PROADI - Hospital Israelita Albert Einstein

Versão preliminar

PROADI - Hospital Israelita Albert Einstein

Qualidade de dados
CNES-LT (Leitos)

Histórico de revisões

Data	Versão	Descrição	Autor	Responsável
29/04/2020	1.0	Versão preliminar	Leandro Furlam, Elias Ribeiro	Alexandre Rodrigues
13/05/2020	1.0	Revisão textual	Wiliam Hisatugu	Wiliam Hisatugu
07/07/2020	1.0	Inclusão de análise de disponibilidade e mudanças de domínio	Leandro Furlam, Elias Ribeiro	Alexandre Rodrigues

Sumário

1 Qualidade de dados

O processo de análise de qualidade de dados está focado na avaliação de conjuntos de dados e na aplicação de ações corretivas, para garantir que estes estejam adequados aos propósitos para os quais foram originalmente destinados (1). Dessa forma, a qualidade de dados está diretamente relacionada a confiabilidade dos dados de entrada. Considerando que os dados têm níveis inadequados de qualidade, é provável que ocorram erros, que podem se propagar acidentalmente e inconscientemente por todo o fluxo da informação, prejudicando a eficiência do sistema. Formas regulares de avaliar a qualidade de dados com modelos clássicos geralmente se destinam a detectar e corrigir erros em fontes conhecidas com base em um conjunto limitado de regras. No ambiente de *Big Data*, a quantidade de regras pode ser enorme e o custo da aplicação para correção de erros pode não ser viável e nem apropriado (*e.g.* o enorme volume de dados ou a volatilidade dos dados de *streaming*). Isso ocorre principalmente porque o *Big Data* não é apenas sobre dados, mas também sobre uma pilha conceitual e tecnológica completa, incluindo dados brutos e processados, armazenamento, formas de gerenciar dados, processamento e análise (1).

Uma dimensão de qualidade de dados é um termo descritor de um recurso de dados, o qual pode ser medido ou avaliado de acordo com padrões definidos, a fim de determinar a qualidade de um conjunto de dados (2). Geralmente, dados só têm valor quando dão suporte a um processo ou a uma tomada de decisão. Em consequência, as regras de qualidade de dados definidas devem levar em consideração o valor que os dados podem fornecer para o sistema.

Neste relatório, seis dimensões de qualidade de dados são analisadas: completude, conformidade, acurácia, consistência e temporalidade (2). A dimensão unicidade, que objetiva mensurar o grau de duplicidade nos dados, foi excluída deste estudo, uma vez que dados de identificação dos pacientes são removidos da planilha de informações.

Completude caracteriza a taxa de preenchimento das variáveis. Para cada variável é calculado o percentual de entradas com informação não nulas, respeitando, quando houver, sua dependência com outras variáveis.

Conformidade detecta concordância nos valores digitados nos campos das variáveis, avaliando se os valores de entrada não nulos estão em conformidade com os padrões descritos pelo dicionário de dados. Para cada variável estudada é calculado o percentual de entradas em conformidade com o padrão adotado.

Acurácia visa detectar se informação registrada reflete o evento ou objeto descrito, isto é, verificar se o dado cadastrado está em concordância com o evento

observado. Devido ao processo de anonimização dos dados, a análise de acurácia se restringe a verificar a possibilidade das informações registradas. Note que acurácia e conformidade são dimensões distintas, pois enquanto conformidade avalia o padrão do dado, acurácia avalia a razoabilidade dos dados. Para cada variável estudada é calculado o percentual de entradas com informações acuradas.

Consistência constitui de testes envolvendo duas ou mais variáveis visando detectar inconsistências entre dados de um mesmo registro. Para cada teste considerado é calculado os percentuais de aprovação e falha.

Temporalidade objetiva efetuar medidas estatísticas nos intervalos de tempos entre eventos, *e.g.* Nascimento de um recém-nascido e inclusão desse registro no sistema. O principal interesse é verificar se o dado é disponibilizado prontamente.

Neste relatório a Seção 2 retrata a base de dados, a Seção 3 define a metodologia, a Seção 4 expõe a disponibilidade dos dados, a Seção 5 analisa as mudanças que as variáveis sofreram em relação ao tempo, a Seção 6 apresenta os respectivos resultados e a Seção 7 estende a análise com considerações finais.

2 Base de dados

O **Cadastro Nacional de Estabelecimentos de Saúde (CNES)** é um sistema de informação oficial de cadastramento de informações acerca de estabelecimentos de saúde do país, independentemente de sua natureza jurídica ou integração com o Sistema Único de Saúde (SUS) (3).

De acordo com a Portaria GM/MS n° 1.646/2015 (3), as finalidades do CNES são as seguintes: cadastrar e atualizar as informações sobre estabelecimentos de saúde e suas dimensões, como recursos físicos, trabalhadores e serviços; disponibilizar informações dos estabelecimentos de saúde para outros sistemas de informação; ofertar para a sociedade informações sobre a disponibilidade de serviços nos territórios, formas de acesso e funcionamento; fornecer informações que apoiem a tomada de decisão, o planejamento, a prorrogação e o conhecimento pelos gestores, pesquisadores, trabalhadores e sociedade em geral acerca da organização, existência e disponibilidade de serviços, força de trabalho e capacidade instalada dos estabelecimentos de saúde e territórios.

O CNES é a base para operacionalização de diversos sistemas, como Sistema de Informação Ambulatorial (SIA), Sistema de Informação Hospitalar (SIH), e-SUS Atenção Básica (e-SUS AB), entre outros. É uma ferramenta auxiliadora, que proporciona o conhecimento da realidade da rede assistencial existente e suas potencialidades, de forma a auxiliar no planejamento em saúde das três esferas de Governo, para uma gestão eficaz e eficiente (4).

