Tratamento dos dados do Sisagua

SISAGUA

SISTEMA DE INFORMAÇÃO DE VIGILÂNCIA DA QUALIDADE DA ÁGUA PARA CONSUMO HUMANO

Dados do monitoramento da qualidade da água realizado pelo prestador de serviço em frequência superior à mensal, contemplando os resultados das análises de qualidade da água de alta complexidade

Fonte dos dados usada de 2018, 2019 e 2020: SISAGUA - Controle Semestral - Conjuntos de dados - Portal Brasileiro de Dados Abertos

Orientações originais de Tiago de Brito Magalhães

Dados do Sisagua, do conjunto de dados de controle semestral, de 2018 e 2019

Grupos de parâmetros agrotóxicos, substâncias orgânicas, substâncias inorgânicas e produtos secundários da desinfeção.

Foram desprezados da análise os resultados de amostras coletadas no ponto de captação

Os registros que se enquadraram em alguma das situações abaixo foram considerados inconsistentes:

- registros com resultado ‘menor que LQ’, sem o valor de LQ;
- registros com resultado ‘menor que LD’, sem o valor de LD;
- registros com valor de LD ou LQ igual a 0;
- registros com valor de LD superior ou igual ao valor de LQ;
- registros com resultado quantificado, igual a 0 (zero);
- registros com valor de resultado quantificado, inferior ao LQ;
- registros com valor de resultado quantificado, igual ou inferior ao LD;

A classificação dos resultados em relação aos valores de referência definidos para cada substância no padrão de potabilidade conforme definições abaixo.

Resultado quantificado com valor menor ou igual ao VMP Abaixo do VMP

Resultado ‘menor que LQ’, sendo o LQ menor ou igual ao VMP Abaixo do VMP

Resultado ‘menor que LD’, sendo o LD menor ou igual ao VMP Abaixo do VMP

Resultado ‘quantificado’ com valor maior que o VMP Acima do VMP

Resultado menor que LQ, sendo o LQ maior que o VMP Inconclusivo

Resultado menor que LD, sendo o LD maior que o VMP Inconclusivo

Inconsistentes N/A

Os registros classificados como inconclusivos são os que não apresentam uma inconsistência analítica, mas não permitem avaliar o atendimento ao valor de referência devido ao uso de um método ou equipamento não apropriado.

A classificação dos municípios para os mapas foi realizada da seguinte maneira.

Sem informação Municípios que não possuem registro de análises.

Abaixo do VMP Municípios que apresentaram pelo menos uma análise, mas nenhum resultado acima do VMP.

Acima do VMP Municípios que apresentaram pelo menos uma análise com resultado quantificado e acima do VMP.

Transposição para limpeza dos dados, criação de novas colunas e análises – algoritmos para o Python

1. Como estão os dados originais:

Essas são as colunas originais dos arquivos de cada ano (em média, um ano tem 1,5 milhão de registros)

Região Geográfica

UF

Regional de Saúde

Município

Código IBGE

Tipo da Instituição

Sigla da Instituição

Nome da Instituição

CNPJ da Instituição

Nome do escritório regional/local

CNPJ do escritório regional/local

Tipo da Forma de Abastecimento

Código Forma de abastecimento

Nome da Forma de Abastecimento

Nome da ETA / UTA

Ano de referência

Semestre de referência

Data de registro

Data de preenchimento do relatório semestral

Data da coleta

Data da análise

Ponto de Monitoramento

Grupo de parâmetros

Parâmetro

LD

LQ

Resultado

Aqui está o dicionário de dados das colunas:
https://sage.saude.gov.br/dados/sisagua/dicionarios/dicionario_controle_semestral.odt

2. Colunas mais importantes para limpeza dos dados e análises futuras

Neste momento do projeto as colunas mais utilizadas são

Município – nome do município com medição da qualidade d'água

Código IBGE – código do munícipio, com seis dígitos, no IBGE – usado para georreferenciamento

Ponto de Monitoramento - Local onde foi realizada a coleta da amostra - dados como "PONTO DE CAPTAÇÃO" têm de ser desprezados porque não têm o registro de medição adequado. O motivo de não utilizarmos as análises no ponto de captação é que se trata de água bruta, e o valor de referência (VMP) se refere à água tratada, que será destinada ao consumo da população.

