the causal effects[4].

Sensitivity analysis can be used to test the robustness of the results to the presence of bias due to an omitted covariable. Here this analysis aims to evaluate the effect of a possibly omitted variable on the decision to grow sugarcane, as well as the results on the social indicators that can change our conclusions.

The probability of participation (growing cane) of an individual *i* is given by[5]:

$$\pi_i = \Pr(D_i = 1 \mid X_i) = F(\beta X_i + \gamma u) \quad (11)$$

---

[4] For a more complete discussion, see Rosenbaum (2002)

[5] Resende and Oliveira (2008).

where $D_i = 1$ if the individual receives the treatment; X_i are observed characteristics of individual i; u_i corresponds to the unobserved variable; and γ represents the effect of u_i on the decision to grow sugarcane. If there is no selection bias, then γ will be zero and the probability of growing cane will be exclusively determined by the observable characteristics. However, in the presence of selection bias, two regions with the same observable covariables, X, will have different probabilities of participating in the production.

Assuming that two individuals, i and j, are matched and that F has a logistic distribution, the relative probability (odds) of the regions' receiving treatment is given be:

$$\frac{\pi_i}{1-\pi_i} \quad \text{and} \quad \frac{\pi_j}{1-\pi_j} \tag{12}$$

and the odds ratio is given by:

$$\frac{\dfrac{\pi_i}{1-\pi_i}}{\dfrac{\pi_j}{1-\pi_j}} = \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} = \frac{\exp(\beta X_j + \gamma u_j)}{\exp(\beta X_i + \gamma u_i)} = \exp[\gamma(u_i - u_j)] \tag{13}$$

If the regions have the same observable characteristics, then the βX terms cancel each other. Therefore, if there are no differences in the unobserved variables ($u_i = u_j$) and these variables do not influence the probability of participation ($\gamma = 0$), the odds ratio will be equal to 1, implying there is no selection bias. It follows, then, that if their odds of participation differ – that is, if the odds ratio is different from 1 – any selection bias can only be due to the presence of unobservable factors. The sensitivity analysis evaluates how much the program's effect is altered by the change in the values of γ and $u_i - u_j$.

In practice this means examining the bounds of the odds ratio of participation. Rosenbaum (2002) showed that (2.13) if bounded as follows:

$$\frac{1}{e^{\gamma}} \leq \frac{\pi_i(1-\pi_j)}{\pi_j(1-\pi_i)} \leq e^{\gamma} \tag{14}$$

The matched regions have the same probability of participation only if $e^{\gamma} = 1$. However, if $e^{\gamma} = 2$, then regions apparently similar in terms of X will differ in their probabilities of receiving treatment by a factor of up to 2. Hence, according to Rosenbaum (2002), e^{γ} is a measure of the degree of departure from a study free of selection bias.

3.5. Database

The data on sugarcane production comes from the Municipal Agricultural Survey (*Pesquisa Agrícola Municipal* - PAM), conducted by the Brazilian Institute of Geography and Statistics (IBGE, the official census bureau). Since during the past decade new municipalities have been created by splitting off from existing ones, we grouped them into minimum comparable areas (MCA). There are a total of 4,248 MCAs in the database. We chose as treated the regions in which the area planted with sugarcane represented an average of 30% of the area farmed during the 1990s. These regions correspond to about 10% of Brazilian MCAs (424 MCAs). We performed robustness tests also for 5%, 15% and 20% of the MCAs[6].

To compose the control group we considered only the MCAs with no sugarcane production in any year in the period from 1991 to 2000. This criterion thus excluded the MCAs with some production, but less than the cutoff point (30% of tilled area for the base group). Because of the very

[6] These percentages correspond, respectively, to 65%, 15% and 9% of the total agricultural area dedicated to growing sugarcane, i.e., 213, 644 and 846 MCAs, in each case in the restricted sample.

small sugarcane production in MCAs in the country's North region and to avoid undue selection bias from these areas in the control group, we excluded the MCAs from this region from the sample. The final control group contains 907 MCAs.

