Rev Saúde Pública 2011;45(6)

Bianca Schmid[I]

Nilza Nunes da Silva[II]

Estimation of live birth underreporting with a capture-recapture method, Sergipe, Northeastern Brazil

ABSTRACT

OBJECTIVE: Estimate the number of live births and, therefore, underreporting of live births.

METHODS: The databases of the Live Birth Information System and the Civil Registry of the Brazilian Institute of Geography and Statistics, from the second and third trimesters of 2006 in Sergipe state (Northeastern Brazil) were paired by deterministic linkage based on the number of the Live Birth Declaration. The geographic disaggregation utilized was mother's microregion of residence. Huggins closed population models were used to estimate the capture probabilities for each database and the total live births during the period, within each geographic subdivision. MARK® software was used for the estimates.

RESULTS: Underregistration during the period studied was 19.3%. Application of the capture-recapture method to estimate underregistration of live births is possible, including for geographic disaggregations smaller than a state. The deterministic linkage was impaired in four microregions, due to non-inclusion of the Live Birth Declaration number in the database of the Brazilian Institute of Geography and Statistics. Maternal age, a heterogeneity characteristic in the population of live births, affected the probability of capture by the civil registry.

CONCLUSIONS: Capture-recapture was a viable method to estimate the underregistration of live births.

DESCRIPTORS: Live Birth. Birth Certificates. Underregistration. Registries. Records as Topic. Vital Statistics.

INTRODUCTION

Underregistration of vital events is still a reality in Brazil.[18,a,b] According to Simões,[c] the lack of coverage by vital statistics is a barrier to the direct calculation of fertility and mortality rates in Brazil.

Calculation of the fertility and child mortality rates with direct methods, without correction for the underreporting of births and deaths, can hide the demographic

[I] Instituto Brasileiro de Geografia e Estatística. Unidade Estadual São Paulo. São Paulo, SP, Brasil

[II] Faculdade de Saúde Pública. Universidade de São Paulo. São Paulo, SP, Brasil

Correspondence:
Bianca Schmid
R. Frei Caneca, 443 – Apto. 102
Consolação
01307-001 São Paulo, SP, Brasil
E-mail: bika.schmid@gmail.com

Received: 12/13/2010
Approved: 7/27/2011

Article available from: www.scielo.br/rsp

[a] Romero DEM. Vantagem e limitações do método demográfico indireto e dos dados da PNAD'98 para a estimativa da mortalidade infantil. In: Anais do XIII Encontro Nacional de Estudos Populacionais, Ouro Preto, BR. São Paulo: Associação Brasileira de Estudos Populacionais; 2002[cited 2007 Jan 03]. Available from: http://www.abep.nepo.unicamp.br/docs/anais/pdf/2002/gt_sau_st3_romero_texto.pdf

[b] Instituto Brasileiro de Geografia e Estatística. Estatísticas do registro civil. Rio de Janeiro; 2005. v. 32.

[c] Simões CCS. Brasil: Estimativas da Mortalidade Infantil por microrregiões e municípios. Brasília: Ministério da Saúde; 1999.

reality of a population.[a] To calculate these indicators, indirect techniques are employed for estimates, with information sources including demographic census and representative studies. Often, violation of the assumptions implicit when implementing such techniques causes distortion of estimates. When estimates are made for smaller geographic disaggregation of federal units, the problem becomes more complex due to the small population size of many Brazilian municipalities.[c]

Various indicators may be calculated using statistics from the Civil Registry, such as fertility rates, mortality coefficients and life expectancy at birth. Efforts, to understand the negligence of civil records, attempt to remedy this situation and adhere to the principals of efficient professional practices contained in the United Nation's Fundamental Principles of Official Statistics.[d]

Estimation through the capture-recapture method seeks to use the overlap between incomplete registries to formally measure the underestimation of these sources. This allows for the correction of statistics and the production of indicators that better approximate reality. These available sources (lists) may include mandatory reportable diseases, statistics from hospitals and other health services and death records, in addition to other sources.[2,11,12]

In summary, the capture-recapture method was utilized to estimate the population of France in 1793. Since the 19th century, the technique was widely used to estimate the population of wild animals,[3,8] and in various other application in medicine, demography and epidemiology.

In 1984, Petersen developed the most simply model for estimation with the capture-recapture method, using two samples.[8] In the 1940s, Sekar & Deming[13] estimated the under registration of live births and deaths in India. Using census data, Shapiro[14] applied the technique to calculate the under registration of live births in the USA. In 1968, Wittes & Sidel[19] introduced a generalized capture-recapture method for epidemiology applications, through use of two or more lists. Interest continued to increase in this method, and since the 1990s there was a considerable increase in its use in epidemiologic research.[8]

The objective of this study was to apply the capture-recapture method to estimate live births.

