



Universidade do Estado do Rio de Janeiro

Centro Biomédico

Faculdade de Ciências Biológicas e Saúde

Vinicius Tavares Martins Germano

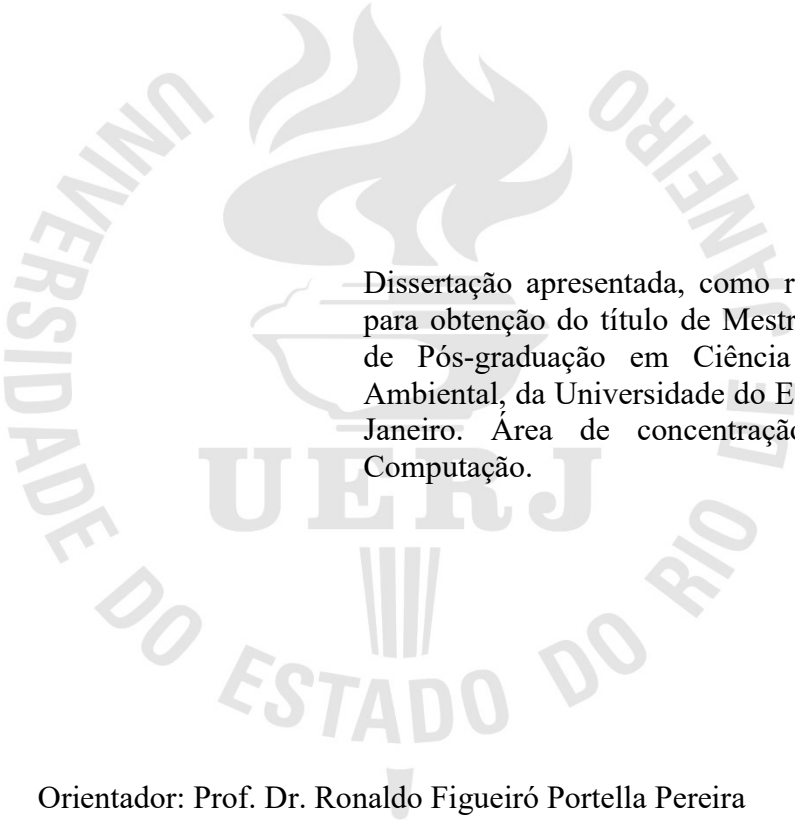
**Construção de um *pipeline* de dados para o Sistema Alerta Rio utilizando
Knime Analytics Platform**

Rio de Janeiro

2023

Vinicius Tavares Martins Germano

**Construção de um *pipeline* de dados para o Sistema Alerta Rio utilizando Knime
Analytics Platform**



Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Ciência e Tecnologia Ambiental, da Universidade do Estado do Rio de Janeiro. Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. Ronaldo Figueiró Portella Pereira

Rio de Janeiro

2023

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CZO

G373

Germano, Vinicius Tavares Martins

Construção de um pipeline de dados para o Sistema Alerta Rio utilizando
Knime Analytics Plataforma / Vinicius Tavares Martins Germano - 2023.
69 f.

Orientador: Ronaldo Figueiró Portella Pereira

Dissertação (Mestrado Profissional) – Universidade do Estado do Rio de
Janeiro, Faculdade de Ciências Biológicas e da Saúde, Programa de Pós-
Graduação em Ciência e Tecnologia Ambiental.

1. Ciência e Tecnologia de Ambiental - Teses. 2. Sistema Alerta Rio -
Teses. 3. Dados abertos - Teses. 4. Big Data. 5. Pipeline - Teses I. Pereira,
Ronaldo Figueiró Portella. II. Universidade do Estado do Rio de Janeiro.
Faculdade de Ciências Biológicas e da Saúde. III. Título.

CDU 502

Bibliotecária: Rosana Maria dos Passos – CRB 7 4908

Autorizo apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta
dissertação, desde que cite a fonte.

Assinatura

Data

Vinicius Tavares Martins Germano

**Construção de um *pipeline* de dados para o Sistema Alerta Rio utilizando Knime
Analytics Platform**

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-graduação em Ciência e Tecnologia Ambiental, da Universidade do Estado do Rio de Janeiro. Área de concentração: Ciência da Computação.

Aprovada em 31 de outubro de 2023.

Banca Examinadora:

Prof. Dr. Ronaldo Figueiró Portella Pereira (Orientador)
Faculdade de Ciências Biológicas e Saúde - UERJ

Prof. Dr. Alexander Machado Cardoso
Faculdade de Ciências Biológicas e Saúde - UERJ

Prof. Dr. Paulo Roberto de Amoretty
Centro Universitário de Volta Redonda

Rio de Janeiro

2023

Os dados são o novo petróleo

Clive Humby

RESUMO

GERMANO, Vinicius Tavares Martins Germano. *Construção de um pipeline de dados para o Sistema Alerta Rio utilizando Knime Analytics Platform*. 2023. 69 f. Dissertação (Mestrado em Ciência e Tecnologia Ambiental) – Faculdade de Ciências Biológicas e Saúde, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

O presente trabalho está inserido no paradigma da Ciência Aberta, o qual, impulsionado pelo desenvolvimento de tecnologias digitais, preconiza o compartilhamento e a colaboração científica. Nesse contexto, dados governamentais abertos são reconhecidos como relevantes insumos para pesquisa. O Sistema Alerta Rio disponibiliza dados meteorológicos e pluviométricos do município do Rio de Janeiro, porém sem uma estruturação adequada. Este estudo empregou técnicas de Engenharia de Dados utilizando o software Knime, com o objetivo de desenvolver um *pipeline* para integração, limpeza e análise desses dados, seguindo princípios de reprodutibilidade, rigor e transparência preconizados pela Ciência Aberta. Os dados foram consolidados em um novo banco de dados (*Data Staging*) e disponibilizados no repositório Mendeley Data. O banco de dados desenvolvido poderá ser utilizado como fonte para estudos futuros, para interpretação e discussão de resultados, bem como para novas etapas de processamento e integração com outras fontes de dados, viabilizando aplicações de aprendizado de máquina (*Machine Learning*).

Palavras-chave: Sistema Alerta Rio. Dados Abertos. *Big Data*. *Pipeline*.

ABSTRACT

GERMANO, Vinicius Tavares Martins Germano. Building a data pipeline for the Alerta Rio System using Knime Analytics Platform. 2023. 69 f. Dissertação (Mestrado em Ciência e Tecnologia Ambiental) – Faculdade de Ciências Biológicas e Saúde, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

The present work is framed within the Open Science paradigm, which, catalyzed by the development of digital technologies, advocates for scientific sharing and collaboration. In this context, open government data are recognized as relevant research inputs. The Sistema Alerta Rio provides meteorological and rainfall data for the municipality of Rio de Janeiro, albeit without proper data structuring. This study employed Data Engineering techniques using Knime software, with the goal of developing a pipeline for integration, cleaning and analysis of these data, following principles of reproducibility, rigor and transparency espoused by Open Science. The curated data were consolidated into a new database (Data Staging) and made available in the Mendeley Data repository. The developed database can be used as a source for future studies, for interpretation and discussion of results, as well as for new processing steps and integration with other data sources, enabling machine learning applications.

Keywords: Sistema Alerta Rio. Big Data. Pipeline. Open Data.

LISTA DE FIGURAS

Figura 1 –	O icebergue da Investigação Científica	16
Figura 2 –	Preprints publicados no medRxiv entre junho de 2019 e setembro de 2020	18
Figura 3 –	Evolução do número de preprints biomédicos	19
Figura 4 –	Mudanças na investigação científica biomédica após a Pandemia COVID-19	20
Figura 5 –	Taxonomia para Ciência Aberta	21
Figura 6 –	Frequência da narrativa Crise (título, resumo ou palavras-chave)	23
Figura 7 –	Processo de Publicação de Artigos na F1000Research	25
Figura 8 –	Metodologia desenvolvida pelo Grupo Executivo do Comitê Interministerial Governo Aberto (GE-CIGA) para a construção dos Compromissos que farão parte dos Planos de Ação Nacionais	27
Figura 9 –	Dado, Informação e Conhecimento	29
Figura 10 –	Pirâmide DIKW	31
Figura 11 –	Modelo DIK2	32
Figura 12 –	Processos P1 a P14 do modelo DIK2	33
Figura 13 –	Ambiente da Tecnologia da Informação	35
Figura 14 –	Cadeia de valor DIKW	35
Figura 15 –	6 Vs do Big Data	37
Figura 16 –	Engenharia de Dados e Ciência de Dados	38
Figura 17 –	Interdisciplinaridade da Ciência de Dados	40
Figura 18 –	Interdisciplinaridade da Bioinformática	40
Figura 19 –	Tomada de Decisão Guiada por Dados	41
Figura 20 –	Relação Big Data, Ciência de Dados e Tomada de Decisão	41
Figura 21 –	Estrutura esquemática de um pipeline de dados	42
Figura 22 –	Representação esquemática de pipeline Big Data	43
Figura 23 –	Exemplo de um fluxo de trabalho (workflow) Knime	45
Figura 24 –	Forum Knime “RULE ENGINE”	46
Figura 25 –	Workflow Knime "BLAST to the Past to Investigate Ancient DNA" com autoria de Jeanette (Jeany) Prinz	47
Figura 26 –	Maturidade em Dados Abertos	60

LISTA DE TABELAS

Tabela 1 –	População das AP em 2010	42
Tabela 2 –	População estimada das AP em 2020	42
Tabela 3 –	Área das AP.....	42
Tabela 4 –	Estações (ativas) Sistema Alerta Rio por AP	43
Tabela 5 –	Áreas das Zonas Pluviométricas por AP	44

LISTA DE MAPAS

Mapa 1 –	Rio de Janeiro – Áreas urbanizadas e não urbanizadas e suas Áreas de Planejamento (AP) e Regiões de Planejamento (RP)	50
Mapa 2 –	Estações Sistema Alerta Rio (ativas e inativas)	54
Mapa 3 –	Estações Sistema Alerta Rio (ativas) e Zonas Pluviométricas.....	55

LISTA DE ABREVIATURAS E SIGLAS

AP	Área de Planejamento
Capes	Coordenação de Aperfeiçoamento de Pessoal do Nível Superior
CGU	Controladoria Geral da União
CINASAMA	Congresso Internacional de Saúde e Meio Ambiente
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COVID	Coronavírus
DIK2	Dado-Informação-Conhecimento-Inteligência-Sabedoria
DIKW	<i>Data-Information-Knowledge-Wisdom</i>
DOI	<i>Digital Object Identifier</i>
Embrapa	Empresa Brasileira de Pesquisa Agropecuária
Fiocruz	Fundação Oswaldo Cruz
GE-CIGA	Grupo Executivo do Comitê Interministerial Governo Aberto
GEO-RIO	Fundação Instituto de Geotécnica do Município do Rio de Janeiro
I4OC	<i>Initiative for Open Citations</i>
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
IPEA	Instituto de Pesquisa Econômica Aplicada
LAI	Lei de Acesso à Informação
Latindex	<i>Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal</i>
MonitorAr Rio	Programa de Monitoramento da Qualidade do Ar
OGP	<i>Open Government Partnership</i>
OLAP	<i>Online Analytical Processing</i>
OMS	Organização Mundial da Saúde
PLOS	<i>Public Library of Science</i>
PPGB	Programa de Pós-Graduação em Bioinformática
Redalyc	Rede de Revistas Científicas da América Latina e Caribe, Espanha e Portugal
RNP	Rede Nacional de Ensino e Pesquisa
RP	Regiões de Planejamento
SciELO	<i>Scientific Electronic Library Online</i>
SSRN	<i>Social Science Research Network</i>
TDIC	Tecnologias Digitais da Informação e Comunicação
TIC	Tecnologias da Informação e Comunicação
TSSMN	<i>Thoracic Surgery Social Media Network</i>
UFPR	Universidade Federal do Paraná
UNIFEI	Universidade Federal de Itajubá
ZP	Zona Pluviométrica

SUMÁRIO

	INTRODUÇÃO	13
1	REVISÃO DA LITERATURA	17
1.1	Ecosistema da Ciência Aberta	17
1.2	Ciência Aberta e a Pandemia COVID-19	17
1.3	Tecnologias e Ferramentas para Comunicação Científica	23
1.4	Governo Aberto e Ciência Aberta no Brasil	27
1.5	Dado, Informação e Conhecimento	29
1.6	Ambiente de Tecnologia da Informação	35
1.7	BigData	37
1.8	Engenharia de Dados e Ciência de Dados	38
1.9	Pipeline de dados	43
1.10	Knime	45
1.11	Rio de Janeiro	49
1.12	Sistema Alerta Rio	52
2	OBJETIVOS	56
3	MATERIAIS E MÉTODOS	57
4	RESULTADOS	58
4.1	DISCUSSÃO	59
4.2	Knime Analytics Platform	59
4.3	Sistema Alerta Rio	60
4.4	Mendeley Data	61
	CONCLUSÃO	63
	REFERÊNCIAS	65
	ANEXO – Capítulo de Livro	70

INTRODUÇÃO

As Tecnologias da Informação e Comunicação (TIC), como televisão, mimeógrafo, xerox, impressão a jato etc., tiveram forte impacto na vida das pessoas, contribuindo para a inovação e a disseminação de conhecimento em diversas áreas. O advento das Tecnologias Digitais da Informação e Comunicação (TDIC), também chamadas de Novas Tecnologias (em contraposição às TIC, consideradas antigas), como a internet, dispositivos móveis e sensores digitais, resultou em uma explosão de dados sem precedentes (CABALLERO-RIVERO, SÁNCHEZ-TARRAGÓ e SANTOS, 2019; COSTA, DUQUEVIZ e PEDROZA, 2015).