To calculate the propensity score, we considered, besides the neighborhood spatial effects, the proximity of the MCA to the center of an MCA with a sugar mill and/or alcohol distillery (just called mills hereafter) and a dummy for those MCAs located in states with the highest densities of producing MCAs (eight states: Alagoas, Espírito Santo, Mato Grosso, Minas Gerais, Paraíba, Pernambuco, Rio de Janeiro and São Paulo). The data on MCAs with mills were obtained from the Ministry of Agriculture[7]. The density of producing states was calculated based on the same sugarcane production data mentioned above.

We calculated the HDI for the MCAs considering the same variables and methodologies employed to calculate the HDI-M (PNUD; IPEA; FJP, 2003), both for the general index and its components, using data from the IPEA database (IPEADATA), the same source used to calculate the official HDI-M.

To construct the neighborhood matrix, we employed the notion of neighborhood by means of an inverse distance matrix (Anselin, 1998; Chagas, 2004). The distances were obtained from the official geographic coordinates of the center of each municipality (latitude and longitude). For the MCAs (aggregations of municipalities), this was the average latitude and longitude, weighted by the average population of each member municipality between 1991 and 2000. Unlike the usual practice, we used geodesic rather than Euclidian distance. By this criterion, more distant neighboring places receive a lower weight than in the case of Euclidian distance (but greater than in the case of neighborhood matrices that consider only places that share borders). This criterion is more suitable since sugarcane production tends to be influenced by the proximity of the areas and not necessarily by contiguity.

We calculated the neighborhood matrix considering a neighborhood radius of 150 kilometers. In the robustness tests we considered radii of 100 and 200 kilometers[8].

4. Results

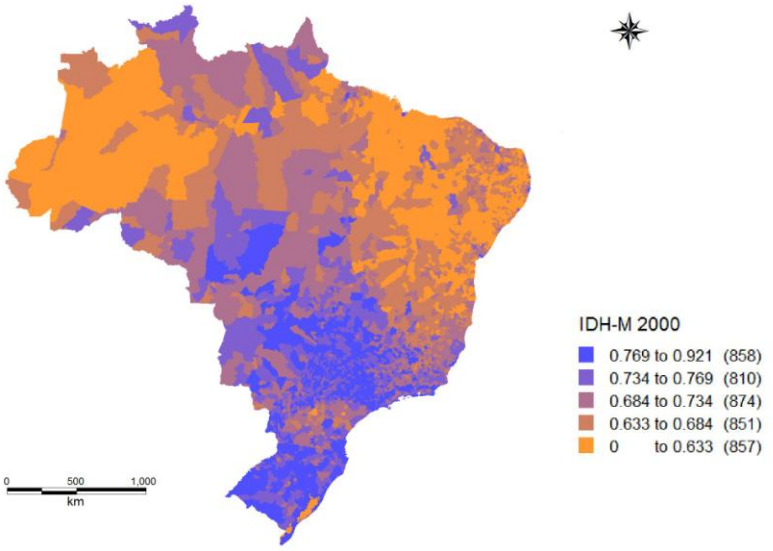

Figure 1 shows the geographic dispersion of the HDI-M for 2000 for Brazilian MCAs. It can be seen that there is a strong spatial concentration of higher HDI-M scores in the South and Southeast regions, while the North and Northeast regions concentrate MCAs with lower HDI-M levels. In the Midwest region, the HDI-M levels are generally in the intermediate range, with some having high indicators.

FIGURE 1 HERE

[7] Ministry of Agriculture: <www.agricultura.com.br>. Accessed on March 15, 2009.

[8] The area supplying a typical sugar mill and/or alcohol distillery is generally small, which makes growing cane in areas very far from them uneconomic. This justifies our working here only with limited radii.

