XXIV SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

Avaliação dos Dados Monitorados de Qualidade da Água usando a detecção de *Outliers* nas bacias Experimentais e Representativa do Rio Piabanha – Região Serrana do Estado do Rio de Janeiro – RJ.

Michele Bruna de Souza do Nascimento [1]; Janaina Pires Gomes da Silva [2]; Mariana Dias Villas Boas [3] ; João Pedro Costa da Silva[4]; Carlos Eduardo da Silva Sacramento[5]; Arthur Moreira de Abreu[6] & Rubens Esteves Kenup[7]

Palavras-Chave – *Outliers*, casos extremos, bacia Experimental, monitoramento

1 - INTRODUÇÃO

O monitoramento da qualidade da água é fundamental no gerenciamento dos recursos hídricos conforme a Lei nº 9.433, de 08 de janeiro de 1997, que ficou conhecida como Lei das Águas, a Política Nacional de Recursos Hídricos (PNRH). Seu planejamento inclui a seleção das variáveis de qualidade da água, localização das estações de amostragem e determinação das frequências de amostragem (KARAMOUZ et al., 2009).

Este monitoramento, para ser adequado, necessita que os dados obtidos, sejam consistidos para a análise do seu comportamento. Em uma modelagem estática dos mesmos, pode-se verificar o surgimento de *outliers*, que são observações atípicas do comportamento esperado dos mesmos, que irão afetar de uma forma negativa, modificando a precisão dos resultados.

Segundo a definição de OLIVEIRA et al. (2014), *Outliers* são os registros que em uma determinada série de números possuem, prevalecendo um determinado elemento do conjunto muito maior ou menor que os restos dos demais números. Com isso, pressupõem uma matriz de dados com determinado padrão e cujas dispersões serão os dados a serem pesquisados e analisados. OSBORNE; OVERBAY (2004) mencionam que geralmente os erros são causados por erro

1) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: michele.nascimento@cprm.gov.br
2) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: janaina.silva@cprm.gov.br
3) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: mariana.villasboas@cprm.gov.br
4) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: joão.pedro@cprm.gov.br
5) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: carlos.sacramento@cprm.gov.br
6) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: arthur.abreu@cprm.gov.br
7) CPRM/SGB: Companhia de Pesquisa de Recursos Minerais –Av. Pasteur 404–DEHID – RJ – (21) 2295-4546 – e-mail: rubens.kenup@cprm.gov.br

humano, como erros de coletas na gravação ou de entrada. Sendo necessário o tratamento destes, para não interferir na acurácia do produto final.

Conforme SILVA, (2004), em uma abordagem estatística, existe duas formas de se tratar *outliers*. A primeira é a acomodação, que consiste em criar métodos estatísticos que permitam inferências válidas da população mesmo que ocorram *outliers* na amostra dos dados. A segunda é conhecida como testes de discordância, que consiste na criação de um teste estatístico com o objetivo de identificar os *outliers*.

Na área de qualidade de água, um dado considerado *outlier, pelas metodologias de detecção* normalmente aplicadas, nem sempre significa um erro ou uma anomalia. Por isso, esses dados devem ser analisados de forma cuidadosa antes de serem excluídos da série de dados consistidos.

Dessa forma, o objetivo desse trabalho consiste na aplicação de metodologia para detecção de *outliers de modo a avaliar os dados e compreender possíveis tendências na qualidade da água* da bacia representativa do rio Piabanha (Projeto Estudo de Bacias Experimentais - CPRM) a partir da delimitação de zonas de normalidade e anormalidade utilizando como base as bacias experimentais e representativas definidas para o Projeto.