METHODS

In ecology the most straightforward method involves sampling the population, marking the individuals, allowing them to mix with the remaining population, and then taking a new sample. The marked and recaptured individuals are counted, and the total population size is estimated based on the number of individuals exclusively contained in the first sample, (n_A), exclusively in the second sample (n_B) and in both samples ($n_{A\cap B}$). To use this technique, the following assumptions are necessary:[4,8]

1. The population is closed, or in other words there are no births, deaths, or migration in the period between samples;

2. Marking is unique, meaning that each individual is identified by the mark and there is no possibility of losing it;

3. In each sample, every individual has the same probability of being sampled (equiprobability);

4. The two samples are independent, i.e. the event of one individual captured in a sample is independent from the event of one individual captured by another sample; and

5. In each sample, any individual is captured (re-captured) independently from others.

The idea is that if the population in a given area is small, a large number of individuals captured by the second sample will have been marked in the first sample. On the other hand, if the population is large, the second sample will have a small number of individuals marked by the first sample.

In epidemiology, each available list is considered a sample of the population and "being registered on the list" is equivalent to "being captured" in the sample. For more details about the development of this method in epidemiology, refer to Coeli et al,[2] Hook & Hegal,[4,5] *International Working Group for Disease Monitoring and Forecasting* (IWGDMF),[8,9] Wittes et al[18] and Wittes & Sidel.[19]

Huggins[6,7] introduced a procedure to estimate the size of a closed population when the capture probabilities are heterogenous, by modeling based on observed variables such as sex, age and capture history The modeling is performed by calculating the conditional likelihood of captured individuals in order to estimate the parameters.

If p_{ij} is the probability of individual *i* being captured in sample *j*, where $i = 1, 2, 3, \ldots, N$ are the individuals in the population (N is the population size) and $j = 1, 2, \ldots, t$ are the individuals sampled. The conditional likelihood of captured individuals can be expressed in terms of:

[d] United Nations. Statistical Division. Official statistics: principles and practices, organization and management. New York; 2006[cited 2009 Oct 26]. Available from: http://unstats.un.org/unsd/methods/statorg/default.htm

$$\gamma_{ij} = \frac{p_{ij}}{1 - (1 - z_{ij}) \prod_{l=j}^{t} (1 - p^{*}_{ij})}$$

Where p^{*}_{ij} equals p_{ij} when $z_{ij} = 0$, where z_{ij} indicates a prior capture of individual *i*. Alternatively:

$$z_{ij} = \begin{cases} 1, \text{ if individual } i \text{ was captured before sample } j \\ 0, \text{ if individual } i \text{ was not captured before sample } j \end{cases}$$

Therefore, γ_{ij} is the probability of individual *i* being captured in sample *j* given its capture history and given it was captured at least once during the study.

If $x_{ij} = 1$ when individual *i* is captured in sample *j* and $x_{ij} = 0$ if the individual is not captured, individuals are renamed as 1, 2, 3, …, *n* and the non-captured individuals renamed *n*+1, *n*+2, *n*+3, ..., *N*, then the conditional likelihood is proportional to:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{t} \gamma_{ij}^{x_{ij}} (1 - \gamma_{ij})^{(1 - x_{ij})}$$

This only depends on the individuals sampled. The formula for the linear adjustment according to individual and/or environmental characteristics is a logistic function $\{\ln[p_{ij}/(1 - p_{ij})]\}$.[7] According to the author, the variables are normally distributed and their variances can be estimated with a secondary derivatives matrix. Various models can be adjusted based on the observed variables and the capture history.

To estimate the population size, the probability of individual i being captured at least once during the study is:[7]

$$p_i(\beta) = 1 - \prod_{j=1}^{t} (1 - p^{*}_{ij})$$

Where β is the vector of the parameters associated with the adjusted model. An estimate that does not depend on the population size is:

$$\hat{N}(\beta) = \sum_{i=1}^{n} p_i(\beta)^{-1}$$

And the variance is:

$$\text{var}[\hat{N}(\beta)] = \sum_{i=1}^{n} p_i(\beta)^{-2} [1 - p_i(\beta)]$$

The standard error of $\hat{N}(\beta)$ is the square root of its variance. The 95% confidence interval is:

$$95\%CI = \hat{N}(\beta) \pm 1{,}96 * ep[\hat{N}(\beta)]$$

Data were obtained from the Ministry of Health (preliminary data for 2006 were from the Live Birth Information System, *Sistema de Informações sobre Nascidos Vivos*, SINASC) and the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística*, IBGE – Civil Registry of Live Births for 2006). In 2006, the collection form used by IBGE included the number from the Live Birth Certificate, which was used to link the two databases and identify unreported live births. Information on data organization, standardization and linkage between the two databases was previously described.[e]

The total number of live births was estimated with Huggins models, adjusted for the second and third trimesters of 2006, in Sergipe state (Northeastern Brazil). Under-reporting was calculated using the estimates for total live births. Each data source was considered as a sample, or occurrence. SINASC was considered the first occurrence (first capture) and the Civil Registry was considered the second occurrence (re-capture). The geographic analysis considered the microregion of mother's residence. Each microregion was considered as a group of individuals.