O crescimento exponencial dos dados ao longo do século XX resultou na cunhagem do termo *Big Data* e aprofundou os estudos em Ciência de Dados e Ciência da Informação. A globalização e a profusão tecnológica fomentou movimentos internacionais em prol de irrestrito acesso à Informação e à Ciência, sem restrições legais, tecnológicas ou sociais.

A demanda social por informação levou Governos, Instituições Públicas e Privadas, a publicarem diversos repositórios de dados em estado bruto, muitas das vezes não estruturados, demandando o processamento e manutenção desses dados.

Os trabalhos científicos em equipe (colaboração científica) remontam ao século XVII após o surgimento da ciência moderna, com o primeiro artigo colaborativo publicado em 1665 por Hooke, Oldenburg, Cassini e Boyle. Os trabalhos colaborativos aumentaram de forma expressiva no final do século XVIII, sobretudo na área da Astronomia, e expuseram as “disparidades entre a pesquisa experimental e a teórica” (AUTRAN, 2020; RODRIGUES, 2020; CABALLERO-RIVERO, SÁNCHEZ-TARRAGÓ e SANTOS, 2019).

No mesmo período, o progresso sócio-tecnológico contribuiu para aumento de volume e velocidade no compartilhamento das informações científicas, antes restritas a manuscritos, cartas, livros etc., e posteriormente em periódicos com sistema “*peer review*” [revisão pelos pares], construindo o *corpus* da literatura científica e conferindo *status* às Universidades e Institutos de Pesquisa. Nesse contexto, surgiram os primeiros periódicos especializados nas diversas áreas do conhecimento (Artes, Letras, Ciências etc.), amplificando a quantidade de informação produzida (AUTRAN, 2020; OLIVEIRA e SOBREIRA, 2020; RODRIGUES, 2020; CABALLERO-RIVERO, SÁNCHEZ-TARRAGÓ e SANTOS, 2019).

As TDIC contribuíram não só para o volume de dados gerados, mas também impuseram transformações na forma com que as informações científicas são disseminadas, com impacto comparável ao surgimento da imprensa. O custo de assinatura dos periódicos

desencadeou uma “crise dos periódicos”, com impacto no mercado editorial e conglomerados internacionais, culminando em uma limitação de acesso às publicações científicas. Como resposta, surgiram repositórios de acesso público (reduzindo a importância dos “intermediários comerciais” na distribuição da informação) e a formação de movimentos internacionais de *Open Access*, *Open Science* e *Open Data* (AUTRAN, 2020; OLIVEIRA e SOBREIRA, 2020; OLIVEIRA, GUIMARÃES e KOSHIYAMA, 2019; RIBEIRO, OLIVEIRA e ARAÚJO, 2019; SILVA e SILVEIRA, 2019).

Rodrigues (2020) e Silva e Silveira (2019) são francos em afirmar que antigamente, as inovações científicas eram financiadas por pessoas ricas e pela Igreja. Em consenso, os autores afirmam que a consequência de a Ciência contribuir para a evolução da sociedade foi que os investimentos científicos foram institucionalizados nos meios governamentais, sociais e empresariais, sobretudo pela Ciência ter se constituído como “ativo competitivo no contexto capitalista”. Esse fenômeno inverte a lógica inicial: “se antes era a economia que fomentava a ciência, agora a ciência é que oferece subsídios para o crescimento da economia”.

Oliveira, Guimarães e Koshiyama (2019) defendem que os Direitos de Propriedade Intelectual em uma sociedade caracterizada por globalização, compartilhamento, colaboração e ambiente digital, devem ser repensados para “um novo fazer científico no cenário da ciência contemporânea”, a conduzir a investigação científica aberta e pública.

Os autores exortam a “compreensão da ciência tradicional para o prisma da ciência do *commons*”, privilegiando a colaboração e a coletividade. Nesse paradigma do *commons*, os bens intelectuais são tidos como constructos¹ e viabilizados por práticas colaborativas como “produção *peer-to-peer* [entre os pares], *wiki* e *crowdsourcing*”. Nos paradigmas econômicos, o *commons* contrapõe-se aos princípios de propriedade, exclusividade e individualidade.

Quintanilha (2019) investigou os desafios da ciência aberta e apontou 4 fragilidades:

- (des)acreditação: a legitimação das publicações e periódicos atrelada ao alto fator de impacto (muitos de acesso fechado) impõe uma “ditadura das métricas e dos fatores de impactos”, levando Eugene Garfield (autor do conceito de fator de impacto) a se referir a uma “Era da *cienciometria* e da *journalologia*”²;

¹“Conceito ou construção teórica, puramente mental, elaborada ou sintetizada com base em dados simples, a partir de fenômenos observáveis, que auxilia os pesquisadores a analisar e entender algum aspecto de um estudo ou ciência” (Michaelis).

²Ver Lei de Goodhart em Rodrigues (2020) p.274-5: “*When a measure becomes a target, it ceases to be a good measure*” [“Quando uma medida se torna um alvo, ela deixa de ser uma boa medida”, tradução própria].

- informalidade: ausência de procedimentos padronizados, como os inerentes ao corpo editorial em um periódico. Autran (2020) menciona o uso do Twitter como “alternativa para postagens científicas e alvo de estudos altmétricos”³;
- comodificação: consequência do primeiro desafio, consiste na criação de periódicos fechados, sob motivação injusta de comercializar prestígio, na forma de publicação de artigos de qualidade duvidosa ou inexistente;
- predação: atrelada à (des)acreditação e comodificação, consiste na produção em massa de publicações, impulsionando prestígio, em detrimento de trabalhos de qualidade.

Por fim, Quintanilha (2019) atribui a atores políticos e instituições acadêmicas a competência para debater, legislar e regular a Ciência Aberta, no sentido de requalificá-la, sugerindo apoio e valorização de periódicos de acesso aberto, “sistema alternativo de recompensa acadêmica baseada na revisão de artigos científicos por pares”, e o debate sobre a hegemonia do fator de impacto na legitimação do conhecimento científico.

Com o uso da Tecnologia, a Ciência Aberta propõe uma comunicação científica “desde o início da pesquisa, por meio das novas práticas de compartilhamento”. Desta forma, a colaboração científica pode ser realizada desde a concepção e delineamento da pesquisa, e não somente após a publicação via periódicos, teses e dissertações. O fator cultural de não compartilhamento em vigor, oculta inúmeras etapas e elementos da pesquisa científica (Figura 1) (SILVA e SILVEIRA, 2019; CABALLERO-RIVERO, SÁNCHEZ-TARRAGÓ e SANTOS, 2019).

³“Medidas de impacto acadêmico não tradicionais baseadas em atividades na web, como o ato de tuitar”, curtir, compartilhar etc. (QUINTANILHA, 2019).

1 REVISÃO DA LITERATURA

1.1 Ecossistema da Ciência Aberta

A Ciência Aberta possui diversas dimensões, também definida como “movimento de movimentos”, com abrangência em diversas áreas, “de economia a práticas culturais”. Em resumo, a Open Knowledge Foundation preconiza que a abertura do conhecimento científico deve garantir seu livre acesso, uso, reuso e distribuição sem restrições legais, tecnológicas ou sociais. Nesse sentido, relaciona-se às práticas de dados abertos, acesso aberto, domínio público, *software* livre, *open source*, *E-science*, entre outras (SILVEIRA, *et al.*, 2021; CLINIO, 2019; SILVA e SILVEIRA, 2019; CABALLERO-RIVERO, SÁNCHEZ-TARRAGÓ e SANTOS, 2019).

O grupo Facilitate Open Science Training for European Research (Foster) elaborou a Taxonomia Foster para Ciência Aberta, apresentando algumas dessas diversas frentes de demandas atreladas à Ciência Aberta agrupadas em 9 facetas principais. Silveira *et al.*, (2021) traduziram a Taxonomia Foster, consideraram 13 definições de Ciência Aberta de autores brasileiros, reformularam a taxonomia, e em colaboração com especialistas brasileiros convidados, propuseram nova Taxonomia para Ciência Aberta, com 11 facetas (**Erro! Fonte de referência não encontrada.**) (SILVEIRA *et al.*, 2021).

1.2 Ciência Aberta e a Pandemia COVID-19

Com a emergência da Pandemia COVID-19, a investigação científica nas Ciências Biomédicas foi fortemente impactada, acentuando experiências de colaboração científica ocorridas em eventos recentes nas emergências de Ebola e Zika (RODRIGUES, 2020).

O repositório de *preprints* medRxiv⁴, especializado em Ciências da Saúde, foi lançado em junho de 2019 e a partir de janeiro de 2020, instalada a crise internacional de saúde

⁴www.medrxiv.org.

pública que levou a Organização Mundial da Saúde (OMS) a declarar estado de emergência,⁵ o medRxiv registrou forte crescimento, saindo de menos de 500 para quase 2.000 *preprints* publicados em um mês (**Erro! Fonte de referência não encontrada.**). Junto com outros repositórios, os *preprints* na literatura biomédica do PubMed saiu de 2,4% em janeiro de 2019 para mais de 8% em junho de 2020 (**Erro! Fonte de referência não encontrada.**). De forma esquemática, Rodrigues (2020) elenca as principais mudanças na investigação científica na área biomédica com a pandemia de COVID-19, explicitando a colaboração como novo novo paradigma (**Erro! Fonte de referência não encontrada.**) (RODRIGUES, 2020).

⁵www.agenciabrasil.abc.com.br/saude/noticia/2020-01/oms-declara-estado-de-emergencia-global-em-razao-do-coronavirus.

1.3 Tecnologias e Ferramentas para Comunicação Científica

Oliveira e Sobreira (2020) afirmam que, atualmente, a “comunicação científica tem a circulação como pilar para a troca de informações, sendo essencial para a ciência”. Segundo os autores, a sistemática do fluxo dessa circulação se dá pelos processos de Descoberta, Análise, Escrita, Publicação, Divulgação e Avaliação. Os autores discorrem sobre formas de apropriação da Ciência Aberta em cada etapa:

- Descoberta: relacionada à investigação bibliográfica, prevê uso de repositórios de acesso aberto, como Google Acadêmico, Mendeley, ResearchGate etc. Os autores citam a Wikipédia como “espaço de reputação por endosso, alternativa [... a]o sistema de prestígio tradicional baseada em dinâmicas de trocas em pares”;
- Análise: associada à metodologia, e portanto, à reprodutibilidade da pesquisa, pode ser alcançada com ferramentas como RStudio⁶, Dh-Box⁷, Projeto Jupyter⁸, e espaços de comunidades como o Github⁹. Para Ribeiro, Oliveira e Araújo (2019), foi um marco histórico à abertura dos dados científicos, as Agências de Fomento terem tornado o planejamento do gerenciamento de dados obrigatório para obter financiamento em projetos de pesquisa. Rodrigues (2020) relata que em 2016 a Nature divulgou que 70% dos pesquisadores entrevistados em uma pesquisa, já teve dificuldade em reproduzir a investigação dos pares e 52% em reproduzir a própria pesquisa, impulsionando o debate acerca de uma Crise de Reprodutibilidade da Ciência. Fanelli (2018), em um artigo de opinião, rebate essa narrativa, que remonta ao ano 1990 **(Erro! Fonte de referência não encontrada.)**;
- Escrita: as TDIC colaborativas como Google Docs já são amplamente difundidas, podendo serem incorporadas nesta etapa. Outros editores, baseados em linguagem LaTeX¹⁰, além de colaborativos, caracterizam-se por definirem