2 - MATERIAIS E MÉTODOS

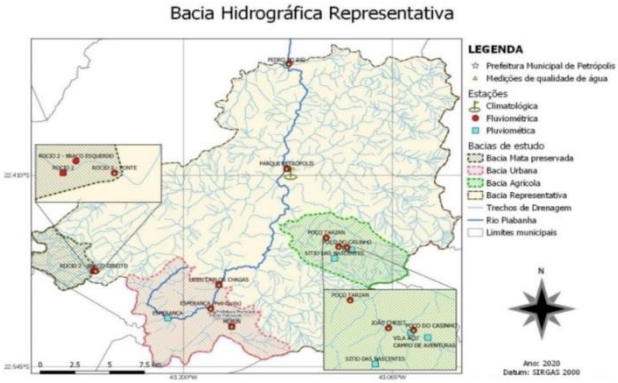
2.1 - Área de estudo

A área de estudo deste trabalho está contida na bacia representativa do rio Piabanha, que é a área de drenagem até a estação fluviométrica de Pedro do Rio (código: 58405000). Instalada em 1930, a estação possui uma série longa de dados consistentes. Atualmente compõe a Rede Hidrometeorológica Nacional (RHN). Sua bacia reúne as principais características de uso do solo e vegetação da bacia do rio Piabanha, localizada na região serrana do Estado do Rio de Janeiro, sendo definida como bacia representativa, de forma a possibilitar a extrapolação, para o restante da bacia, de resultados de estudos para ela realizados (SILVA, 2020). Dentro dessa área, foram definidas três bacias experimentais, onde predominam diferentes usos do solo como área de Mata Atlântica preservada, área relevante de uso agrícola e área de intensa ocupação urbana.

A rede hidrometeorológica quali-quantitativa do projeto EIBEX(Estudos Integrados em Bacias Experimentais e Representativas do Rio Piabanha, Região Serrana do RJ) apoia a avaliação do comportamento hidrológico nas bacias sob os diferentes usos. Para a bacia representativa, foram desenvolvidos estudos hidrológicos, com ênfase em climatologia, geomorfologia, umidade do solo e qualidade da água, sendo elaborados mapas temáticos e também realizados testes de diferentes tecnologias de medição de dados (VILLAS BOAS, 2018).

A figura 1 ilustra a localização das bacias experimentais e da bacia representativa, bem como os seus usos e cobertura vegetal.

Figura1: Mapa da bacia representativa as bacias experimentais monitoradas no projeto EIBEX – Fonte: SILVA, 2020



O monitoramento da qualidade de água bacia aconteceu com o total de 59 campanhas de medição *in situ* e coletas de amostras para análise em laboratório, ao longo do período 2009-2019 nas estações ilustradas no quadro 1.

Quadro 1 - Estações da rede EIBEX – localização e datas de início da operação

ESTAÇÃO	SIGLA	TIPO	Codigo FLU	Codigo PLU	BACIA EXPERIMENTAL	CURSO D'AGUA	LATITUDE	LONGITUDE	Data de início da operação - PLU/FLU	Data de início da operação - QA
Pedro do Rio	PR	PPRFDFrQT	58405000	2243012	-	Rio Piabanha	22º 24' 19"	43º 08' 00"	01/08/2009	27/08/2009
Pq.Petropolis	PP	CFDFrQT	58400250	2243286	-	Rio Piabanha	22º 30' 39"	43º 12' 37"	28/04/2007	27/08/2009
Esperança	E	PPRFDFrQ	58400010	2243287	URBANA	Rio Piabanha	22º 29' 14"	43º 10' 38"	24/04/2007	27/08/2009
Liceu	L	PPRFDFrQT	58400050	2243289	URBANA	Rio Palatinado	22º 31' 00"	43º 10' 08"	22/04/2007	27/08/2009
Morin	M	PPRFDFrQ	58400030	2243288	URBANA	Rio Bonfim	22º 27' 14"	43º 06' 28"	23/04/2007	27/08/2009
Poço Tarzan	PT	PPRFDFrQT	58400110	2243290	AGRÍCOLA	Rio Açu	22º 27' 39,6"	43º 05' 40,8"	31/10/2007	27/08/2009
Poço do Casinho	PC	FDFrQ	58400104	****	AGRÍCOLA	Rio Alcobaça	22º 27' 37,19"	43º 05' 59,76"	28/10/2007	27/08/2009
Joao Christ	JC	FDQ	58400108	****	AGRÍCOLA	Rio Piabanha	22º 19' 56"	43º 08' 01"	01/08/1930	27/08/2009
Rocio 2 - Ponte	R	FDQ	58400212	****	PRESERVADA	Rio da Cidade	22º 28' 38,70"	43º 15' 24,60"	28/04/2010	27/08/2009

