Estimating regional poverty lines with scarce data: an application to Brazilian regions*

Fernando G. da Silveira[1], Alexandre X. Carvalho[2], Carlos R. Azzoni[3], Bernardo Campolina[4], Antonio Ibarra[5]

Abstract

The recent emphasis on fighting poverty in Brazil makes the determination of the size of the targeted population an important issue (What is the right poverty line? What is the real size of the poor population? How much money should be given to each poor family?). The application of poverty lines based on national income levels tends to produce important distortions at the regional level. Using data from a Household Expenditure Survey (HES) that covered some regions in Brazil, the paper develops and applies a methodology to define poverty lines for all regions and urban areas. These lines are based on nutritional requirements, thus avoiding the purchasing power parity problem, and take into account non-monetary income and in-kind consumption, aspects that are very important at the rural level. The HES results are matched with Census data, allowing for the estimation of rural and urban poverty lines for Brazilian regions.

JEL Classification: I 320; R 290

Key Words: Regional Poverty; nutrition; poverty lines; non-monetary income; in-kind consumption

1. Introduction

Brazil is a large middle-income country with an impressive number of poor people. Over time, economic growth has not been able to significantly improve this situation (Ferreira and Barros, 1999), and government administrations started social programs targeted to poor families in recent years. The Fome Zero (Zero Hunger) program brought attention of the media in Brazil and abroad. After a bad initial reaction, the Brazilian government unified different income transfer programs into the Bolsa Família Program. In September 2005, this program reached 7.8 million families in

* This paper is base on a study developed at Fipe – Fundação Instituto de Pesquisas Econômicas, for NEAD – Núcleo de Estudos Agrários e Desenvolvimento Rural, with support from IICA – Instituto Interamericano de Cooperação para a Agricultura. The authors acknowledge support from these institutions, as well as from CNPq – Bolsa Produtividade. Technical support from Guilherme Moreira is also acknowledged. The ideas and opinions expressed in this paper do not represent the views of these institutions.

[1] Ph. D. Student, Dept of Economics, Unicamp, and IPEA
[2] Head of the Spatial Studies Group, DIRUR, IPEA
[3] Professor of Economics, University of Sao Paulo
[4] Economist, Ph. D. Student, Dept of Geography, USP
[5] Social Scientist, Consultant IPEA and Fipe

5.561 cities in the country, transferring US$ 220 million. Other programs, such as Bolsa Escola (to keep children in school), Bolsa Alimentação, Cartão Alimentação (for food), and Auxílio Gás (for stove gas) are also in action, reaching 7.5 million families, and transferring another US$ 5 million.

There are different ways to measure poverty. Some consider infant malnutrition (height for the age) and adult chronic malnutrition (corporal mass below 18 kg/m^2), based on anthropometric measurements. In Brazil, infant malnutrition was only 6% in 1999, for an indigent population of 14.5% (Barros et al, 2000); recent surveys indicate that the ratio of people with weight deficit was below 5%, except for women in the rural Northeast region (7.6%). Other indicators consider lack of income, some considering a fixed monetary value (number of minimum wages; US$ 1 PPP a day, etc.). In this study we focus on the consumption of food. From this point of view, poverty is a situation in which the consumption of food is not enough to provide the minimum supply of calories. This brings in the question of defining what is the minimum caloric need, how to transform quantities of goods in calories, how to consider non-monetary income, how to take into account in-kind consumption, etc. On the other hand, this approach avoids the problems caused by different purchasing parities of income, and the fact that people in rural areas tend to rely more on self-consumption.

The contributions of this paper for the measurement of poverty in Brazil are manifold. First of all, we use data from a recent Household Expenditure Survey (HES) that covered all urban and rural areas in the country. We include non-monetary income and in-kind consumption, which is a critical issue in rural areas. We use semi-parametric methods to regress consumption on income and on number of adult-equivalents in the household. We match the HES data with Census data, producing measurements for all cities and regions in the country, differentiating rural and urban areas. The paper is organized as follows: Section 2 describes the HES data base, used for caloric consumption estimation. Section 3 presents a detailed description of the semi-parametric method employed to estimate the relationship between caloric consumption and per capita income. This relationship is then employed to calculate the number of poor individuals in different sub regions. In Section 4, we describe the methodology to combine both HES and Census data, so as to estimate poverty level in all Brazilian municipalities. The conclusions are presented in Section 5.

2. Database

The basic data come from a HES developed by IBGE, the Brazilian statistics office, between July 2002 and June 2003 (POF – Pesquisa de Orçamentos Familiares). The database includes information on household conditions (availability of water, sanitation, number of rooms, condition of the buildings, etc.); characteristics of the individuals inhabiting each household in the sample (gender, education, age, school attendance, weight and height, etc.); expenditure on public services, rent, etc.; expenditure on food, personal items, transportation, etc.; income (wage and non-wage) for each household member; and a subjective assessment of living conditions. It includes non-monetary expenditure and income. A total of 48.470 households, representing 0.1% of households in the country, were visited. The results are representative of the following 70 geographical domains: 27 state totals, 11 metropolitan areas (including the capital cities), 27 state urban areas (average of all urban areas in the state, including the capital city and the metropolitan cities), and 5 rural regions.

In order to expand the results from the 70 domains, for which the consumption structures are statistically representative, into the 5,507 municipalities in the country, a data base matching process is performed. For that, data from the Population Census 2000, produced by IBGE, are used. The socio-economic characteristics of households in the HES are matched with the same characteristics in the census, leading to estimates of number of poor people in each municipality.

3. Determining the number of poor households in each region

It is usual in the literature to define two cut-off lines for defining the number of poor people, one for indigence and one for poverty. In this paper we are dealing with a sole definition of poverty, which is "a situation in which a family finds difficulties to accommodate the cost of the ingestion of the appropriate amount of nutrients within the household budget". We start with the necessary caloric requirement for an adequate diet. FAO/WHO/ONU recommends 2,236 kilo-calories per day per person. It does not make any differentiation between urban and rural areas, or between age cohorts. A recommendation is made that each country should do a study to determine solutions for specific situations. Some studies were performed in order to define a scale of equivalent-adult, suggesting that a child should ingest, on average, 70% of the energetic needs of an adult, and elderly should ingest 90%.

We consider the actual consumption preferences of households in different areas of the country. We take their purchase of food items and verify if their income levels is large enough to buy them. Thus, we are not dealing with the cheapest (lowest cost bundle) or healthiest (best combination of goods considering health effects) basket of goods that could provide the minimum caloric need.

