Centro de Estudos da Metrópole

Nota técnica:

Instruções para o uso dos bancos de microdados das amostras dos Censos Demográficos Brasileiros (1960 a 2010)

Rogério Jerônimo Barbosa

antrologos@gmail.com

Outubro de 2013

Sumário

1. Introdução

Neste ano de 2013, o Centro de Estudos da Metrópole (CEM) disponibilizou em seu site (http://www.fflch.usp.br/centrodametropole/) os bancos de dados completos das amostras dos Censos Demográficos do IBGE produzidos entre 1960 e 2010. Este documento reúne um conjunto de dicas, orientações e instruções para facilitar o uso dos microdados[1].

O tópico a seguir apresenta brevemente a diferença entre as noções de amostra e universo. Uma vez que divulgamos as informações das amostras, é importante especificar a natureza dos dados que divulgamos. As seções subsequentes são dedicadas a cada uma das edições, trazendo particularidades de cada banco de dados e indicando quando são necessárias transformações de variáveis ou seleções de casos. A última seção traz algumas diretrizes que devem ser observadas para maximizar a validade das comparações longitudinais que fazem uso de mais de um Censo.

1.1. Amostra e Universo

Censos demográficos são pesquisas extensivas e caras. A população brasileira é muito grande e o território nacional bastante extenso; o que eleva muito as dificuldades de coordenação e articulação para a coleta e processamento das informações. Justamente por esse fato, os questionários dos Censos brasileiros sempre foram restritos, contendo entre 9 e 30 questões, ao longo do período entre 1940 e 2010 (CAVENAGHI, 2010). Tais restrições de ordem prática limitam intensamente o escopo dos temas investigados – e certamente enseja inúmeras disputas (acadêmicas, mas principalmente políticas) com respeito aos itens que devem ou não ser perguntados.

Para contornar esse problema, desde 1960, o Censo Demográfico consiste de duas pesquisas realizadas simultaneamente. Naquele ano, o IBGE introduziu uma pesquisa amostral: uma fração da população responde um questionário mais extenso, que abarca e contém as mesmas perguntas feitas à população como um todo, isto é, ao universo ou não amostra. Com isso, se tornou possível realizar estudos mais aprofundados que não seriam logística e economicamente viáveis na escala populacional. Em contrapartida, com o advento da pesquisa amostral, o número de tópicos investigado no universo foi se tornando mais reduzido.

[1] Microdados são bancos de dados em que os casos/registros (linhas) são as próprias unidades de coleta de informação. No caso de pesquisas sociais por questionários, trata-se dos indivíduos respondentes. Nos bancos dos Censos disponibilizados pelo CEM, as pessoas são a unidade de análise. A noção de microdados se contrapõe à de dados agregados, nos quais os casos são coletividades ou unidades-resumo das informações obtidas a partir dos microdados. Tais agregações podem ser, por exemplo, domicílios, setores censitários, áreas de ponderação, distritos, municípios, setores de atividade econômica etc. A partir dos microdados é possível produzir dados agregados.

A fração amostral era de 25% nos anos de 1960, 1970, 1980 e 1991 e passou para 10% em 2000 e 2010. Trata-se de um contingente bastante extenso, que abarca em torno de 25 milhões de pessoas em cada uma das edições. Deste modo, possui uma ínfima margem de erros. Assim, desde 1960, os dados do universo apenas se revelam interessantes para análises de níveis geográficos muito desagregados (setores censitários, que reúnem cerca de 230 domicílios ou aproximadamente 700 indivíduos). Desde 2000, não disponibiliza mais os microdados do universo – sob a justificativa de que o detalhamento das informações poderia violar o sigilo dos respondentes. Deste modo, o que os pesquisadores têm à mão basicamente são os microdados da amostra e os dados agregados por setor censitário. No site do CEM (www.centrodametropole.org.br), disponibilizamos ambos – mas neste documento tratamos apenas dos microdados da amostra.

2. Censo de 1960

2.1. Histórico

O Censo de 1960 tem uma história muito peculiar e que teve consequências importantes e graves sobre a qualidade dos dados (ver BARBOSA *et al*, 2013).

Essa seria a primeira edição em que o processamento de dados seria realizado com auxilio de computadores modernos, com vistas a aumentar a velocidade do processo, bem como sua confiabilidade e validade (minimizando erros). No entanto, por razões não muito claras, o emprego desses processos computadorizados não foi possível e os relatórios oficiais da época se baseiam em análises realizadas do modo tradicional, como havia sido feito para os censos de 1940 e 1950.