Com o conjunto de parâmetros iniciais previstos para análise na RMQAP-EIBEX, alguns não figuraram em concentrações detectáveis ou não puderam ser analisados na maioria das campanhas em todas as estações.

Os 12 parâmetros selecionados foram: oxigênio dissolvido (OD); demanda bioquímica de oxigênio (DBO); demanda química de oxigênio (DQO); alumínio (Al); ferro (Fe); zinco (Zn); nitrato (NO_3); nitrogênio amoniacal (NH_4^+); fosfato (PO_4); sulfato (SO_4); sólidos em suspensão; e turbidez.

2.2 – Metodologia

Para aplicação da metodologia foi selecionada a bacia do rio Piabanha na qual está sendo desenvolvido um projeto institucional da CPRM – Serviço Geológico do Brasil intitulado "Estudos Integrados em Bacias Experimentais e Representativa – Região Serrana – RJ – EIBEX".

É importante explorar os dados disponíveis para avaliar a viabilidade de aplicação dos procedimentos que compõem a metodologia e suas possíveis restrições. Esse tipo de avaliação deve

ser feito com todos os dados fundamentais para apoiar as análises e a interpretação dos resultados. Essas análises permitem o conhecimento da variação estatística, como das concentrações dos parâmetros de qualidade da água e das tendências e dos comportamentos de cada um dos parâmetros.

Na literatura, encontramos diversos autores, que mostram algumas definições sobre o que são *outliers*. Abaixo, no quadro 2, temos alguns deles e suas descrições sobre o assunto.

Quadro 2 – Diferentes definições de *outliers* – Fonte: elaborado pelos autores (2016). FILHO et al., (2016)

AUTORES/ANO	DESCRIÇÃO
GRUBBS (1969)	*Uma observação periférica, ou outlier, é aquela que parece se desviar acentuadamente de outros membros da amostra em que ocorre.*
HAWKINS (1980)	*Uma observação que se desvia muito de outras observações a ponto de levantar suspeitas de que foi gerado por um diferente mecanismo.*
FOX (1991)	*Um outlier é uma observação cujo valor de variável dependente é incomum dado o valor da variável independente.*
JOHNSON (1992)	*Uma observação em um conjunto de dados que parece ser inconsistente com o restante desse conjunto de dados.*
MENDENHALL et al (1993)	*Observações cujos valores estão muito distantes do meio da distribuição em qualquer direção.*
ROSS (1996)	*Outlier são pontos de dados que parecem seguir o padrão dos outros casos.*
PYLE (1999)	*Um outlier é uma ocorrência única ou de frequência muito baixa do valor de uma variável que está longe da maioria dos valores da variável.*
MOORE e McCABE (1999)	*Um outlier é uma observação que está fora do padrão geral de um distribuição.*
RAMASMAWY, RASTOGI e SHIM (2000)	*Um outlier em um conjunto de dados é uma observação ou um ponto que é consideravelmente diferentes ou inconsistentes com o restante dos dados.*
BLUMAN (2000)	*Um "outlier" é um valor de dados extremamente alto ou extremamente baixo quando comparado com o resto dos valores de dados.*

A metodologia proposta tem como base método de Tukey, ou mais conhecido como boxplot (FILHO, 2016), **onde são definidos os limites inferior (L_{inf}) e superior (L_{sup})**, a partir dos quais um dado é considerado *outlier, utilizando o intervalo interquartil (IQR)*. O intervalo interquartil dá-se pela diferença entre o primeiro (Q1) e terceiro quartil (Q3), onde Q1 , representa um quarto ou 25% de todos os dados e o Q3 representa três quartos ou 75% de todos os dados considerados na série temporal.

$$IQR = Q3 - Q1 .$$

É utilizado um fator de 1,5 sobre o IQR para determinar os limites a partir das equações abaixo:

$$Linf = Q1 - (1.5 * IQR)$$

$$Lsup = Q3 + (1.5 * IQR)$$

Assim, qualquer valor que der fora deste intervalo (Q1>x>Q3), será considerado um *outlier*.