⁶<https://posit.co/>

⁷<https://dhbox.commons.gc.cuny.edu/>

⁸<https://jupyter.org/>

⁹<https://github.com/>

¹⁰“Uma linguagem de programação simplificada para diagramação de documentos científicos”, destinada a aumentar “o foco no texto e não na formatação”. A Revista Acústica e Vibrações da Sociedade Brasileira de Acústica (Sobrac), destinada a publicação de temas de acústica, vibrações, áudio, audição e

a estrutura (a depender se artigo, livro ou carta) e a formatação (negrito, itálico), sendo exigido que o usuário insira o texto não formatado (*plain text*). Os autores também citam ferramentas de inteligência artificial, denominados “algoritmos de escrita acadêmica”, mencionando o livro *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research*¹¹, o primeiro livro de pesquisa gerado por máquina;

- Publicação: após endossarem as críticas ao oligopólio do mercado editorial formado por ACS, Reed-Elsevier, Springer, Wiley-Blackwell, Taylor & Francis, e Sage, os autores citam plataformas, sobretudo na América Latina, como Scientific Electronic Library Online (SciELO), Rede de Revistas Científicas da América Latina e Caribe, Espanha e Portugal (Redalyc) e Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal (Latindex). Clinio (2019) se debruçou especificamente sobre a Ciência Aberta na América Latina e menciona uma postura de “internacionalização”, em oposição aos “imperativos europeus e norte-americanos e a sistemas de avaliação ‘objetivos’ que, paradoxalmente, são opacos e orientados para o mercado”. Oliveira e Sobreira (2020) também citam repositórios de *preprints*, como os mantidos pela Open Science Framework, organizados em disciplinas. Esse tipo de repositório também despertou interesse na Elsevier, que comprou a SSRN em 2016¹²¹³;
- Divulgação: devido à estreita relação da mídia com a divulgação científica,¹⁴ as mídias digitais (Facebook, Twitter, ResearchGate, Academia.edu, Vimeo, TedX, YouTube etc) se tornam “espaços privilegiados” à divulgação científica, embora os autores alertem para o viés mercantil dessas redes sociais (nesse aspecto, a Wikipédia se diferencia pela cultura colaborativa). Os autores alertam que nessas plataformas, pelo fato de o engajamento ser uma importante métrica, há risco de uma “caça por cliques” [*clickbait*], privilegiando “títulos

áreas correlatas, tem sua diagramação finalizada em LaTeX, e por isso, não só recomenda o uso pelos autores, como disponibiliza um modelo de artigo (template) em LaTeX:

<https://pt.overleaf.com/read/mnmwhwcsykjh> (FONSECA, 2020).

¹¹<https://doi.org/10.1007/978-3-030-16800-1>

¹²<https://www.elsevier.com/connect/ssrn-the-leading-social-science-and-humanities-repository-and-online-community-joins-elsevier>.

¹³<https://www.elsevier.com/pt-br/solutions/ssrn>.

¹⁴Ver item “2 Da hélice tríplice à hélice quádrupla: a ciência aberta enquanto campo em disputa” em Oliveira e Sobreira (2020) para compreender a evolução da tríade [academia-governo-indústria] em [academia-governo-indústria-mídia-sociedade].

instigantes e jogos de palavras”. Investigando a divulgação científica no Twitter, Karami *et al.* (2020) mapearam 38 tópicos em mais de 18 mil manuscritos publicados entre 2006 e 2019;

- Avaliação: como alternativa ao fator de impacto medido pela Clarivate Analytics e Scopus¹⁵, os autores citam a Initiative for Open Citations (I4OC),¹⁶ co-fundada por Open Citations, Fundação Wikimedia, PLOS (Public Library of Science), eLife, DataCite e Centro de Cultura e Tecnologia da Curtin University, com mais de 60 organizações associadas, sendo 29 dessas, editoras científicas dispostas a compartilhar seus metadados de citação abertamente (como Springer, Nature, Taylor & Francis e Wiley). Com relação à avaliação por pares, considerado o “maior entrave do fluxo editorial científico”, Oliveira e Sobreira (2020) citam, como alternativa, a F1000Research¹⁷ (**Erro! Fonte de referência não encontrada.**), co-fundada pela Organização Mundial de Saúde com editores do oligopólio científico.

¹⁵Juntos com Google Scholar, Scopus e Clarivate são “os indicadores bibliométricos utilizados no Qualis-Referência” da CAPES (Documento Técnico do Qualis Periódicos disponível em <https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/avaliacao/avaliacao-quadrinal-2017/DocumentotecnicoQualisPeridicosfinal.pdf>).

¹⁶www.i4oc.org.

¹⁷www.f1000research.com

1.4 Governo Aberto e Ciência Aberta no Brasil

Em 2011, o Brasil, junto com outros 7 países, co-fundou a Open Government Partnership (OGP), com objetivo de “difundir e incentivar, globalmente, práticas governamentais relacionadas à transparência dos governos, ao acesso à informação pública e à participação social”. Em 2023 a OGP contabiliza 76 países associados e teve orçamento médio anual de US\$ 2,8 mi no período 2018-2022¹⁸ (BERTIN *et al.*, 2019).

Ao associar-se à OGP, o país membro deve atender a critérios de elegibilidade de “transparência fiscal, acesso à informação, participação cidadã e divulgação de declarações patrimoniais das autoridades”, subscrevendo a Declaração de Governo Aberto.¹⁹ Após a elegibilidade, um Plano de Ação é construído para um período de execução de 2 anos. Nesse Plano de Ação, o Estado se compromete com princípios de Governo Aberto de “transparência, *accountability*²⁰, participação cidadã e tecnologia e inovação” (BERTIN *et al.*, 2019).

No Brasil, o Governo Aberto está a cargo da CGU, através do Grupo de Trabalho (GT) Inovação e Governo Aberto na Ciência, que tem como membros a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), a Coordenação de Aperfeiçoamento de Pessoal do Nível Superior (Capes), o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Fiocruz (Fundação Oswaldo Cruz), a Open Knowledge Foundation Brasil, a Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação da UnB, a Equipe de Monitoramento (da CGU – Controladoria Geral da União) e a Rede Nacional de Ensino e Pesquisa (RNP) (SILVA e SILVEIRA, 2019).

Neste ano de 2023, a CGU elabora o 6º Plano de Ação²¹, constituído de (**Erro! Fonte de referência não encontrada.**):

1. Definição dos Temas
 - Temas indicados pela sociedade civil
 - Temas indicados pelo Governo
2. Oficinas de Cocriação

¹⁸ <https://www.opengovpartnership.org/about/transparency-financial-information/government-contributions/>

¹⁹ <https://www.gov.br/cgu/pt-br/governo-aberto/central-de-conteudo/documentos/arquivos/declaracao-governo-aberto.pdf>

²⁰ <https://portal.tcu.gov.br/governanca/governancapublica/componentes/accountability/>

²¹ <https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/6deg-plano-de-acao-brasileiro/elaboracao-do-6o-plano-de-acao-brasileiro/>

3. Aprovação do Plano

Ainda em 2011, foi publicada a Lei nº 12.527/2011 (Lei de Acesso à Informação – LAI),²² destinada a “assegurar o direito fundamental de acesso à informação” com 5 diretrizes:

1. observância da publicidade como preceito geral e do sigilo como exceção;
2. divulgação de informações de interesse público, independentemente de solicitações;
3. utilização de meios de comunicação viabilizados pela tecnologia da informação;
4. fomento ao desenvolvimento da cultura de transparência na administração pública;
5. desenvolvimento do controle social da administração pública.

Em 2012 o Movimento Ciência Cidadã lançou no Brasil, o Manifesto pela Ciência Cidadã²³ com o lema “A Pesquisa é um bem público. Pertence ao público a escolha de seus benefícios”, denunciando uma “acumulação de crises provocada por uma sociedade inspirada em uma única visão produtivista”, citando, dentre tantas, o “aquecimento global”, hoje compreendido como as Emergências Climáticas²⁴.

Em 28 de setembro de 2016, Dia Internacional do Acesso Universal à Informação, o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) lançou o Manifesto de Acesso Aberto a Dados da Pesquisa Brasileira para Ciência Cidadã²⁵ (Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica²⁶), propondo o “paradigma do acesso livre à informação” baseado nos termos da Declaração de Berlim sobre Acesso ao Conhecimento nas Ciências e Humanidades²⁷. No Manifesto Brasileiro, o tópico Recomendações à Comunidade Científica elenca um total de 20 recomendações de compromissos endereçadas às Instituições Acadêmicas, aos Pesquisadores (Autores), às Agências de Fomento, às Editoras Comerciais e às Editoras Não Comerciais.

Em decorrência da implementação do Governo Aberto, as instituições públicas vem disponibilizando e incrementando repositórios de dados na internet, a exemplo do Governo

²² www.gov.br/capes/pt-br/aceso-a-informacao/servico-de-informacao-ao-cidadao/sobre-a-lei-de-aceso-a-informacao

²³ www.movimentocienciacidadada.org/manifesto

²⁴ www.unep.org/pt-br/explore-topics/climate-change/fatos-sobre-emergencia-climatica

²⁵ www.acesoaberto.usp.br/ibict-lanca-manifesto-de-aceso-aberto-a-dados-da-pesquisa-brasileira-para-ciencia-cidada

²⁶ www.livroaberto.ibict.br/Manifesto.pdf

²⁷ www.repositorium.sdum.uminho.pt/about/DeclaracaoBerlim.htm

Federal²⁸, da Câmara Federal²⁹, da Prefeitura do Rio de Janeiro³⁰, do Instituto de Pesquisa Econômica Aplicada (IPEA)³¹, do Instituto Nacional de Meteorologia³², entre outros.

No ranking mundial de Dados Abertos da Open Knowledge Foundation de 2018, o Brasil ficou em 8º lugar (SILVA e SILVEIRA, 2019).

1.5 Dado, Informação e Conhecimento

O crescimento exponencial dos dados ao longo do século XX resultou na cunhagem do termo *Big Data* e aprofundou os estudos em Ciência de Dados e Ciência da Informação, tornando imperativo a distinção dos conceitos Dado, Informação e Conhecimento.

Dentre os conceitos para Dado, Informação e Conhecimento, o mais difundido é o que considera uma ordem sequencial ou hierárquica, operacionalizada pela decodificação humana ou processamento tecnológico: o Dado é o estado bruto e inicial, matéria-prima da Informação, que por sua vez é matéria-prima do Conhecimento. Nesse processo de transformação, os dados não estruturados

“ganham uma carga semântica, transformando-se em dados estruturados, e, posteriormente, em informação ou conjuntos de informação contextualizados em unidades maiores, de modo a formar uma rede de conhecimento” (Figura 9) (SANTOS-D’AMORIM et al., 2020).

Uma exemplificação desse processo é a percepção pelos sentidos do corpo (visão, audição) os dados nuvens carregadas e trovões, que ao serem decodificados pelo cérebro (semântica), nos dá a informação de que poderá chover. Essa informação, por sua vez, quando processada pelo cérebro (semântica), nos faz perceber que poderemos ficar molhados, permitindo a ativação de potenciais soluções, como usar guarda-chuvas ou buscar abrigo, baseadas em uma rede de conhecimento prévio (memória) que respalda a tomada de decisão (ação/intencionalidade).

A decodificação dos dados, revestindo-os de semântica, eleva a abstração desses dados, uma vez que as experiências pessoais influenciam o próprio processo de decodificação,

²⁸ www.dados.gov.br

²⁹ www.dadosabertos.camara.leg.br

³⁰ www.data.rio

³¹ www.ipeadata.gov.br

³² www.bdmep.inmet.gov.br

de forma que um mesmo dado terá diferentes interpretações para diferentes pessoas. Ao processarmos um dado inicial, seja a palavra “manga”, algumas pessoas podem imediatamente associa-la a uma fruta, outras à manga de uma camisa. Ou ainda, “Emergências Climáticas”, pode assumir diferentes significados, uma vez que as pessoas podem associa-las a diversos contextos, como elevação de temperaturas, mudança no padrão de temperaturas, elevação do nível do mar, perda de habitat e biodiversidade, mudança de hábitos etc.

O Conhecimento se distingue em (ANGELONI, 2006):

- Conhecimento Explícito: formal, concreto, sistemático, de fácil codificação, comunicação e compartilhamento em linguagem comum, tais como fórmulas matemáticas, procedimentos operacionais padrões, manuais, partituras etc.;
- Conhecimento Tácito: informal, subjetivo, intuitivo, baseado em experiências vividas, valores e emoções, possuindo duas dimensões:
 - dimensão técnica: capacidade informal, habilidade, por exemplo habilidades manuais de um artesão ou um pianista;
 - dimensão cognitiva: esquemas e modelos mentais, crenças e percepções tidas como certas, por exemplo ao fazer um bolo ou pedalando em uma bicicleta.