Grupo de parâmetros – São as subdivisões de substâncias que podem existir na água: Parâmetros Organolépticos, Substâncias Inorgânicas, Produtos de Desinfecção, Substâncias Orgânicas, Agrotóxicos e Radioatividade. São retirados os dados com Parâmetros Organolépticos porque há discordância entre químicos sobre sua relevância

Parâmetro – É a lista de todas as substâncias dos grupos acima – são 85 diferentes. E cada um tem escrito o VMP (valor máximo permitido) para ser encontrado nas medições d´água

LD – Limite de Detecção pode ter um número, 0 ou estar vazio

LQ - Limite de Quantificação pode ter um número, 0 ou estar vazio

Resultado - Resultado de análise - pode ter um número, 0, MENOR_LQ, MENOR_LD, ou estar vazio

Junto o valor do VMP, essas três últimas colunas andam juntas para fazer a limpeza dos dados – ver se são válidos ou não naquela medição da água.

Há casos que existe apenas Limite de Quantificação (LQ), ou seja, expressam o resultado considerando o menor resultado que pode ser quantificado com precisão e exatidão, sendo este relacionado ao tipo de metodologia e ao equipamento utilizado. Ou também valores divergentes do Resultado mostrado ou células em branco. São dezenas de regras para considerar abaixo

3. Criação de novas para análise e filtro dos dados de interesse

Com esses dados e as regras de validação do Ministério da Saúde, as colunas novas são criadas - “tipo_de_resultado”, “vmp”, “consistencia” e “atendimento_ao_padrao”

“tipo_de_resultado” indica se o resultado foi MENOR_LD, MENOR_LQ, QUANTIFICADO ou TUDO_VAZIO

“vmp” captura o Valor Máximo Permitido que está no Parâmetro

“consistencia” indica se a medição naquela linha foi Consistente ou Inconsistente

“atendimento_ao_padrao” coloca o rótulo final na medição: Abaixo do VMP, Acima do VMP, Inconclusivo ou not applicable

4. Algoritmos das quatro novas colunas

A - “tipo_de_resultado”

Faço primeiro um comando para retirar espaços em branco na coluna "resultado", antes ou depois dos registros

Substituo na coluna "resultado" as vírgulas "," por ponto ".". Isso serve para a coluna ser transformada em numérica por linguagens de programação - elas consideram decimal com ponto

1 - Primeiro bloco de teste. Em cada linha primeiro vejo se o resultado é diferente de "MENOR_LQ" ou de "MENOR_LD" - isso significa que é uma linha com número ou vazia

Depois faço um teste para transformar esse valor em número float (com decimais). Se apontar erro já considero essa linha vazia - "vazio" - "TUDO_VAZIO"

Caso não aponte erro ainda testo se o valor é NAN (not a number) ou "". Caso for deixo como valor "vazio" - "TUDO_VAZIO"

2 - Segundo bloco de teste. Se o valor for igual "MENOR_LQ" ou "MENOR_LD", apenas deixo esse valor em “tipo_de_resultado”

Se não, se os testes acima tiveram dado “vazio” deixa como "TUDO_VAZIO"

E se passar por esses testes sem detectar vai considerar a coluna tipo_resultado com número original -- "QUANTIFICADO"

B - “vmp”

Faz um teste nas colunas “parâmetro” para retirar o valor do VMP. No caso, já tenho um dicionário com todas as substâncias e valores respectivos – então quando na linha ela aparece o programa busca a chave e o valor já previamente cadastrado. Com o decimal com ponto “.”

C - “consistencia”

Faço primeiro um comando para retirar espaços em branco nas colunas "resultado", “lq” e “ld”, antes ou depois dos registros

Substituo nessas colunas também as vírgulas "," por ponto ".". Isso serve para a coluna ser transformada em numérica por linguagens de programação - elas consideram decimal com ponto

1 - Primeiro bloco de teste. Em cada linha primeiro vejo se o lq e ld são diferentes de vazio. Faço por meio de um teste para transformar esse valor em número float (com decimais). Se apontar erro já considero essa linha vazia - "vazio"

Caso não aponte erro ainda testo se o valor é NAN (not a number) ou "". Caso for deixo como valor "vazio"

Em lq e ld isso é feito com uma variável de cópia e caso passe sem erros vai ser numérica

2 – Segundo bloco de teste. Agora só vejo se a linha tem tipo_resultado igual a "TUDO_VAZIO". Caso sim, a coluna “consistencia” já é "Inconsistente"

3 – Terceiro bloco de teste. Agora se o tipo_resultado é igual a "QUANTIFICADO", faz uma cópia do valor de resultado em float (decimal):

3.a – Se o resultado for 0 ou 0.0 a coluna “consistencia” já é "Inconsistente"

3.b – Caso não, serão feitos outros vários subtestes, todos com linhas "QUANTIFICADO":