O intervalo interquartil mostra como os dados são espalhados sobre a mediana. É menos suscetível do que o intervalo a valores discrepantes e pode, portanto, ser mais útil.

A metodologia desenvolvida teve como base:

- Análise estatística/comparativa dos dados observados e da bibliografia;

- Organização dos dados disponíveis da bacia experimental e representativa para cada estação.
- Cálculos dos limites inferiores e superiores (25% e 75%) de cada parâmetro em três diferentes níveis: estação de monitoramento, bacia experimental e representativa. Assim para o nível de estação são utilizados a série de dados da estação, para o nível de bacia experimental são considerados os dados de todas as estações inseridas na bacia experimental e para o nível de bacia representativa são considerados os dados de todas as estações, ou seja, das nove estações consideradas no estudo
- Avaliação de cada dado de acordo com os três critérios em normal e anormal para cada estação.

RESULTADOS E DISCUSSÃO

Neste trabalho foram observados 12 (doze) parâmetros de qualidade de água, de 9 (nove) estações fluviométricas do Projeto EIBEX com características apresentadas na figura 1, referentes a cada uma das bacias do estudo, sendo três na bacia urbana (Morin, Esperança e Liceu Carlos Chagas), três na bacia agrícola (Poço do Tarzan, Poço Casinho e João Christ), duas nas estações representativas de controle (Pedro do Rio e Parque Petrópolis) e uma na bacia preservada (Rocio 2 - ponte).

Analisando os dados através dos gráficos e sabendo-se que a qualidade das águas muda ao longo do ano, devem ser considerados fatores importantes, como eventuais sazonalidades, entre eles, chuva, despejam de efluentes e vazões. Assim, para a análise exploratória comparativa da presença de cada parâmetro ao longo do tempo na bacia, representada pelas nove estações de monitoramento do Projeto EIBEX foram construídos gráficos relacionando as concentrações dos parâmetros de cada estação e os limites calculados através da metodologia.

As Figuras 2, 3 e 4 apresentam os gráficos das estações de monitoramento para os 12 parâmetros selecionados agrupados por bacia experimental. Os pontos representam os valores das concentrações OD parâmetro e as linhas os limites calculados, nos três níveis propostos, para orientar a definição de dados espúrios e identificar tendências de alteração na qualidade da água.

Vê-se que alguns parâmetros, tendem a mostrar dados fora dos limites das bacias, como DBO e Sólidos em Suspensão, principalmente na agrícola e urbana. Podemos atribuir este resultado, ao fato desses parâmetros, estarem ligados diretamente ao despejo de efluentes não tratados.

Temos um exemplo também, da estação João Christ, onde no parâmetro Nitrato, ele está fora do limite máximo da bacia agrícola, mas dentro do limite da bacia representativa.

Encontraram-se alguns pontos muito discrepantes. Na mesma estação de João Christ, por exemplo, no parâmetro de NH_4^+, observa-se apenas um ponto fora de todos os limites, sendo este quase 800 vezes maior que os outros. Ao observa-se todo o comportamento da estação, vemos que este dado é um *outlier*, que pode ter ocorrido a um erro na conversão de unidade.

Vemos outros casos semelhantes, como na estação de Poço Tarzan nos parâmetros Sulfato e Zinco, na estação Rocio no parâmetro Sólido em Suspensão e na estação Morin, no parâmetro Nitrato. Estes dados podem ser retirados para análises mais profundas, como para modelagem e aplicações de técnicas estatísticas multivariadas.

