**Processo de Validação de Dados - Estabelecimentos de Saúde em Santa Catarina**

**1. Introdução**

Este documento descreve o processo de validação dos dados sobre estabelecimentos de saúde em Santa Catarina para a plataforma Cidade Única do Observatório FIESC. A validação é uma etapa crucial para garantir a confiabilidade dos dados que serão disponibilizados aos gestores públicos.

**2. Metodologia de Validação**

O processo de validação foi estruturado nas seguintes etapas:

1. **Compreensão dos dados**: Análise inicial dos diferentes conjuntos de dados (bronze, silver e gold) e entendimento de suas estruturas.
2. **Validação estrutural**: Verificação da consistência das estruturas das tabelas, como presença de colunas esperadas e formatos adequados.
3. **Validação cruzada**: Comparação entre os dados processados (pasta gold) e os dados oficiais do TABNET.
4. **Validação de consistência interna**: Verificação da coerência entre os diferentes arquivos dentro da pasta gold.
5. **Detecção e documentação de inconsistências**: Registro sistemático de discrepâncias encontradas.
6. **Propostas de correção**: Indicação de soluções para os problemas identificados.

**3. Dados Analisados**

**3.1 Dados Processados (pasta gold)**

* **gold\_micro**: Agregação de estabelecimentos de saúde por microrregião (20 microrregiões)
* **gold\_municipio**: Agregação de estabelecimentos de saúde por município (295 municípios em SC)

**3.2 Dados Oficiais (TABNET)**

* **cnes\_microrregiao**: Dados oficiais do CNES agregados por microrregião
* **cnes\_municipio**: Dados oficiais do CNES agregados por município

**4. Validação Estrutural**

**4.1 Verificação de Colunas**

* Verificado que os arquivos da pasta gold contêm as colunas necessárias:
  + gold\_micro: [microrregiao, nu\_quantidade]
  + gold\_municipio: [municipio, nu\_quantidade]

**4.2 Verificação de Formatos**

* Os valores numéricos estão representados como números
* Foram encontrados problemas de codificação em caracteres especiais:
  + 11 microrregiões no Gold com problemas de acentuação
  + 109 municípios no Gold com problemas de acentuação
* Exemplo de problemas encontrados: "AraranguÃ¡", "ChapecÃ³", "ConcÃ³rdia", "ItajaÃ­"

**5. Validação Cruzada com Dados TABNET**

**5.1 Comparação de Totais**

* **Total de estabelecimentos no gold\_micro**: 24.829 (soma das quantidades por microrregião)
* **Total de estabelecimentos no gold\_municipio**: 24.829 (soma das quantidades por município)
* **Total de estabelecimentos no TABNET**: 24.829 (conforme indicado na linha "Total")
* **Conclusão**: O total geral está consistente entre todas as fontes

**5.2 Comparação por Microrregião**

| **Microrregião** | **Gold** | **TABNET** | **Diferença** | **% Diferença** | **Status** |
| --- | --- | --- | --- | --- | --- |
| Blumenau | 2.726 | 2.398 | +328 | +13,68% | ❌ |
| Joinville | 2.647 | 2.908 | -261 | -8,98% | ❌ |
| São Miguel do Oeste | 693 | 693 | 0 | 0% | ✓ |
| Chapecó | 1.886 | 1.886 | 0 | 0% | ✓ |
| Xanxerê | 714 | 714 | 0 | 0% | ✓ |

**Inconsistências encontradas em microrregiões**:

1. Blumenau: Excesso de 328 estabelecimentos em relação aos dados oficiais
2. Joinville: Déficit de 261 estabelecimentos em relação aos dados oficiais

**5.3 Comparação por Município (Amostra de Verificação)**

| **Município** | **Gold** | **TABNET** | **Diferença** | **% Diferença** | **Status** |
| --- | --- | --- | --- | --- | --- |
| Blumenau | 1.470 | 1.142 | +328 | +28,72% | ❌ |
| Abdon Batista | 7 | 7 | 0 | 0% | ✓ |
| Abelardo Luz | 65 | 65 | 0 | 0% | ✓ |
| Água Doce | 26 | 26 | 0 | 0% | ✓ |

**Observação**: A inconsistência em Blumenau (328 estabelecimentos) é a mesma tanto na microrregião quanto no município, indicando um problema sistemático com este local específico.