Ainda apresentando conceitos difusos na literatura, a Sabedoria completa o modelo DIKW (*Data-Information-Knowledge-Wisdom*). Assim como o Conhecimento, possui forte relação com a intencionalidade e incremento de valor agregado, porém distanciando-se de valores pessoais, enquanto aproxima-se de valores éticos, sendo comumente representada como o topo de uma hierarquia piramidal (Figura 10). Outros modelos incluem a Inteligência como estágio intermediário entre Conhecimento e Sabedoria (RIBEIRO e SANTOS, 2021).

Ribeiro e Santos (2021), em seu trabalho intitulado “Isso não é uma pirâmide: revisando o modelo clássico de dado, informação, conhecimento e sabedoria”, propõe o Modelo DIK2, considerando o valor agregado e o nível de complexidade de vários aspectos, como processo cognitivo, utilidade, intencionalidade, contexto, presença de crenças/valores etc (Figura 11 e **Erro! Fonte de referência não encontrada.**).

Neste modelo mais complexo, tomemos como exemplo um rabisco feito por um bebê. Este rabisco ininteligível, desprovido de qualquer significado, é um dado não percebido (t-1). Consideremos outro rabisco: “C”. Desta vez, ao parecer-se com a letra “C”, o rabisco é notado (t0). O trânsito entre t1 e t2 (P3 e P13) se dá pelo processo cognitivo (ou por processamento tecnológico), aumentando a utilidade, dando significado e interpretação dentro de um contexto, extraindo-se informação (P4 e P5).

Em diferentes contextos, “C” pode ser: letra do alfabeto; linguagem de programação C; acorde Dó; elemento químico Carbono. As informações, quando processadas, resultam em Conhecimento (P6), que como vimos, leva a um estado de alta intencionalidade (aprendo: letras; programação; teclado; química).

As ações embasadas em conhecimento elevam a capacidade de julgamento, porém com forte motivação subjetiva, pessoal e de crenças (quero ser: tradutor; programador; tecladista; químico). A Inteligência é o uso do Conhecimento em ações tidas como “certas”, sem relação com os valores pessoais (usar o termo mais apropriado em uma tradução; tornar um código mais eficiente; empregar métricas musicais; corrigir o pH de uma substância).

Por fim, a sabedoria é o uso do conhecimento “de forma eficaz e eficiente”, se distinguindo da inteligência por estar mais “intimamente ligada a valores, especialmente a ética, [...] abrangendo preocupações pertinentes aos interesses” coletivos (RIBEIRO E SANTOS, 2021).

Os autores salientam que “a sabedoria não retrocede, nem o dado que já foi notado (t0) volta a ser não percebido (t-1)”, ou seja, o rabisco “C” nunca mais será desprovido de significado.

1.6 Ambiente de Tecnologia da Informação

Como exposto, obter conhecimento leva a um estado de alta intencionalidade/ação. O dado, a informação, o conhecimento, a comunicação e o processo decisório se inter-relacionam de forma a criar uma interdependência entre essas etapas (**Erro! Fonte de referência não encontrada.**). A partir do processamento do dado e obtenção da informação, uma teia de eventos é disparada, ora amparando uma decisão, ora comunicando uma decisão. Essa inter-relação evidenciou às organizações que deter conhecimento não é sinônimo de poder. O conhecimento ganha valor a partir da comunicação e decodificação por outras

peças. É importante compreender que a comunicação (através de um *paper*, por exemplo) passa por um processo de codificação e, portanto, comunica um dado não estruturado, que será decodificado por outras pessoas (**Erro! Fonte de referência não encontrada.**) (ANGELONI, 2006).

1.7 *Big Data*

O termo *Big Data* ainda hoje é abstrato, tendo como consenso, que são conjuntos de dados que os tradicionais softwares de processamento são incapazes ou extremamente ineficientes em processá-los. Essa dificuldade conceitual levou à concepção de características inerentes a esses dados e inicialmente foram propostos 3 Vs (Volume, Velocidade e Variedade). No decorrer dos estudos, outros autores acrescentaram mais Vs (Veracidade, Variabilidade, Valor) (**Erro! Fonte de referência não encontrada.**), conforme Rautenberg e Carmo (2019):

- Volume: grande volume de dados gerados em redes sociais, sensores digitais, organizações governamentais etc.;
- Velocidade: a geração dos dados é em alta velocidade, potencializada pelo incremento tecnológico de produzir, capturar e processar dados;
- Variedade: a origem dos dados é de diversas fontes (sensores, instituições, redes sociais etc);
- Veracidade: integridade e precisão dos dados, evitando ruídos e incertezas;
- Variabilidade: compreensão dos dados a nível de detectar padrões de comportamento insustentáveis no tempo, devido a sazonalidade ou eventos específicos;
- Valor: resultado da análise de dados precisos, fornece informação e insights para Tomada de Decisão.

Outra sugestão inclui Vocabulário, relacionando os dados a um contexto (RIBEIRO, LIVEIRA e ARAÚJO, 2019).

Dentre as fontes de dados para *Big Data*, se destacam sensores digitais/remotos, experimentos científicos (astronomia, nuclear etc), dados de saúde, economia, transporte etc. O volume e a variedade exigem a organização e estruturação desses dados, que darão suporte à tomada de decisão (previsão, monitoramento etc.) (RIBEIRO, OLIVEIRA e ARAÚJO, 2019).

1.8 Engenharia de Dados e Ciência de Dados

Engenharia de Dados e Ciência de Dados são duas áreas relacionadas aos dados que podem causar confusão, pois ambas lidam com a coleta, processamento e análise de grandes conjuntos de dados, mas têm abordagens e objetivos diferentes (**Erro! Fonte de referência não encontrada.**).

Engenharia de Dados é a área dedicada a construir e gerenciar a infraestrutura e arquitetura para coletar, armazenar, processar e disponibilizar grandes volumes de dados (*Big Data*). Utilizando Ciência da Computação, Banco de Dados, Processamento Distribuído e linguagens como Java, Python e ferramentas como Hadoop e MapReduce, desenvolve-se sistemas capazes de tratar uma quantidade massiva de dados em tempo real ou para posterior análise. A infraestrutura envolve *Data Lakes*, *Data Warehouses*, *pipelines*, entre outros, garantindo modularidade, escalabilidade, performance e confiabilidade no tratamento dos dados (DATA SCIENCE ACADEMY, 2022).

Ciência de Dados consiste em um “suporte metodológico ao Processo de Tomada de Decisão” (também denominado Tomada de Decisão Guiada por Dados) (**Erro! Fonte de referência não encontrada. e Erro! Fonte de referência não encontrada.**), dedicada:

“a obtenção de informação contextualizada; a explicitação de fenômenos subliminares contidos nos dados; ou a refutação/confirmação de hipóteses previamente estabelecidas” (RAUTENBERG e CARMO, 2019).

Quanto à interdisciplinaridade, cabe mencionar a proximidade, em suas origens, da Ciência de Dados com a Bioinformática (**Erro! Fonte de referência não encontrada.** e Figura 18), que caracteriza-se por ter a Biologia (mais especificamente as Ciências Ômicas) como Conhecimento de Domínio. Não à toa, um dos tipos de algoritmos de Aprendizado de Máquina é denominado Algoritmo Genético, proposto por Holland, inspirado no princípio darwiniano de seleção natural e reprodução genética, “para encontrar solução de problemas de otimização”, hoje empregado em inúmeras áreas do conhecimento³³ (MATOS *et al.*, 2023; FARIA, 2020).

³³Sequenciamento de Ações Pedagógicas por Algoritmo Genético Utilizando Taxonomia de Bloom e ASSIST (COSTA e FERNANDES, 2020). Algoritmos Genéticos: uma abordagem visual para reduzir as perdas por transporte em canteiros de obra (PÉREZ, COSTA E IRIZARRY, 2020). Um Algoritmo Genético para a Reconstrução de Fontes Concentradas via Método das Soluções Fundamentais (FARIA, 2020).

1.9 Pipeline de dados

Pipeline de dados é uma sequência de etapas ou processos que são executados em conjunto para transformar e processar dados, desde a obtenção, análise e entrega de *insights* úteis (Figura 21). É uma abordagem sistemática para lidar com o fluxo de dados, permitindo organização, limpeza, integração, transformação e análise de forma eficiente (QUEMY, 2019).

Em sistemas complexos, envolvendo *Big Data*, após a coleta dos dados (*Data Collection*), o *pipeline* envolve processos de criação de novos banco de dados (*Data Stagin*), a partir dos dados ingeridos e processados. Em seguida, com a criação de *Data Staging* utilizando outras fontes de dados, esses *Data Staging* são integrados em um novo banco de dados (*Data Warehouse*), para serem processados por outras metodologias (**Erro! Fonte de referência não encontrada.**) (MUNAPPY, BOSCH, HOLMSTRÖM OLSSON *et al.*, 2020).

Romero, Wrembel e Song (2020) elencam 8 etapas relacionadas ao *pipeline* de dados:

1. integração dos dados heterogêneos e distribuídos;
2. limpeza e padronização;
3. eliminação de duplicatas;
4. centralização dos dados tratados;

Relacionadas à análise:

5. centralização (em outro banco de dados) de visualizações dos dados (tabelas, gráficos etc);
6. pré-processamento específico para determinadas análises (alguns algoritmos exigem normalização dos dados ou transformação de variável categórica em numérica, por exemplo);
7. criação de conjuntos de dados para teste e validação, incluindo rotulagem de dados para aprendizado de máquina;

8. análise estatística descritiva (relatórios, OLAP – Online Analytical Processing³⁴ – etc) e modelos preditivos (mineração de dados, aprendizado de máquina etc).

Segundo os autores, até 80% do tempo é gasto nas etapas de 1 a 3 e na 7, devido à falta de padronização (estrutura universal) dos dados, levando à construção de pipelines *ad hoc* (para uma finalidade específica).

1.10 Knime

Knime Analytics Platform³⁵ é um software de código aberto, para *Big Data*, de interface intuitiva, abordagem *low-code*³⁶ e conceito *pipeline* modular (**Erro! Fonte de referência não encontrada.**), cujo lema é “Open for Innovation”. Knime (acrônimo de Konstanz Information Miner) tem origem em 2004 na Universidade de Konstanz, no sul da Alemanha. Possui ambientes colaborativos para os usuários:

- Community Forum: em formato de fórum, permite compartilhamento de informações, onde ocorrem as práticas colaborativas entre os usuários (**Erro! Fonte de referência não encontrada.**);
- Community Hub: espaço para os usuários disponibilizarem *workflows* prontos para reuso (Figura 13).

³⁴Processamento Analítico Online (OLAP), permite a visualização dos dados em múltiplas perspectivas (de forma dinâmica), sem alteração na estrutura dos dados (navegação entre os dados, seleção de conjunto de dados etc). Ver Han (1998), Youtube/IBM Technology (2022) em www.youtube.com/watch?v=iw-5kFzIdgY.

³⁵www.knime.com

³⁶*Low-code* consiste no paradigma de se reduzir tempo e custo na implementação de softwares, exigindo menor treinamento de pessoal. Trata-se de viabilizar ferramentas para programadores e não programadores (ALVES e ALCALÁ, 2022).

1.11 Rio de Janeiro

Segundo a Lei Orgânica da cidade do Rio de Janeiro:

“o território do Município está situado a vinte e dois graus, quarenta e cinco minutos e cinco segundos S de latitude extremo Norte; vinte e três graus, quatro minutos e dez segundos S de latitude extremo Sul; quarenta e três graus, seis minutos e trinta segundos Wgr. de longitude extremo Leste e quarenta e três graus, quarenta e sete minutos e quarenta segundos Wgr. de longitude extremo Oeste; na direção Norte-Sul tem distância angular de zero grau, dezenove minutos e cinco segundos e distância linear de trinta e cinco quilômetros; na direção Leste-Oeste, distância angular de zero grau, quarenta e um minutos e dez segundos e distância linear de setenta quilômetros”.

Com 1.204 km² de área e mais de 160 bairros, uma das formas de divisão territorial, é a que divide o território em 16 Regiões de Planejamento (RP), cada uma dentro de uma das 5 Áreas de Planejamento (AP) (**Erro! Fonte de referência não encontrada.**).

O Censo 2010 calculou a população do município em 5.857.904 em 2010, distribuída entre as Áreas de Planejamento (Tabela 1). Para 2021, o IBGE estima que a população seja de 6.775.561 habitantes, um crescimento de 15,67% com relação ao último Censo realizado. Uma parcela significativa da cidade reside na AP 3 (40,18%) e na AP 5 (26,57%). O Instituto Pereira Passos (IPP), vinculado à Prefeitura do Rio, estima a população de 2020 em 6.516.508 habitantes, distribuída entre as Áreas de Planejamento (Tabela 2).

Quanto à área das AP, quase metade do território municipal (47,5 %) está contido na AP 5. A AP 4 é a segunda maior, representando 24,4% da área (Tabela 3).