Apenas parte dos questionários da amostra (cuja fração era 25%) foi digitalizada, referentes a algumas unidades da federação. Com esse conjunto de dados não é possível fazer inferências para o país como um todo. Já no início da década de 1970, para viabilizar análises verdadeiramente nacionais, alguns pesquisadores interessados nas tendências das desigualdades de renda, extraíram uma amostra aleatória estratificada e autoponderada de domicílios a partir do conjunto completo de questionários, que representava uma fração de aproximadamente 1,25% da população (quase 900 mil casos de indivíduos). Os questionários selecionados foram digitalizados. Este é o banco de dados que disponibilizamos. O próprio IBGE, até agora, não disponibilizava esses dados para venda em sua loja virtual – e a circulação do banco, em boa medida, se fazia entre pesquisadores. O professor e pesquisador Carlos Antônio Costa Ribeiro foi quem compartilhou esses dados com o Centro de Estudos da Metrópole.

2.2. Uso dos dados

Devido ao fato de que a fração amostral é mais reduzida, os microdados do Censo de 1960 apenas são representativos para as unidades da federação (em contraposição às amostras de 25% dos outros censos, que permitem análises intramunicipais).

Devido ao fato de que sua amostra probabilística é autoponderada, não é necessária a aplicação de pesos amostrais. Todas as proporções, médias e estatísticas pontuais são não enviesadas. No entanto, desconhecemos detalhes dos procedimentos empregados na construção da amostra e, deste modo, não sabemos se os erros amostrais podem ser calculados sob a suposição de amostragem aleatória simples ou se requerem especificação de desenho de amostra complexa. Se o caso for este último, as medidas dos erros amostrais para estatística inferencial são subestimadas – ou seja, testes de hipótese apresentam significância estatística quando, de fato, não deveriam.

O censo de 1960 entrevistou todos os presentes no domicílio no momento da aplicação do questionário. Deste modo, há recenseamento de moradores presentes, moradores ausentes (cujas informações foram fornecidas pelos presentes) e não moradores presentes (isto é, indivíduos visitantes que estavam no domicílio durante a aplicação). Quaisquer análises utilizando esses microdados devem excluir necessariamente os não moradores presentes.

3. Censo de 1970

A amostra do Censo de 1970 que disponibilizamos é aquela com fração de 25%. Deste modo, é representativa inclusive para áreas intramunicipais. A amostra não é autoponderada, deste modo, são necessários os pesos amostrais para estimação de estatísticas não enviesadas. Os pesos contêm também fatores de expansão, ou seja, todas as frequências de variáveis revelam totais populacionais estimados a partir da amostra. Contudo, pesos com fatores de expansão sempre levam à subestimação dos erros amostrais – deste modo, os resultados de testes de hipótese não são válidos. Uma alternativa é a aplicação de pesos analíticos, que podem facilmente ser construídos a partir da equação:

$$Peso\ analítico = n \times \frac{Peso\ Expandido}{N}$$

Os pesos analíticos apenas reduzem a subestimação dos erros, sem resolver completamente o problema. Não se alteram as estimativas pontuais. O IBGE não

especifica como incorporar o plano amostral para a correção da autocorrelação e para a estimação não enviesada dos erros padrão.

O censo de 1970 também entrevistou não moradores presentes no domicílio durante o momento da entrevista e, por isso, também requer seleção apenas dos moradores para a realização de quaisquer análises.

Os microdados originais do Censo de 1970 não continham um número identificador dos domicílios – para que pudéssemos calcular características agregadas ou parear o banco de indivíduos com o banco de domicílios. No entanto, a ordenação original dos casos dentro do banco de dados agrupava e hierarquizava os indivíduos dentro dos domicílios. A partir desse ordenamento, a equipe de análise de dados do CEM criou IDs e incorporou as características dos domicílios como variáveis no banco de indivíduos. Os valores dessas variáveis são sempre idênticos para pessoas que vivem juntas num mesmo domicílio.

4. Censo de 1980

A amostra do Censo de 1980 também tem fração de 25%, sendo representativa inclusive para áreas intramunicipais. É necessário o uso de pesos amostrais para estimação não enviesada dos dados descritivos. Na ausência de informações oficiais do IBGE sobre um método de incorporação do desenho de amostragem complexa, recomendamos o uso de pesos analíticos sem fator de expansão para testes de hipótese, conforme exposto no tópico anterior.

A partir do Censo de 1980, apenas são entrevistados os moradores (presentes ou não) do domicílio. Não é necessário excluir casos das análises.