- 3.b1 – Se lq e ld forem iguais, é "Inconsistente"

- 3.b2 - Se ld não for "vazio" e for resultado <= ld, é "Inconsistente"

- 3.b3 - Se lq não for "vazio" e for resultado < lq, então é "Inconsistente"

- 3.b4 - Se lq e ld forem diferentes de "vazio", e ao mesmo tempo ld >= lq, então é "Inconsistente"

- 3.b5 – Testa se ld ou lq é 0 ou 0.0, se sim é "Inconsistente"

- 3.b6 - Se lq for diferente de "vazio" e ld for igual a "vazio", é "Consistente"

- 3.b7 - Se lq for igual a "vazio" e ld for diferente de "vazio", então é "Consistente"

- 3.b8 – Se existir outra condição ainda, marca como "Consistente"

4 – Quarto bloco de teste. Agora se o tipo_resultado é diferente de "QUANTIFICADO", faz vários testes:

4.a1 - Se ld for igual a lq, então é "Inconsistente"

4.a2 - Se lq e ld forem diferentes de "vazio", e ao mesmo tempo ld >= lq, então é "Inconsistente"

4.a3 - Se o resultado for "MENOR_LQ" e o lq for "vazio", então é "Inconsistente"

4.a4 - Testa se ld ou lq é 0 ou 0.0, se sim é "Inconsistente"

4.a5 - Se resultado for igual a "MENOR_LD" e ld for igual a "vazio", então é "Inconsistente"

4.a6 - Se resultado for "MENOR_LQ" e lq diferente de "vazio", então é "Consistente"

4.a7 - Se resultado for "MENOR_LD" e copia_ld diferente de "vazio", então é "Consistente"

4.a8 - Se existir outra condição ainda, marca como "Consistente"

D - “atendimento_ao_padrao”

Faço primeiro um comando para retirar espaços em branco nas colunas "resultado", “lq” e “ld”, antes ou depois dos registros

Substituo nessas colunas também as vírgulas "," por ponto ".". Isso serve para a coluna ser transformada em numérica por linguagens de programação - elas consideram decimal com ponto

1 - Primeiro bloco de teste. Em cada linha primeiro checa se "consistencia" é igual a "Inconsistente", se sim é "atendimento_ao_padrao" como "not applicable"

2 – Segundo bloco de teste. Nos casos em que tipo_de_resultado é "QUANTIFICADO", faz subtestes:

2.a1 - Checa de resultado <= vmp, então é atendimento_ao_padrao como "Abaixo do VMP"

2.a2 - Checa de resultado > vmp, então é atendimento_ao_padrao como "Acima do VMP"

3 – Terceiro bloco de teste. Nos casos em que tipo_de_resultado diferente de "QUANTIFICADO", primeiro faz testes em lq e ld.

Em cada linha primeiro vejo se o lq e ld são diferentes de vazio. Faço por meio de um teste para transformar esse valor em número float (com decimais). Se apontar erro já considero essa linha vazia - "vazio"

Caso não aponte erro ainda testo se o valor é NAN (not a number) ou "". Caso for deixo como valor "vazio"

Em lq e ld isso é feito com uma variável de cópia e caso passe sem erros vai ser numérica

Aí parte para os subtestes:

3.a1 - Se resultado é "MENOR_LQ" e lq <= vmp, então é atendimento_ao_padrao = "Abaixo do VMP"

3.a2 - Se resultado é igual "MENOR_LD" e ld <= vmp, então é atendimento_ao_padrao = "Abaixo do VMP"

3.a3 - Se resultado é igual "MENOR_LQ" e lq > vmp, então é atendimento_ao_padrao = "Inconclusivo"

3.a4 - Se resultado é igual a "MENOR_LD" e ld > vmp, então é atendimento_ao_padrao = "Inconclusivo"

3.a5 - Se existir outra condição ainda, marca como "not applicable"

Criação de base de dados finais para o site

Com a limpeza final dos dados originais do Sisagua, incluindo a criação de novas colunas, é possível criar uma classificação dos municípios de acordo com a gravidade dos testes de qualidade da água encontrados em cada município

Depois de debates com técnicos do Ministério da Saúde e especialistas, o projeto considera então os testes mais graves aqueles como "Consistente" e "Acima do VMP", para estas substâncias mais perigosas. São substâncias com risco de gerar doenças crônicas, como câncer, se acima do limite, e também diversos outros problemas de saúde (substâncias que aumentam risco de câncer, mutação genética e doenças endócrinas / substâncias que aumentam risco de alterações no sistema nervoso, cerebral, imune - entre outras doenças):