At present, Brazilian government is developing a methodology which considers the same data basis as ours, but with simplifying steps. Households are ordered by income and percentiles are formed. The average values of per capita income and expenditure for each group is calculated; rolling 20-group moving averages based on income levels are formed, in order to avoid unnecessary oscillations in values. Based on the price of kilo-calories in each region, the corresponding amount of money necessary to buy the minimum basket is identified, constituting the cut-off income value. Based on these cut-off values, the number of poor households is defined in each region. This per capita cost is the reference amount of money used to define a family as poor in these studies. It is still necessary to adjust income to include non-monetary gains, and to input a cost for owned housing, as well as to adjust for under declaration of income.

Although this is an interesting methodology, it has many limitations. It does not take into consideration possible economies of scale within families, meaning that a 5-individual household does not have necessarily to spend 5 times as much as a 1-individual household. Choosing 100 groups is also arbitrary. Even with moving averages, monotonicity in the consumption-income curve is not automatic. This procedure is not a standard regression statistical estimator. For small samples (rural areas, for example), it could not work well. Non-monotonicity could be another problem. The technique does not allow the analysis of censored data, hypothesis testing, the definition of confidence intervals and standard errors, given the sampling procedures of the HES from which the raw data comes.

3.1. A semi-parametric approach

In order to avoid some of the above-mentioned limitations, a methodology based on a semi-parametric regression model is used[6]. Consumption is regressed against income and number of equivalent-adults through a function such as

$$c_i / n_i = g(r_i / m_i) + \varepsilon_i, \qquad (1)$$

in which c_i is total consumption in household i, r_i is household total income, n_i is the number of adult-equivalents in the household, and m_i is the number of individuals in the household. The ratio r_i/m_i corresponds to household per capita income and c_i/n_i corresponds to the caloric consumption per equivalent-adult; ε_i is a random variable with zero mean and unknown variance. The functional form of $g(r_i/m_i)$ is unknown, and must be estimated from the data. For that we employ a semi-parametric estimation, using an expansion of basis functions. This expansion is based on an approximation of the unknown $g(r_i/m_i)$ function using the flexible parametric form

$$g(r_i / m_i) \cong \sum_{l=1}^{L} b_l \times u_l(r_i / m_i), \qquad (2)$$

in which $u_l(r_i / m_i)$ are the basis functions with known functional form. The expression in (2) represents a series of models commonly found in the data-mining literature, such as neural network regressions and wavelets regressions (Hastie, Tibshirani and Friedman, 2001).

We use an expansion of the type B-splines of order q. Since the $u_l(x)$ basis functions in the B-splines expansion depend on q, we explicitly write $u_{l,q}(x)$. It is assumed that the range of the independent variable x is $[x_{min}, x_{max}]$[7]. Consider a vector of w points $(x_1, x_2, \ldots, x_w)$ sharing the interval $[x_{min}, x_{max}]$, in which $x_{min} < x_1 < x_2 < \ldots < x_w < x_{max}$. The idea of the B-splines expansion is to adjust a $(q-1)$ degree polynomial in each interval defined by consecutive points within the set $x_{min}, x_1, x_2, \ldots, x_w, x_{max}$. In general, the choice if for $q = 3$ or $q = 4$, such that the polynomials used are of degree 2 or 3.

Consider now the vector of nodes $x_{min}, \ldots x_{min}, x_1, x_2, \ldots, x_w, x_{max}, \ldots, x_{max}$, in which the values x_{min} and x_{max} in the extremes are repeated q times. To facilitate the discussion, we write the vector $(x_{min}, \ldots x_{min}, x_1, x_2, \ldots, x_w, x_{max}, \ldots, x_{max})$ in the form $(t_1, t_2, \ldots, t_{w+2q})$. Thus, $t_1 = x_{min}, \ldots, t_q = x_{min}$. From the vector of nodes $(t_1, t_2, \ldots, t_{w+2q})$ and the order q, the basis functions $u_{l,q}(x)$ can be constructed recursively as

$$u_{l,1}(x) = \begin{cases} 1, t_l \le x < t_{l+1} \\ 0, otherwise \end{cases} \qquad (3)$$

---

[6] Non-parametric regression models are those in which the form of the response function is unknown, being estimated with the data. In many cases the estimation uses polynomial expansions of basis functions, and the non-parametric problem becomes the choice of a parametric model in which the transformations of the explanatory variables are appropriately constructed. In this case, the non-parametric regression model can be labeled semi-parametric regression. For details, see Hastie, Tibshirani and Friedman (2001).

[7] In this paper, the independent variable is per capita income r_i/m_i.

$$u_{l,p}(x) = \frac{x - t_l}{t_{l+p-1} - t_l} u_{l,p-1}(x) + \frac{t_{l+p} - x}{t_{l+p} - t_{l+1}} u_{l+1,p-1}(x), \; for \; p = 2,\ldots,q. \quad (4)$$

In the second term of expression (4) t_{l+p} appears in the denominator. Therefore, we have a B-splines of order q, and the number of basis functions is given by the number of nodes minus q. The number L of basis functions is $L = w + q$, and the functions are $u_{1,q}(x)$, ... , $u_{L,q}(x)$. Expression (2) can thus be rewritten, specifically for the B-splines expansion, as

$$g(r_i / m_i) \cong \sum_{l=1}^{L} b_l \times u_{l,q}(r_i / m_i). \quad (5)$$

The degree of flexibility of (5) is regulated by the number L of basis functions, which is directly related to the number of cutting points w, and to the order q. The flexibility of the semi-parametric expansion increases as the value of L increases. Given L, the estimation of the parameters b_l, l = 0,1,2,...,L can be obtained by OLS or Weighted LS with correction for heteroskedasticity or outliers. The adjustment of the $g(r_i/m_i)$ curve is obtained by the estimation of the regression model (linear in the parameters)

$$(c_i / n_i) = \sum_{l=1}^{L} b_l \times u_{l,q}(r_i / m_i) + \varepsilon_i. \quad (6)$$

The choice of the number of basis functions (L) can be performed by employing some in-sample criterion for model selection, such as AIC or BIC[8], or by some out-of-sample cross-validation rule. These procedures aim at avoiding over fitting of the semi-parametric model. If the number of basis functions is increased indefinitely, the adjustment of the model will be perfect, but its predictive power (out-of-sample) will be questionable. A lower value for L will produce less flexibility for the basis function expansion, and will also cause lower predictive power. The model selection criteria help choosing L such as to maximize the trade-off between model flexibility and the excessive number of unknown parameters (Hastie, Tibshirani and Friedman, 2001).

It is possible to estimate the function $g(r_i,/ m_i)$ directly from the micro data, without any previous grouping of households. This reduces arbitrariness in the definition of such groupings and uses the richness of information present in the individual data. Moreover, by choosing different values of L for different regions, it is possible to find semi-parametric expansions best adjusted for the regional specificities.