Various factors can influence under-reporting of live births, and some characteristics can be incorporated in models, including mother's education, race, number of previous children and the existence of piped water and sewage connection in the residence. Nonetheless, to use this technique the individual variables included in the linear model must be available at all captures,[7] in this case, the two data sources. Only the sex of the child and maternal age were available in the two databases and were considered in the estimation models. The lack of an official document center (*cartório*) for the registration of people in the municipality is an institutional factor that can hinder civil registration of live births; in Sergipe, only two municipalities did not have this type of official document center, in 2006.[f] Therefore, this factor was not considered, since microregions were used in the geographic disaggregation and official document centers were located in all sub-divisions.

Between 4/1/2006 and 9/30/2006, SINASC captured 19,502 live births and the Civil Registry captured 17,254. The creation of pairs from the databases through use of the birth certificate number generated 15,532 pairs. Based on this pairing, the two databases included 21,224 registrations of live births from mothers residing in Sergipe. During the study period, the Civil Registry had 808 registrations with a missing birth certificate number, approximately 4.7% of the database. When the registrations with a birth certificate number are compared to the ones without a number, there is no statistical difference in the average age of the mother (p = 0.992) and in the proportion of the sex of the child (p = 0.510).

[e] Schmid B. Aplicação do método de captura- recaptura para estimar sub-registro de eventos vitais [doctoral thesis]. São Paulo: Universidade de São Paulo; 2010.

[f] Instituto Brasileiro de Geografia e Estatística. Estatísticas do registro civil. Rio de Janeiro; 2006. v. 33.

The distribution of registrations with a missing birth certificate number in the Civil Registry revealed that some microregions have more than 5% of registrations with a missing birth certificate number: Agreste de Lagarto (31.9%), Tobias Barreto (9.2%), Boquim (8.2%) and Japaratuba (7.9%). In the other microregions, the percentage missing varied from 0.4% in Nossa Senhora das Dores to 4.0% in Carira.

For the sex of child, the null hypothesis was that the proportion of girls in each health microregion was the same as the proportion for the state as a whole. For maternal age, the null hypothesis was that the mean age in the health microregion was the same as in Sergipe. The proportion of female children was not statistically different from the mean of state. Mother's age was statistically different. Therefore, only maternal age was considered for inclusion in the estimation models.

Considering $i = 1, 2, 3, ..., N$ as live births and b = the databases (SINASC, Civil Registry), the full linear model for the capture probabilities of SINASC and the Civil Registry were calculated as:

$$1n\left(\frac{p_{ib}}{1-p_{ib}}\right)=\beta_0+\beta_1 g_1+\beta_2 g_2+...+\beta_{12} g_{12}+\beta_{13}\text{idmae}_i+ \beta_{13}(g_1 * \text{idmae}_i)+\beta_{15}(g_2 * \text{idmae}_i)+...+\beta_{25}(g_{12} * \text{idmae}_i) \quad (1)$$

Where,

p_{ib} is the probability of individual i being in database b (SINASC or Civil Registry);

β_0 is the intercept;

β_k is the parameter estimate for group k (k is the microregion);

g_k are the individuals that belong to group k;

β_{13} is the parameter estimate for maternal age;

idmae_i is the age of the mother of individual i, in years, and

β_{k+13} is the parameter estimate for the interaction between group k and maternal age.

In this model the probability of capture varies according to individual characteristics. The notation adopted for the full model was [p(g+idmae+g* idmae) c(g+ idmae +g* idmae)].

Each sub-model generated specific parameter estimates in accordance with the terms specified. For example, in one model the probability of capture by SINASC depends on the group (geographic disaggregation) and maternal age and the probability of capture by the Civil Registry depends only on maternal age, [p(g + idmae) c(idmae)], have different parameter estimates of another model where the probabilities of capture by SINASC and the Civil Registry do not vary across microregions and are independent of maternal age, [p(.) c(.)]. Models with more parameters respond better to the data but the precision of parameter estimates decrease.

One method to evaluate responsiveness and precision is to evaluate the models according to information criteria. One of these methods is the Akaike Information Criteria (AIC), which relates the conditional likelihood of the model to the number of parameters estimated:

$$AIC = -2\ln(L) + 2k$$

Where L is the conditional likelihood of the model and k is the number of parameters. Models with greater responsiveness have higher conditional likelihoods, decreasing the value [– 2ln(L)]. The additive term [+ 2k] penalizes the AIC value. Additional parameters decrease the AIC, since the conditional likelihood increases. However, the sum of the term [2k] balances the AIC value. Therefore, the model with a smaller AIC is the most parsimonious in relation to its likelihood and number of parameters.[g]