-> Agrotóxicos

Aldrin + Dieldrin

Alaclor

Atrazina

Carbendazim + benomil

Clordano

DDT + DDD + DDE

Diuron

Glifosato + AMPA

Lindano (gama HCH)

Mancozebe

Trifluralina

-> Inorgânicos

Arsênio

Cádmio

Chumbo

Cromo

Níquel

Nitrato (como N)

Nitrito

Selênio

-> Orgânico

Acrilamida

Benzeno

Cloreto de Vinila

Diclorometano

Estireno

Pentaclorofenol

Tetracloroeteno

Tricloroeteno

Benzo[a]pireno

-> Radioatividade

Atividade beta total

Atividade alfa total

Rádio-228

Rádio-226

-> Subprodutos da desinfecção

2, 4, 6 Triclorofenol

Assim o Mapa da Água considera os testes mais perigosos nessa ordem:

1 – TIPO COR 1 – Testes como 'Consistente', diferentes de 'PONTO DE CAPTAÇÃO' no ponto de monitoramento, diferente de 'Parâmetros Organolépticos', e que sejam no atendimento_ao_padrao igual a 'Acima do VMP' e sejam com as substâncias mais perigosas listadas acima

2 – TIPO COR 2 – Testes como 'Consistente', diferentes de 'PONTO DE CAPTAÇÃO' no ponto de monitoramento, diferente de 'Parâmetros Organolépticos', e que sejam no atendimento_ao_padrao igual a 'Acima do VMP' e que tenham outras substâncias diferentes das mais perigosas listadas acima

3 – TIPO COR 3 – Testes como 'Consistente', diferentes de 'PONTO DE CAPTAÇÃO' no ponto de monitoramento, diferente de 'Parâmetros Organolépticos', e que sejam no atendimento_ao_padrao igual a 'Abaixo do VMP' e sejam com as substâncias mais perigosas listadas acima

4 – TIPO COR 4 – Testes como 'Consistente', diferentes de 'PONTO DE CAPTAÇÃO' no ponto de monitoramento, diferente de 'Parâmetros Organolépticos', e que sejam no atendimento_ao_padrao igual a 'Abaixo do VMP' e que tenham outras substâncias diferentes das mais perigosas listadas acima

O usuário do projeto Mapa da Água vê essas 4 classificações contadas para cada município em que aconteceram os testes de qualidade da água, isto é, a mesma substância pode ser encontrada diversas vezes no mesmo ano nos testes, com resultados diferentes de qualidades. Esses resultados são contados e mostrados

Com o cruzamento do Código IBGE é possível plotar um mapa de todas as cidades do país com essas características e mostrar as cidades onde não há nenhum dado “Consistente” ou mesmo não houve nenhum teste na água no período estudado

Resumindo, temos essas classificações nos milhões de dados analisados – sendo que algumas delas não são mostradas neste momento do projeto por serem dados sem relevância científica agora:

tipo1 = Consistente, diferente de ponto de captação, Acima do VMP e dentro das substâncias perigosas

tipo2 = Consistente, diferente de ponto de captação, Acima do VMP e fora das substâncias perigosas

tipo3 = Consistente, diferente de ponto de captação, Abaixo do VMP e dentro das substâncias perigosas

tipo4 = Consistente, diferente de ponto de captação, Abaixo do VMP e fora das substâncias perigosas

tipo5 = municípios sem dados – depois cruzar com o código IBGE

tipo6 = Inconclusivo ou N/A ou vazio no atendimento_ao_padrao

tipo7 = tudo que sobrou depois de rodar 1,2,3,4,6 e 8, no caso pt_monitoramento igual a PONTO DE CAPTAÇÃO

tipo8 = Parâmetros Organolépticos

Os passos para fazer tudo isso foram assim:

1 - Com dados do IBGE, criar um arquivo com os códigos de seis e sete dígitos de cada um dos 5.570 munícipios, nomes oficiais, incluindo colunas com sua latitude e longitude. Também foram incluídos depois as 31 unidades administrativas do Distrito Federal

2 - Fazer a união dos dados tratados acima do Sisagua de 2018, 2019 e 2020. O dataframe único somado foi de 5.126.181 linhas ou testes - dados baixados em 2 de novembro de 2021