Os arquivos deste cadastro são compostos por diferentes naturezas de informação, cada qual possuindo uma nomenclatura com uma sigla correspondente, conforme é demonstrado na seguir. Neste relatório são analisados os arquivos de **Leitos**.

Natureza da informação	**Sigla utilizada**
Estabelecimentos	ST
Dados Complementares	DC
Profissional	PF
Leitos	**LT**
Equipamentos	EQ
Serviço Especializado	SR
Equipes	EP
Habilitações	HB
Regras Contratuais	RC
Gestão e Metas	GM
Estabelecimento de Ensino	EE
Estabelecimento Filantrópico	EF
Incentivos	IN

2.1 Informações gerais

Base de dados	CNES-LT
Fonte	<ftp://ftp.datasus.gov.br/dissemin/publicos/CNES/200508_/Dados/LT/>
Data de obtenção dos dados	06 de julho de 2020
Período	ago/2005 a abr/2020
Região geográfica	Todas as 27 Unidades Federativas
Volume	184,2 MB
Número máximo de variáveis	28
Número de registros	8.393.737

3 Métodos

A análise dos dados constitui-se de um esquema cíclico, iniciando no mapeamento da documentação e do comportamento dos dados, através da observação

de trechos das bases. Em seguida, são definidas as variáveis de teste. Após, ocorre a obtenção e avaliação dos resultados obtidos, recorrendo, e se necessário retificando, conclusões obtidas nos passos anteriores. O maneio dos dados ocorreu através dos serviços **Amazon Athena** e **Amazon S3**, assim como testes e análises se deu utilizando linguagem R. Os *scripts* utilizados estão disponíveis no repositório de qualidade de dados no **GitHub**.

Enfatiza-se que esses dados podem sofrer alterações, caso ocorram atualizações.

Para analisar a disponibilidade dos dados e as mudanças ocorridas nas variáveis e nos respectivos domínios, foram avaliados os microdados e informações obtidas do DATASUS e do Ministério da Saúde, através de informes técnicos; e os arquivos auxiliares e de tabulação, aplicados principalmente no levantamento de domínio e definição de valores ignorados e sem informação.

Os resultados apresentados neste relatório consideram a inclusão/retirada de variáveis ao longo do tempo. Mudanças no domínio das variáveis também são detectadas, relatadas e consideradas no cálculo de medidas de qualidade dos dados.

O Cômputo dos resultados numéricos ocorre de modo cascata, isto é, os registros submetidos ao teste de conformidade devem ser não nulos, os registros submetidos ao teste de acurácia devem estar conformes, os registros submetidos aos testes de consistência devem estar acurados, e quando não for possível, conformes e o mesmo se aplica aos registros submetidos aos testes de temporalidade. Em prosseguimento, os resultados numéricos são avaliados nas dimensões analisadas calculando-se a média ponderada dos testes realizados, utilizando como peso o total de registros por variável. Para a consistência, é realizado um ajuste em que todas as variáveis testadas devem existir simultaneamente. Objetivando avaliar a base de dados, o conjunto de resultados representando cada dimensão foi classificada como excelente (> 90%), ótimo (75% - 89,9%), regular (50% - 74,9%) ou ruim (< 49,9%), baseado nos relatórios do livro *Saúde Brasil*, organizado pela Secretaria de Vigilância em Saúde (5). Em decorrência do método cascata utilizado, é realizado o produto dos resultados obtidos, na Seção 7, caracterizando a qualidade da base de dados como um todo, que também pode ser classificada considerando as classes definidas em *Saúde Brasil* (5).

4 Disponibilidade dos dados

Esta seção tem o objetivo de dissertar acerca da disponibilidade dos dados em todo o período representado pela base de dados e em todas as Unidades Federativas.

Após realização de testes, averiguou-se que para todo o período representado pela base de dados a informação está disponível, ou seja, existem registros relativos a todos os anos, meses e Unidades Federativas.

5 Variáveis existentes e mudanças ocorridas

Esta seção tem por objetivo identificar as variáveis existentes na base de dados e relatar as mudanças ocorridas ao longo do tempo. Nesse sentido, o Gráfico 1 apresenta um resumo do quantitativo de variáveis no banco de dados ao longo dos anos analisados, onde podemos identificar que houve apenas o acréscimo de uma variável em 2012. A saber, *nat_jur*, que representa a natureza jurídica, é encontrada em todos os estados somente a partir de jun/2012.

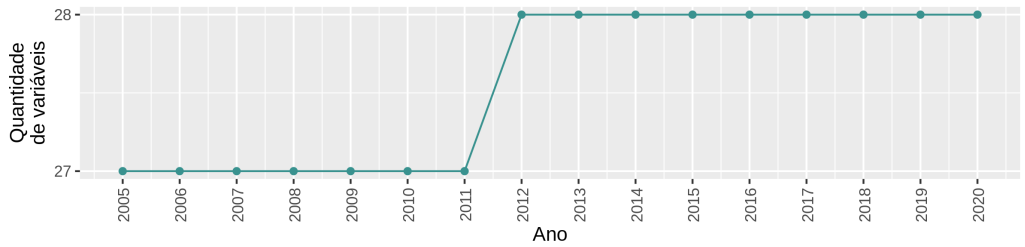

Gráfico 1: evolução do número de variáveis na base de dados.

Mesmo que presentes na base de dados, identificou-se que algumas variáveis não possuem qualquer registro em um determinado ano, isto é, estão totalmente vazias em um período específico. Esse fato torna-se um problema quando deseja-se realizar análises sob uma perspectiva anual, visto que ocorrerá lacunas. Nesse contexto, a Tabela 1 apresenta as variáveis nesta situação.

