



Construção da Coorte de 100 milhões de brasileiros

Adriana Viriato Ribeiro, David Gorender, Luan Menezes

Para a construção da Coorte de 100 milhões de brasileiros foi necessária a utilização dos *backups* do sistema do Cadastro Único (versões 6, 7 e 7.1) cedidos pelo Ministério de Desenvolvimento Social. Esses *backups* contém cópias das bases em diferentes momentos entre os anos de 2006 a 2015. Para cada ano de backup existem tabelas associadas às informações de família e do indivíduo. Na versão 6 existem três tabelas, as tabelas A, B e C. Enquanto na versão 7 existem as tabelas de 1 a 17. Nem todas as tabelas foram utilizadas na construção da Coorte. Na versão 6 foram utilizadas as tabelas A e B e na versão 7 foram usadas as tabelas 1 a 9, 10, 11, 13 e 15.

A lista com todas as bases utilizadas pode ser observada na Tabela x.

Ano referente	Data da extração	Nome diretório	Versão
2006	Março de 2007	MARCO_2007	6
2007	Dezembro de 2007	DEZEMBRO_2007	
2008	Dezembro de 2008	DEZEMBRO_2008	
2009	Dezembro de 2009	DEZEMBRO_2009	
2010	Setembro de 2010	SETEMBRO_2010	
2011	Janeiro de 2012		7
2012	Janeiro de 2013		
2013	Dezembro de 2013		
2014	Dezembro de 2014		

2015	Novembro de 2015		
------	------------------	--	--

Essas bases são consideradas as bases “originais” e representam as cópias cedidas sem nenhum tipo de modificação ou tratamento. No entanto, para a construção da Coorte e para preparar o banco final para análise por parte dos pesquisadores, é necessário realizar algumas tarefas em relação à limpeza e tratamento dos dados. Desta forma, algumas etapas foram definidas para a criação da Coorte baseada no Cadastro Único, a saber:

1. Conversão para csv
2. Criação de um dicionário do CIDACS
3. Padronização das variáveis
4. Harmonização das variáveis
5. Split por data de atualização
6. Bind por ano

Cada uma dessas etapas será explicada nas próximas seções, considerando algumas subtarefas, as diferentes características entre as versões e as avaliações de integridade e qualidade dos dados.

1 Conversão para csv

A conversão para csv é uma etapa realizada para tratamento inicial dos dados. A ideia é que os dados sejam convertidos para csv para que as

variáveis categóricas, inteiros, datas, entre outras, sejam tratadas antes de ser aplicado o *schema* em parquet. Desta forma, quando a cópia da base cedida ao Cidacs está em formato csv, é feita uma verificação de inconsistências e ela é salva novamente no mesmo formato adequando-se aos padrões de codificação e outras premissas. Quando a cópia da base chega em outros formatos, é realizada a conversão de acordo com as mesmas verificações de inconsistências. Em relação ao Cadastro Único, as bases relacionadas à versão 6 tem formato *fixed width*, enquanto as bases derivadas das versões 7 e 7.1 tem formato csv.

1.1 Conversão da Versão 6 para csv

A versão 6 do Cadastro Único é composta por duas tabelas: tabela A e tabela B. A tabela A refere-se às informações de família, enquanto a tabela B mantém informações dos indivíduos. Ambas tabelas são utilizadas na construção da Coorte, portanto, é preciso fazer a conversão de todas as tabelas em todos os anos (de 2006 a 2010). Desta forma, fez-se a conversão de *fixed width* para csv dos seguintes arquivos:

```
tbas = ['MARCO_2007/CNT.ICO.RJ.BHA1.ICOD884A.D070402',  
        'DEZEMBRO_2007/CNT.ICO.RJ.BHA1.ICOD884A.D080102',  
        'DEZEMBRO_2008/CNT.ICO.RJ.BHA1.ICOD884A.D090102',  
        'DEZEMBRO_2009/CNT.ICO.RJ.BHA1.ICOD884A.D100101',  
        'SETEMBRO_2010/CNT.ICO.RJ.BHA1.ICOD884A.D100901']
```

```
tbbs = ['MARCO_2007/CNT.ICO.RJ.BHA1.ICOD884B.D070402',  
        'DEZEMBRO_2007/CNT.ICO.RJ.BHA1.ICOD884B.D080102',  
        'DEZEMBRO_2008/CNT.ICO.RJ.BHA1.ICOD884B.D090102',  
        'DEZEMBRO_2009/CNT.ICO.RJ.BHA1.ICOD884B.D100101',  
        'SETEMBRO_2010/CNT.ICO.RJ.BHA1.ICOD884B.D100901']
```