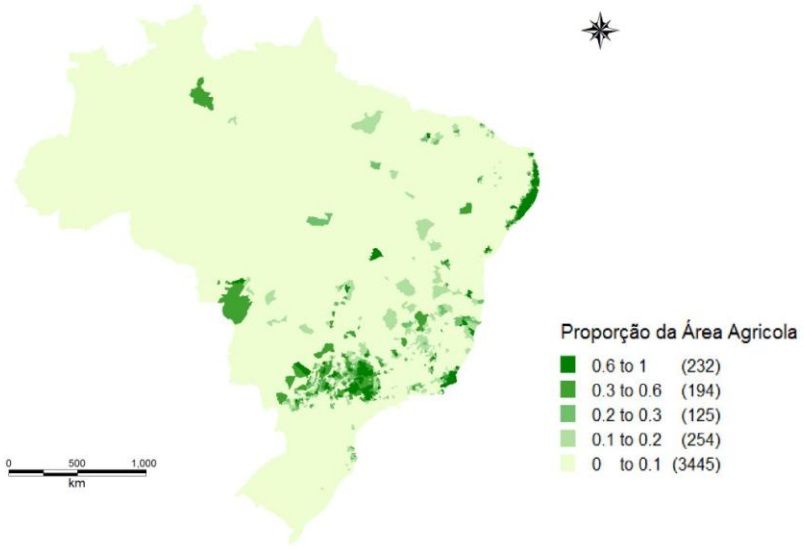

Figure 2 shows that the Southeast, especially the state of São Paulo, is the country's main producing region. Note also that the MCAs in the Midwest and Northeast, along with some in states in the South and Southeast (e.g., Paraná, Minas Gerais and Espírito Santo) also have a considerable concentration of cane growing in relation to total agricultural output.

FIGURE 2 HERE

As mentioned before, the MCAs in the North region do not have significant cane production, and some regions stand out as well due to the small participation of agricultural in general (making a small plantation for production of cane liquor, for example, appear large in proportional terms).

A comparison of the two figures shows an apparent relation between the production of sugarcane and the HDI-M, except for MCAs in the Northeast. Indeed, a comparison of the producing and non-producing regions shows that the HDI-M of the producing MCAs is always greater (on average) than that of non-producing regions. Table 1 details this fact.

TABLE 1 HERE

In the base treatment group (MCAs where sugarcane represents 30% of total agricultural output), the average HDI-M is 0.028 point higher than in the control group (MCAs where no sugarcane is grown). This difference is even greater in the case of the HDI-Income (0.04 point), while the smallest difference occurs for the HDI-Education (0.01 point). No caso do HDI-Longevity, the difference is positive at 0.034 point.

When we used other classifications for the treatment group, the qualitative results were not significantly different. It was higher when including regions where cane production represents less than 30% of total agricultural output. The greatest difference occurred when we considered the 646 producing regions with less than 15% of tilled area given over to cane (0.031 point difference in the HDI-General). The smallest difference was when only considering the 213 MCAs where 65% or more of the tilled area is used to grow sugarcane. In this group there was also a negative difference between the treated and control groups for the HDI-Education.

Restricting the sample by excluding the MCAs from the North region did not produce any significant qualitative differences in the treated group, while the average HDI-M value for the control group increased. This result is further justification for the restriction of the sample adopted (Table 2).

TABLE 2 HERE

Although the HDI-M of the cane producing regions is on average higher than that in non-producing regions, giving support to the results seen previously on the sector's positive impacts on local social conditions, this is not a correct comparison, since it considers different regions as equal. The ideal would be to compare the producing regions against themselves if they did not produce

cane. Since this is not possible, we propose to use the spatial propensity score to compare producing and non-producing regions that are most "similar" (in terms of probability of production).

4.1. Spatial Propensity Scoring

To calculate the spatial propensity score, we considered the neighborhood spatial effects, the proximity of the center of one MCA with that of the nearest MCA with a mill and a dummy for those MCAs located in states dense in producing MCAs. The spatial effects capture both the fact that the an MCA with neighboring MCAs that are producers is more likely to produce cane itself (dependence or spatial autocorrelation) and the soil and climate specificities of each region, for example, controlled by the spatial dependence specification in the error term.

The purpose of the second aspect is to control for the probability that production will occur in regions near mills (potential buyers of output). The last variable, in turn, captures the effects related to possible specific attractions of the state, such as legislation, ease of transporting production, access to tax incentives, etc. It is thus necessary to take care in analyzing the result of the estimated coefficients, due to the potential colinearity between the spatial neighborhood effects and the other variables.

The omission of other covariates is justified on two counts. First, observable factors that are important to explain the HDI-M results of a given place may not be so significant to explain the production of sugarcane. Therefore, in this study our approach for estimation of the propensity score is more parsimonious from the outset, by restricting as much as possible the number of independent variables. The second reason is the application of the Rosenbaum bounds sensitivity tests, which verifies the impact a possible omitted variable would have on the identification of the treatment effect on the treated. If the effect is very small, our strategy is more suitable, since the model is more parsimonious.