Tabela 1: População das AP em 2010

Área de Planejamento	População (2010)	%
1	268280	0,046
2	997478	0,17
3	2353590	0,402
4	682051	0,116
5	1556505	0,266

Fonte: Autor, com dados do data.rio.

Tabela 2: População estimada das AP em 2020

AP	População estimada (2020)	%
1	316547	4,9
2	1017788	15,6
3	2290190	35,1
4	1077930	16,5
5	1814053	27,8

Fonte: Autor, com dados do data.rio.

Tabela 3: Área das AP

AP	m²	%
1	34393641	2,9
2	100432028	8,3
3	203498383	16,9
4	293822112	24,4
5	572197999	47,5

Fonte: Autor, com dados do data.rio.

1.12 Sistema Alerta Rio

O Sistema Alerta Rio³⁷ é um programa gerido pela Prefeitura do município do Rio a partir da Fundação GEO-RIO (Fundação Instituto de Geotécnica do Município do Rio de Janeiro). Possui atualmente um conjunto de 33 estações de monitoramento dispersas pelo território municipal (Tabela 4) que realizam medições a cada 15 minutos, sendo 25 apenas pluviométricas. As estações entraram em operação em diferentes datas, marcadas por investimentos em novas estações e sensores ao longo dos anos. Os dados para *download* são atualizados diariamente.

Os parâmetros meteorológicos medidos são: velocidade do vento, direção do vento, temperatura, pressão e umidade, porém nem todas as estações estão equipadas com todos os sensores. A estação Vidigal, por exemplo, que era somente pluviométrica, em 2020 passou a monitorar velocidade e direção do vento.

Tabela 4: Estações (ativas) Sistema Alerta Rio por AP

AP	Quantidade de Estações	%
1	3	9,1
2	10	30,3
3	7	21,2
4	7	21,2
5	6	18,2

Fonte: Autor, com dados do data.rio.

³⁷<http://www.sistema-alerta-rio.com.br>

As estações ativas são: Vidigal, Urca, Rocinha, Tijuca, Santa Teresa, Copacabana, Grajaú, Ilha do Governador, Penha, Madureira, Irajá, Bangu, Piedade, Jacarepaguá/Tanque, Saúde, Jardim Botânico, Barra/Barrinha, Jacarepaguá/Cidade de Deus, Barra/Riocentro, Guaratiba, Est. Grajaú/Jacarepaguá, Santa Cruz, Grande Méier, Anchieta, Grotta Funda, Campo Grande, Sepetiba, Alto da Boa Vista, Av. Brasil/Mendanha, Recreio dos Bandeirantes, Laranjeiras, São Cristóvão e Tijuca/Muda (**Erro! Fonte de referência não encontrada.**).

Cada estação abrange Zona Pluviométrica (ZP), com delimitações próprias. Considerando a localização da respectiva estação como sede, as Zonas Pluviométricas foram associadas a uma AP (Tabela 5 e **Erro! Fonte de referência não encontrada.**). A diferença no somatório das áreas se deve à não contabilização da área de ilhas desabitadas nas Zonas Pluviométricas.

Tabela 5: Áreas das Zonas Pluviométricas por AP

AP	Área Total das ZP (m²)	Área da Menor ZP (m²)	Área Média das ZP (m²)	Área da Maior ZP (m²)
1	43.919.025	7.901.084	14.639.675	25.248.796
2	113.672.064	2.553.417	11.367.206	21.390.442
3	209.892.404	9.979.112	29.984.629	47.354.632
4	300.080.385	33.338.297	42.868.626	67.969.063
5	546.291.582	70.225.597	91.048.597	127.922.516

Fonte: Autor, com dados do data.rio, área calculada com QGIS.

2 OBJETIVOS

Objetivo geral: Construir um *pipeline* de dados para os dados do Sistema Alerta Rio utilizando Knime.

Objetivos específicos:

- Coletar e integrar os dados pluviométricos e meteorológicos do Sistema Alerta Rio;
- Desenvolver um fluxo de trabalho (*pipeline*) no *software* Knime para pré-processamento, limpeza e integração desses dados;
- Gerar visualização dos dados por meio de gráficos em diferentes recortes de tempo (dia, mês, ano etc);
- Agregar valor aos dados brutos, classificando as leituras a partir de conceitos consagrados como a Escala Beaufort (intensidade do vento), direção do vento em escalas ordinal, cardinal e subcolateral, classificação de nível de umidade segundo a OMS etc.

3 MATERIAIS E MÉTODOS

Foram utilizados Knime (4.7), incluindo algumas extensões (KNIME Interactive R Statistics Integration³⁸ e outras), CRAN R Project³⁹ (4.2.0), um computador com Sistema Operacional Linux Ubuntu 20.04 LTS⁴⁰. Os parâmetros do Knime foram personalizados no arquivo .ini aumentando a quantidade de memória *heap* (130 Gb). Para suprir a memória *heap*, foi criado um total de 140 Gb de memória *swap* (equivalente à memória de paginação do Windows).

O programa QGIS (3.24.3) foi utilizado para gerar mapas.

Os dados do Sistema Alerta Rio são disponibilizados em 2 conjuntos de dados distintos, e foram baixados no período de 1997 a 2022:

- Pluviométricos (113,2 Mb zipado; 1,8 Gb descompactado);
- Meteorológicos (23,5 Mb zipado; 269,1 Mb descompactado).

Foram construídos dois *pipeline* de dados, um para os conjuntos de dados Pluviométricos e outro para Meteorológicos.

Na etapa de agregação de valor dos dados, foram agregados: estação do ano (primavera, verão, outono, inverno), quadrimestre do ano, semana do ano, dia do ano, período do dia (em 4 fases – madrugada, manhã, tarde, noite; e em 2 fases – diurno, noturno), classificação de velocidade do vento (Escala Beaufort⁴¹ e classificação própria do Sistema Alerta Rio), direção do vento (ordinal, cardinal, subcolateral), ocorrência de noites quentes (temperaturas de 25° C, 30° C, 35° C e 40° C), classificação de umidade segundo OMS (ótimo, inadequado, atenção, alerta, emergência), entre outras.

Após processamento dos dados, as leituras com valor agregado foram consolidadas em um único arquivo estruturado csv. Em redundância, cada estação foi consolidada em um arquivo csv com suas respectivas leituras.

³⁸ www.hub.knime.com/knime/extensions/org.knime.features.r/latest

³⁹ www.cran.r-project.org

⁴⁰ www.releases.ubuntu.com/focal

⁴¹ Classificação de velocidade do vento quanto ao potencial impacto de destruição, variando de 0 (calmo) a 12 (furação ou ciclone) (FONSECA e FERENTZ, 2020).

4 RESULTADOS

Como resultados, serão apresentados:

- Capítulo de livro publicado e apresentado no V Congresso Internacional de Saúde e Meio Ambiente (CINASAMA) (Germano e Figueiró. Caracterização Climatológica no Município do Rio de Janeiro/Rj Utilizando dados do Sistema Alerta Rio. In: ONE, G. M. da C. (org.). **Saúde e Meio ambiente: Pesquisa, tecnologia e aplicabilidade**. João Pessoa – PE, IMEA, 2023. p. 261-284. ISBN: 978-65-89069-23-2 <https://cinasama.com.br/wp-content/uploads/2023/05/Livro-Saude-e-meio-ambiente-2023.pdf>). (ANEXO A)
- Repositório de dados (Germano, Vinicius e Figueiró (2023), “Sistema Alerta Rio (Data Staging Meteorologia)”, **Mendeley Data**, V1, <https://doi.org/10.17632/2bdnntz5nn.1>).

5 DISCUSSÃO

A discussão será apresentada a seguir, inicialmente debruçando-se no uso do Knime Analytic Platform como ferramenta apropriada à análise dos dados do Sistema Alerta Rio, à luz das técnicas de *Big Data*. Em um segundo momento, serão discutidos aspectos do Sistema Alerta Rio quanto aos Dados Abertos, *Big Data* e Governo Aberto.

5.2 Knime Analytic Platform

O uso do Knime neste trabalho aumentou o valor agregado dos dados do Sistema Alerta Rio, processando mais de 12.500 arquivos (totalizando 2,1 Gb).

Foi realizada a consolidação em diferentes perspectivas (hora, dia, mês, ano) dos dados (estação do ano, turno do dia, quadrimestre, Escala Beaufort, intensidade pluviométrica, noites com temperatura acima de 25° C etc).

Conforme proposto por Munappy, Bosch, Holmström Olsson *et al.* (2020) (Figura 22), os dados consolidados foram alocados em um novo banco de dados (*Data Staging*), composto por 1 único arquivo csv contendo as leituras de todas as estações e, de forma redundante, individualização das leituras de cada estação em arquivos separados, facilitando a busca dos dados de uma estação específica.

O modelo DIK2 (Figura 11) se aplica, conceitualmente, ao *workflow* Knime. Cada arquivo é carregado no *workflow* (dado percebido, t0) e após sucessivos processamentos (t1 e t2), foram obtidas diversas informações (estação do ano, turno do dia, média horária, média anual etc). No workflow Knime, a complexidade de diversos nós de processamento pode ser encapsulada em um *unicometanode*, uma representação menos complexa (abstração de vários processos de transformação dado → informação (Figura 9)).

A abordagem *low-code* de Knime além de facilitar o uso de métodos de programação por não programadores, reduziu o tempo de desenvolvimento, uma vez que substitui a digitação de várias linhas de código por configuração de parâmetros em uma interface gráfica.

5.3 Sistema Alerta Rio

Os dados do Sistema Alerta Rio são disponibilizados de forma semi estruturada, implicando em maior processamento dos dados para estrutura-los (1 estrela de Maturidade em Dados Abertos, Figura 26). Essas etapas de processamento são amplificadas pela quantidade de arquivos a serem processados. A título de comparação, as leituras das 8 estações do Programa de Monitoramento da Qualidade do Ar (MonitorAr Rio⁴²), com mais de 700 mil registros, são disponibilizadas em 1 único arquivo csv (4 estrelas, Figura 24), de forma estruturada o que, com relação ao Sistema Alerta Rio, dispensa várias etapas de processamento (sobretudo a etapa 2. *limpeza e padronização* citada por Romero, Wrembel e Song (2020)), facilitando seu uso de forma imediata.

Considerando as características do *Big Data* apresentadas por Rautenberg e Carmo (2019), os dados do Sistema Alerta Rio se relacionam da seguinte forma: *Volume* (elevada quantidade de arquivos e volume total dos dados brutos); *Velocidade* (os dados são atualizados diariamente); *Veracidade* (dados oficiais; realizado tratamento para eliminação de ruídos); *Variabilidade* (dados provenientes de diversas estações, com 2 conjuntos distintos – meteorologia e pluviometria); *Variedade* (além das leituras em horário de verão, destaca-se a velocidade do vento com 2 unidades de medidas em períodos diferentes, exigindo tratamento de padronização e uniformidade); *Valor* (estruturação, consolidação, agregação de valor e análise exploratória). A continuação do projeto tem potencial de agregar mais valor aos dados, com a integração a outras fontes (outros conjuntos de dados pluviométricos/meteorológicos; dados de saúde humana, qualidade ambiental etc.), elevando a relação com outras características do *Big Data* (*Variedade* e *Valor*).

As políticas de Governo Aberto não se resumem à simples publicização dos dados, mas requer *Transparência*, ou seja, os dados devem ser apresentados de forma acessível. Nesse aspecto, o site do Sistema Alerta Rio⁴³ se destaca em apresentar os dados em tempo real (últimas leituras realizadas), a previsão do tempo, a sensação térmica (do dia corrente), temperatura máximas/mínimas (do dia corrente), chuva (do dia anterior), climatologia mensal (relatório técnico, apenas o mais recente), temperaturas máximas/mínimas (mensais). Há

⁴²www.data.rio/datasets/PCRJ::dados-hor%C3%A1rios-do-monitoramento-da-qualidade-do-ar-monitorar

⁴³www.sistema-alerta-rio.com.br consultado em 16/07/2023.

serviços com *link* quebrado, descontinuados ou defasados: maiores chuvas, estatísticas de escorregamento, relatórios anuais de chuva, relatórios anuais de escorregamento.

Em resumo, a *Transparência* dos dados conferida pelo Sistema Alerta Rio tem um viés imediatista, predominando a exibição de dados recentes, últimas leituras e últimos relatórios, restringindo o acesso a estudos anteriores. Em comparação com a *Transparência* promovida pelo curso de Ciências Atmosféricas (UNIFEI/MG) de sua única estação meteorológica, o site do Sistema Alerta Rio está muito aquém para uma cidade da relevância do Rio de Janeiro.