Variável	**Ano**															
	2005	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
esfera_a																
natureza																
niv_hier																
retencao																
terceiro																

Tabela 1: Variáveis contendo apenas valores não disponíveis (em cinza) por período.

Em relação à modificação no tamanho das variáveis, foi elaborada uma descrição detalhada do domínio por ano, conforme é descrito na Tabela 2.

Variável	Mudança	Período de ocorrência
distrsan	Diminuiu de tamanho: passou de um tamanho máximo de 5 para 4, onde o tamanho estabelecido pelo dicionário de dados é 5	2019
	Aumentou de tamanho: passou de um tamanho mínimo de 1 para 2	2020
micr_reg	Aumentou de tamanho: passou de um tamanho máximo de 6 para 9, onde o tamanho estabelecido pelo dicionário de dados é 9	2014

Tabela 2: mudanças ocorridas nas variáveis por período.

Em virtude da análise realizada na Seção 4 e nas Tabelas 1 e 2, conclui-se que apenas 2013 e 2015 não apresentaram qualquer tipo de problema relacionado a disponibilidade dos dados e ocorrência de alterações nas variáveis.

6 Resultados

Esta seção apresenta e avalia os resultados dos testes aplicados. As considerações são apresentadas nas subseções a seguir, uma para cada dimensão.

Descrições das variáveis são apresentadas no Apêndice A. Resultados numéricos dos testes de completude, conformidade e acurácia são exibidos no Apêndice B, onde estão organizados em três tabelas: resultado geral, resultado agregado por ano e resultado agregado por Unidade Federativa. Já o Apêndice C expõe uma tabela contendo valores atípicos[1] de variáveis quantitativas, isto é, registros numéricos que apresentam grande afastamento em relação aos demais, dentro do universo de uma única variável. Descrições dos testes de inconsistência realizados, bem como seus respectivos resultados numéricos estão descritos no Apêndice D. Os resultados agregados por ano foram obtidos em relação ao nome do arquivo.

6.1 Completude

Nesta dimensão são detectados valores faltantes através da busca pelas constantes representando valores ausentes. Nesse sentido, considerou-se como incompletos os registros contendo os valores *NA*, constante lógica que indica valor ausente em linguagem R, e *NULL*, que representa objetos nulos.

Ressalta-se que foram realizados ajustes no cômputo da completude da variável *cnpj_man*, uma vez que esta relaciona-se no que tange ao grau de dependência à variável *niv_dep*; e da variável *nat_jur*, que é encontrada em todos os estados somente a partir de 2012.

[1] <https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/boxplot.stats>

A respeito da distribuição da completude por ano agregada por Unidade Federativa, observa-se, através do Gráfico 2, que há pioras segundo o avanço do tempo. De 2005 a 2015, os resultados distribuem-se em torno de 88%, sendo o maior valor de 89,00% em 2005, o início do período. Em seguida ocorre queda, permanecendo em torno de 71% até o fim do período representado pela base de dados, sendo o menor valor de 71,16% em 2017.

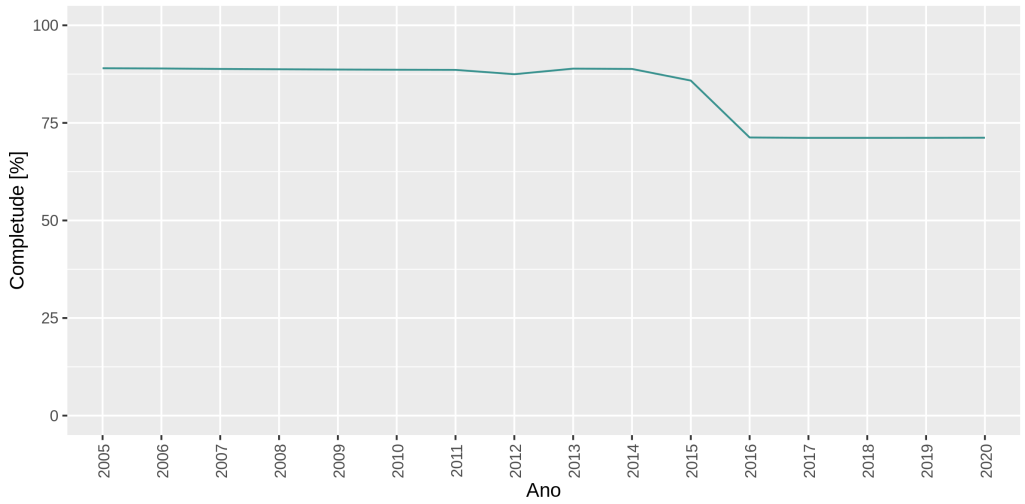

Gráfico 2: distribuição da completude por ano agregada por Unidade Federativa.

Sobre a distribuição espacial da completude por Unidade Federativa agregada no tempo, nota-se pelo Gráfico 3 que os estados permanece em torno de 80% e 87%, com os menores valores encontrados na região norte e os maiores na região nordeste. Rondônia e Roraima, únicos abaixo de 81%, apresentaram 79,31% e 80,30%, respectivamente, enquanto com 85,05% e 87,31%, Paraíba e Ceará ultrapassaram 85%.