Table 3 presents the result of estimating the spatial propensity score by means of spatial logit estimation, as suggested by LeSage (1999).

TABLE 3 HERE

The general model (SAC model) produced weaker results than the SAR model. The model's fit (pseudo-R^2) is lower – 26.5% in the case of the SAC and 35.7% in the SAR case. The efficiency is also lower – the variance (σ^2) is greater in the first case than in the second. Besides this, the acceptance rate (particularly the coefficient associated with the spatial errors component) is very low (fewer than half the cases). Finally, there is the counterintuitive result that the spatial errors are negative. For these reasons, the SAR model is more indicated than the SEM.

Note that both models produce the result (also counterintuitive) that proximity to mills reduces the probability of producing sugarcane. This parameter should also be interpreted with caution, given the existence of multicollinearity among the variables. Conditional on the fact that neighboring regions are cane producers and that the MCA is located in a densely producing state, the fact of having a mill or not is less important.

4.2. Treatment Effect on the Treated

Table 4 reports the analysis of sugarcane production on the HDI and its components of producing regions[9]. As suggested by Härdle and Linton (1994), we used a bandwidth of 0.2 for the kernel function.

TABLE 4 HERE

Unlike seen before, on considering the most similar MCAs to producing ones – given by the propensity scores – the effect of the treatment on the treated appears to be negative, except for the HDI-Education. The greatest difference (-0.012 point) occurs for the HDI-Income, coincidentally this indicator that appeared to have the most positive effect in the analysis without matching. The other effects are also far lower.

[9] The estimates were processed according to Leuven and Sianesi (2003).

However, considering the test statistics, all the effects are statistically insignificant, even at the 10% level. In other words, there are no statistically significant differences between the average HDI values of the producing and control regions.

This result suggests that the sector's presence in a given area is not significant to determine the local social conditions, for better or worse. Possibly public policies, particularly to improve education and health conditions, as well as to improve production and income distribution, have more evident impacts on the HDI-M.

4.3. Sensitivity and Robustness Tests

Table 5 reports the sensitivity analysis of the models, using Rosenbaum bounds. The purpose is to test the selection bias necessary to invalidate the results of the estimates. As formulated by Diprete and Gangl (2004), the method starts with estimating the effect of the treatment on the treated, assuming the hypothesis of no selection bias. Then this assumption is relaxed. According to the potential impact of the omitted variable on the probability of the cane-producing region (expressed in terms of the odds ratio) becoming stronger, the confidence interval of the estimated effects increases, and the level of significance of the null hypothesis – that *D* does not affect Y – diminishes (that is, the *p*-value falls).

TABLE 5 HERE

For the HDI, the critical level of Γ ranges from 1.1 to 1.2, that is, if the presence of the unobserved variables leads to a difference in the odds ratio of receiving treatment between producing and control regions by a factor of from 1.1 to 1.2, then this casts doubt on the previous result that there is a significant impact of the sector on the HDI. The result for the HDI-Longevity is also robust to the presence of selection bias. In this case, the critical level is greater than 1.1 and can reach 1.3. However, the results for the HDI-Education and HDI-Income appear to be less robust to the presence of unobservable factors, given that their critical values are nearer one.

According to Diprete and Gangl (2004), it is important to note that these results are considered the "worst scenarios". The range from 1.1 to 1.2, for example, for the HDI does not necessarily suggest the sector has a negative effect on this index. It only implies that the confidence interval for the effect of the treatment will no longer include zero if the presence of selection bias causes the odds of participation to differ between the treatment and control groups by a factor of 1.2. If an omitted variable has a strong influence on the probability of production, but only a weak influence on the outcome variable, the confidence interval will still contain zero. Although Rosenbaum sensitivity analysis presents results for the worst-case scenario, it shows how large the influence of an unobserved variable must be to cast doubt on the conclusions obtained by matching methods.

As a further measure of the quality of the reported estimates, we implemented robustness tests considering different measures for the bandwidth used in the kernel and the neighborhood radius for the spatial weights matrix, along with different criteria for inclusion in the treatment group. The results are reported in Table 6.