Outro dado interessante é o comportamento de Fosfato nas bacias. A bacia preservada e a agrícola, embora esteja tendo um valor discrepante, onde existe pouca influência industrial e doméstica, tem quase que todos seus valores iguais, variando muito pouco. Apesar de uma das fontes de fosfato ser na cultura agrícola. Já na bacia urbana, os valores variados, já que a região sofre influencia de efluentes domésticos, e industriais. Sabe-se que o esgoto é uma das principais fontes de fosfato, devido aos detergentes e sabões.

Para análise exploratória comparativa da presença de cada parâmetro ao longo do tempo na bacia, representada pelas nove estações de monitoramento do Projeto EIBEX, os resultados das análises de todos os parâmetros divididos por bacias.

Figura 2: Gráfico de monitoramento DBO e Turbidez

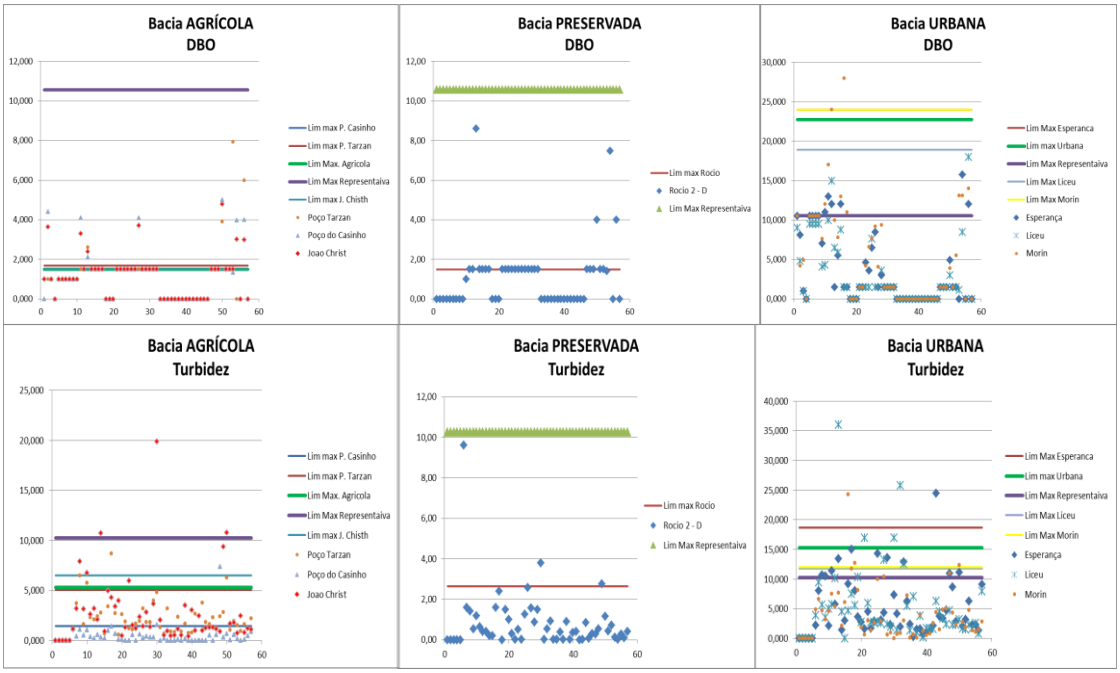


Figura 3: Gráfico de monitoramento Sulfato, Nitrato, DQO, OD e Sólidos em Suspensão