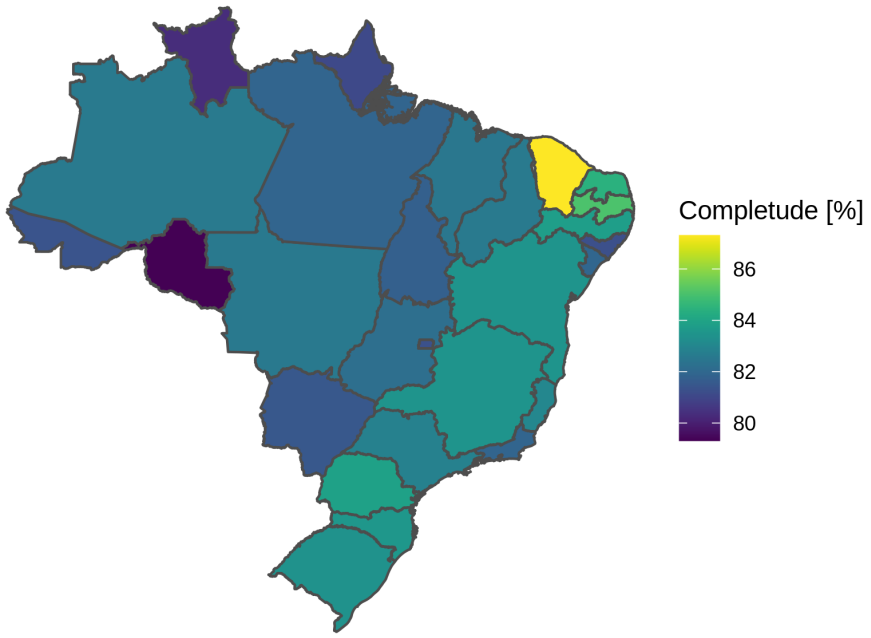

Gráfico 3: distribuição espacial da completude por Unidade Federativa agregada no tempo.

No geral, os resultados de completude das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 4. O resultado percentual por variável está descrito no Apêndice B. O cômputo da média ponderada dos resultados obtidos é de **83,04%**, ou seja, a **completude é ótima**.

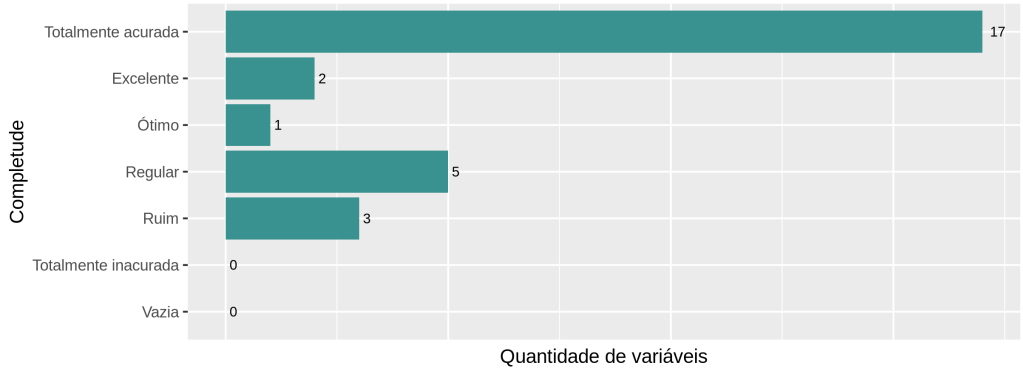

Gráfico 4: distribuição dos resultados de completude.

6.2 Conformidade

Verificou-se se os dados apresentam os padrões descritos no dicionário de dados adotado como referência (Apêndice A) a respeito da quantidade de caracteres e presença de variantes definidos. O resultado percentual por variável está descrito no Apêndice B, onde nota-se que todas as variáveis apresentaram conformidades totais em seus registros, **100,00%**, ou seja, a **conformidade é excelente** para a base

analisada.

6.3 Acurácia

Inicialmente a métrica de acurácia foi aplicada a dois tipos de registros: datas, ao verificar se o dado configura-se uma data válida e condizente ao período representado pela base de dados; nomes e códigos de municípios, ao verificar se estão contidos na tabela de códigos de municípios e estados do IBGE[2]. Após esta análise de datas e municípios é realizada investigação acerca do preenchimento das variáveis com o objetivo de detectar a presença de preenchimentos sem informações relevantes. Em seguida, são verificados os registros representando informações numéricas, a respeito do sinal (*e.g.* número de filhos deve ser positivo) e do conjunto ao qual pertence (*e.g.* número de filhos deve ser um número inteiro).

A respeito de registros representando datas, para todas variáveis verificou-se se existiam datas superiores ao mês e ano referência da base de dados analisada, isto é, último dia representado pela tabela de dados. Não houve ocorrência de falhas.

Acerca de registros representando informações numéricas, ressalta-se a variável *qt_nsus*, representando a quantidade de leitos não SUS, onde há 72 valores negativos variando de -880 à -1.

Ainda em relação as várias quantitativas, o Apêndice C expõe uma tabela contendo valores atípicos, os quais implicam, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados. Neste, observa-se valores exorbitantes para todas as variáveis analisadas, como por exemplo estabelecimentos de saúde com 2220 leitos existentes.

Acerca dos demais registros, pode-se mencionar a ocorrência de valores indicando a ausência de informações ou que foram ignorados, além de sequências finitas do numeral zero. Este fato pode representar um problema, visto que estará de acordo ao tamanho estabelecido pelo dicionário de dados, porém não estará acurado, não representando informação alguma. Por exemplo, na variável *cnpj_man*, representando o CNPJ da mantenedora do Estabelecimento, 58,63% dos registros apresentam o valor *00000000000000*, completo e conforme ao dicionário de dados.