TABLE 6 HERE

An upward or downward variation of the bandwidth of 50% (bandwidths of 0.15 and 0.25, respectively) does not produce a statistically significant difference. Reductions increase the test statistic (in absolute value), but not to the point of making it significant. With a larger bandwidth, the estimated effects on the HDI become positive, but still not statistically significant.

Variations in the neighborhood radius for calculating the spatial weights also do not affect the results. For the case of the smallest radius tested (100 kilometers), the test statistic is higher (in absolute value), but still not significant. For the largest radius (200 kilometers), the results are very similar to those obtained with the basic hypotheses.

Finally, the results are more sensitive with variation of the definition of the treatment group. For regions where sugarcane production represents over 65% of the farmed area, the sector appears to have a negative impact on the HDI, particularly from the negative effect on HDI-Income. However, the number of treated MCAs in this case is only 213. Since these regions are mainly agricultural ones, one can question the need in this case to exclude non-agricultural regions from the

control group (or include variables that control for this fact in the calculation of the propensity score). For the other classifications, however, the results are very similar to those of the basic scenario.

5. Conclusion

The expanding production of sugarcane in Brazil in the recent past has prompted the need to evaluate the sector's economic, social and environmental impacts on the country in general and on producing regions in particular. The benefits for the producing regions may not be as evident as for the country at large, because the negative impacts might outweigh the benefits in these regions.

In this paper we sought to verify the impact of growing sugarcane on the social indicators of producing municipalities. We chose the municipal human development index (HDI-M) as the metric that synthesizes local social conditions, along with its sub-indicators for education, longevity and income.

We implemented a spatial propensity score matching test, an original contribution to this type of study. This methodology is useful because it deals with the fact that one cannot immediately compare average indicators of cane producing regions with those of non-producing ones, since the probability of production is not a random variable. Thus, spatial factors need to be considered to control for the probability of producing or not.

Although there are arguments in favor and against the sector's impacts on local social conditions in growing regions, our results suggest that the presence of the sector in these places is not relevant to determine their social conditions, for better or worse. Possibly public policies, especially those aimed directly at improving education, health, production and income distribution, have greater impacts on the HDI-M.

REFERENCES

ALVES, F. J. C. Por que morrem os cortadores de cana? **Saúde e Sociedade**, vol.15 no. 3. São Paulo, pp. 90-98, Sept./Dec. 2006.

_____. Migração de trabalhadores rurais do Maranhão e Piauí para o corte de cana em São Paulo - será este um fenômeno casual ou recorrente da estratégia empresarial do Complexo Agroindustrial Canavieiro? In NOVAES, R.; ALVES, F. J. C. (Orgs.). **Migrantes: trabalho e trabalhadores no complexo agroindustrial canavieiro** - os heróis do agronegócio brasileiro. São Carlos: EDUFSCar, 2007, pp. 21-54.

ANSELIN, L. **Spatial Econometrics**: Methods and models. Dordecht: Kluwer, 1988.

BACCARIN, J. G.; ALVES, F. J. C.; GOMES, L. F. C. Emprego e condições de trabalho dos canavieiros no centro-sul do Brasil, entre 1995 e 2007. **Anais do XLVI Congresso da Sober**. Rio Branco: Sociedade Brasileira de Economia e Sociologia Rural, 2008.

BECKER, S. O.; ICHINO, A. Estimation of average treatment effects based on propensity score. **Stata Journal**, v. 2, n. 4, pp. 358-377, 2002. Available at: <http://www.sobecker.de/pscore.html.>. Accessed on April 12, 2009.

BNDES; CGEE. **Bioetanol de cana-de-açúcar**: energia para o desenvolvimento. Rio de Janeiro: BNDES, 2008.

CAMARGO-JR., A. S.; TONETO-JR, R. Indicadores sócio-econômicos e a cana-de-açúcar no estado de São Paulo. **Anais do I Workshop do Observatório do Setor Sucroalcooleiro**. Ribeirão Preto, 2008.

CHAGAS, A. L. S. **Externalidades da aglomeração**: microfundamentação e evidências empíricas. University of São Paulo, master's dissertation, 2004.