3.2. Imposing monotonicity to the income-consumption curve

It is reasonable to admit that the $g(r_i/m_i)$ function is monotonically increasing, that is, caloric ingestion by equivalent-adult always grows as per capita income increases, although not necessarily at constant rates. Therefore, the choice of the basis functions expansion must grant monotonicity (Chen and Conley, 2001; Leitenstorfer

[8] Burnhan e Anderson (1998).

and Tutz, 2005a and 2005b). The restriction $b_1 \leq b_2 \leq \cdots \leq b_L$ is thus imposed on (6), and the estimation process uses Restricted Weighted OLS. This procedure deals with the problem of different weights for different observations in the sample, and allows for the consideration of outliers.

Let Y be a vector with all the piled observations for the independent variable c_i/n_i, for the N households in the sample. Let X be the N x L matrix in which j corresponds to the piling of the N values $u_{j,q}(r_i / m_i)$, for the basis function j, in (6). Let β be the L x 1 vector of unknown parameters b_1, b_2, ... , b_L. It should be noted that Y and X are fully observed, since the functional form of the basis functions in the B-splines expansion is known. The problem of the estimation of β with OLS can be written as a quadratic maximization problem. In fact, consider the sum of the squared errors

$$\begin{aligned} S(\beta) &= [Y - X\beta]^T \times [Y - X\beta] \\ &= Y^T Y - 2[Y^T X]\beta + \beta^T [X^T X]\beta, \end{aligned} \tag{7}$$

in which A^T corresponds to the transpose of matrix (vector) A. Thus, minimizing the sum of the squared errors $S(\beta)$ corresponds to the maximizing the quadratic form $2[Y^T X]\beta - \beta^T [X^T X]\beta$. Introducing the monotonicity restriction $b_1 \leq b_2 \leq \cdots \leq b_L$, one gets the restricted quadratic maximization problem

$$\begin{aligned} &\max_{\beta} [Y^T X]\beta - \frac{1}{2}\beta^T [X^T X]\beta \\ &\text{subject to } M\beta \geq z, \end{aligned} \tag{8}$$

in which matrix M and vector z are given by

$$M = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ 0 & 0 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}_{(L-1)\times L}, \quad z = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \ldots \\ 0 \\ 0 \end{bmatrix}_{L-1} \tag{9}$$

The maximization situation in (8) corresponds to a classical quadratic programming problem, and can be solved with the algorithms described in Winston (2003) or Hillier and Lieberman (2002) for example.

In the estimation with Restricted Weighted LS, each observation i is weighted by its weight, w_i. Let W be the diagonal matrix of the N squared weights w_i^2. Estimating with Restricted Weighted LS corresponds to maximizing

$$\begin{aligned} &\max_{\beta} [Y^T WX]\beta - \frac{1}{2}\beta^T [X^T WX]\beta \\ &\textit{subject to } M\beta \geq z. \end{aligned} \tag{10}$$

From (8) and (10) one can get point-estimators for the parameters $b_1, b_2, \ldots, b_L$. From these values and from the known functional forms of the basis functions, it is possible to construct point-estimators for the average caloric consumptions per equivalent-adult c_i/n_i for the corresponding per capita income values r_i/m_i. By continuously varying r_i / m_i, it is possible to obtain the $\hat{g}(r_i / m_i)$ estimates of the $g(r_i / m_i)$ curve.

Once the functional form of $g(r_i / m_i)$ is chosen, the next step is to determine the per capita income value r_i/m_i that distinguishes poor and non-poor. Let C be the minimum amount of calories by adult-equivalent acceptable[9]. The cut-off per capita income value is given by the intersection of the estimated $\hat{g}(r_i / m_i)$ curve and the horizontal line equivalent to the per capita income value $c_i/n_i = C$. Thus, the cut-off per capita income level $\hat{r}_{cut-off}$ is just the solution to $\hat{g}(\hat{r}_{cut-off}) = C$. Finally, from $\hat{r}_{cut-off}$ it is possible to determine the number of households below the minimum caloric consumption level. It is also possible to determine which households in the sample are below the poverty line, in the sense that they face difficulties to accommodate that minimum expenditure on food within their budget.

One advantage of this methodology is the possibility to construct confidence intervals, using, for example, bootstrap methods (Davison and Hinkley, 1997; Hall, 1992) or first order approximations (Lehmann, 1999). Thus, one can also obtain confidence intervals for the cut-off per capita income level and for the number of poor households. However, given the computational effort involved in estimating monotonic expansions of B-splines, the use of bootstrap for the definition of confidence intervals was avoided. Besides, the complexity imposed by the restriction in the Weighted LS makes it difficult to obtain analytical results for the estimators asymptotic distribution, based on first-order approximations. Therefore, the option was made to use confidence intervals estimated the traditional LS. This alternative is computationally attractive, given its simplicity, and allows for the incorporation of observations with different weights (Draper and Smith, 1998).

3.3. Results

The estimation procedure considered 22 regions, or geographical domains, as described in Table 1. Although the database provides information for 70 geographical domains, analysis of the number of observations in each case, as well as of the dispersion of consumption values, recommended working with a smaller number of regions. Table 1 also shows the minimum caloric needs determined by Cepal (1996) and Lustosa (1999) for each of the 22 domains. In the first study, the minimum caloric need varies between 2,400 kcal/day per capita, in the colder Southern rural area, and 2,191 kcal/day per capita in the warmer urban areas of the North. In the second study, the rural area in the South also shows the highest value (2,408), but the lowest value refers to the Fortaleza Metropolitan Area, in the Northeast region. Although there are some differences between the results of the two studies, they show in general the same North-South pattern. In what follows we use the Cepal values.

[9] Although all estimations are made from region-specific samples, subscripts k are omitted in the text for ease of exposition.