Although it is easy to interpret and to select the model the best fits the data, sometimes there can be models with very similar AIC values, which makes selection difficult. Then the models can be calibrated in a way to provide a relative plausibility index, using the normalized Akaike weights. The weights, w_i, are calculated for each model of the group of I candidate models, according to the formula:

$$w_i = \frac{\exp\left(\frac{-2\Delta AIC_i}{2}\right)}{\sum_{i=1}^{I}\left\{\exp\left(\frac{-2\Delta AIC_i}{2}\right)\right\}}$$

Where ΔAIC_i is the difference between the AIC value of model i and the model with the smallest AIC. The weight w_i is considered as evidence that model i is the best model of all the candidate models. Greater model weight can be interpreted to better support the data.[f]

The models were adjusted with MARK,® which estimates the capture-recapture probabilities in accordance with the linear model one desires to adjust for each probability. This way various models were adjusted where the probability of capture by SINASC was adjusted starting with a constant through the full model described in the equation (1). The same procedure was adopted for the capture probability of the Civil Registry, totaling 30 sub-models for the microregions. In addition, four models were adjusted for the entire

[g] Cooch E, White G, organizers. Program MARK: a gentle introduction. Ithaca: Cornell University; 2008 [cited 2009 Mar 02]. Available from: http://www.phidot.org/software/mark/docs/book

state of Sergipe, in which no group was considered and to investigate the influence of maternal age on capture by the databases.

The MARK® program calculated the number of parameters for each model, as well as the conditional likelihood, the AIC value, ΔAIC, w_i, the probabilities for capture (SINASC) and recapture (Civil Registry), and also the derived estimate for total live births ($\hat{N}$).

After obtaining the estimates for total live births, the civil underreporting was calculated as a percentage l:

$$\widehat{SR}_s = \frac{(N_s - n_{RCs})}{\hat{N}_s} \times 100 \quad (2)$$

Where,

$\widehat{SR}_s$ = The percentage of underreporting in subdivision s (microregion).

$\hat{N}_s$ = Estimate of total live births for subdivision s.

n_{RCs} = Number of live births captured by the Civil Registry, in subdivision s.

The cartogram was created with Tabwin® application.

RESULTS

Among the four adjusted models for Sergipe state as a whole, the model with greatest weight {p(idmae) c(.)} was the model where maternal age interferes in the capture of live births by the Civil Registry, with approximately 66% of the total weight among the models.

This model estimated 21,391 (95%CI 21,363;21,423) live births in the second and third trimesters of 2006 in Sergipe, with a probability of capture by SINASC ($\hat{p}$) estimated at 0,912 and for the Civil Registry, 0.804. By deriving the estimates from the number of live births, civil underreporting was calculated at 19.3%. By using the estimate given in (2), with a 95%CI for total live births, the variation in underreporting was estimated between 19.2% and 19.5%.

When including the microregions of maternal place of residence as groups of live births, only 5 of the 30 models fitted to the data demonstrated any relative weight when evaluating the AIC and conditional likelihood criteria. The model with the greatest weight {p(g) c(g + idmae + g*idmae)}, 67%, was selected (Table 1).

The probability of capture by SINASC was high in all microregions, varying from 0.69 in Agreste de Lagarto to 0.95 in Estância and Nossa Senhora das Dores. In Agreste de Lagarto there are a large number of records in the Civil Registry with a missing birth certificate number (more than 31%), which limited the pairing of the databases. Also, in Tobias Barreto, Japaratuba and Boquim, the percentage of records with a missing birth certificate number was greater than 5%. Besides these microregions, where matching was affected by the lack of a birth certificate number, only the Propriá microregion had a SINASC capture probability less than 0.90 (Table 2).

The capture probabilities for the Civil Registry were noticeably smaller than for SINASC. In the Civil Registry, the greatest capture probability estimated was in the Aracaju microregion (0.85) and the smallest was in Sergipana do Sertão do São Francisco (0.71), excluding the mircoregions with problematic matching (Table 2).

The total estimated live births ($\hat{N}$) was very close to the total measured in all the microregions, due to the high overlap of the lists. Again in Agreste de Lagarto, the high percentage of Civil Registry records with a missing birth certificate number created low overlap (relatively low $n_{S \cap RC}$) and therefore inflated the total estimate of live births. The absolute difference between the estimated live births and captured live births was less than 20 in almost all the microregions where the percentage of missing birth certificate numbers in the Civil Registry was less than 5%, and reached only 2 live births in Carira and Nossa Senhora das Dores (Table 3).

Underreporting across microregions varied from slightly more than 12% in Baixo Contiguiba to almost 27% in Sergipana do Sertão do São Francisco. Estimated underreporting in Agreste de Lagarto exceeded 40%, although this finding should be interpreted with caution. Civil underreporting was less in mircoregions located in the central part of the state, Aracaju, Baixo Contiguiba and Agreste de Itabaiana (< 15% of live births). As the microregions of maternal residence increase in distance from the central area, civil underreporting of live births increases (Figure 1).