DIPRETE, T.; GANGL, M. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. **Sociological Methodology**, v. 34, n. 1, pp. 271-310, Dec. 2004.

FAO. **The State of Food and Agriculture 2008**: Biofuels: Prospects, risks and opportunities. Rome, 2008.

FRANZESE-JR, R. J.; HAYS, J. C. The spatial probit model of interdependent binary outcomes: Estimation, interpretation, and presentation. **24th Annual Summer Meeting of the Society for Political Methodology**, July 20, 2007. Available at: <http://polmeth.wustl.edu/retrieve.php?id=715>. Accessed on May 31, 2009.

GREENE, W. H. **Econometric Analysis**. 4th ed. New Jersey: Prentice Hall, 2000.

HÄRDLE, W; LINTON, O. Applied nonparametric methods. In ENGLE, R. F.; MACFADDEN, D. L. (eds.) **Handbook of Econometrics**, vol. 4, Amsterdam: Elsevier Science, 1994, pp. 2295-2339.
HECKMAN, J.J.; LALONDE, R.; SMITH, J. The economics and econometrics of active labor market programs. In ASHENFELTER, O.; CARD, D. (eds.) **Handbook of Labor Economics**, vol. 3. Amsterdam: Elsevier Science, 1999, pp. 1865-2097.
HECKMAN, J.J; ICHIMURA, H; TODD, P. Matching as an econometric evaluation estimator. **Review of Economic Studies**, vol. 65, No. 2, April 1998, pp. 261-294.
_____. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. **Review of Economic Studies**, vol. 64, no. 4, Oct. 1997, pp. 605-654.
HOFFMAN, R.; OLIVEIRA, F. C. R. Remuneração e características das pessoas ocupadas na agro-indústria canavieira no Brasil, de 2002 a 2006. **Anais do XLVI Congresso da Sober**. Rio Branco: Sociedade Brasileira de Economia e Sociologia Rural, 2008.
LESAGE, J.P. **Spatial Econometrics**. Mimeo, 1999. Available at: <http://www.rri.wvu.edu/WebBook/LeSage/spatial/wbook.pdf>. Accessed on March 1, 2009.
_____. Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models. **Geographical Analysis,** 32(1):19-35, 2000.
LEUVEN, E.; SIANESI, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Available at: <http://ideas.repec.org/c/boc/bocode/s432001.html>. Accessed on June 1, 2009.
NORONHA, S. et al. **Agronegócio e biocombustíveis**: uma mistura explosiva - Impactos da expansão das monoculturas para a produção de Bioenergia. Rio de Janeiro: Núcleo Amigos da Terra, 2006.
PIKETTY, M. G.; MENEZES, T. M; DUARTE, J. B. N. A. Sugar cane in Brazil, poverty and equity: Evidence for the 1992-2006 period. **Anais do XXXIV Congresso da ANPEC**. Salvador: ANPEC, 2008.
PNUD; IPEA; FJP. **Atlas de desenvolvimento humano no Brasil**. Brasília, 2003.
RESENDE, A. C. C.; OLIVEIRA, A. M. H. C. Avaliando resultados de um programa de transferência de renda: o impacto do Bolsa-Escola sobre os gastos das famílias brasileiras. **Estudos Econômicos**, vol. 38, n.2, 2008, pp. 235-265.
ROSENBAUM, P.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. **Biometrika**, 70, pp. 41-55, 1983.
ROSENBAUM, P. **Observational Studies**, Springer, New York. 2002.
SILVA, R. Setor sucroalcooleiro no estado de São Paulo: mensurando impactos sócio-econômicos. **Anais do I Workshop do Observatório do Setor Sucroalcooleiro**. Ribeirão Preto, 2008.
TODD, P. **A practical guide to implementing matching estimators**. Mimeo. 1999. Available at: <http://athena.sas.upenn.edu/~petra/papers/prac.pdf>. Accessed on May 31, 2009.
TONETO-JR., R.; LIBONI, L.B. Mercado de Trabalho da cana-de-açúcar. **Anais do I Workshop do Observatório do Setor Sucroalcooleiro**. Ribeirão Preto, 2008.
WOOLDRIDGE, J. M. **Econometric Analysis of Cross-Section and Panel Data**. Cambridge: MIT Press, 2001.