5. Censo de 1991

A amostra do Censo de 1991 tem fração de 25% e, igualmente, é representativa de áreas intramunicipais. São necessários pesos amostrais para a estimação não enviesada de estatísticas pontuais e, pelas mesmas razões já apontadas anteriormente, recomendamos o uso de pesos analíticos para testes de hipótese.

6. Censo de 2000

A amostra do Censo de 2000 tem fração de 10% e é representativa de áreas intramunicipais. São necessários pesos amostrais para a estimação não enviesada de estatísticas pontuais e, pelas mesmas razões já apontadas anteriormente, recomendamos o uso de pesos analíticos para testes de hipótese.

Em censos demográficos, é comum que nem todos os indivíduos sejam alcançados e respondam o questionário. Há populações de difícil acesso, como é o caso de moradores de rua, há domicílios que permaneceram fechados durante todo o período de aplicação. Isto produz um contingente de "casos perdidos" (*missing cases*), levando à subestimação dos totais populacionais. Desde o Censo de 2000, o IBGE adotou procedimentos de estimação desses casos através de procedimentos estatísticos estratificados que aproximam um número de moradores para os domicílios fechados e imputa nesses casos algumas características, com base em uma distribuição de probabilidades. Todos os casos que foram imputados são identificados através de variáveis dicotômicas (0/1) denominadas "marcas de imputação". Há uma variável adicional, marca de imputação, para quase todos os quesitos originais do banco de dados, praticamente duplicando o volume de dados.

7. Censo de 2010

A amostra do Censo de 2000 tem fração de 10% e, igualmente, é representativa de áreas intramunicipais. São necessários pesos amostrais para a estimação não enviesada de estatísticas pontuais e, pelas mesmas razões já apontadas anteriormente, recomendamos o uso de pesos analíticos para testes de hipótese.

O banco de dados do Censo de 2010 que disponibilizamos contém três tipos de registro ou casos:

(1) **Pessoas:** casos de indivíduos residentes nos domicílios entrevistados
(2) **Mortalidade:** Casos relativos a indivíduos falecidos que foram moradores permanentes do domicílio no último ano, juntamente com alguma das famílias que nele habitam (podendo ser parentes ou não).
(3) **Emigração:** Casos de indivíduos que residiram no domicílio e emigraram para outros países.

Análise usual dos dados requer apenas a utilização dos casos de pessoas. Deste modo, é necessário não incluir os outros tipos de registro.

Assim, como o Censo de 2000, a edição de 2010 também contém variáveis que indicam marcas de imputação de casos perdidos.

8. Comparação entre censos e padronizações

A comparação de dados provenientes de diferentes Censos não é automática. Nem todas as perguntas estão presentes ao longo de todo o escopo de tempo. Além disso, muitas das perguntas "sempre presentes" tiveram enunciados, alternativas e escopo de aplicação alterado. Em alguns casos, bastam algumas transformações ou padronizações de variáveis. Noutros, a comparabilidade é apenas aproximada. E, em

raras ocasiões, não é viável. Seria necessário um trabalho específico para da conta de todas as possibilidades e formas de garantir comparabilidade – o que está fora dos propósitos deste documento. Mas algumas diretrizes e direções podem ser apontadas.

A primeira delas, já mencionada, refere-se à estabilidade das perguntas, alternativas e seus significados. Antes de produzir análises, é sempre importante conhecer os questionários, dicionários de variáveis e documentação auxiliar. Deve-se observar a quais grupos ou subgrupos da amostra as questões foram dirigidas (questões sobre educação geralmente são feitas apenas a pessoas com 5 anos de idade ou mais; questões sobre trabalho e rendimento, à pessoas com 10 anos ou mais; sobre fecundidade, a mulheres com 15 anos ou mais; e assim por diante). Uma mesma questão, porém em Censos diferentes, pode ter diferentes escopos de aplicação. O segundo ponto a observar é sobre a estabilidade da forma de enunciação: estímulos diferentes aos entrevistados produzem diferentes respostas. Por fim, quanto às alternativas de resposta, para além da semelhança nominal que possa existir entre rótulos e descrições, é preciso verificar, nos manuais e documentações oficiais, se não há outros significados subsumidos e não explicitados. Um exemplo importante é o caso da alternativa "pardo" na questão sobre cor/raça: os censos de 1960 e 1980 subsumem os indígenas dentro da categoria "pardo"[2], apenas em 1991 os indígenas são captados por uma opção de resposta específica (Osório, 2003). Excelentes fontes de informação sobre esses pontos são os próprios Questionários da Amostra dos Censos (documento cujo código identificador no IBGE é CD 1.02) e o Manual do Recenseador (CD 1.09), disponíveis no site do IBGE (dentro da seção 'Produtos e Serviços > Biblioteca > Instrumentos de Coleta'). Em especial, esse último documento pode ser bastante útil, por conter definições explícitas e completas dos significados das alternativas, enunciados, bem como dos próprios propósitos dos quesitos.