Acerca dos testes de acurácia no que diz respeito aos resultados distribuídos por ano, é possível observar através do Gráfico 5, que a acurácia dos registros piorou um pouco com o decorrer do tempo. De 2005 a 2015, período este que apresentou melhores resultados de completude (Subseção 6.1), os resultados distribuem-se em torno de 96,4%, enquanto no restante do período, estes distribuem-se em

[2] <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>

torno de 95,5%, sendo o menor valor encontrado em 2020, de 95,54%.

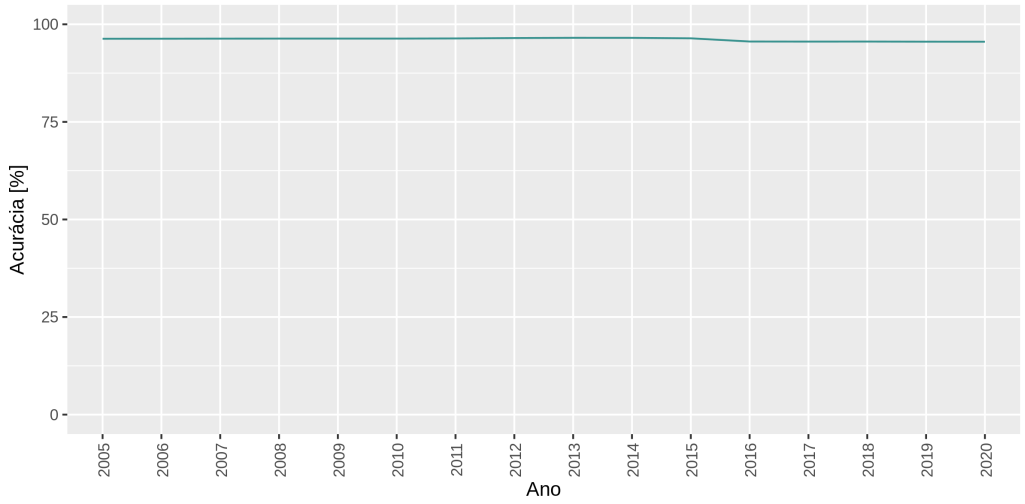

Gráfico 5: distribuição da acurácia por ano agregada por Unidade Federativa.

Acerca dos resultados distribuídos por Unidade Federativa, observa-se, pelo Gráfico 6, que apenas Goiás, Sergipe e Paraná não ultrapassaram 99% dos registros acurados.

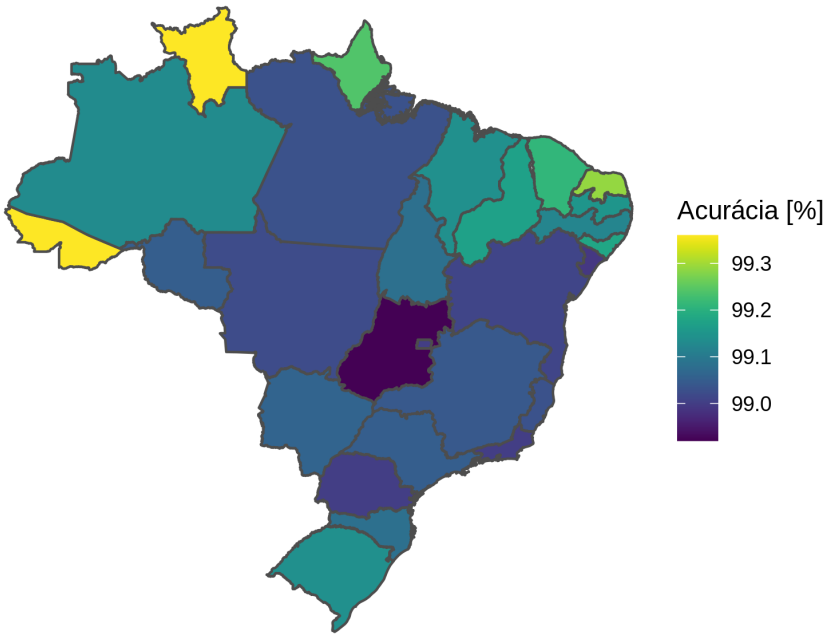

Gráfico 6: distribuição espacial da acurácia por Unidade Federativa agregada no tempo.

No geral, os resultados de acuracidade das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 7. O resultado percentual por variável está descrito no Apêndice B. O cômputo da média ponderada, conside-

rando todos os testes realizados, é **94,98%**, ou seja, a **acurácia é excelente** para a base analisada.

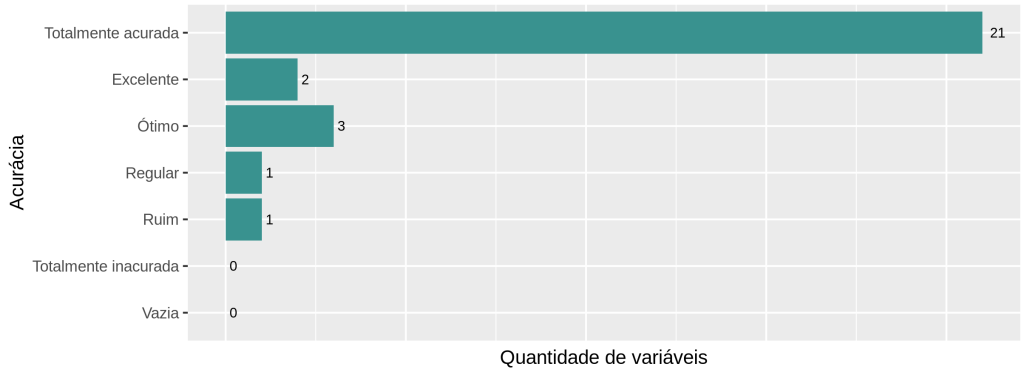

Gráfico 7: distribuição dos resultados de acurácia.