5.3 Mendeley Data

Além de garantir Maturidade 4 estrelas aos dados consolidados, Mendeley Data permite a organização e citação apropriada dos dados compartilhados. Os dados recebem DOI (Identificadores de Objetos Digitais) permanentes, o que facilita o rastreamento de sua utilização posterior por outros trabalhos acadêmicos, inclusive por outros pesquisadores. Contando com controle de versionamento, permite a atualização do repositório, com incremento e/ou atualização dos dados.

Além do benefício direto de tornar mais dados disponíveis publicamente, o Mendeley Data promove a reprodutibilidade e o rigor científico. Ao divulgar os dados brutos (e no caso deste trabalho, a metodologia empregada no processamento dos dados) por trás de descobertas de pesquisa, outros cientistas podem verificar os resultados e as conclusões, identificando erros ou vieses não intencionais.

CONCLUSÃO

Os paradigmas da Ciência Aberta, baseados no *commons*, possibilitados pelas TDIC, vão ao encontro de uma sociedade cada vez mais tecnológica, globalizada e colaborativa.

Na área de Biociências, após a pandemia de COVID-19, foi observado o alto crescimento da publicação de *preprints*, motivo pelo qual as pesquisas em andamento para desenvolvimento de vacinas em diversos países se beneficiaram com o compartilhamento dos trabalhos em andamento. O *commons* demanda a cultura do compartilhamento e colaboração, desde o delineamento da Pesquisa até a publicação.

Os Dados Abertos Governamentais (e não governamentais) implementados por diversas esferas de governo e instituições (privadas), fruto da política de Governo Aberto, configuram-se como inúmeras fontes de dados de relevante valor científico. Esses Dados Abertos, entretanto, necessitam de processamento para que as informações sejam extraídas, e devido às suas qualidades relacionadas com *Big Data*, ferramentas específicas para Ciência de Dados e *Big Data* são adequadas.

O Sistema Alerta Rio é uma importante ferramenta de monitoramento e previsão das condições climáticas na cidade do Rio de Janeiro. Seus dados brutos disponibilizados para *download* são semi-estruturados (motivo pelo qual não atendem aos requisitos de Dados Abertos estabelecidos pela LAI), demandando mais etapas de processamento para adequação/estruturação dos dados.

Knime se mostrou uma ferramenta de Ciência de Dados fortemente alinhada aos preceitos da Ciência Aberta. Compara-se a outras ferramentas utilizadas na etapa de Análise da Pesquisa, possibilitando o uso de inúmeras metodologias de processamento e análise dos dados, destacando-se pelas abordagens *low-code* (facilitando o uso por pessoas com pouco ou nenhum conhecimento de programação) e *pipeline* modular (configurando-se como conhecimento explícito, representa a metodologia empregada, assegura um *Fluxo de Trabalho* documentado e compartilhável, garantindo sua *Reprodutibilidade e Integridade*). Por ter a cultura *Open* como lema, Knime insere-se em inúmeras facetas da Taxonomia para Ciência Aberta.

A publicação dos dados processados em repositório (Mendeley Data) permite que os dados sejam Abertos, de Maturidade 4 estrelas (Figura 26), podendo serem acessados *online* por meio de link único, com DOI, servindo como fonte para novos estudos.

Ao promover a acessibilidade, reprodutibilidade, rigor e reconhecimento no compartilhamento de dados de pesquisa, Mendeley Data e Knime se colocam como uma solução TDIC para acelerar o progresso científico colaborativo e aberto.

REFERÊNCIAS

- ALVES, FÁBIO RAMOS; ALCALÁ, SYMONE GOMES SOARES. Análise da abordagem LOW-CODE como facilitador da transformação digital em indústrias. **Revista e-TECH: Tecnologias para Competitividade Industrial**. Florianópolis, v. 15 n. 2 (2022). <https://doi.org/10.18624/etech.v15i1.1186>
- ANGELONI, MARIA TEREZINHA. **Gestão do conhecimento em TI**: livro didático. Cipriani Pandini. – Palhoça : UnisulVirtual, 2006. ISBN 85-60694-37-4. ISBN 978-85-60694-37-2.
- AUTRAN, MARYNICE M. M. Gênese, Evolução e Tendências da Comunicação da Ciência. IN: BORGES, MARIA MANUEL; CASADO, ELIAS SANZ (org.). Sob a lente da ciência aberta: olhares de Portugal, Espanha e Brasil. **Conferências e debates interdisciplinares**. Imprensa da Universidade de Coimbra, 2020. <https://doi.org/10.14195/978-989-26-2022-0>
- BERTIN, PATRÍCIA ROCHA BELLO; FORTALEZA, JULIANA MEIRELES; SILVA, ADRIANA CRISTINA DA; OKAWACHI, Massayuki FRANCO. A parceria para Governo Aberto como plataforma para o avanço da Ciência Aberta no Brasil. **TransInformação**, Campinas, 31: e190020, 2019. <http://dx.doi.org/10.1590/2318-0889201931e190020>
- CABALLERO-RIVERO, ALEJANDRO; SÁNCHEZ-TARRAGÓ, NACY; SANTOS, RAIMUNDO NONATO MACEDO DOS. Práticas de Ciência Aberta da comunidade acadêmica brasileira: estudo a partir da produção científica. **TransInformação**, Campinas, 31: e190029, 2019. <http://dx.doi.org/10.1590/2318-0889201931e190029>
- CLINIO, ANNE. Ciência Aberta na América Latina: duas perspectivas em disputa. **TransInformação**, Campinas, 31: e190028, 2019. <http://dx.doi.org/10.1590/23180889201931e190028>
- COSTA, Newarney Torrezão da; FERNANDES, Marcia Aparecida. Sequenciamento de Ações Pedagógicas por Algoritmo Genético Utilizando Taxonomia de Bloom e ASSIST. In: **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 1273-1282. <https://doi.org/10.5753/cbie.sbie.2020.1273>.

COSTA, SANDRA REGINA SANTANA; DUQUEVIZ, BARBARA CRISTINA; PEDROZA, REGINA LÚCIA SUCUPIRA. Tecnologias Digitais como instrumentos mediadores da aprendizagem dos nativos digitais. **Revista Quadrimestral da Associação Brasileira de Psicologia Escolar e Educacional**, SP. Volume 19, Número 3, Setembro/Dezembro de 2015: 603-610. DOI: <http://dx.doi.org/10.1590/2175-3539/2015/0193912>

DATA SCIENCE ACADEMY. A Diferença Entre Cientistas de Dados, Engenheiros de Dados, Estatísticos e Engenheiros de Software. Postado em 30 jun. 2022. Disponível em <https://blog.dsacademy.com.br/a-diferenca-entre-cientistas-de-dados-engenheiros-de-dados-estatisticos-e-engenheiros-de-software/>. Acessado em: 23 ago. 2023.

Erskine N, Hendricks S. **The Use of Twitter by Medical Journals: Systematic Review of the Literature**. **J Med Internet Res** 2021;23(7):e26378. URL: <https://www.jmir.org/2021/7/e26378> DOI: 10.2196/26378

FANELLI, DANIELE. Is science really facing a reproducibility crisis, and do we need it to? **PNAS (Proceedings of the National Academy of Sciences)**, March 13, 2018 | vol. 115 | no. 11. <https://doi.org/10.1073/pnas.1708272114>

FARIA, JAIRO ROCHA de. Um Algoritmo Genético para a Reconstrução de Fontes Concentradas via Método das Soluções Fundamentais. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 7, n. 1, 2020. <https://doi.org/10.5540/03.2020.007.01.0447>

FONSECA, M. N. da, & FERENTZ, L. M. da S. (2020). Ocorrência de Vendavais no Estado do Paraná: um panorama entre 2013 e 2017. **Revista Nacional De Gerenciamento De Cidades**, 8(57). <https://doi.org/10.17271/2318847285720202281>

FONSECA, William D'A. Introdução ao LaTeX e como iniciar um novo projeto no Overleaf. **Acústica e Vibrações: Seção Chamadas**, v. 35, n. 52, p.133–140, 2020. ISSN 2764-3611, 1983-442X. <https://doi.org/10.55753/aev.v35e52.40>

GERMANO, V. T. M.; PEREIRA, R. F. P. Caracterização Climatológica no Município do Rio de Janeiro/Rj Utilizando dados do Sistema Alerta Rio. In: ONE, G. M. da C. (org.). **Saúde e Meio ambiente: Pesquisa, tecnologia e aplicabilidade**. João Pessoa – PE, IMEA, 2023. p. 261-284. ISBN: 978-65-89069-23-2 <https://cinasama.com.br/wp-content/uploads/2023/05/Livro-Saude-e-meio-ambiente-2023.pdf>.

GERMANO, VINICIUS (2023), “Pipeline de dados para Sistema Alerta Rio em Knime Analytics Platform”, **Mendeley Data**, V1, <https://doi.org/10.17632/2bdnntz5nn.1>

Han, J. (1998). OLAP Mining: An Integration of OLAP with Data Mining. In: Spaccapietra, S., Maryanski, F. (eds). **Data Mining and Reverse Engineering: Searching for Semantics. IFIP — The International Federation for Information Processing. Springer**, Boston, MA. https://doi.org/10.1007/978-0-387-35300-5_1

KARAMI, A; LUNDY, M; WEBB, F; DWIVEDI, MY K. Twitter and Research: Systematic Literature Review Through Text Mining. **IEEE Access**, vol. 8, pp. 67698-67717, 2020. <https://doi.org/10.1109/ACCESS.2020.2983656>

LUC, J G Y; ARCHER, M A; ARORA, R C; BENDER, E M; BLITZ, A; COOKE, D T; HLICI, T N; KIDANE, B; OUZOUNIAN, M; VARGHESE JR, T K; ANTONOFF, M B. Does Tweeting Improve Citations? One-Year Results From the TSSMN Prospective Randomized Trial. **The Annals of Thoracic Surgery**. Volume 111, Issue 1, January 2021, Pages 296-300. <https://doi.org/10.1016/j.athoracsur.2020.04.065>

MATOS, D F; ALVES, M C S; RUZENE, D S; BATISTA, M V de A; SILVA, D P. Bioinformática aplicada às Ciências Ômicas: uma breve revisão bibliográfica. <https://dx.doi.org/10.37885/230312569>. In: . **Tecnologia e Inovação na Agricultura: aplicação, produtividade e sustentabilidade em pesquisa** - ISBN 978-65-5360-317-2- Vol. 1 - Ano 2023 – Editora Científica Digital.

MUNAPPY, A., BOSCH, J., HOLMSTRÖM OLSSON, H. *et. al.* (2020). Modelling Data Pipelines. Proceedings - **46th Euromicro Conference on Software Engineering and Advanced Applications**, SEAA 2020: 13-20. <http://dx.doi.org/10.1109/SEAA51224.2020.00014>

OLIVEIRA, ADRIANA CARLA SILVA DE; GUIMARÃES, PATRÍCIA BORBA VILAR; KOSHIYAMA, DÉBORA COSTA ARAÚJO DI GIACOMO. A ciência aberta e os Direitos de Propriedade Intelectual: um olhar a partir da economia criativa e da ciência do commons. **Revista de Direito da Cidade**, vol. 11, nº 1. ISSN 2317-7721 pp. 663-681. <https://doi.org/10.12957/rdc.2019.32031>

OLIVEIRA, THAIANE; SOBREIRA, RAFAEL. Transformações, disputas e circuitos de inovação nas publicações científicas frente à Ciência Aberta. In: **Ciência Aberta: visão e contribuição a partir dos periódicos científicos**. MIRANDA, ANGÉLICA C. D.;

DAMASIO, EDILSON; FIRME, SIMONE MACHADO (org.). Rio Grande, RS: Ed. da FURG, 2020. 176 p. ISBN: 978-65-5754-006-0 (eletrônico):

<http://repositorio.furg.br/handle/1/8792>

PÉREZ, Cristina T.; COSTA, DayanaB.; IRIZARRY, Javier. Algoritmos genéticos: uma abordagem visual para reduzir as perdas por transporte em canteiros de obra. In: **Anais Do XVIII Encontro Nacional de Tecnologia do Ambiente Construído**, 18., 2020, Porto Alegre. Anais... Porto Alegre: ANTAC, 2020. <https://doi.org/10.46421/entac.v18i.1171>

QUEMY, ALEXANDRE. Data Pipeline Selection and Optimization. **Workshop Proceedings of the EDBT/ICDT 2019 Joint Conference**. March 26, 2019, Lisbon, Portugal. <https://ceur-ws.org/Vol-2324/Paper19-AQuemy.pdf>

QUINTANILHA, TIAGO LIMA. Os quatro grandes desafios ao modelo de Ciência Aberta:(des)acreditação, informalidade, comodificação e predação. **Texto Livre: Linguagem e Tecnologia**. Belo Horizonte, v. 12, n. 2, p. 202-213, mai.-ago. 2019 – ISSN 1983-3652. <https://doi.org/10.17851/1983-3652.12.2.202-213>