When considering maternal age from the Civil Registry, there was a subtle decreasing trend in underreporting as maternal age increased (Figure 2). The Agreste de

Table 1. Results of the Huggins models for closed populations of live births, which show some weight, according to microregion of maternal residence. Sergipe state (Northeastern Brazil), second and third trimesters of 2006.

Model	w_i	No. Parameters
{p(g) c(g + idmae + g* idmae)}	0.66837	39
{p(g + idmae) c(g + idmae + g* idmae)}	0.27672	40
{p(g + idmae + g* idmae) c(g + idmae + g* idmae)}	0.05485	52
{p(g) c(g + idmae)}	0.00004	27
{p(g + idmae) c(g + idmae)}	0.00002	28

Table 2. Probability estimates for capture by SINASC ($\hat{p}$) and the Civil Registry ($\hat{c}$) for the model {p(g) c(g + idmae + g*idmae)}, according to the microregion of maternal residence. Sergipe state (Northeastern Brazil), second and third trimesters of 2006.

Microrregion	$\hat{p}$	95%CI	$\hat{c}$	95%CI
Sergipana do Sertão do São Francisco	0.912	0.896;0.926	0.714	0.692;0.735
Carira	0.937	0.912;0.956	0.822	0.787;0.852
Nossa Senhora das Dores	0.949	0.926;0.965	0.823	0.789;0.852
Agreste de Itabaiana	0.932	0.917;0.944	0.861	0.843;0.878
Tobias Barreto[a]	0.874	0.849;0.896	0.719	0.690;0.746
Agreste de Lagarto[a]	0.694	0.655;0.731	0.516	0.486;0.546
Propriá	0.876	0.849;0.898	0.807	0.779;0.832
Cotinguiba	0.914	0.883;0.938	0.829	0.793;0.860
Japaratuba[a]	0.874	0.839;0.903	0.775	0.738;0.809
Baixo Cotinguiba	0.921	0.900;0.938	0.870	0.847;0.890
Aracaju	0.944	0.938;0.949	0.848	0.840;0.856
Boquim[a]	0.886	0.868;0.902	0.749	0.728;0.770
Estância	0.949	0.935;0.960	0.799	0.777;0.820

[a] These microregions had more than 5% of records in the Civil Registry missing the birth certificate number, which impaired matching.

Table 3. Distribution of capture by the two databases [n_S, n_{RC}, $n_{(S\cap RC)}$ and r] and derived estimates for the model {p(g) c(g + idmae + g*idmae)}, according to the microregion of maternal residence. Sergipe state(Northeastern Brazil), second and third trimesters of 2006.

Microrregion	n_S	n_{RC}	$n_{(S\cap RC)}$	r	$\hat{N}$	95%CI
Sergipana do Sertão do São Francisco	1,692	1,359	1,210	1,841	1,855	1,849;1,867
Carira	541	479	445	575	577	576;583
Nossa Senhora das Dores	565	493	464	594	596	594;601
Agreste de Itabaiana	1,505	1,399	1,296	1,608	1,616	1,611;1,625
Tobias Barreto[a]	986	833	709	1,110	1,128	1,120;1,142
Agreste de Lagarto[a]	1,056	868	545	1,379	1,521	1,482;1,575
Propriá	885	821	711	995	1,011	1,003;1,024
Cotinguiba	491	448	406	533	537	534;545
Japaratuba[a]	549	493	424	618	628	623;640
Baixo Cotinguiba	949	901	826	1,024	1,030	1,027;1,039
Aracaju	7,286	6,590	6,183	7,693	7,717	7,708;7,730
Boquim[a]	1,661	1,434	1,245	1,850	1,874	1,865;1,890
Estância	1,336	1,136	1,068	1,404	1,408	1,405;1,414

n_S = total live births captured by SINASC. n_{RC} = total live births captured by the Civil Registry. r = total distinct live births identified ($r = nS + nRC - n_{(S\cap RC)}$)
[a] These microrregions had more than 5% of Civil Registry records with missing birth certificate numbers, which impaired matching

Lagarto microregion was not included in the analysis due to the large number of missing birth certificate numbers in the Civil Registry.

DISCUSSION

The findings suggest that the probability of civil underreporting increases as maternal age decreases (Figure 2); while underreporting is also influenced by proximity to the central area of the state. The microregion of Baixo Contiguiba, did not show high underreporting, although it had the lowest mean maternal age; the opposite occurred in Tobias Barreto, the microregion farthest from Aracaju, which has the highest average maternal age and also high underreporting. Nonetheless, matching in Tobias Barreto was harmed since almost 10% of records in the IBGE database were missing the birth certificate number.