Table 1 – Minimum caloric requirements by region *(Kcal/day per capita)*

Macro Region	Geographical Domain	CEPAL (1996)	Lustosa (1999)
North	Belém Metropolitan Region	2.191	2.160
	North Non-Metropolitan Urban	2.191	2.125
	North Rural	2.258	2.125
Northeast	Fortaleza Metropolitan Region	2.200	2.098
	Recife Metropolitan Region	2.200	2.126
	Salvador Metropolitan Region	2.200	2.127
	Northeast Non-Metropolitan Urban	2.200	2.169
	Northeast Rural	2.207	2.142
Southeast	Rio de Janeiro Metropolitan Region	2.288	2.233
	São Paulo Metropolitan Region	2.288	2.233
	São Paulo Non-Metropolitan Urban	2.288	2.246
	Rural São Paulo	2.318	2.309
	Belo Horizonte Metropolitan Region	2.288	2.233
	MG+ES+RJ Non-Metropolitan Urban	2.288	2.246
	MG+ES+RJ Rural	2.318	2.309
South	Curitiba Metropolitan Region	2.313	2.282
	Porto Alegre Metropolitan Region	2.313	2.284
	Urbano Não Metropolitano South	2.313	2.287
	Rural South	2.400	2.408
Center-West	Distrito Federal	2.259	2.198
	Center-West Urbano Não Metropolitano	2.259	2.220
	Center-West Rural	2.328	2.229

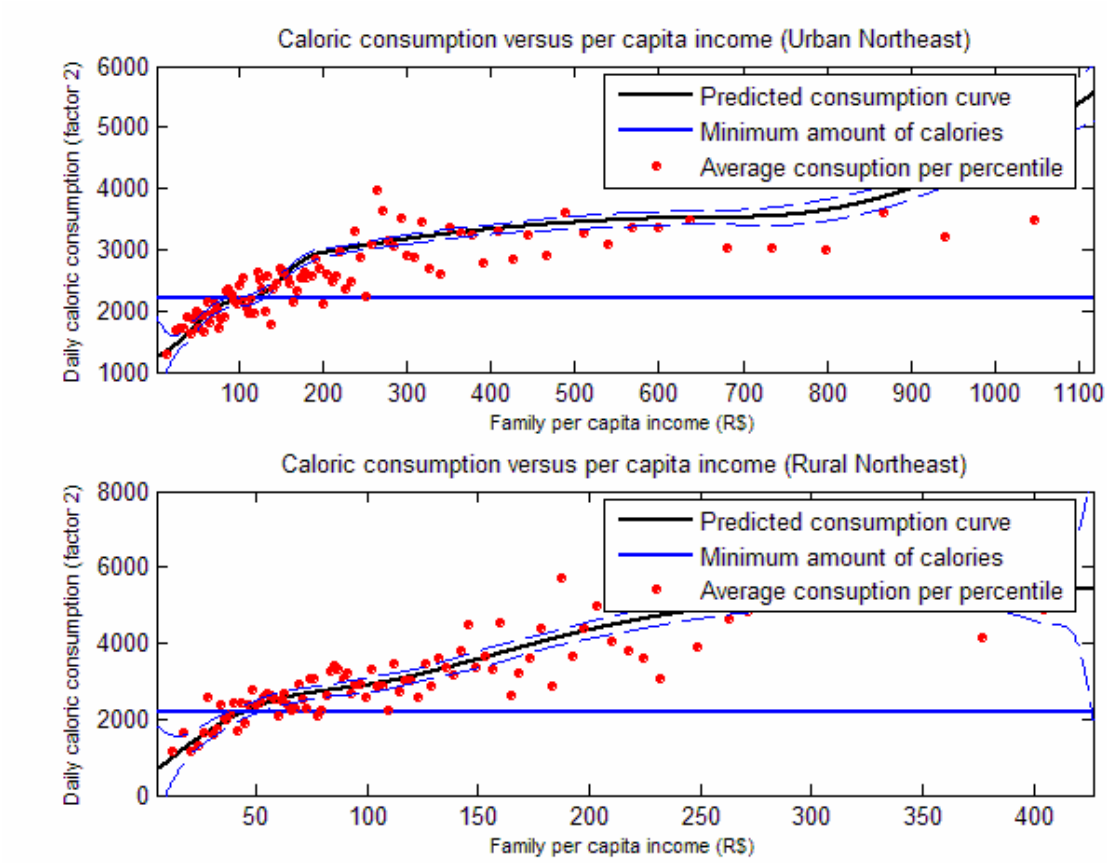
Another aspect to be considered is consumption of food outside the household. The HES data informs the amount of food items purchased and prepared in the household. It does not inform about consumption in restaurants, schools, work place, etc. This might introduce an important bias, for these forms of food ingestion are important and are becoming more and more popular. According to HES, in 2002-2003, 17% of household income was spent with food in Brazil, ranging from 10% in Brasília to 41% in the rural Northeast. Expenditure on food outside the household amounted on average to almost one fourth of total expenditure on food, ranging from 9% in the Rural North, to 38% in Brasília. The problem is that there is information only on the amount of income spent, not on the number of calories obtained. Thus, there is a need to transform income spent on food outside the household into calories. In this work we make two alternative suppositions: the first considers that one monetary unit buys half the amount of calories outside home, as compared to food prepared within the household; the other admits a 1 to 1 ratio.

By applying the model presented in Sub-sections 3.1 and 3.2 to each one of the 22 geographical domains, we arrived at the results presented in Table 2. In Figure 1 we present the estimated caloric consumption-per capita income estimated lines for Rural and Urban Northeast[10]. Per capita income is displayed in the horizontal axis, and caloric consumption per adult-equivalent is presented in the vertical axis. The black line is the estimated $\hat{g}(r_i / m_i)$ of the $g(r_i / m_i)$ curve. The confidence intervals are presented by the dotted lines (90% significance level). The solid horizontal blue line represents the

[10] Results for the other 21 geographical domains are available on request from the authors

minimum caloric consumption for that specific geographical domain (Table 1). For ease of visualization, averages for percentiles are presented in the red dots[11]. Since the interest of the study is in determining the number of poor households and persons, the 5% richest families were left out of the estimations[12]. The intersection of the estimated $\hat{g}(r_i / m_i)$ line with the horizontal line defines the cut-off per capita income level. The intersections of the dotted lines with the horizontal lines define the lower and upper limit for the cut-off per capita income level. It can be observed that the estimated $\hat{g}(r_i / m_i)$ functions are monotonically increasing, as imposed by the model.

Figure 1 – Income x consumption curves for the Northeast region



(Price of calories consumed outside home = 2 price of calories consumed at home)

The numbers in Table 2 reveal much higher cut-off income levels for metropolitan regions, a result which is compatible with the higher cost and importance of other items in these areas (housing, transportation, urban services). The cut-off per capita income level in the Rural North is R$ 30.80, while in Metropolitan São Paulo it is R$ 743.40. This brings into the discussion the concept of poverty adopted in this study, which is a situation in which a family has difficulties to accommodate in the budget the

[11] Off course, we used information for all households in estimating the functions.
[12] The problem of under declaration of income is well-known for rich people. The variance in consumption values increases sharply as income increases beyond a certain range, imposing considerable difficulties for the estimation.

minimum consumption of food. In Metropolitan São Paulo, other items of expenditure are more important, making it more difficult to buy the necessary calories. Given this high cut-off income level, this area presents the largest share of poor people, 68.01%, contrasting with only 9.08% in the Rural North.