A segunda diretriz diz respeito às comparações territoriais. O número de municípios variou muito ao longo do tempo, passando de 3951 em 1970 para 5565 em 2010. Os novos municípios são fruto não apenas de divisões de municípios maiores, como também de fusões e redelineamentos que envolveram diversas localidades. Deste modo, as unidades políticas "constantes no tempo" não são exatamente constantes. Parcelas mais ricas podem ter se emancipado ou parcelas mais pobres podem ter sido segregadas – o que certamente varia imensamente em cada um dos casos. Algumas estratégias envolvem a demarcação de áreas minimamente comparáveis no tempo (REIS *et al*, 2005; RESENDE, CARVALHO, SAKOWSKI, 2013) – ou seja, aqueles municípios cujas fronteiras permaneceram constantes ao longo do tempo (que eram em número de 3657, entre 1970 e 2000). Mas outros caminhos são possíveis, desde que se leve em conta a heterogeneidade implicada nas redefinições territoriais. O número de Unidades da Federação também variou: o estado da

[2] O censo de 1970 não contém questão sobre classificação racial.

Guanabara (antigo Distrito Federal) foi incorporado ao Rio de Janeiro; Mato Grosso e Mato Grosso do Sul se separaram; Tocantins se emancipou de Goiás; os territórios federais foram abolidos: Fernando de Noronha foi anexado à Pernambuco; Rondônia, Roraima e Amapá se tornaram estados.

O terceiro ponto diz respeito às moedas e seus valores. Durante o período entre 1960 e 2010, o Brasil implementou diversas políticas monetárias que alteraram a unidade de medida e de valor das moedas – não raro, em meio a períodos de intensa inflação e crise. É preciso fazer as conversões monetárias e deflacionar os valores. Sugerimos a adoção dos deflatores de Corseuil e Fogel (2002), bastante utilizados e testados. É possível encontrar dados atualizados periodicamente para esses índices no site do IpeaData (http://www.ipeadata.gov.br/).

O quarto ponto diz respeito aos períodos de captação da informação de determinadas variáveis. Um exemplo ilustrativo é o caso da informação ocupacional. Em 1960 e 1970, pergunta-se sobre a ocupação habitual do indivíduo. Ou seja, é possível que um indivíduo declare ser "cozinheiro", por exemplo, mesmo que não estiver desenvolvendo, naquele momento, as funções a essa atividade. Ele "é" cozinheiro, mesmo que não "esteja" trabalhando nesse ramo. O Censo de 1980 indaga sobre a ocupação habitual nos últimos 12 meses (que tem certo paralelo com os quesitos dos dois censos anteriores) e também a ocupação desempenhada na semana da aplicação do questionário. O censo de 1991 apenas indaga sobre a ocupação nos últimos 12 meses – e os censos de 2000 e 2010 apenas sobre a ocupação na semana de referência. Essas alterações têm implicações importantíssimas sobre o formato da estrutura ocupacional identificada em cada período, bem como sobre taxas de atividade, ocupação e desemprego. É sempre necessário observar os períodos de referência e captação das questões.

O quinto ponto diz respeito aos sistemas de classificação empregados em determinadas variáveis. Alguns casos são particularmente importantes: ocupação, setores de atividade econômica, níveis de ensino, cursos realizados no sistema formal de educação e religião. Os sistemas ocupacionais e setoriais se alteraram profundamente, em parte como resultado do aprimoramento das medidas, em parte como consequência da própria diversificação do mercado de trabalho, que contém crescentemente funções, posições e empresas mais diversificadas. Algumas ocupações se tornaram mais especializadas, requerendo agora ensino superior, como é o caso da Fisioterapia (anteriormente uma profissão de nível técnico) – ou seja, mudam de caráter, o que tem consequência para a definição do perfil dos ocupantes, da o patamar esperado de rendimentos e sua dispersão, para as possibilidades de organização em torno de associações, sindicatos etc. Há também profissões "novas", que decorrem da especialização (como Gerentes de RH) ou do desenvolvimento tecnológico (técnicos e profissionais de TI). Inversamente, houve ocupações que