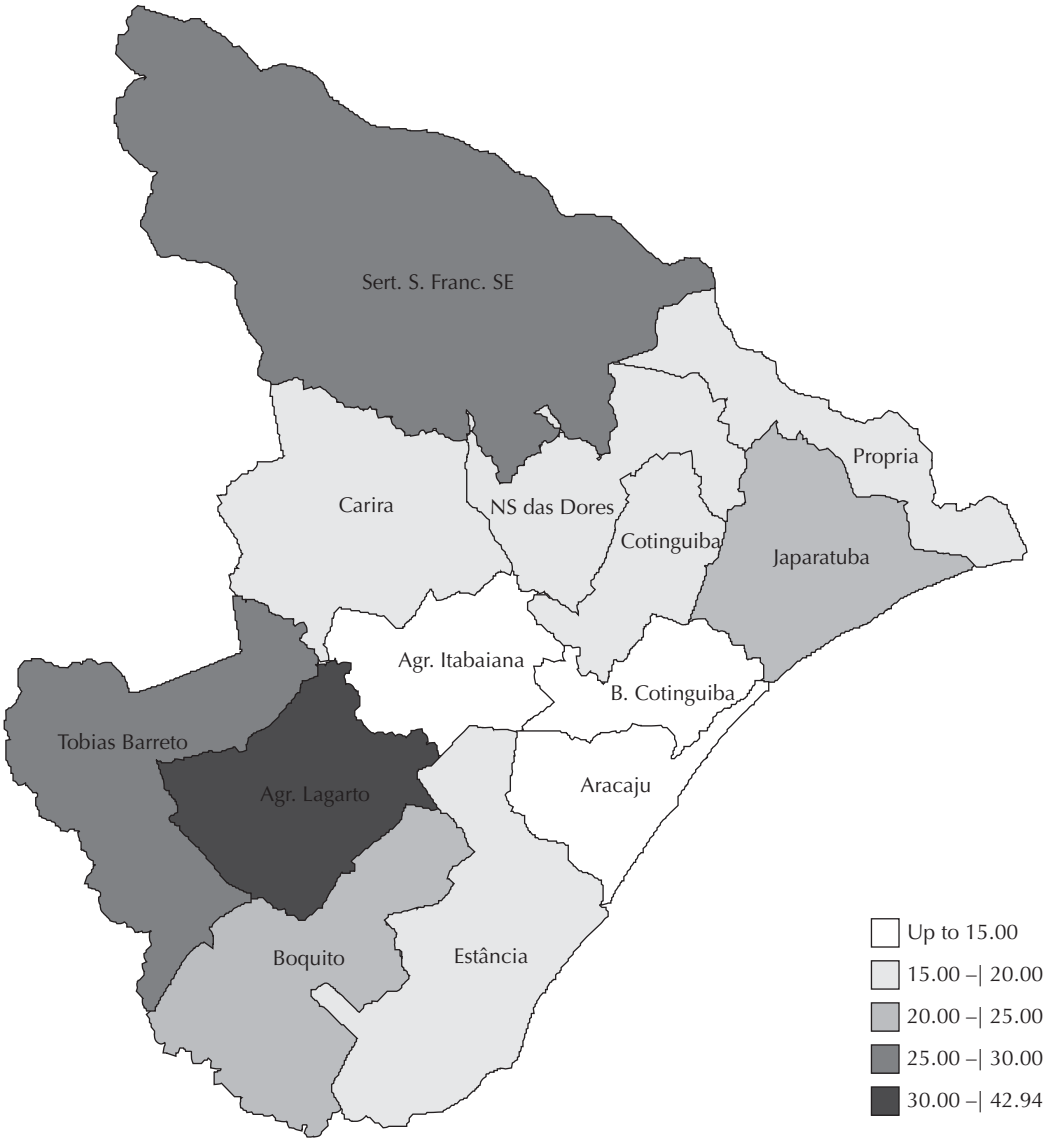
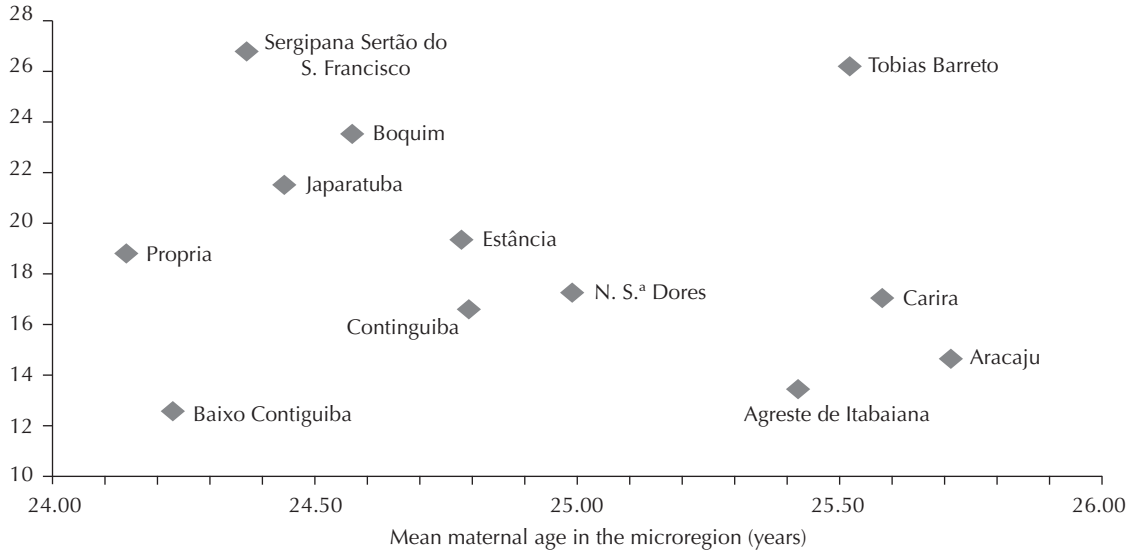