Considering this result for Metropolitan São Paulo, but also for Brasília and Rio de Janeiro, it is clear that poverty has a very important urban face. Usually, it is believed that the poorest would be in the rural areas of North and Northeast in Brazil. Considering self-consumption and non-monetary income, as we do in this study, and the burden presented by transportation and housing in the urban areas, it is clear that the incidence of poverty is more important in urban areas, especially so in large metropolitan areas.

For the country as a whole, the number of people with difficulty to accommodate the expenditure on food within their budget is impressive, over 50 million, representing 28.5% of total population. Considering the computed confidence intervals, the number could go up to 72 million. If the 2 to 1 ratio of food prepared within the household or outside it is replaced by a 1 to 1 ratio, the number of poor drops to 35 million, and the share goes down to 20%. This change affects more intensively the large urban areas, as expected. In the case of Metropolitan São Paulo, the share of poor people drops from 68% to 47%. As for the confidence intervals, they vary with the number of households in the sample for each specific geographical domain and with the dispersion of the caloric consumption of those families. In the case presented in Figure 1, the adjustment is the best, since the sample presented 11,896 urban and 4,007 rural households. For the metropolitan regions of São Paulo (770 households), Rio de Janeiro (812) and Brasília (934), precision is not so good.

4. Expanding the results to other geographical domains

The results presented in the previous section are important, for they provide another way of estimating poverty levels in Brazil. The methodology allows for the estimation of confidence intervals, highlighting the variance in results that result from the variability in the data, as well as from the suppositions made about consumption outside the household. By considering information from all households in the estimation, it also improves the estimation of the consumption-income function.

As discussed above, after we obtained the cut-off income value $\hat{r}_{cut-off}$, it is possible to specify in the micro data which households are below the poverty line $\hat{r}_{cut-off}$ and which ones are above. Therefore, one can assign a binary variable z_i to household i, where $z_i = 1$ whenever income of i lies below $\hat{r}_{cut-off}$ and $z_i = 0$ otherwise. This approach allows us to identify the characteristics of the households below the poverty line. More specifically, we can estimate a proper regression model to predict the probability $\Pr[household\ is\ poor]$ of a certain household i being below the poverty line, based on its characterizing variables $x_1, x_2, \ldots, x_p$ (gender, education, number of residents, income, etc.). Therefore, we can write

$$\Pr[household\ is\ poor] = h(x_1, x_2, x_3, ..., x_p), \qquad (11)$$

and we can use any a Logit binary response regression model (Greene, 1993; Wooldridge, 2002).

Table 2 – Estimated cut-off per capita income levels and number of poor people

Geographical Domain	Cut-off *per capita* income (R$)		Number of poor individuals		% of poor individuals		Upper limit for the cut-off per capita income level (R$)		Upper limit for the number of poor individuals	
	2 to 1 (*)	1 to 1 (*)	2 to 1	1 to 1	2 to 1	1 to 1	2 to 1	1 to 1	2 to 1	1 to 1
Belém	108,62	96,09	343.140	273.581	18,50%	14,75%	199,54	131,40	798.475	465.084
Urban North	62,63	57,05	1.137.497	979.165	13,76%	11,85%	73,15	65,91	1.482.492	1.210.258
Rural North	38,78	37,55	321.254	285.711	9,08%	8,08%	49,85	48,67	571.145	546.993
Fortaleza	142,96	130,52	1.105.041	971.085	36,49%	32,07%	185,81	161,80	1.458.440	1.238.944
Recife	135,01	62,48	855.713	158.723	25,62%	4,75%	204,57	185,60	1.476.252	1.348.739
Salvador	190,72	135,84	916.951	638.598	29,49%	20,54%	280,15	212,80	1.486.779	1.121.164
Urban	118,66	80,18	10.490.616	6.272.113	40,84%	24,42%	130,72	109,26	11.575.079	9.680.360
Rural Northeast	45,67	42,51	3.056.619	2.815.597	21,90%	20,17%	53,28	48,55	3.859.150	3.380.418
Belo Horizonte	136,05	122,97	491.582	358.837	11,03%	8,05%	218,96	173,40	1.116.612	725.559
MG+ES+RJ Urban	161,93	149,91	4.408.063	3.917.444	26,14%	23,23%	180,80	160,97	5.167.124	4.355.065
Rural MG+ES+RJ	70,07	50,67	563.588	233.594	13,81%	5,73%	138,53	97,87	1.687.424	1.107.138
Rio de Janeiro	309,41	217,98	4.909.126	3.119.474	44,26%	28,12%	471,65	320,71	6.597.007	5.088.050
São Paulo	743,43	438,63	12.063.946	8.328.998	68,01%	46,95%	1,159,60	675,45	14.376.456	11.464.357
Urban São Paulo	84,06	64,44	709.650	363.257	3,90%	1,99%	221,68	205,80	4.464.808	4.187.803
Rural São Paulo	65,48	56,53	39.353	19.830	1,57%	0,79%	187,57	117,22	818.515	351.226
Curitiba	223,16	184,91	516.147	340.101	19,43%	12,80%	339,74	270,00	987.266	705.033
Porto Alegre	244,33	204,53	944.812	747.231	25,72%	20,34%	443,86	411,00	1.950.289	1.694.864
Urban South	131,28	115,00	2.263.170	1.784.529	14,97%	11,80%	238,04	156,02	5.452.126	2.951.678
Rural South	98,88	95,55	529.723	470.110	11,93%	10,59%	133,71	128,47	979.806	926.235
Brasília	562,04	351,49	1.301.577	991.701	59,99%	45,71%	844,15	501,68	1.561.035	1.238.122
Urban Center-West	164,26	125,65	2.939.039	2.026.015	33,92%	23,38%	206,67	159,90	3.860.177	2.873.913
Rural Center-West	72,31	67,81	142.367	126.444	10,29%	9,14%	103,62	97,26	287.099	248.057
Brazil			50.048.974	35.222.135	28,46%	20,03%			72.013.553	56.909.060

(*) 2 to 1 and 1 to 1 refer to the price of calories consumed outside the household in relation to calories consumed within the household

Given the calculated probability of a household *i* being below the poverty line, we can estimate the total number of poor individuals by using a more geographically detailed data base. In fact, the HES micro data used to estimate the consumption curves above do not have enough observations to estimate the number of poor people in each of the 5,507 Brazilian municipalities. However, one can resort to the Census data to obtain this regional level of detail[13]. This is possible because, even though the Census does not contain consumption information, we have identifies 52 variables common both to Census and HES[14]. We use these variables to predict the probability of each household in the Census data base to be below the poverty line. Therefore, by using an appropriate regression model to predict the probability values, based on the set of common variables to both data bases, one can use the model to estimate the number of poor families in each Brazilian municipality.