RAUTENBERG, SANDRO; CARMO, PAULO RICARDO VIVIURKA DO. Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Studies: Research Trends**. 13:1 (2019) p.56-p.67. ISSN 1981-1640. <https://doi.org/10.36311/1981-1640.2019.v13n1.06.p56>

RIBEIRO, ANNA CAROLINA MENDONÇA LEMOS; SANTOS, CARLOS DENNER DOS. Isso não é uma pirâmide: revisando o modelo clássico de dado, informação, conhecimento e sabedoria. **Ciência Da Informação**, 49(2),p. 67-87 . Brasília, DF,maio/ago. 2020. <https://doi.org/10.18225/ci.inf.v49i2.5066> <https://revista.ibict.br/ciinf/article/view/5066>

RIBEIRO, N. C., OLIVEIRA, D. A., & ARAÚJO, R. F. de. (2019). Conjecturas da Ciência Aberta na contemporaneidade do Big data. **BIBLOS: Revista do Instituto de Ciências Humanas e da Informação**, 33(2), 163–179. <https://doi.org/10.14295/biblos.v33i2.9648>

RODRIGUES, ELOY. A pandemia e a emergência da Ciência Aberta. 2020. <https://doi.org/10.21814/uminho.ed.24.12>. In: RODRIGUES, Eloy; MARTINS, Manuela. **A Universidade do Minho em tempos de pandemia: Tomo II: (Re)Ações**. UMinho Editora. Braga, 2020. ISBN digital 978-989-8974-28-0 DOI: <https://doi.org/10.21814/uminho.ed.22> DOI Tomo II: <https://doi.org/10.21814/uminho.ed.24>

ROMERO, OSCAR; WREMBEL, ROBERT; SONG, IL-YEOL. An Alternative View on Data Processing Pipelines from the DOLAP 2019 Perspective. **Information Systems**. 92 (2020) 101489. <https://doi.org/10.1016/j.is.2019.101489>

SANTOS-D'AMORIM, KAREN; SILVA, MARCELA LINO DA; CRUZ, RÚBIA WANESSA DOS REIS; CORREIA, ANNA ELIZABETH GALVÃO COUTINHO. Dos dados ao conhecimento: tendências da produção científica sobre Big Data na Ciência da Informação no Brasil. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 25, p. 01-23, 2020. Universidade Federal de Santa Catarina. ISSN 1518-2924. <https://doi.org/10.5007/1518-2924.2020.e70527>

SILVA, FABIANO COUTO CORRÊA DA; SILVEIRA, LÚCIA DA. O ecossistema da Ciência Aberta. **TransInformação**, v.31, e190001, 2019. <http://dx.doi.org/10.1590/2318-0889201931e190001>

SILVEIRA, LÚCIA DA; RIBEIRO, NIVALDO CALIXTO; SANTOS, SARAH RÚBIA DE OLIVEIRA; SILVA, FERNANDA MIRELLE DE ALMEIDA; SILVA, FABIANO COUTO CORRÊA DA; CAREGNATO, SÔNIA ELISA; OLIVEIRA, ADRIANA CARLA SILVA DE; OLIVEIRA, DALGIZA ANDRADE; GARCIA, JOANA COELI RIBEIRO; ARAÚJO, RONALDO FERREIRA. Ciência Aberta na perspectiva de especialistas brasileiros: proposta de taxonomia. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, vol. 26, e79646, 2021. Programa de Pós-graduação em Ciência da Informação - Universidade Federal de Santa Catarina. ISSN: 1518-2924. <https://doi.org/10.5007/1518-2924.2021.e79646>

ANEXO – Capítulo de Livro

Instituto Medeiros de Educação Avançada - IMEA

Editor Chefe

Giselle Medeiros da Costa One

Corpo Editorial

Giselle Medeiros da Costa One
 Maria Luiza Souto Porto
 Roseanne da Cunha Uchôa
 Iara Medeiros de Araújo

Revisão Final

Ednice Fideles Cavalcante Anízio

FICHA CATALOGRÁFICA

Dados de Acordo com AACR2, CDU e CUTTER

One, Giselle Medeiros da Costa.

O59s Saúde e Meio ambiente: Pesquisa, tecnologia e aplicabilidade /
 Organizador: Giselle Medeiros da Costa One. IMEA. 2023.
 648 fls. PDF

Vários Colaboradores

ISBN: 978-65-89069-23-2 (on-line)

Modelo de acesso: Word Wibe Web

<<http://www.cinasama.com.br>>

Instituto Medeiros de Educação Avançada – IMEA – João Pessoa -

PB

1. Biodiversidade 2. Meio Ambiente 3. Resíduos sólidos 4. Biotecnologia.
 Giselle Medeiros da Costa One II. Saúde e Meio ambiente: Pesquisa,
 tecnologia e aplicabilidade

CDU: 577

Laureno Marques Sales, Bibliotecário especialista. CRB -15/121

Direitos desta Edição reservados ao Instituto Medeiros de Educação Avançada –
 IMEA

Impresso no Brasil / Printed in Brazil

SUMÁRIO

SAÚDE E MEIO AMBIENTE	15
CAPÍTULO 1	16
ABORDAGEM ECOSISTÊMICA NA EXTENSÃO COMUNITÁRIA: INTEGRAÇÃO DE SABERES PARA UMA MELHOR PERCEPÇÃO DA SAÚDE AMBIENTAL EM ÁREA URBANA DE RECIFE/PE.	16
CAPÍTULO 2	52
CONDIÇÕES DE TRABALHO E SAÚDE DOS CATADORES DE MATERIAIS REICLÁVEIS NO MUNICÍPIO DE AMÉRICA DOURADA-BA	52
CAPÍTULO 3	71
DESCARTE DE MEDICAMENTOS – O QUE PENSAM OS ALUNOS E PROFESSORES DE ESCOLA PÚBLICA NA PARAÍBA	71
CAPÍTULO 4	91
DOENÇAS RELACIONADAS AO SANEAMENTO AMBIENTAL INADEQUADO (DRSAI) NO ESTADO DA PARAÍBA	91
CAPÍTULO 5	116
ESTUDO DA VIABILIDADE DO REUSO DA ÁGUA REJEITADA PELO SISTEMA DE OSMOSE REVERSA EM UM SERVIÇO DE HEMODIÁLISE	116
CAPÍTULO 6	138
EVOLUÇÃO DO ACESSO AO SANEAMENTO BÁSICO NAS PERIFERIAS DE CIDADES AMAZÔNICAS: ESTUDO DE CASO NO BAIRRO MARIA MAGDALENA, ITAITUBA/PA (2012-2022).	138

CAPÍTULO 7	157
METAL(LOÍDE)s DE REJEITOS MINERÁRIOS EM SOLOS: IMPLICAÇÕES A SAÚDE HUMANA	157
CAPÍTULO 8	176
O ACÚMULO DE MICRO E NANOPLÁSTICOS NO MEIO AMBIENTE E SUAS POSSÍVEIS IMPLICAÇÕES PARA A SAÚDE HUMANA	176
CAPÍTULO 9	198
PREDIÇÃO IN CHEMICO E IN SILICO DE EFEITOS TÓXICOS DO ANTIPARASITÁRIO DE USO VETERINÁRIO MOXIDECTINA EM PEIXES: TOXICIDADE AGUDA E NEUROTOXICIDADE	198
CAPÍTULO 10	219
PREDIÇÃO IN CHEMICO E IN SILICO DE EFEITOS TÓXICOS DO HERBICIDA ÁCIDO 2,4 DICLOROFENOXIACÉTICO (2,4-D) EM PEIXES: TOXICIDADE AGUDA E HEPATOTOXICIDADE	219
CAPÍTULO 11	240
SAÚDE E ACESSO À ÁGUA: ANÁLISE BIBLIOMÉTRICA DOS ARTIGOS INDEXADOS NA BASE WEB OF SCIENCE (2012-2022)	240
PLANEJAMENTO E GESTÃO AMBIENTAL	260
CAPÍTULO 12	261
CARACTERIZAÇÃO CLIMATOLÓGICA NO MUNICÍPIO DO RIO DE JANEIRO/RJ UTILIZANDO DADOS DO SISTEMA ALERTA RIO	261
SAÚDE E MEIO AMBIENTE	285



PLANEJAMENTO E GESTÃO AMBIENTAL

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

CAPÍTULO 12

CARACTERIZAÇÃO CLIMATOLÓGICA NO MUNICÍPIO DO RIO DE JANEIRO/RJ UTILIZANDO DADOS DO SISTEMA ALERTA RIO

Vinicius Tavares Martins GERMANO^{1, 2}

Ronaldo Figueiró Portella PEREIRA³

¹ Bacharel em Ciências Biológicas - Modalidade Biotecnologia e Produção, UEZO, RJ; ² Bacharel em Sistemas de Informação, Estácio, RJ; ³ Doutor em Ecologia, UFRJ, RJ.
vinicius.tm.germano@gmail.com

RESUMO: O monitoramento climático é de extrema importância para o meio rural, florestal e urbano. Em 1992, a Rio-92 foi o gatilho para diversas ações estatais, inclusive no Brasil, em prol do Meio Ambiente. Na cidade do Rio de Janeiro, o Sistema Alerta Rio é composto por 33 estações de monitoramento automáticas, todas pluviométricas, porém somente 8 com sensores meteorológicos. Seus dados são públicos, porém disponibilizados de forma que dificulta seu manuseio. Neste trabalho foi utilizado Knime Analytics Platform, software voltado à Ciência de Dados e *BigData* para leitura e processamento dos dados brutos. A partir das leituras (com intervalos de 15 minutos) foram geradas séries temporais mensais dos valores mínimos, médios e máximos. Em cada série temporal calculou-se a média móvel com janela de 60 meses, e depois submetidas à correlação linear. Das correlações com significância estatística ($p < 0,05$) obteve-se a equação da reta por regressão linear. As estações São Cristóvão, Guaratiba e Alto da Boa Vista apresentam maiores séries temporais de temperatura e umidade. Knime se mostrou eficiente para lidar com o volume de arquivos e dados (leituras). As estações Guaratiba e São Cristóvão apresentam tendência de aumento das médias

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

móveis das temperaturas médias e mínimas. As mesmas estações apresentam tendência de queda nas médias móveis da umidade média e mínima mensal.

Palavras-chave: Monitoramento Ambiental. Estação Climatológica. Mudanças Climáticas. Meteorologia.

INTRODUÇÃO

No meio rural, florestal e na atividade agrícola, o monitoramento de variáveis climáticas é fundamental para a gestão do desenvolvimento vegetal (OLIVEIRA-JÚNIOR, 2015). No meio urbano a discussão sobre o clima se volta às mudanças climáticas e ilhas de calor, tornando imperativo o monitoramento do clima de forma extensiva nas cidades (AMORIM, 2010; FIALHO et al., 2019).

Essas preocupações são fruto das discussões sobre Meio Ambiente em âmbito internacional iniciadas em 1972 na Conferência de Estocolmo, realizada pela ONU (Organização das Nações Unidas), em que o conceito de Justiça Ambiental começou a ser delineado (SOFFIATI, 2021). Em prosseguimento à discussão ambiental na esfera internacional, em 1992 o Rio de Janeiro sediou a Conferência Rio-92, também organizada pela ONU, projetando, desta vez, o conceito de desenvolvimento sustentável. O Brasil, que aprovara a Política Nacional do Meio Ambiente em 1981, criou o Ministério do Meio Ambiente logo após o encerramento da Rio-92 (SILVEIRA, 2021).

O Rio de Janeiro, por sua vez, inaugurou em 1996 o Sistema de Alerta de Chuvas Intensas e Deslizamentos em Encostas. Hoje, o Sistema Alerta Rio é composto por 33 estações de monitoramento automáticas, todas pluviométricas,

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

porém apenas 8 com sensores meteorológicos (Temperatura, Umidade, Pressão, Velocidade do vento, Direção do vento), e destas, somente 2 completas (São Cristóvão e Guaratiba). As estações fazem leituras com intervalos de 15 minutos. Os dados são disponibilizados em diversos arquivos (1 por mês de cada estação), fragmentando a informação em um volume expressivo de arquivos.

Com o desenvolvimento tecnológico das últimas décadas, observamos o crescimento exponencial de informações armazenadas no mundo, culminando no advento do *BigData*, conceito ainda abstrato e subjetivo, mas pacífico quanto à impossibilidade de perceber, adquirir, administrar e processar *datasets* por meio de softwares tradicionais dentro de um tempo tolerável (MOLINARO; LEAL; 2018).