Figure 1. Civil underregistration, according to the microregion of maternal residence. Sergipe state (Northeastern Brazil), second and third trimesters of 2006.

It is important to discuss the assumptions in estimation by capture-recapture. It assumes a closed population where there is no migration nor births or deaths during the study period. In this study, the population to estimate was the total number of live births. The event of a birth happens once, and the number of live births is constant in the period and the geographic area used. Clearly, neonatal and/or child deaths can occur during the period analyzed and the families may have moved to another municipality of federative unit after the birth. Nonetheless, these factors do not alter the size of the "population of live births of mothers residing in Sergipe", since a death and change of address do not change the fact that the baby was born alive and the mother resided at the given location.

In regards to the unique marking, the failings in filling out the birth certificate number harmed the deterministic linkage utilized. Of the 17,254 live births present in the Civil Registry, 808 (4.7%) had a missing birth certificate number, which generates questions concerning the extent that these 808 records are able to be matched with the SINASC records.

When using the results of the model for the entire state of Sergipe, the estimated capture probability for SINASC was 0.912 and for the Civil Registry 0.804. The probability for one live birth to be included in the two databases would therefore be, 0.912 * 0.804 = 0.733. Of the 808 records with a missing birth certificate number in the Civil Registry, 0.733 * 808



Note: Excludes the microregion of Agreste de Lagarto, due to impaired matching

Figure 2. Civil underregistration in the microregions of maternal residence. Sergipe state (Northeastern Brazil), second and third trimesters of 2006.

= 592 would also have been captured by SINASC. Using these numbers, one can assume that the correct number of pairs formed by matching between the two databases would be 15,532 + 592 = 16,124 live births. Although it is possible to estimate the probable number of pairs, performing this relation after estimation would be imprecise, since the available variables have little discriminatory power, which implies finding more than one record with the same characteristics. The year 2006 was the first time that the birth certificate number was collected on the IBGE questionnaires, and in subsequent years data quality may improve, with less discrepancy between SINASC and the Civil Registry.

There are two issues related to the assumption of equal probability, where each individual has the same probability of capture in a given sample. The first issue concerns the influence of maternal age on capture by the Civil Registry. This source of heterogeneity can be included in the models and was found significant. The second issue concerns the time interval used for matching the databases. The Civil Registry database at the IBGE is organized by the year of data collection. The records and their respective dates of birth are not lost, although they are organized according to the year collected. Therefore, the total number of live births in 2006 will vary as the year of data collection progresses. Evidently the variation will not be large, since the late records will be residual with the passage of time. As addressed by Oliveira & Simões,[h] the coverage by the Civil Registry increases when analyzing records one year after the live birth. Among those born in the second trimester of 2006, close to 94% were registered in the third trimester. Since access to data was restricted to collection year 2006, births in the third trimester of 2006 and registered beginning in 2007 were not detected in this study. Therefore, there was not equal probability of capture in the Civil Registry among individuals born in the third trimester of 2006 in relation to those born in the second trimester. Since access to the IBGE database was restricted to collection in 2006, the matching between SINASC and Civil Registry for live births in the third trimester has additional limitations, in addition to missing birth certificate numbers. It can therefore be deduced that the number of pairs presented in the 2 x 2 table for capture by $n_{(A \cap B)}$, would be larger than measured, demonstrating the large overlap between the two databases.

In relation to SINASC, the data are consolidated by sending information from state health secretaries to the Ministry of Health and are eventually updated.[i] The number of live births available on the internet site differs from the preliminary data available for this study, in 2007. The dynamism of the two databases should be considered, characterized by continual changes in the total number of records.

The assumption that the capture-recapture probability of one individual does not affect the probabilities for others was not violated because the birth of a baby does

[h] Oliveira ATR, Simões CCS. Perfil dos municípios com informações precárias sobre eventos vitais. Brasília: Rede Interagencial de Informações para a Saúde, Organização Pan-Americana da Saúde; 2005.
[i] Ministério da Saúde. Departamento de Informática do SUS – DATASUS. Informações em Saúde. Brasília; 2009[cited 2009 May 20]. Available from: http://www.datasus.gov.br/

not interfere with the identification of another baby by SINASC, as well as the Civil Registry. In regards to multiple births – the fact that the live births are or are not registered together – does not meet this assumption. Nonetheless, the occurrence of multiple gestations is very rare, and the number of twins born alive does not harm this assumption.