praticamente se extinguiram, como é o caso dos datilógrafos. Todas essas alterações fizeram com que não existisse um pareamento unívoco entre os sistemas ocupacionais e setoriais adotados em cada um dos censos (apesar da grande semelhança entre as classificações de 1960 a 1991). Qualquer comparação longitudinal apenas se torna válida se fizer uso de categorias mais agregadas e construídas de forma customizada pelo pesquisador – mas que são, no entanto, também mais imprecisas. Para uma reflexão mais completa sobre esse ponto, ver Jannuzzi (2003). Com respeito aos níveis educacionais, destacamos que houve durante o período considerado, diversas reformas federais importantes do sistema educacional (destacando-se as de 1961, 1968, 1971, 1982 e 1996). Algumas delas alteraram a organização e o conteúdo do currículo básico e também a própria duração do período de escolarização formal obrigatória. Em 1960, por exemplo, o 1º Grau (hoje denominado Ensino Fundamental) era constituído de apenas 6 anos, nos censos de 1970 a 2000 o nível equivalente de ensino se compunha de 8 anos e, em 2010, de 9 anos. Isto implica que o escopo de variação das possibilidades de realização educacional não permaneceu constante, logo, médias de anos de estudo não são, por exemplo, comparáveis. Um indivíduo com 6 anos de estudo em 1960 é equivalente a um indivíduo com 9 anos em 2010. Uma boa referência sobre a mensuração dos níveis educacionais nos Censos é Rigotti (2004). Os Censos também captam o curso superior realizado pelo indivíduo. Porém as classificações das áreas e a quantidade de cursos arrolados na listagem variam bastante e são poucos aqueles que são estritamente comparáveis ao longo dos anos. O mesmo, com respeito à variável sobre religião. A despeito da recente pluralização e diversificação do campo religioso brasileiro – que justifica o aumento de categorias e alternativas – é possível dizer que o reduzido número de opções nos censos de 1960 e 1970 não faz jus completamente à heterogeneidade já existente à época. Em todos esses casos é preciso definir categorias mais agregadas que se tornam minimamente comparáveis.

9. Referências Bibliográficas

BARBOSA, R. J.; MARSCHNER, M., FERRARI, D.; SILVA, P.; PRATES, I.; BARONE, L. S. "Ciências sociais, censo e informação quantitativa no Brasil: entrevista com Elza Berquó e Nelson do Valle Silva". Novos Estudos Cebrap, n.95, 2013.

CAVENAGHI, Suzana. "A experiência do Brasil nos módulos de domicílio e família nos Censos Demográficos". Seminario-taller "Los censos de 2010 y las condiciones de vida". CEPAL - Comisión Económica para América Latina y el Caribe, Centro Latinoamericano y Caribeño de Demografía (CELADE) – División de Población. Santiago, 7 al 9 de junio de 2010

CORSEUIL, Carlos Henrique; FOGUEL, Miguel N. "Uma sugestão de deflatores para rendas obtidas a partir de algumas pesquisas domiciliares do IBGE". Rio de Janeiro: Ipea, 2002. (Texto para Discussão, n. 897).

IBGE. "Censo Demográfico 2010 – Metodologia de Estimação do Número de Moradores em Domicílios Fechados". Nota técnica, Novembro de 2010. Disponível em:

http://www.ibge.gov.br/home/estatistica/populacao/censo2010/primeiros_resultados/nota_tecnica.pdf

JANNUZZI, Paulo de Martino. “Os quesitos de mão de obra nos censos demográficos brasileiros de 1960 a 2000”. Encontro transdisciplinar espaço e população, Campinas: NEPO/Unicamp, 2003.

OSORIO, Rafael Guerreiro. “O sistema classificatório de ‘cor ou raça’ do IBGE”. IPEA - Texto para Discussão Nº 996, Brasília, novembro, 2003.

REIS, E. J. “O PIB dos municípios brasileiros: metodologia e estimativas, 1970-1996”. Brasília: Ipea, 2005 (Texto para Discussão, 1064)

RESENDE, G.; CARVALHO, A.; SAKOWSKI, P. “Avaliando o crescimento econômico no Brasil em múltiplas escalas espaciais com a utilização de modelos de painel espacial (1970-2000)”. Brasília: Ipea, 2013 (Texto para Discussão, 1830).

RIGOTTI, J.I.R. Variáveis de educação dos censos demográficos brasileiros de 1960 a 2000. In: RIOS-NETO, E.L.G., RIANI, J.L.R. (org.). Introdução à Demografia da Educação. Campinas, ABEP, 2004.