6.4 Consistência

Os resultados descritos na Tabela 3 são de testes aplicados a um mesmo registro, ou seja, mesma linha do conjunto de dados. Estes testes detectam principalmente problemas na entrada de dados envolvendo condições específicas de inconsistências. A descrição dos testes realizados, bem como os resultados, se encontram no Apêndice D, onde também é destacada a quantidade de inconsistências por teste para cada ano.

Teste realizado	**Falhas [partes por mil]**
niv_dep == 1 & cnpj_man != NULL	0,000
(qt_contr + qt_exist) != (qt_nsus + qt_sus)	1,521

Tabela 3: resultados de inconsistência.

Acerca dos resultados de consistência distribuídos temporalmente, nota-se, através do Gráfico 8, que ocorreram significativas quedas com relação ao inicio do período representado pela base de dados. Partindo de um máximo de 3,3 em 2005, a taxa de registros inconsistentes chega, em 2015, à 0,3, permanecendo assim até o fim do período.

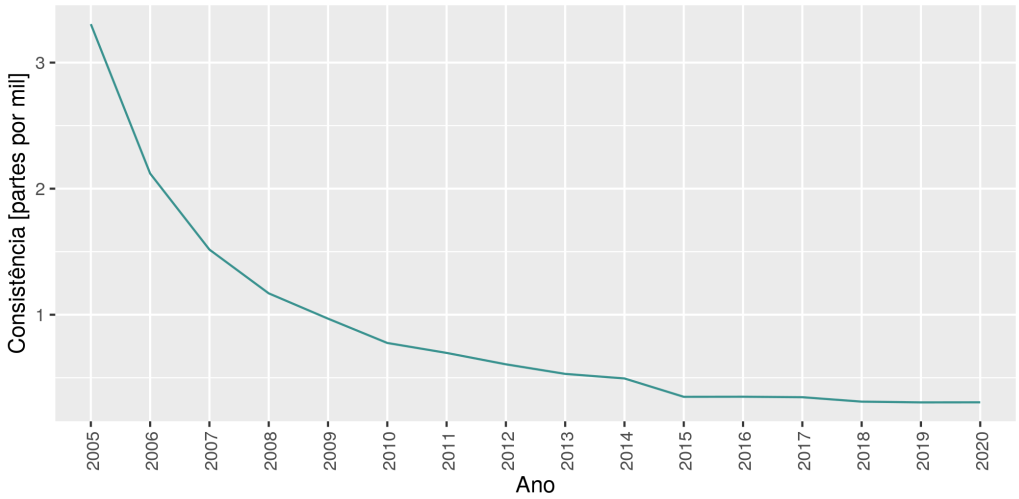

Gráfico 8: distribuição da consistência por ano agregada por Unidade Federativa.

Já acerca da distribuição espacial, o estado do Mato Grosso foi estado que mais se destacou negativamente entre os demais, atingindo cerca de 8,896 partes por mil de registros inconsistentes. Em contrapartida, Acre, Alagoas, Amazonas, Amapá, Distrito Federal, Roraima, Rio Grande do Sul e Sergipe sequer apresentaram algum registro inconsistente.

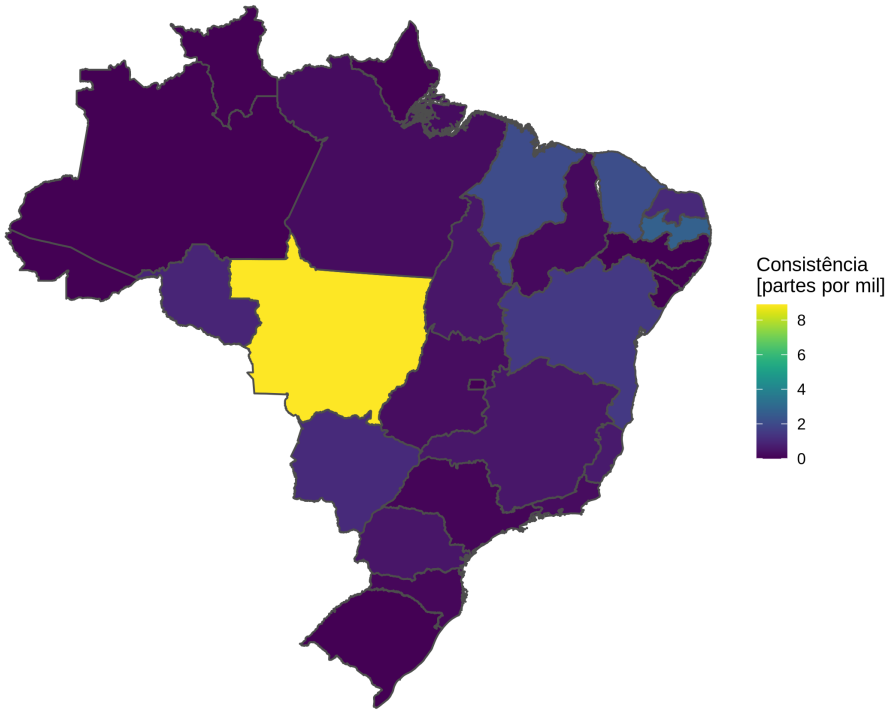

Gráfico 9: distribuição espacial da consistência por Unidade Federativa agregada no tempo.

Dentro deste contexto, a média ponderada dos resultados dos testes de con-

sistência é **99,92%**, ou seja, a **consistência é excelente** para a base analisada.

6.5 Temporalidade

Para mensurar esta dimensão é calculada a quantidade de dias entre duas variáveis representando datas que estejam conformes, acuradas e consistentes. Para esta base não há ocorrência de variáveis que retratam esta dimensão.