We use a maximum of 52 explanatory variables (x), as described in the appendix. Estimation was made separately for each of the 22 geographical domains, and the sub-set of variables changes from case to case, depending on the quality of adjustment. It is likely that, when the estimated values for all municipalities are summed-up, the overall number of poor individuals in Brazil, or in each geographical domain, will be different from the overall numbers estimated by the model. To make sure that the totals match, we linearly rescale the municipality estimates so as to have the total number of poor individuals equal to the overall sums obtained from HES micro data in the previous section, in each geographical domain. Table 3 presents information on the quality of adjustment. The smallest percentage of concordant cases was 96.3% in the case expenditure with food outside home is considered equivalent, in caloric terms, to expenditure with food at home, and 95.8% in the case in which expenditure with food outside home buys only half of calories. Except for a few cases, the percentages of concordant cases are all above 98%.

Given these results, the next step is to estimate the number of poor people for all municipalities in the country. The coefficients of the Logit regressions for each geographical domain are applied to the equivalent census data for each municipality, leading to the estimated number of poor people in each case. Results are presented in Maps 1 and 2, for the two alternative suppositions about the caloric equivalent of expenditure with food outside home. In order to provide a background for analyzing the results, Maps 3 and 4 present per capita GDP and the percentage of households with access to public sewage across all municipalities.

First of all, it is clear that using the supposition that expenditure on food outside the household is equivalent to half the caloric consumption within the household increases the estimated number of poor everywhere, as expected. It does not change the geographical distribution, though. The highest incidence is in the poor Northeast region, followed by the agricultural frontier Center-West region. The rich Southeast and South regions present very low incidence, which is concentrated in the capital city and or in the metropolitan regions; in the rich state of São Paulo, the incidence is only present in its metropolitan region.

[13] The population census sample is composed of 5,221,467 households (12%).
[14] The 52 variables are described in Table A1, in the appendix, which also shows descriptive statistics

Table 3 - Quality of adjustment of the Logit regressions

Macro Region	Geographical Domain	Food outside = food home				Food outside = 1/2 food home			
		Percent Concordant	Somers' D	Gamma	Tau-a	Percent Concordant	Somers' D	Gamma	Tau-a
North	Belém Metropolitan Region	98,3	0,966	0,966	0,141	98,2	0,964	0,965	0,164
	North Non-Metropolitan Urban	96,3	0,932	0,933	0,078	95,8	0,918	0,918	0,086
	North Rural	98,4	0,968	0,969	0,348	98,6	0,973	0,973	0,391
Northeast	Fortaleza Metropolitan Region	98,8	0,976	0,977	0,054	98,1	0,962	0,963	0,293
	Recife Metropolitan Region	99,3	0,987	0,987	0,234	98,7	0,974	0,975	0,331
	Salvador Metropolitan Region	98,4	0,968	0,968	0,294	98,4	0,969	0,970	0,426
	Northeast Non-Metropolitan Urban	96,9	0,939	0,940	0,229	96,9	0,938	0,939	0,248
	Northeast Rural	98,8	0,976	0,976	0,100	98,3	0,966	0,966	0,132
Southeast	Belo Horizonte Metropolitan Region	98,2	0,965	0,966	0,269	98,3	0,966	0,967	0,300
	MG+ES+RJ Non-Metropolitan Urban	98,0	0,960	0,961	0,066	96,2	0,924	0,924	0,146
	MG+ES+RJ Rural	98,6	0,971	0,971	0,314	98,2	0,966	0,968	0,436
	Rio de Janeiro Metropolitan Region	98,8	0,977	0,978	0,451	99,0	0,984	0,988	0,479
	São Paulo Metropolitan Region	99,7	0,995	0,995	0,021	99,2	0,983	0,983	0,044
	São Paulo Non-Metropolitan Urban	99,1	0,982	0,982	0,018	98,6	0,972	0,972	0,028
	Rural São Paulo	99,5	0,991	0,991	0,155	99,5	0,990	0,991	0,228
South	Curitiba Metropolitan Region	99,2	0,984	0,984	0,230	99,1	0,982	0,982	0,285
	Porto Alegre Metropolitan Region	98,7	0,975	0,975	0,134	98,6	0,972	0,972	0,170
	South Non-Metropolitan Urban	97,5	0,951	0,951	0,123	97,4	0,947	0,948	0,135
	Rural South	98,8	0,977	0,978	0,437	98,6	0,974	0,976	0,487
Center-West	Federal District	98,4	0,968	0,968	0,275	98,2	0,965	0,966	0,373
	Center-West Non-Metropolitan Urban	97,7	0,955	0,956	0,126	97,7	0,955	0,956	0,139
	Center-West Rural	98,8	0,976	0,976	0,140	99,5	0,989	0,990	0,192

A visual association with the geographical distribution of per capita GDP reveals that municipalities with high incidence of poverty are those with low per capita GDP, as it would be expected. It is interesting to notice that the Logit regressions use per capita income as explanatory variable, that is, household disposable income, not per capita GDP, which is associated with production. That is, poverty is more intense in places in which per capita production is lowest. The last map provides information on the geographical distribution of access to public sewage, which is another indicator of well-being of population. It is clear from a visual analysis of the maps that there is a negative correlation between incidence of poverty and access to public sewage collecting systems.