Antes de realizar a conversão, é preciso substituir os caracteres que podem interferir na quantidade de colunas em um arquivo csv. A saber: contra barra, vírgula e aspas duplas. Para isso, antes da conversão é realizada uma alteração na cópia do arquivo original (nomedoarquivooriginal_sed). Essa alteração é feita no próprio terminal do linux utilizando o seguinte comando:

```
sed -i 'y/,,\"\\\/.../' nomedoarquivo
```

Depois dessa etapa, inicia-se o processo de conversão para csv. Esse processo consiste na utilização das chaves cedidas pelo MDS para mapeamento das variáveis. Existem duas chaves, uma para a tabela A e uma para tabela B, e as chaves consistem do nome da variável, sua posição inicial e sua extensão. Por exemplo:

```
tbla = {  
    'CD_DOMICILIAR':[1,9],  
    'CD_FAMILIAR':[10,15],  
    'CD_IDENTIFICACAO_DOMICILIO':[25,15],  
    'DT_PESQUISA_CADASTRO':[40,8],  
    'NU_NIS_ENTREVISTADOR':[48,11]  
    ...}
```

Desta forma, CD_DOMICILIAR é o nome da variável iniciada na posição 1, cuja extensão é de 9 caracteres. Assim, basta aplicar as chaves das tabelas A e B para os arquivos anuais que representam cada tabela. Como a intenção é realizar a conversão para csv, o procedimento feito é o mapeamento de cada coluna, que é armazenada em um arquivo de texto que utiliza vírgula como separador (csv), todo esse procedimento pode ser observado no código **01_v6-to-csv**.

Depois de salvar todos os arquivos de todas as tabelas, é preciso adicionar o cabeçalho com o nome das variáveis. Para isso, basta executar os seguintes procedimentos:



1 - Criar arquivos com os cabeçalhos das tabelas A e B:

```
vim header_tbla_lower.csv
```

```
#Digite:
```

```
ds_logradouro,cd_destino_lixo,cd_tipo_domicilio,cd_construcao,  
cd_eas_ms,cd_familiar,nm_bairro_logradouro,cd_ddd_logradouro,c  
d_identificacao_domicilio,cd_escoamento_sanitario,nm_entrevist  
ador,nm_estabelecimento_saude,dt_cadastro_domicilio,cd_ilumina  
cao,cd_situacao_domicilio,cd_identificacao_caixa,cd_abastecime  
nto_agua,cd_tipo_cobertura,dt_pesquisa_cadastro,qt_pessoas_inf  
ormada,cd_tipo_localidade,nu_telefone_logradouro,qt_mulheres_g  
ravidas,nu_comodos,in_domicilio_excluido,in_domicilio_ativo,nu  
_nis_entrevistador,cd_ibge_logradouro,qt_maes_amamentando,nu_c  
npj_pref_orgao,dt_alteracao_domicilio,nm_logradouro,in_complem  
entado_bes,cd_tratamento_agua,nu_logradouro,dt_inclusao_domici  
lio,qt_deficientes,cd_cep_logradouro,sg_uf_logradouro,cd_orige  
m_cadastro,cd_domiciliar,ds_complemento_logradouro
```

```
vim header_tblb_lower.csv
```

```
#Digite:
```

```
cd_folha_doc_certidao,nu_ordem_rl,dt_chegada_brasil,dt_nascime  
nto,cd_ocupacao,nm_outro_benef_pessoa,tp_documento_certidao_ci  
vil,vl_despesa_prest_habitacional,vl_remuneracao_emplo,vl_re  
nda_pensao_alimenticia,cd_parentesco,cd_censo_inep,cd_benefici
```