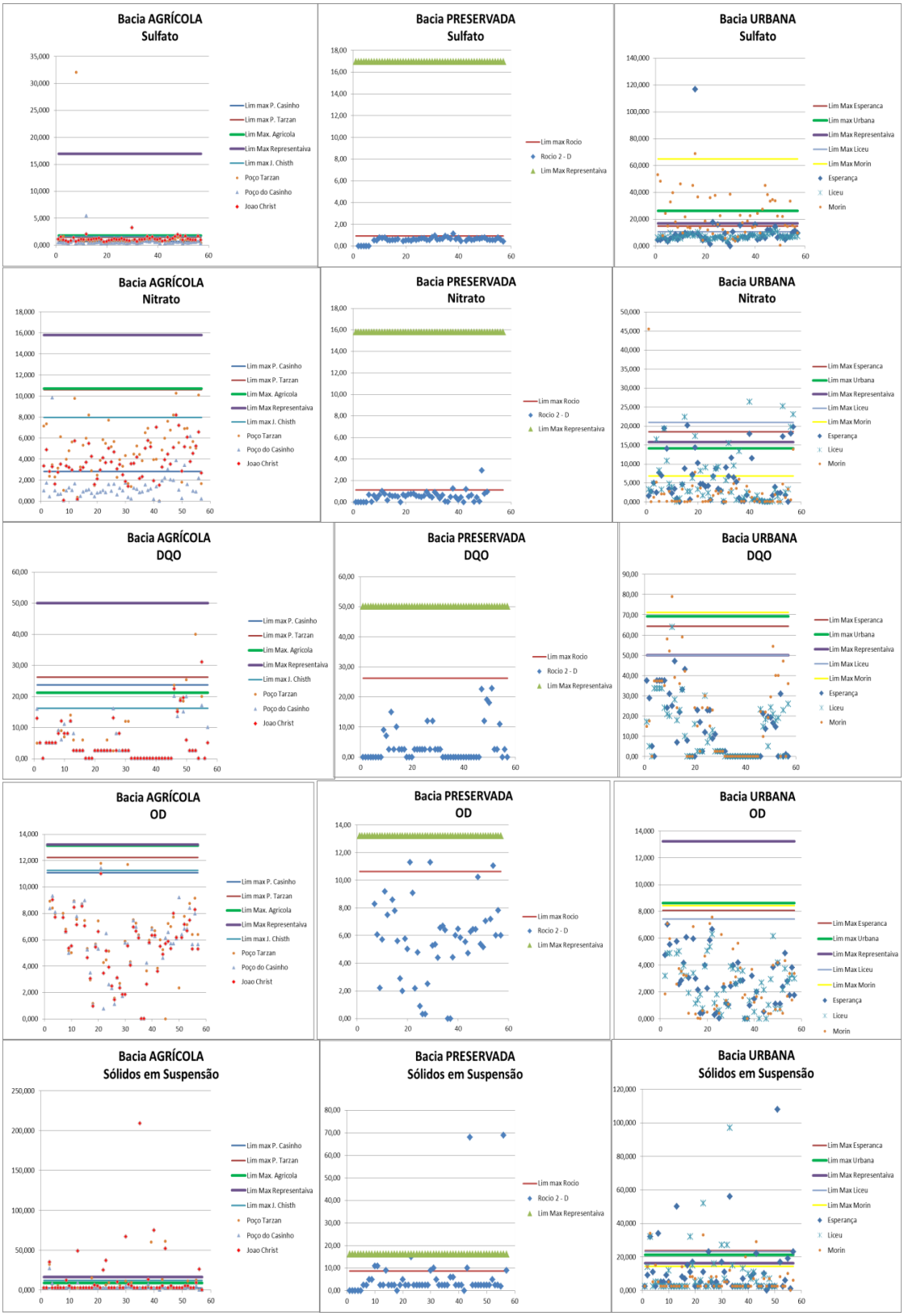