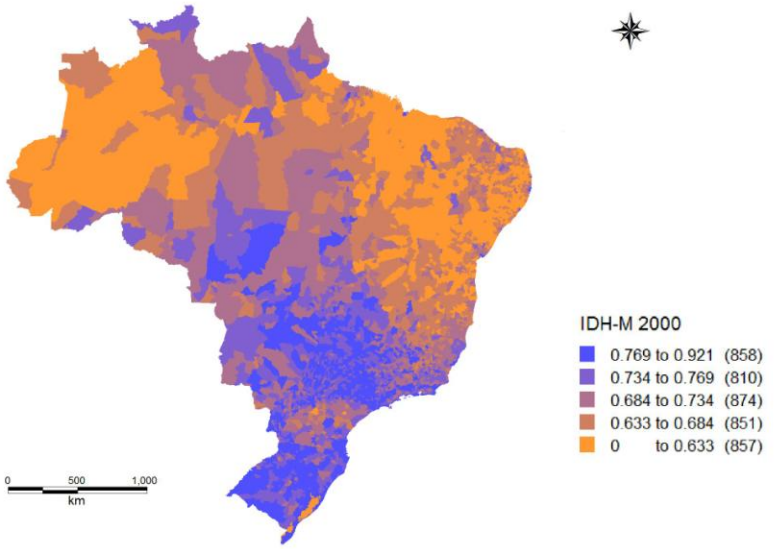

Figure 1: HDI-M for the year 2000 by MCA

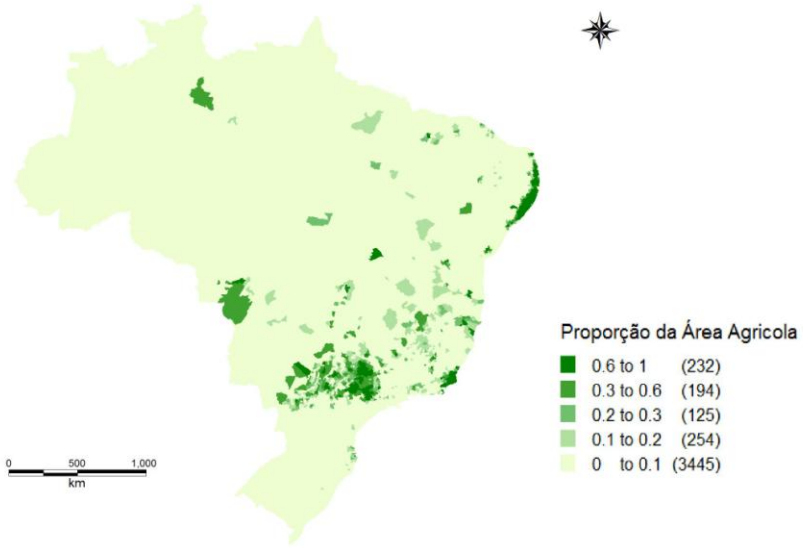

Figure 2: Production of sugarcane as a proportion of total tilled area in the MCA – average for 1991 to 2000

Table 1: Average HDI-M for 2000 for the treatment and control regions – total sample

Group		HDI				Number of MCAs
		General	Eduation	Longevity	Income	
Treatment	(Cane Growing ≥ 30% of the agricultural area)	0,716	0,787	0,730	0,632	426
Control		0,688	0,777	0,696	0,592	1000
Difference		0,028	0,010	0,034	0,040	
Treatment	(Cane Growing ≥ 9% of the agricultural area)	0,718	0,791	0,731	0,632	851
Control		0,688	0,777	0,696	0,592	1000
Difference		0,029	0,014	0,035	0,040	
Treatment	(Cane Growing ≥ 15% of the agricultural area)	0,719	0,790	0,731	0,636	646
Control		0,688	0,777	0,696	0,592	1000
Difference		0,031	0,014	0,035	0,044	
Treatment	(Cane Growing ≥ 65% of the agricultural area)	0,697	0,770	0,713	0,609	213
Control		0,688	0,777	0,696	0,592	1000
Difference		0,009	-0,007	0,017	0,017	

Source: Prepared by the authors.