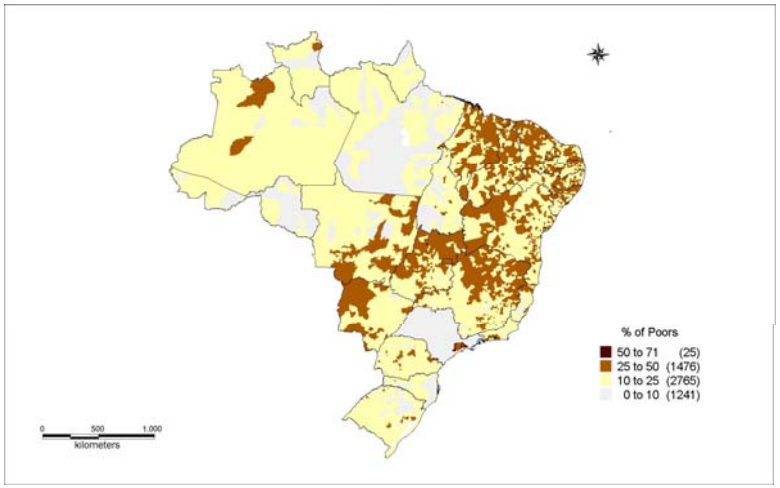
4. Conclusions

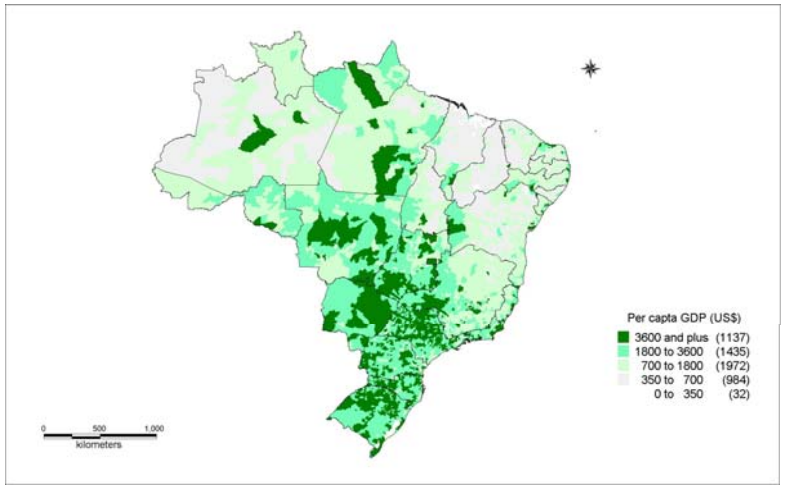
This paper provides an estimation of the number of poor people in Brazilian municipalities. This information is relevant for the design of social policies, especially in reference to focusing such policies on the relevant target population. We have used data from a recent Household Expenditure Survey (HES) that covered all urban and rural areas in the country. We have included non-monetary income and in-kind consumption, which is a critical issue in rural areas. We have used semi-parametric methods to regress caloric ingestion on income, and have matched the HES data with Census data, producing measurements for all cities and regions in the country, differentiating rural and urban areas.

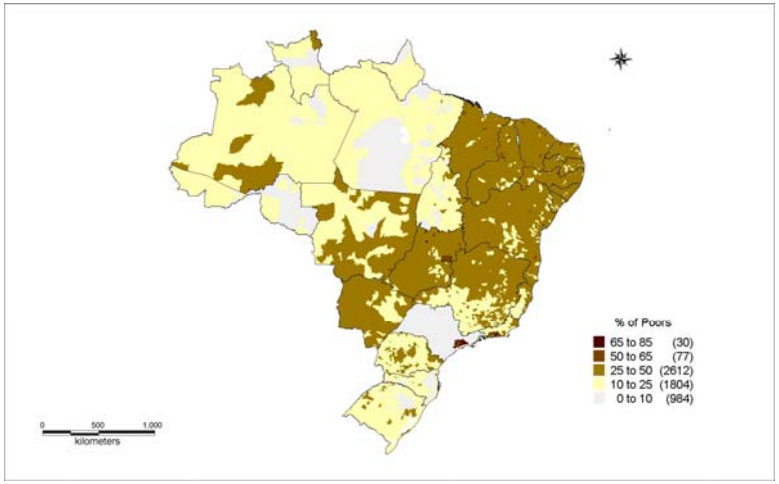
The results indicate that, for the country as a whole, the number of people with difficulty to accommodate the expenditure on food within their budget is over 50 million, representing 28.5% of total population. Considering the computed confidence intervals, the number could go up to 72 million. If the 2 to 1 ratio of food prepared within the household or outside it is replaced by a 1 to 1 ratio, the number of poor drops to 35 million, and the share goes down to 20%. This change affects more intensively the large urban areas, as expected. In the case of Metropolitan São Paulo, the share of poor people drops from 68% to 47%.

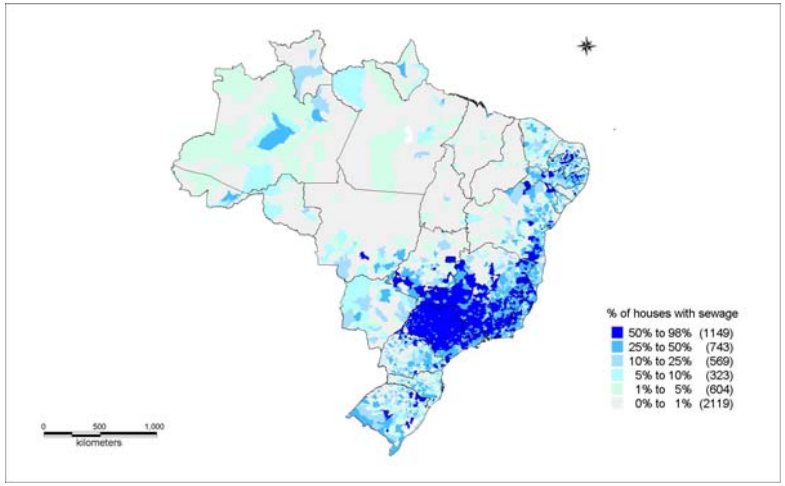
In order to match the HES data with census data, we have used Logit regressions, in which we have estimated the probability of a person with a set of characteristics to be classified as poor. Based on these characteristics, we have used census data, available at the municipality level, to estimate the number of poor at that geographical level. The results indicate that municipalities with high incidence of poverty are those with low per capita GDP, and to poor access to public utilities, as proxied by access to the public sewage collecting system.

Maps 1 to 4 – Percentage of poor people, per capita GDP and access to public sewage in Brazilian municipalities









References

Arias, A. *Proposta sobre a utilização do método da renda na preparação das medições de indigência e pobreza baseadas em linhas preparadas através da POF 2002-2003*. Brasília, IPEA, dezembro 2003. mimeo (publicação restrita).

Barreto, S. A. J. and Cyrillo, D. C. Análise da composição dos gastos com alimentação no município de São Paulo (Brasil) na década de 1990. *Revista de Saúde Pública,* São Paulo, v. 35, n. 1, 2001.

Burnan, K. and Anderson, D. *Model Selection and Inference. A Practical Information-Theoretic Approach*. Springer, 1998.

CEPAL. *Medición de la pobreza en Brasil: una estimación de las necesidades de energía y proteínas de la población*. Santiago, CEPAL, 1996.

Chen, X. and Conley, T. A New Semiparametric Spatial Model for Panel Time Series. *Journal of Econometrics*. No. 105, pp 59—83, 2001.

Cochran, W. *Sampling Techniques*. Wiley, 1977.

Davison, A. and Hinkley, D. *Bootstrap Methods and their Applications*. Cambridge University Press, 1997.

FAO/WHO/UNU. Necessidades de Energia e Proteínas. Genebra, FAO/WHO, 1985. *Série Informes Técnicos nº 724*. Revisão do mesmo estudo elaborado em 1973 e disponibilizado em 1974.

Ferreira, F. H. G. and Barros, R. P. (1999) "The slippery slope: explaining the increase in extreme poverty in urban Brazil, 1976-1996". The World Bank, Policy Research Working Paper No. 2210, October

Greene, W. *Econometric Analysis*. Prentice Hall, 1993.