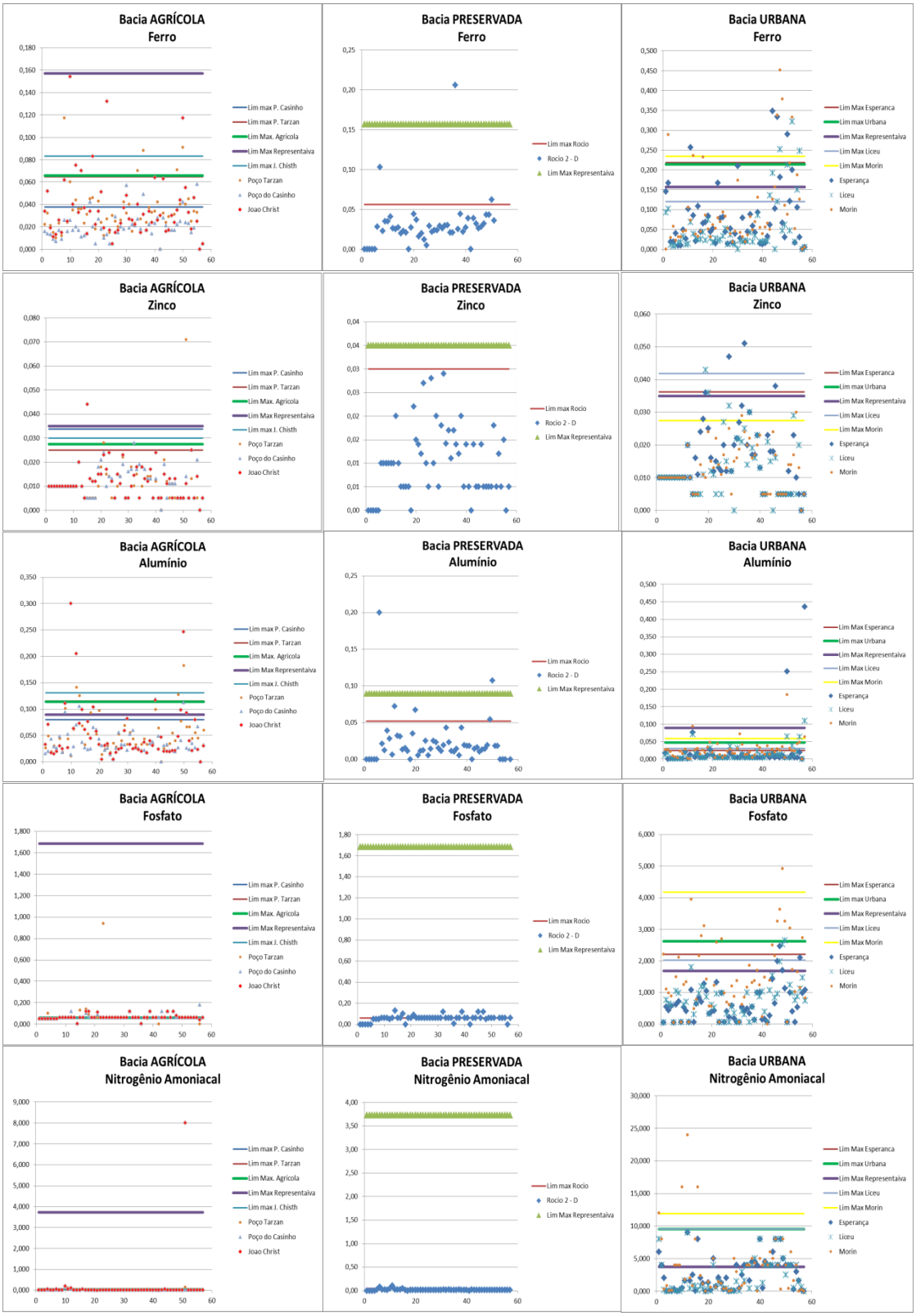