3 - Cria uma lista com os nomes das substâncias perigosas para comparação depois nas linhas. São todas em letras maiúsculas e no formato adequado para comparar: exemplo, as substâncias aparecem originalmente com seu nome e valor ("Chumbo - VMP:0,01 mg/L" ou "Rádio-228 - VMP: 0,1 Bq/L"), mas para comparar o nome eu preciso separar a string pelo valor "-", então a lista de nomes tem que ter sempre o primeiro valor separado para fazer uma comparação correta

4 - Cria também um dicionário das substâncias, que tem como chave o nome completo da substância e como valor o nome isolado (exemplo, 'Aldrin + Dieldrin - VMP: 0,03 µg/L': 'Aldrin + Dieldrin'). Isso será usado para fazer a primeira camada de isolar a string. A última, se for necessária, é feita logo abaixo, no caso de nomes que têm mais de um "-"

5 – É iniciado então o loop principal no dataframe. Como dito acima, em cada linha as colunas são olhadas: consistencia, pt_monitoramento, atendimento_ao_padrao, ano, semestre, grupo e substancia. A primeira tarefa é fazer o isolamento do nome da substância, pelo dicionário criado e pelo símbolo (-). Depois é essa string que será comparada na lista criada de substâncias perigosas

6 – São iniciados então os testes para classificação de cada linha:

6.1 - Se o grupo é igual 'Parâmetros Organolépticos', então o "tipo_cor" é igual a "8"

6.2 - Se atendimento_ao_padrao é igual 'Inconclusivo' ou atendimento_ao_padrao é igual a 'N/A' ou 'not applicable', ou ainda atendimento_ao_padrao é vazio ou "", então "tipo_cor" é igual a '6'

6.3 – Se na linha a substância testada está entre as mais perigosas, e é 'Consistente' e diferente de 'PONTO DE CAPTAÇÃO', e o atendimento_ao_padrao é igual a 'Acima do VMP', então é tipo cor 1

6.4 – Se na linha a substância testada não está entre as mais perigosas, e é 'Consistente' e diferente de 'PONTO DE CAPTAÇÃO', e o atendimento_ao_padrao é igual a 'Acima do VMP', então é tipo cor 2

6.5 – Se na linha a substância testada está entre as mais perigosas, e é 'Consistente' e diferente de 'PONTO DE CAPTAÇÃO', e o atendimento_ao_padrao é igual a 'Abaixo do VMP', então é tipo cor 3

6.6 – Se na linha a substância testada não está entre as mais perigosas, e é 'Consistente' e diferente de 'PONTO DE CAPTAÇÃO', e o atendimento_ao_padrao é igual a 'Abaixo do VMP', então é tipo cor 4

6.7 – As demais linhas que sobraram são classificadas como tipo cor 7 – o que sobra são as linhas com ponto de monitoramento igual a 'PONTO DE CAPTAÇÃO'

7 – Os dados do Sisagua têm informações também das 31 unidades administrativas do Distrito Federal, porém na nossa visualização de mapa não é possível ver esse espaço pequeno, como se fossem munícipios. Por isso foi decidido transformar todos os códigos IBGE do DF no código IBGE de Brasília (530010).

8 – Os grupos de parâmetros sobre Desinfecção têm diferentes nomenclaturas. Resolvemos unificá-los como "Subprodutos da desinfecção"

9 – Para fins da visualização inicial dos municípios – que destaca entre todos os testes, se existir, aqueles piores primeiro – fazemos um ordenamento na base de dados por ['tipo_cor', 'cod_ibge']. Existem então duas bases de dados, a principal com todos os detalhes e mais de 5 milhões de linhas, e a secundária apenas com a pior a aparição em cada município. As duas bases são usadas no mapa

10 – Por fim, são separados os munícipios com o chamado alerta máximo:

- 'tipo_cor' igual a 1 ou 2 em algum momento de seus testes
- Aí verificar nesse grupo se uma mesma 'substancia' aparece três 'ano' seguidos

Isso foi feito assim:

10.1 - Separa apenas os testes tipo_cor 1 ou 2, e faz uma contagem por 'cod_ibge', 'substancia', 'tipo_cor', 'ano' em grupo

10.2 Faz uma cópia desses dados e retira ['cod_ibge'] duplicados

10.3 Faz uma iteração na cópia em cada ['cod_ibge']

10.4 Em cada passagem filtra o dado original agrupado pelo ['cod_ibge'] da vez e pega cada lista de substâncias

10.5 Com essa iteração procura também o 'ano' e o 'tipo_cor', e armazena em outras listas

10.6 Depois checa se 'ano' aparece ao mesmo tempo como '2020', '2019' e '2018', se sim registra numa base dados a parte

10.7 O resultado é uma base dados apenas dos municípios de alerta máximo