7 Considerações finais

A avaliação realizada é especialmente oportuna, tendo em vista o cenário nacional e o atual empenho em fomentar o debate em torno da qualidade das informações acerca de estabelecimentos de saúde do país.

Avaliando a disponibilidade dos dados, ressalta-se que a falta de informações de determinados períodos reflete em problemas na análise aqui realizada, visto que não irá representar a abrangência total dos registros. Ainda, algumas variáveis apresentaram problemas de estarem totalmente vazias em determinado ano e merecem serem analisadas, com o intuito de incrementar a qualidade da informação dos dados e incentivar o preenchimento correto pelos profissionais de saúde envolvidos no processo. Sobre estas alterações numa escala temporal, conclui-se que apenas 2013 e 2015 não apresentam qualquer tipo de deficiência.

Analisado os resultados obtidos pela métrica de completude, observa-se menores taxas de preenchimento entre 2016 a 2020, enquanto o maior valor é encontrado em 2005, o início do período representado. Isto significa que a mesma piorou segundo o avanço do tempo. Ao que diz respeito a distribuição espacial, o preenchimento dos registros é menor em alguns estados da região norte e maior na região nordeste.

Sobre os resultados de consistência, houve excelência no primeiro teste realizado e ocorrência de muitas falhas no segundo teste. Espacialmente, os resultados de Mato Grosso foram discrepantes em relação aos demais, já que sete estados sequer apresentaram falhas. Temporalmente, destaca-se, de maneira positiva, que as inconsistências diminuíram com o decorrer do período.

Finalmente, a média ponderada dos resultados de completude é 83,04%, de conformidade é 100,00%, de acurácia é 94,98% e de consistência é 99,92%. Realizando o produto destes resultados, obtêm-se **78,81%**, caracterizando a **base de dados como ótima**.

Referências

1 MERINO, J. et al. A data quality in use model for big data. *Future Generation Computer Systems*, v. 63, p. 123–130, 2016. ISSN 0167-739X. Modeling and Management for Big Data Analytics and Visualization. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X15003817>.

2 DAMA, U. The six primary dimensions for data quality assessment-defining data quality dimensions. *Bristol: np URL: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf*, v. 3, p. 2017, 2013.

3 SAúDE., B. M. da. *Portaria nº 1.646, de 2 de outubro de 2015*: Institui o cadastro nacional de estabelecimentos de saúde (cnes). 2015. Np URL: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2015/prt1646_02_10_2015.html>.

4 CADASTRO Nacional de Estabelecimentos de Saúde (CNES). 2020. <https://wiki.saude.gov.br/cnes/index.php/P\%C3\%A1gina_principal>. Acessado em: 21 de abril de 2020.

5 SAúDE., B. M. da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação em. *Saúde Brasil 2019: uma análise da situação de saúde com enfoque nas doenças imunopreveníveis e na imunização*. [S.l.]: Ministério da Saúde, 2019.

Apêndice A Descrição das variáveis

Quando **.CNV*, refere-se ao arquivo de tabulação correspondente.

Variável	Descrição / Observações	Tam.	Valores válidos
atividad	Código da atividade de ensino	2	01: unidade universitária, 02: unidade escola superior isolada, 03: unidade auxiliar de ensino, 04: unidade sem atividade de ensino, 05: hospital de ensino, 99: atividade ensino não informada
clientel	Código de fluxo da clientela	2	01: atendimento de demanda espontânea, 02: atendimento de demanda referenciada, 03: atendimento de demanda espontânea e referenciada, 00: fluxo de clientela não exigido, 99: fluxo de clientela não informado
cnes	Número nacional do estabelecimento de saúde	7	
cnpj_man	CNPJ da mantenedora do estabelecimento	14	
codleito	Especialidade do leito	2	ESP_LEIT.CNV
codufmun	Código do município do estabelecimento UF+ MUNIC (sem dígito)	6	
competen	Ano e mês de competência da informação (AAAAMM)	6	
cpf_cnpj	CPF do Estabelecimento, caso pessoa física ou CNPJ, caso pessoa jurídica	14	
distradm	Código do distrito administrativo	5	
distrsan	Código do distrito sanitário	5	
esfera_a	Código da esfera administrativa	2	01: federal, 02: estadual, 03: municipal, 04: privada, 99: esfera não informada
micr_reg	Código da micro-região de saúde	9	
nat_jur	Natureza jurídica	4	NATJUR.CNV, NATJURC.CNV, ESFERAJUR.CNV, ESFERAJURC.CNV
natureza	Código da natureza da organização	2	NATUREZA.CNV
niv_dep	Grau de dependência	1	1: individual, 3: mantida

Variável	Descrição / Observações	Tam.	Valores válidos
niv_hier	Código do nível de hierarquia	2	01: NH PAB-PABA, 02: NH média M1, 03: NH média M2 e M3, 04: NH alta complexidade ambulatorial, 05: NH baixa M1 e M2, 06: NH média M2 e M3, 07: NH média M3, 08: NH alta complexidade hos/amb., 00-99: NH não informado
pf_pj	Indicador de pessoa	1	1: física, 3: jurídica
qt_contr	Quantidade de leitos contratados	4	
qt_exist	Quantidade de leitos existentes	4	
qt_nsus	Quantidade leitos não SUS	4	
qt_sus	Quantidade de leitos para o SUS	4	
regsaude	Código da região de saúde	7	
retencao	Código de retenção de tributos	2	00-99: retenção estab. não informada, 10: estabelecimento publico, 11: estabelecimento filantropico, 12: estabelecimento sem fins lucrativos, 13: estabelecimento privado lucrativa simples, 14: estabelecimento privado lucrativa, 15: estabelecimento sindical, 16: estabelecimento pessoa fisica
terceiro	O estabelecimento é terceiro	1	1: sim, 2: não
tp_leito	Tipo do leito	1	1: cirúrgico, 2: clínico, 3: complementar, 4: obstétrico, 5: pediátrico, 6: outras especialidades, 7: hospital/dia
tp_unid	Tipo de unidade (estabelecimento)	2	TP_ESTAB.CNV
tpgestao	Gestão de saúde	1	Z: não informado, D: dupla, E: estadual, M: municipal, S: sem gestão
turno_at	Código de turno de atendimento	2	01: atendimento turnos intermitentes, 02: atendimento contínuo 24 horas/dia (pl sab dom fer), 03: atendimento turnos manhã/tarde/noite, 04: atendimento somente pela manhã, 05: atendimento somente à tarde, 06: atendimento turnos manhã/tarde, 07: atendimento somente à noite, 99: turno não informado