Hall, P. *The Bootstrap and the Edgeworth Expansion*. Springer, 1992.

Hastie, T., Tibshirani, R. and Frieman, J. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, 2001.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Pesquisa de Orçamentos Familiares 2002-2003: primeiros resultados – Brasil e Grandes Regiões* Rio de Janeiro, 2004.

IPEA. *Contribuições para a Construção de Linhas de Indigência para o Brasil*. Brasília, 06 de outubro de 1997, mimeo, Comissão Mista IBGE/CEPAL/IPEA

Lehmann, E. *Elements of Large-Sample Theory*. Springer, 1999.

Leitenstorfer, F. e Tutz, G. Generalized Monotonic Regression Base don B-Splines with an Application to Air Pollution Data. *Ludwig-Maximilians-Universitat*. Discussion Paper n. 444. Maio 2005.

Lustosa, Tânia. *Cálculo de las necesidades energéticas de la población brasileña para la construcción de una línea de pobreza*. Santiago, CEPAL, 1999 (presented at 4º Taller Regional del MECOVI).

Barros, R. P. et al. A estabilidade inaceitável: desigualdade e pobreza no Brasil. In: Henriques, R. (org.) *Desigualdade e pobreza no Brasil.* Rio de Janeiro: IPEA, 2000.

Rocha, Sônia. *Pobreza no Brasil: afinal do que se trata?* Rio de Janeiro: Editora FGV, 2003.

Tutz, G. and Leitenstorfer, F. Generalized Smooth Monotonic Regression. *Ludwig-Maximilians-Universitat*. Discussion Paper n. 417. March 2005.

Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2002.

Table A1 - Explanatory variables for the Logit regressions

Variable	Unit of measurement	Census				HES			
		Min	Max	Mean	Std Dev	Min	Max	Mean	Std Dev
Radio set	(1,0)	0	1	0,9	0,3	0	1	0,3	0,5
Refrigerator	(1,0)	0	1	0,8	0,4	0	1	0,8	0,4
VCR	(1,0)	0	1	0,3	0,5	0	1	0,2	0,4
Washing machine	(1,0)	0	1	0,3	0,5	0	1	0,3	0,5
Microwave	(1,0)	0	1	0,2	0,4	0	1	0,1	0,3
Computer	(1,0)	0	1	0,1	0,3	0	1	0,1	0,3
TV sets	Number	0	9	1,3	0,9	0	8	1,2	0,8
Cars	Number	0	9	0,4	0,6	0	3	0,0	0,1
Air Conditioners	Number	0	9	0,1	0,5	0	9	0,1	0,5
Type of building	House=1, Apt=2, Room=3	1	3	1,1	0,3	1	1	1,0	0,0
Rooms	Number	1	30	5,7	2,3	1	30	5,8	2,3
Bedrooms	Number	1	9	2,0	0,9	1	9	2,0	0,9
Bathrooms	Number	0	9	1,1	0,8	0	13	1,2	0,7
Water connection	1=connected; 6= oor acess	1	6	2,0	1,7	1	6	1,9	1,6
Sewage connection	1=connected; 7=unavailable	1	7	2,5	1,8	1	7	2,6	1,7
Quality of occupation	1=owned+payed; 6=precarious	1	6	1,8	1,3	1	6	1,8	1,4
Electricity connection	(1,0)	0	1	0,9	0,3	0	1	0,9	0,2
Persons in the Household	Number	1	43	3,8	2,0	1	20	3,8	1,9
Persons in HH Adjusted	Excludes servants and others	1	43	3,8	2,0	1	20	3,7	1,9
Males in HH	Number	0	30	1,9	1,3	1	12	2,0	1,1
Women in HH	Number	0	22	1,9	1,2	1	13	2,0	1,1
Age of HH Head	Years	10	130	45,6	15,4	13	102	45,8	15,5
Gender of HH Head	(1,0)	0	1	0,7	0,4	0	1	0,7	0,4
Race of HH Head	1 = white; 5 = indian	1	5	2,2	1,4	1	5	2,6	1,5
Schooling HH Head	Years of schooling	0	17	5,4	4,5	0	17	5,3	4,6
Children 7-14 in School	Number	0	17	0,6	0,9	0	7	1,5	0,8
Children 7-14	Number	0	17	0,6	0,9	1	7	1,6	0,8
Members age 15-18 in School	Number	0	8	0,2	0,5	0	5	1,0	0,6
Members age 15-18	Number	0	8	0,3	0,6	1	5	1,3	0,5
Members 15 and +	Number	0	23	2,7	1,3	1	8	1,6	0,8
Literate Members 15 and +	Number	0	22	2,3	1,3	0	8	1,5	0,8
Sum years study members 15 and +	Years of schooling	0	177	15,7	12,2	0	63	11,4	7,8
Members 25 and +	Number	0	18	1,9	0,9	1	11	1,9	0,8
Literate Members 25 and +	Number	0	15	1,6	1,0	0	11	1,6	0,9
Sum years study members 25 and +	Years of schooling	0	166	10,5	9,4	0	84	10,5	9,1
Members age 0-5	Number	0	23	0,4	0,7	1	8	1,3	0,6
Members age 5-10	Number	0	18	0,4	0,7	1	7	1,3	0,6
Members age 10-20	Number	0	16	0,8	1,1	1	8	1,7	0,9
Members age 20-30	Number	0	13	0,7	0,9	1	7	1,4	0,7
Members age 30-40	Number	0	9	0,6	0,7	1	5	1,3	0,5
Members age 40-50	Number	0	6	0,4	0,7	1	4	1,3	0,5
Members age 50-60	Number	0	6	0,3	0,6	1	5	1,2	0,4
Members age 60-70	Number	0	6	0,2	0,5	1	4	1,2	0,4
Members age 70-80	Number	0	5	0,1	0,3	1	4	1,1	0,4
Members age 80 and +	Number	0	5	0,0	0,2				
Members 60 and +	Number	0	7	0,3	0,6	1	3	1,1	0,3
Members age 65 and +	Number	0	7	0,2	0,5	1	4	1,2	0,5
Family monetary income	R$/Month	0	1011958	1042,3	2970,4	0	384676	1307,1	2980,3
Per Capita family monetary income	R$/Month	0	701000	345,0	1229,0	0	54954	431,7	957,2
Family labor income	R$/Month	0	1010500	841,7	2715,8	0	367555	961,7	2478,3
Per Capita family labor income	R$/Month	0	700000	265,4	1073,8	0	52508	305,0	715,6
Family retirement and pension income	R$/Month	0	50000	165,2	548,9	0	27250	173,4	666,3