Table 2: Average HDI-M for 2000 for the treatment and control regions – restricted sample

Group		HDI				Number of MCAs
		General	Eduation	Longevity	Income	
Treatment	(Cane Growing ≥ 30% of the agricultural area)	0,716	0,787	0,730	0,632	424
Control		0,691	0,779	0,697	0,596	907
Difference		0,026	0,008	0,033	0,036	
Treatment	(Cane Growing ≥ 9% of the agricultural area)	0,718	0,791	0,731	0,632	846
Control		0,691	0,779	0,697	0,596	907
Difference		0,027	0,012	0,034	0,036	
Treatment	(Cane Growing ≥ 15% of the agricultural area)	0,719	0,790	0,731	0,636	644
Control		0,691	0,779	0,697	0,596	907
Difference		0,029	0,012	0,034	0,040	
Treatment	(Cane Growing ≥ 65% of the agricultural area)	0,697	0,770	0,713	0,609	213
Control		0,691	0,779	0,697	0,596	907
Difference		0,007	-0,009	0,016	0,013	

Source: Prepared by the authors.

Table 3: Logit model for spatial propensity score estimation

		SAC Mod.	SAR Mod.
Pseudo-R^2	=	0,2652	0,3569
σ^2	=	4,3848	1,1443
no. of obs., no. of var.	=	1331 , 2	1331 , 2
no. 0, 1 y-values	=	907 , 424	907 , 424
accept rate ρ	=	0,5853	0,9998
accept. rate. λ	=	0,5301	
Posterior Estimates			
Variable		Coefficient	Coefficient
distance to distillery		-0,0205 *	-0,0099 *
		(-4,365)	(-13,025)
State		1,1806 *	0,5210 *
		(2,778)	(7,552)
ρ		0,9422 *	0,2382 *
		(49,782)	(4,056)
λ		-0,9302 *	
		(-17,727)	

* Significant at 1%.

T-statistic between parentheses.

Source: Prepared by the authors.

Table 4: Estimate of the treatment effect on the treated

Group	**HDI**				**Number of MCAs**
	General	**Eduation**	**Longevity**	**Income**	
Treatment	0,7163	0,7870	0,7299	0,6319	424
Control	0,7222	0,7863	0,7361	0,6441	907
Difference	-0,0059	0,0007	-0,0062	-0,0122	
t-statistic	-0,6861	0,0964	-0,6603	-1,0742	

Source: Prepared by the authors.

Table 5: Sensitivity analysis (Rosenbaum bounds) for the HDI and its components

Variável	Γ	p-crítico
HDI	1	0,2423
	1,05	0,1283
	1,1	0,0605
	1,15	0,0257
	1,2	0,0099
HDI-Education	1	0,0410
	1,05	0,0148
	1,1	0,0048
	1,15	0,0014
HDI-Longevity	1	0,4068
	1,05	0,2511
	1,1	0,1386
	1,15	0,0689
	1,2	0,0311
	1,25	0,0128
	1,3	0,0049
HDI-Income	1	0,0096
	1,05	0,0027
	1,1	0,0007

For the variables HDI, HDI-Longevity and HDI-Income, the p-critical value is p^+. For the variable HDI Education. the p-critical falue is p^-.
If $\Gamma^+ = e^\gamma = 1$, there is no selection bias due to unobservables.
Source: Prepared by the authors.

Table 6: Robustness tests of the estimation

Parameter	HDI				Number of MCAs	
	General	Eduation	Longevity	Income		
	t-statistic				Treated	Control
Bandwidth = 0.15	-1,8233	-1,2821	-1,3130	-2,2053	424	907
Bandwidth = 0.25	0,1988	1,0686	-0,0068	-0,2456	424	907
Neighborhood = 100km	-0,8695	-0,1942	-0,7059	-1,2607	424	907
Neighborhood = 200km	-0,6992	0,0959	-0,7206	-1,0544	424	907
Cane growing = 9% of the agricultural area	-0,7458	0,6165	-0,5898	-1,6023	846	907
Cane growing = 15% of the agricultural area	-0,1626	0,7864	-0,2611	-0,6635	644	907
Cane growing = 65% of the agricultural area	-2,3414	-1,9308	-1,9015	-2,4837	213	907

Source: Prepared by the authors.