Apêndice B Resultados numéricos

B.1 Resultados gerais

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
atividad	100,00	100,00	100,00
clientel	99,82	100,00	100,00
cnes	100,00	100,00	100,00
cnpj_man	100,00	100,00	41,37
codleito	100,00	100,00	100,00
codufmun	100,00	100,00	99,41
competen	100,00	100,00	100,00
cpf_cnpj	100,00	100,00	76,59
distradm	4,71	100,00	70,79
distrsan	11,78	100,00	85,81
esfera_a	68,42	100,00	100,00
micr_reg	7,59	100,00	87,69
nat_jur	95,05	100,00	100,00
natureza	68,42	100,00	100,00
niv_dep	100,00	100,00	100,00
niv_hier	68,18	100,00	100,00
pf_pj	100,00	100,00	100,00
qt_contr	100,00	100,00	100,00
qt_exist	100,00	100,00	100,00
qt_nsus	100,00	100,00	100,00
qt_sus	100,00	100,00	100,00
regsaude	75,02	100,00	99,97
retencao	68,42	100,00	100,00
terceiro	62,76	100,00	100,00
tp_leito	100,00	100,00	100,00
tp_unid	100,00	100,00	100,00
tpgestao	100,00	100,00	100,00
turno_at	100,00	100,00	100,00

B.2 Resultados por ano

Ano	Completude [%]	Conformidade [%]	Acurácia [%]
2005	89,00	99,09	96,30
2006	88,94	99,09	96,31
2007	88,82	99,09	96,33
2008	88,74	99,10	96,34
2009	88,67	99,10	96,34
2010	88,62	99,10	96,34
2011	88,58	99,10	96,38
2012	87,47	99,12	96,48
2013	88,90	99,14	96,54
2014	88,82	99,15	96,53
2015	85,84	99,12	96,41
2016	71,26	98,94	95,59
2017	71,16	98,93	95,57
2018	71,16	98,93	95,58
2019	71,17	98,92	95,55
2020	71,20	98,94	95,54

B.3 Resultados por Unidade Federativa

UF	Completude [%]	Conformidade [%]	Acurácia [%]
AC	81,37	100,00	99,36
AL	81,25	100,00	99,18
AM	82,58	100,00	99,13
AP	81,06	100,00	99,24
BA	83,44	100,00	99,01
CE	87,31	100,00	99,21
DF	81,30	100,00	99,00
ES	83,01	100,00	99,03
GO	82,21	100,00	98,92
MA	82,48	100,00	99,14

UF	**Completude [%]**	**Conformidade [%]**	**Acurácia [%]**
MG	83,45	100,00	99,04
MS	81,47	100,00	99,06
MT	82,54	100,00	99,02
PA	81,89	100,00	99,03
PB	85,05	100,00	99,15
PE	83,78	100,00	99,12
PI	82,58	100,00	99,17
PR	83,88	100,00	99,00
RJ	81,88	100,00	99,00
RN	84,41	100,00	99,29
RO	79,31	100,00	99,05
RR	80,30	100,00	99,36
RS	83,39	100,00	99,14
SC	83,56	100,00	99,08
SE	81,90	100,00	98,99
SP	82,79	100,00	99,05
TO	81,74	100,00	99,08

Apêndice C Valores atípicos

Variável	Valores atípicos
qt_contr	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 24, 25, 27, 28, 30, 31, 34, 35, 37, 39, 40, 48, 50, 54, 57, 66, 67, 70, 76, 120, 128, 220, 295
qt_exist	25, 26, 27, 28, 29, 30, 31, 32, 33, 34, ...[3], 817, 820, 822, 824, 826, 980, 999, 1000, 1730, 2220
qt_nsus	8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...[3], 323, 369, 413, 438, 439, 483, 598, 803, 870, 2210
qt_sus	21, 22, 23, 24, 25, 26, 27, 28, 29, 30, ...[3], 584, 600, 630, 639, 640, 680, 692, 870, 980, 1730

[3] Os valores foram comprimidos devido a alta quantidade.

Apêndice D Testes de inconsistências

D.1 Testes realizados

- **Teste 1:** Se o grau de dependência é *1* (individual, não mantida), então o CNPJ da mantenedora do estabelecimento deve ser nulo;
- **Teste 2:** A quantidade de leitos existentes somada à quantidade de leitos contratados deve ser igual à soma da quantidade de leitos destinados e não destinados ao SUS.

Ano	**Teste 1**	**Teste 2**
2005	0	864
2006	0	2244
2007	0	1650
2008	0	1294
2009	0	1105
2010	0	897
2011	0	812
2012	0	708
2013	0	625
2014	0	588
2015	0	408
2016	0	408
2017	0	408
2018	0	368
2019	0	360
2020	0	180