o_peti,in_participa_proger,in_deficiencia_mental,rf_anos_moradia,cd_nacionalidade,ds_orgao_emissor_identidade,dt_ultima_alteracao_domic,vl_despesa_luz,in_outro_programa,in_nenhum_programa,sg_uf_emissao_ctps,dt_ultima_alteracao_pessoa,nu_ordem_pessoa,dt_inclusao_peti,in_participa_pronaf,dt_emissao_doc_identidade,nu_natural_pessoa,cd_grau_instrucao,vl_renda_seguro_desemprego,nu_termo_doc_certidao,vl_beneficio_peti,in_outra_deficiencia,in_complementado_bes,vl_despesa_gas,nm_pai,in_beneficiario_bal,rf_mes_gravidez,sg_uf_doc_certidao,in_pessoa_duplicado,in_deficiencia_fisica,nu_nis_pessoa,dt_emissao_ctps,in_pessoa_excluida,vl_despesa_transporte,cd_pais_origem,in_deficiencia_cegueira,dt_inclusao_agente_jovem,sg_uf_emissao_identidade,dt_admissao_empresa,nm_empresa,cd_sexo,vl_despesa_aluguel,cd_ocupacao_exercida_peti,in_participa_agente_jovem,in_ocupacao_peti,nu_ordem_esposo,cd_escola,nu_cnpj_empresa,vl_renda_aposentadoria,cd_livro_doc_certidao,nu_ctps,in_participa_juv_cidada,nu_ordem_mae,in_domicilio_excluido,in_participa_prev_rural,nu_documento_identidade,nm_pessoa,in_participa_loas_bpc,dt_emissao_doc_certidao,in_participa_peti,nu_ordem_pai,cd_ibge_nascimento,vl_outras_despesas,vl_despesa_agua,cd_estado_civil,nu_serie_ctps,in_beneficiario_bes,nu_cpf,vl_despesa_medicamentos,cd_mercado_trabalho,ds_complemento_doc_identidade,cd_familiar,in_amamentando,in_participa_bolsa_alimentacao,cd_raca_cor,vl_despesa_alimentacao,nu_secao_eleitor,cd_crianca_quem_fica,in_deficiencia_surddez,nu_nis_original,in_deficiencia_mudez,vl_outras_rendas,nu_zona_eleitor,nm_mae,nm_escola,in_domicilio_ativo,cd_tipo_ocupacao_peti,dt_inclusao_proger,in_nenhuma_deficiencia,nu_pessoa_vivem_renda,in_liberto_trab_escravo,in_participa_bolsa_escola,nu_titulo_eleitor,rf_mes_moradia,cd_serie_escolar

2 - Adicionar o cabeçalho aos arquivos das tabelas (exemplo para tabela A de 2011):

```
cat header_tbla_lower.csv > tbla2011_csv/tba.csv  
cat tbla2011_csv/part* >> tba.csv
```

Depois de realizado o procedimento de conversão para csv, foi feita uma descritiva para avaliar a integridade e consistência dos dados de ambas as tabelas. Nesse processo, foi identificada a existência de 12 registros inconsistentes nas tabelas B dos diversos anos:

2006	2
2007	2
2008	4
2009	2
2010	2

A inconsistência refere-se ao número de colunas desses registros. Enquanto os registros normais da tabela B tem 107 variáveis, esses 12 registros tinham apenas 42. Como 42 é a quantidade de atributos encontrados na tabela A, tentou-se aplicar a chave da tabela A nesse registros para recuperá-los, no entanto, ainda assim os registros permaneceram inconsistentes. Desta forma, esses registros foram removidos da base.