Figura 4: Gráficos de monitoramento Ferro, Zinco, Alumínio, Fosfato e Nitrogênio Amoniacal



CONCLUSÃO

Através dessa análise, conseguimos demonstrar como podemos encontrar os *outliers* nos dados obtidos, o que pode ser expandindo pra diversos tipos de dados, não se limitando a qualidade de água. O intervalo interquartil mostra como os dados são espalhados sobre a mediana.

É um método de fácil aplicação, rigoroso, que traz uma certeza maior aos resultados. Com os gráficos gerados, examinam-se os valores que estão fora dos limites esperados, principalmente quando estão em grande quantidade.

Então, conclui-se que tais dados, podem ocorrer por variadas maneiras, entre eles, erro humano (i.e. erro de digitação) e defeito de equipamento. Neste caso, observa-se, também, que algum comportamento anormal pode ser o responsável pelos *outliers* tais como: comportamento anormal de despejo de efluentes, novas construções, chuvas fora do padrão ou mesmo, a falta de manutenção do equipamento utilizado para medição.

Observar essas discrepâncias, além de melhorar a geração de resultados, podem trazer melhores práticas das equipes no campo, mudanças de técnicas de obtenção de dados e modernização da obtenção dos dados brutos, que são a base para qualquer estudo e trabalho.

Outro elemento importante que se destaca, é que através deste método, pode-se enxergar até que ponto esses *outliers* são mesmo uma anomalia, e devem ser excluídos da série de dados, ou se eles estão apontando um novo comportamento da qualidade a água do local de referência, já que traz uma linha espaço-temporal e comparativa de fácil visualização.

REFERÊNCIAS BIBLIOGRÁFICAS

BLUMAN, Allan. Elementary Statistics, brief version, New York: McGraw-Hill, 2000.

FILHO, D. B. F., SILVA, L. E. O. O *Outlier* que perturba o seu sono: como identificar casos extremos? - 10º Encontro Ciência Política e a Política: Memória e Futuro. Associação Brasileira de Ciências Política, 2016.

FOX, John. Regression diagnostics: An introduction. Sage, 1991.

GRUBBS, Frank E. Procedures for detecting outlying observations in samples. Technometrics, v. 11, n. 1, p. 1-21, 1969.

HAWKINS, Douglas M. Identification of outliers. London: Chapman and Hall, 1980.

Johnson R. Applied Multivariate Statistical Analysis. 1992. Prentice Hall.

KARAMOUZ, M. et al. Design of river water quality monitoring networks: a case study. Environmental Modeling & Assessment, v. 14, n. 705, 2009. DOI: https://doi.org/10.1007/s10666-008-9172-4.

VILLAS BOAS, M.D. Ferramenta para avaliação da rede de monitoramento de qualidade de água da bacia do rio Piabanha – RJ com base em redes neurais e modelagem hidrológica. 2018. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018.

Mendenhall, W., &, Sincich, T. (1996). A second course in statistics.

MOORE, David S.; MCCABE, George P. Introduction to the Practice of Statistics. WH Freeman/Times Books/Henry Holt & Co, 1989.

PYLE, Dorian. Data preparation for data mining. Morgan Kaufmann, 1999.

RAMASWAMY, Sridhar; RASTOGI, Rajeev; SHIM, Kyuseok. Efficient algorithms for mining outliers from large data sets. In: ACM SIGMOD Record. ACM, 2000. p. 427-438.

Ross, Sheldon (1996), *Introductory Statistics*, New York: McGraw-Hill.

SILVA, JANAINA. G. P. Avaliação da Influência do Usos e Ocupação do Solo em Bacias Experimentais sobre a Qualidade de Água no Rio Piabanha – Região Serrana do Rio de Janeiro-RJ. 2020 Dissertação (Mestrado) – Universidade Federal Fluminense, Niterói, 2020. DOI: https://dx.doi.org/10.22409.

SILVA, FLÁVIO. R. Uma abordagem para detecção de outliers em dados categóricos / Flávio Roberto Silva – Campinas/SP. s.n.J, 2004. Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

OLIVEIRA, C. D.; DE CAROLI, A. A.; AMARAL, A. S; VILCA, O. L. Detecção de fraudes, anomalias e erros em análise de dados contábeis: um estudo com base em outliers. REDECA – Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos da FEA-PUC/SP. Redeca, v.1, n. 1. Jan- Jun. 2014 p. 102-127.

OSBORNE, J. W., & Overbay, A. The Power of Outliers (and Why Researchers Should Always Check for Them). Practical Assessment, Research & Evaluation. North Carolina State University. Raleigh, USA, V. 9, N. 6, 2004.