Regarding the independence of the samples, here databases, the large overlap between them (large $n_{A\cap B}$) suggests a positive dependence. This indicates that the number of estimated live births would not be much larger than the distinct individuals identified in the two databases. In order to quantify the dependence between two epidemiologic sources (lists), Brenner[1] included a correction factor, for the probability of an individual to be captured in the two lists. The author simulates situations where both the capture probabilities for each source (n_A and n_B) and the correction factors that modify the probability for inclusion in the two sources ($n_{A\cap B}$) vary, creating positive and negative dependence in order to observe the behavior of the under- and over-estimation factor. In the case of negative dependence, the investigator concluded that overestimation of the total population size would be more serious when the lists have low coverage of individuals and a small probability for including the individuals. In cases of positive dependence, the author affirms that lists with a high inclusion probability have smaller underestimation factors. When considering this type of dependence, estimates of population size will still be closer to reality than simple aggregation of the sources.[1]

In regards to the model selected, Tilling & Sterne[16] and Tilling et al[19] applied the Huggins model for estimating epidemiologic data, demonstrating the viability of the model for these types of data. Also, use of the conditional likelihood for the observed individuals allows flexibility to include covariates to model the capture probabilities with adjusted linear models. We believe the Huggins model can continue to be applied to estimate total live births through linkage of SINASC and the Civil Registry, as long as future studies resolve the problem encountered with the IBGE database regarding the equal probability of capture during the study period used to identify records.

In conclusion, the results of the present study suggest minimal values for civil underreporting and SINASC coverage and that it is possible to apply the capture-recapture methodology to estimate underreporting of live births. In the case of large overlap between two databases, the International Working Group for Disease Monitoring and Forecasting[9] recommends aggregation of the sources and turning them into one source. An alternative may be the deterministic linkage of SINASC and the Civil Registry and probabilistic association between this new database and other sources that can be used when applying capture-recapture, such as enrollment in the Family Health Programs and the Hospital Information System, for example.

REFERENCES

1. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology*. 1995;6(1):42-8. DOI:10.1097/00001648-199501000-00009
2. Coeli CM, Veras RP, Coutinho ESF. Capture-recapture methodology: an option for surveillance of non-communicable diseases in the elderly. *Cad Saude Publica*. 2000;16(4):1071-82. DOI:0.1590/S0102-311X2000000400025
3. Dunn J, Andreoli SB. Método de captura e recaptura: nova metodologia para pesquisas epidemiológicas. *Rev Saude Publica*. 1994;28(6):449-53. DOI:10.1590/S0034-89101994000600009
4. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev.* 1995;17(2):243-64. Erratum in: *Am J Epidemiol.* 1998;148(12):1219.
5. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol.* 1999;52(10):917-26. discussion 929-33.
6. Huggins RM. On the statistical analysis of capture experiments. *Biometrika*. 1989;76(1):133-40. DOI:10.1093/biomet/76.1.133
7. Huggins RM. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*. 1991;47(2):725-32. DOI:10.2307/2532158
8. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. *Am J Epidemiol.* 1995;142(10):1047-58.
9. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: applications in human diseases. *Am J Epidemiol.* 1995;142(10):1059-68.
10. Laporte RE. Assesing the human condition: capture-recapture techniques. *BMJ.* 1994;308(6920):5-6.
11. Mccarty DJ, Tull ES, Moy CS, Kwoh CK, Laporte RE. Ascertainment corrected rates: applications of the capture-recapture methods. *Int J Epidemiol.* 1993;22(3):559-65. DOI:10.1093/ije/22.3.559
12. Pradel R. Utilization of capture-recapture for the study of recruitment and population growth rate. *Biometrics*. 1996;52(2):703-9. DOI:10.2307/2532908
13. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration. *J Am Stat Assoc.* 1949;44(245):101-15.
14. Shapiro S. Estimating birth registration completeness. *J Am Stat Assoc.* 1949;45:261-4. DOI:10.2307/2280684
15. Tilling K, Sterne JAC. Capture-recapture models including covariate effects. *Am J Epidemiol.* 1999;149(4):392-400.
16. Tilling K, Sterne JAC, Wolfe CDA. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol.* 2001;30(6):1351-9. DOI:10.1093/ije/30.6.1351
17. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis.* 1974;27(1):25-36. DOI:10.1016/0021-9681(74)90005-8
18. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *J Chronic Dis.* 1968;21(5):287-301. DOI:10.1016/0021-9681(68)90038-6

Article based on the doctoral thesis by Schmid B, presented to the Faculdade de Saúde Pública at Universidade de São Paulo in 2010.
The authors declare no conflicts of interests.