1.1 Conversão da Versão 7 para csv

As bases da versão 7 estão no formato csv. Algumas operações foram realizadas nas tabelas para evitar erros no tratamento das bases. No cabeçalho de cada uma foi necessário realizar a substituição do caractere “_” pelo “.” e todas as letras foram convertidas em minúsculas. Foi necessário padronizar a codificação de caracteres de cada tabela, convertendo de UTF-8 para ASCII.

As bases originais possuem inconsistências com a quantidade de colunas, e quantidade total de registros, o que traz problemas para os procedimentos futuros. Foi identificado que o caractere “\” levava a quebra de linhas resultando nessas inconsistências citadas. Então foi necessário a remoção do mesmo para que as bases da versão 7 pudessem ser devidamente utilizadas.

2 Merge entre as tabelas

As bases do Cadastro Único são compostas por tabelas anuais que podem representar informações de família, domicílio e do indivíduo, por exemplo. Para a construção dos bancos anuais, é necessário realizar um join entre as tabelas de cada ano.

2.1 Merge da Versão 6

Como foi citado anteriormente, as tabelas da versão 6 que foram utilizadas na construção da Coorte foram as tabelas A e B. A tabela A corresponde às informações de família, enquanto a tabela B corresponde às informações individuais. Assim, foi realizado um merge¹ entre as tabelas A e B a partir da variável CD_FAMILIAR, resultando em bancos compostos por todos os indivíduos de cada ano, na qual cada linha representa as informações individuais e familiares.

O merge realizado foi do tipo *full*

Na versão 7 do Cadastro Único, foram utilizadas as tabelas de 1 a 15 (com exceção das tabelas 9, 10 e 14) para construção dos bancos anuais. As tabelas 1 e 3 contêm informações da família, a tabela 2 mantém informações do domicílio e as tabelas de 4 a 8 são compostas por informações do indivíduo. As tabelas de 11 a 13 e 15 referem-se a informações suplementares de pendência, referente à família e indivíduo. Assim, o merge entre as tabelas de 1 a 4, 11 e 13f foi realizado apenas com a variável cod_familiar_fam, enquanto o merge com as outras tabelas foi realizado com as variáveis cod_familiar_fam e num_membro_fmila.

TABELA 1: REPRESENTAÇÃO DA TRANSFORMAÇÃO ENTRE TABELAS PARA BANCOS ANUAIS NA VERSÃO 6

¹ Merge ou join é uma função que associa dois bancos de acordo com uma ou mais variáveis em comum.



Tabela A_2006	2006
Tabela B_2006	

TABELA 2: REPRESENTAÇÃO DA TRANSFORMAÇÃO ENTRE TABELAS PARA BANCOS ANUAIS NA VERSÃO 7

Tabela 1_2015	2015
Tabela 2_2015	
Tabela 3_2015	
Tabela 4_2015	
Tabela 5_2015	
Tabela 6_2015	
Tabela 7_2015	
Tabela 8_2015	
Tabela 11_2015	
Tabela 12_2015	
Tabela 13_2015	
Tabela 15_2015	

2.1 Inconsistência no merge da versão 7

Após a realização do merge entre as tabelas da versão 7 foi identificado mudanças na quantidade de registros, das variáveis categóricas quando comparados, tabela original e base pós merge. Aumento esse na quantidade de registros nulos e também em algumas categorias.

Foi realizado um merge full entre as tabelas, o que significa que são combinados registros iguais entre elas e também os divergentes. A combinação dos registros divergentes (*cod_familiar_fam* e *num_membro_fmlla*) explica o acréscimo de registros nulos. As tabelas de 1 a 13 (com exceção das tabelas 9 e 10) não possuem registros duplicados, porém foi identificado registros duplicados na tabela 15 explicando o acréscimo de registros nas categorias. A tabela 15 possui dois tipos duplicação, onde foi realizado deduplicação para os registros que possuem valores iguais nas variáveis (*cod_familiar_fam*, *num_membro_fmlla*, *cod_prog_prohab_memb*) porém quando há duplicação apenas nas variáveis (*cod_familiar_fam*, *num_membro_fmlla*) foi realizado a criação de uma nova coluna (*cod_prog_prohab_memb_comp*) para armazenar os duplicados.