**6. Validação de Consistência Interna**

**6.1 Soma por Microrregião vs. Soma dos Municípios por Microrregião**

| **Microrregião** | **Valor Declarado** | **Soma dos Municípios** | **Diferença** | **% Diferença** | **Status** |
| --- | --- | --- | --- | --- | --- |
| Blumenau | 2.726 | 2.519 | -207 | -7,59% | ❌ |
| Campos de Lages | 1.460 | 1.320 | -140 | -9,59% | ❌ |
| Canoinhas | 778 | 520 | -258 | -33,16% | ❌ |
| Curitibanos | 456 | 400 | -56 | -12,28% | ❌ |
| Ituporanga | 151 | 111 | -40 | -26,49% | ❌ |
| Joinville | 2.647 | 1.879 | -768 | -29,01% | ❌ |
| Rio do Sul | 835 | 695 | -140 | -16,77% | ❌ |
| Tabuleiro | 62 | 34 | -28 | -45,16% | ❌ |
| Tijucas | 231 | 172 | -59 | -25,54% | ❌ |

**Observação**: Em todas as inconsistências, o valor declarado da microrregião é maior que a soma dos municípios pertencentes a ela, indicando possíveis municípios faltantes ou mal classificados.

**6.2 Municípios sem Correspondência na tabela silver**

* Total de 106 municípios sem correspondência clara com microrregiões
* Exemplos: "AgrolÃ¢ndia", "AgronÃ´mica", "AnitÃ¡polis", "AntÃ´nio Carlos"
* Este problema está diretamente relacionado aos problemas de codificação identificados

**7. Inconsistências Identificadas**

**7.1 Problemas de Contagem**

1. **Blumenau**: Excesso de 328 estabelecimentos em relação aos dados oficiais (+28,72%)
2. **Joinville**: Déficit de 261 estabelecimentos em relação aos dados oficiais (-8,98%)
3. **Microrregiões com valor declarado maior que soma dos municípios**: 9 microrregiões apresentam esta inconsistência, com a maior discrepância em Tabuleiro (-45,16%)

**7.2 Problemas de Codificação**

1. **Caracteres especiais**: Acentuação incorreta nos nomes de municípios (109 ocorrências) e microrregiões (11 ocorrências)
2. **Impacto no mapeamento**: 106 municípios não puderam ser associados corretamente às suas microrregiões devido a problemas de normalização de texto

**7.3 Problemas de Completude**

1. **Municípios não mapeados**: Impossibilidade de validar completamente a consistência interna entre microrregiões e municípios
2. **Dados faltantes por microrregião**: Em várias microrregiões, a soma dos municípios é significativamente menor que o valor declarado

**8. Proposta de Correções**

**8.1 Correções de Contagem**

1. **Blumenau**: Corrigir a contagem para 1.142 estabelecimentos conforme dados oficiais do TABNET
2. **Joinville**: Atualizar o valor para 2.908 estabelecimentos conforme os dados oficiais do TABNET
3. **Microrregiões com inconsistências**: Revisar e corrigir as 9 microrregiões com diferenças significativas entre valor declarado e soma dos municípios

**8.2 Correções de Codificação**

1. **Padronização UTF-8**: Aplicar codificação UTF-8 consistente em todos os arquivos para corrigir problemas de acentuação
2. **Normalização de nomes**: Implementar rotina de normalização de nomes (remoção de acentos, padronização de caixa) para facilitar correspondências

**8.3 Correções de Mapeamento**

1. **Atualização da tabela silver**: Revisar e atualizar o mapeamento entre municípios e microrregiões, garantindo que todos os municípios estejam corretamente associados
2. **Verificação de consistência**: Implementar validação para garantir que a soma dos estabelecimentos por município corresponda ao total da microrregião

**9. Recomendações Adicionais**

1. **Documentação das fontes**: Incluir metadados completos sobre a origem dos dados e data de extração
2. **Versionamento**: Implementar controle de versão para rastrear alterações nas bases de dados
3. **Validação automatizada**: Incorporar os scripts de validação desenvolvidos no pipeline ETL para validação automática em futuros carregamentos
4. **Monitoramento contínuo**: Estabelecer um processo de monitoramento regular para detectar inconsistências após atualizações de dados

**10. Conclusão**

O processo de validação revelou que, embora os totais gerais estejam consistentes entre as bases Gold e TABNET, existem inconsistências significativas em nível de microrregião e município, especialmente em Blumenau e Joinville. Os principais problemas encontrados foram:

1. Discrepâncias nas contagens de estabelecimentos em microrregiões específicas
2. Problemas de codificação de caracteres que dificultam o mapeamento correto
3. Inconsistências internas onde o valor declarado da microrregião não corresponde à soma dos municípios

As correções propostas visam garantir a integridade e confiabilidade dos dados que serão disponibilizados na plataforma Cidade Única, permitindo que os gestores públicos tenham acesso a informações precisas para a tomada de decisões.

A implementação de um processo de validação automatizado, como o desenvolvido neste trabalho, é fundamental para manter a qualidade dos dados ao longo do tempo e garantir a confiabilidade das informações fornecidas pelo Observatório FIESC.