3 Compatibilização da versão 7

As versões 7.1 e 7.2 do Cadastro Único são muito similares, com exceção de algumas variáveis que foram adicionadas/removidas na versão 7.2. Todas as variáveis que foram adicionadas/removidas na versão 7.2 foram excluídas dos registros em ambas as versões. Enquanto as variáveis com nomenclatura diferente tiveram seus nomes harmonizados. Desta forma, foi preciso mapear e tratar essas variações.

As diferenças mapeadas foram:

Tabela 1:

Removida	vazio
	num_reg_arquivo
Adicionada	qtde_pessoas
Nomenclatura diferente	dat_atual_fam (v 7.1) e dat_alteracao_fam (v 7.2)

Tabela 2:



Removida	vazio
	num_reg_arquivo

Tabela 3:

Removida	vazio
	num_reg_arquivo
	vazio2
Nomenclatura diferente	nu_estbo_saude (v 7.1) e cod_eas_fam (v 7.2)

Tabela 4:

Removida	vazio
	num_reg_arquivo
	ind_transferencia_pessoa
Adicionada	matchcode
	matchcode_mae

Tabela 5:

Removida	vazio
	num_reg_arquivo

Tabela 6:

Removida	vazio
	num_reg_arquivo

Tabela 7:

Removida	vazio
	num_reg_arquivo

Tabela 8:

Removida	vazio
	num_reg_arquivo

4 Padronização

4.1 Criação dos dicionários das versões 6 e 7

A criação dos dicionários é necessária para exposição das variáveis para os pesquisadores e para a automatização do processo de padronização. O dicionário é caracterizado pelos seguintes campos:

- 1) Nome das variáveis.
- 2) Descrição original, incluindo as categorias e formatos.
- 3) Descrição das categorias e formatos padronizados.
- 4) Uma coluna utilizada para a marcação das categorias que será utilizada no script para geração do código de padronização (esse campo não precisa estar presente na versão pública do dicionário).
- 5) Tipo da variável, a saber: integer, long, byte e string.
- 6) Uma coluna de transformação para os casos em que as variáveis mudam de categoria.

- 7) Um campo de comentários para identificar possíveis problemas, características e verificações.
- 8) Campos com a quantidade de missings de cada variável por ano. Esses campos devem ser preenchidos.

Desta forma, foram criados os dicionários para as versões 6 e 7 do Cadastro Único. Esses arquivos podem ser vistos em.

4.2 Padronização da versão 6 e 7

Uma vez definido o dicionário, é inicializado o processo de padronização das variáveis. Como foram criados campos nos dicionários para automatizar o processo de padronização, esses campos são lidos e utilizados como insumo para o script de geração automática das funções de padronização.

As funções de padronização dependem dos tipos de variáveis utilizadas:

- 1) Date: são verificadas inconsistências do ponto de vista sintático, a depender de algumas regras: tamanho do campo, verificação se as partes do campo correspondem a dia, mês e ano válidos, entre outras verificações.
- 2) Integer e Long: garantia de que existem apenas números nesses campos.
- 3) Byte: verificação e transformação das categorias de acordo com o que foi definido no dicionário.

Desta forma, verificou-se o tipo da variável de acordo com o campo “tipo” do dicionário e para cada tipo foram aplicadas as transformações/verificações de acordo com as funções de padronização. Depois disso, foram realizadas as chamadas das funções.

Além disso, na versão 6, as variáveis que indicam valor (renda e despesa), além de serem transformadas em inteiro, tiveram seus valores divididos por 100. Visto que na versão 6 o valor cem, por exemplo, está no formato 10000, com duas casas decimais. Esse processo foi realizado para que as variáveis de valor da versão 6 fossem compatíveis com a versão 7.

É importante ressaltar, que nenhuma variável original foi substituída, para cada uma dessas alterações foi criada uma variável com a tag “_original”, que foi mantida sem alterações. Por



exemplo, foi criada uma variável `sexo_original` que não sofre alterações, enquanto a variável `sexo` foi padronizada.