Dentre softwares *open source* (código-fonte aberto, não comercial e gratuito) voltados à Ciência de Dados e *BigData*, o Knime Analytics Platform tem se destacado por ter interface gráfica intuitiva (dispensando, muitas das vezes, necessidade de programação), integração com outros softwares e conexão com dados de diversas fontes e tipos, além de ter suporte para os Sistemas Operacionais mais utilizados (Windows, Linux e Mac). Hora *et al.* (2018) avaliaram subjetivamente (Modelo de Aceitação de Tecnologia – TAM) Knime e outras três ferramentas em uma turma de discente do curso de Sistemas de Informação e identificaram aceitação do Knime (além de RapidMiner e Weka). Metelo, Bernardino e Pedrosa (2021) avaliaram qualitativa e quantitativamente (Metodologia OSSpal) Knime e outras 3 ferramentas (Dataiku, Orange e RapidMiner) e identificaram Knime como a opção mais robusta e completa.

R é uma linguagem de programação com vasto e robusto recursos estatísticos e gráficos e já vem sendo amplamente

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio utilizado em análises de dados e testes estatísticos em inúmeras e variadas áreas de estudo (COSTA, 2022).

O objetivo deste trabalho foi sistematizar e compilar os dados do Sistema Alerta Rio, produzir séries temporais mensais e submetê-las à análise de correlação e regressão lineares utilizando Knime Analytics Platform e R.

MATERIAIS E MÉTODO

Knime Analytics Platform e R Project (CRAN) foram instalados em um computador com Linux Ubuntu 20.04. Os dados do Sistema Alerta Rio foram baixados em seu site eletrônico (até 12/2021) e totalizaram 2.680.544 leituras em 1.202 arquivos *.txt*. Os arquivos foram lidos e tratados com Knime. Os dados foram normalizados para descompensar o horário de verão (as leituras registradas em sua vigência foram atrasadas em 1 hora). Ao compilar os dados a cada mês, auferiu-se os valores máximo, médio e mínimo. Os dados mensais, por sua vez, foram submetidos ao cálculo de média móvel com janela de 60 meses (MedMov60M). Em todas as séries temporais produzidas (compilados mensais e médias móveis) verificou-se a correlação linear e nas de significância estatística ($p < 0,05$) obteve-se a equação da reta por regressão linear. Os meses foram convertidos em números inteiros subsequentes, sendo 0 = 01/1997; 100 = 05/2005; 200 = 09/2013; e 299 = 12/2021. Valores extremos foram verificados nos dados brutos para validação, ocasionando, em alguns casos, na remoção pontual da leitura ou mesmo do mês inteiro, devido a reiteradas leituras não confiáveis. A representatividade dos dados mensais foi definida em 70% de

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio
leituras válidas e os meses não representativos foram excluídos da série temporal.

RESULTADOS E DISCUSSÃO

Para interpretação dos gráficos, cada estação foi identificada por uma cor (Figura 1). Serão apresentados e discutidos os resultados de Temperatura e Umidade a seguir.

Figura 1. Esquema de cores das estações nos gráficos.



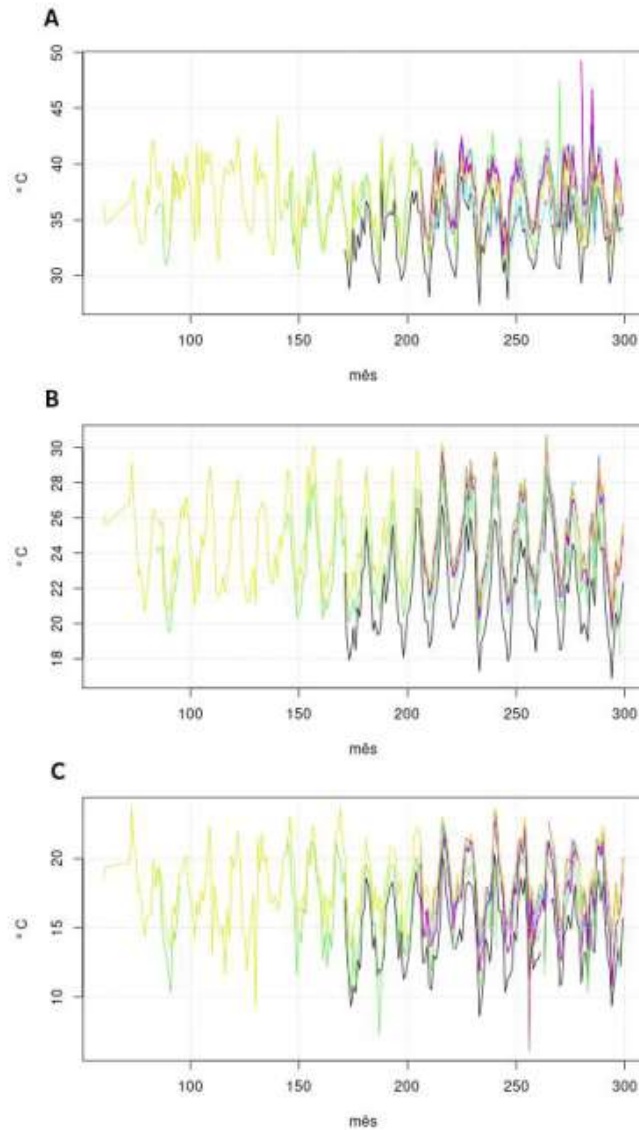
Fonte: Autores.

Temperatura

Nos dados mensais da Temperatura nota-se o comportamento de curva senoidal, característica da mudança de estações do ano e o sincronismo entre as estações. Em meses mais recentes dos gráficos, onde o número de estações é maior, é possível identificar microclimas bem distintos na cidade, com predomínio de Santa Cruz e São Cristóvão no topo e Alto da Boa Vista na parte inferior dos gráficos (Figura 2). Santa Cruz e São Cristóvão são áreas bem urbanizadas, e por isso tendem a reter calor, devido à menor cobertura vegetal e grande área construída, enquanto Alto da Boa Vista é uma região montanhosa, dentro do Parque Nacional da Tijuca e de

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio
 baixa densidade demográfica com predomínio de Mata Atlântica.

Figura 2. Temperatura (máxima, média e mínima mês).



Temperaturas máximas mensais das estações oscilam entre 27,4 e 49,2 °C (Figura 2A); as médias entre 16,9 e 30,7 °C (Figura 2B); enquanto as mínimas registram entre 6,1 e 23,7 °C (Figura 2C).

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Fonte: Autores.

As MedMov60M atenuaram a variação sazonal ditada pelas estações do ano, tornando a curva mais retilínea e indicando com mais precisão movimentos de tendência (Figura 3).

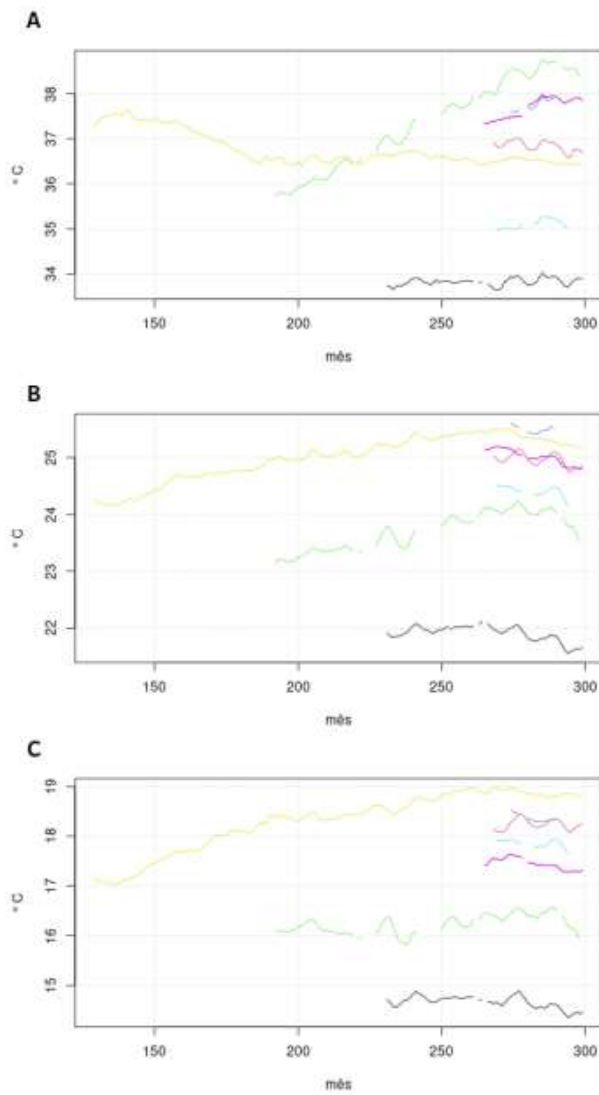
As MedMov60M das máximas mensais de 2 estações apresentam tendência de redução e 6 de alta (Tabela 1). As estações com mais dados, São Cristóvão e Guaratiba, apresentam tendência de queda e aumento, respectivamente (Figura 4).

As MedMov60M das médias mensais (Tabela 2) de 4 estações apresentam tendência de redução, sendo Santa Cruz a mais significativa, e 2 estações com tendência de alta: Guaratiba e São Cristóvão (Figura 5).

As MedMov60M das mínimas mensais de 4 estações apresentam tendência de redução e 2 de alta (Tabela 3), destacando-se São Cristóvão com acréscimo de aproximadamente 2 °C (Figura 6).

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Figura 3. Temperatura (MedMov60M máxima, média e mínima mês).



Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

MedMov60M das temperaturas máximas oscilam entre 33,6 e 38,8 °C (Figura 3A); as MedMov60M das temperaturas médias entre 21,6 e 25,6 °C (Figura 3B); enquanto as MedMov60M das temperaturas mínimas registram entre 14,3 e 19,0 °C (Figura 3C).

Fonte: Autores.

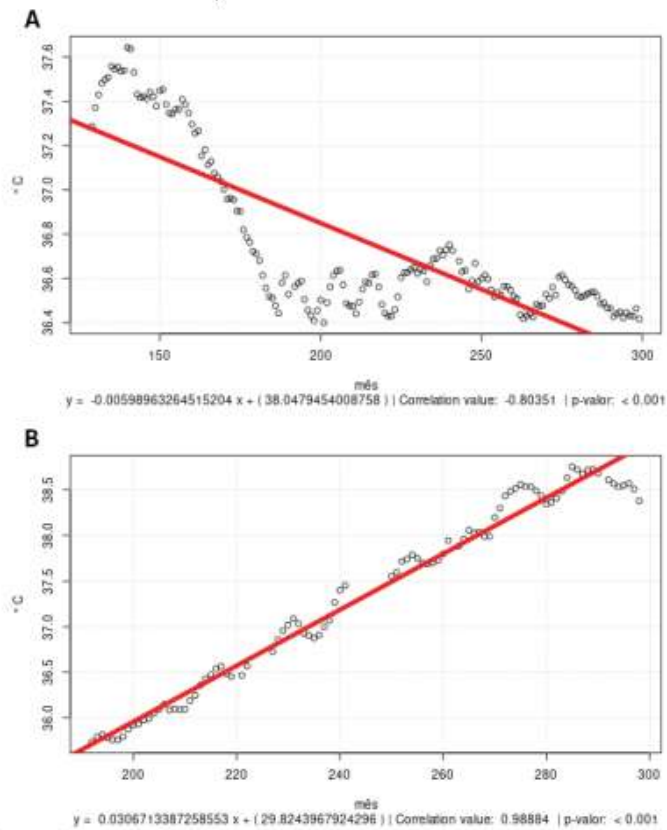
Tabela 1. Correlações lineares de Temperatura (MedMov60M máximas) em função dos meses.

Estação	Correlação linear	p-valor
São Cristóvão	-0.80351	< 0.001
Barra_Riocentro	-0.56478	< 0.001
Alto da Boa Vista	0.33923	< 0.01
Jardim Botânico	0.67327	< 0.001
Santa Cruz	0.89155	< 0.001
Irajá	0.92262	< 0.001
Guaratiba	0.98884	< 0.001

Fonte: Autores.

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Figura 4. São Cristóvão e Guaratiba – Temperatura (MedMov60M máximas)



Em São Cristóvão a MedMov60M das temperaturas máximas reduz em aproximadamente 1 °C (Figura 4A) e em Guaratiba é majorada em mais de 2,5 °C (Figura 4B).

Fonte: Autores.

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Tabela 2. Correlações lineares de Temperatura (MedMov60M médias) em função dos meses.

Estação	Correlação linear	p-valor
Santa Cruz	-0.93803	< 0.001
Jardim Botânico	-0.64168	< 0.001
Alto da Boa Vista	-0.57805	< 0.001
Barra_Riocentro	-0.50171	< 0.001
Guaratiba	0.87804	< 0.01
São Cristóvão	0.92164	< 0.001

Fonte: Autores.

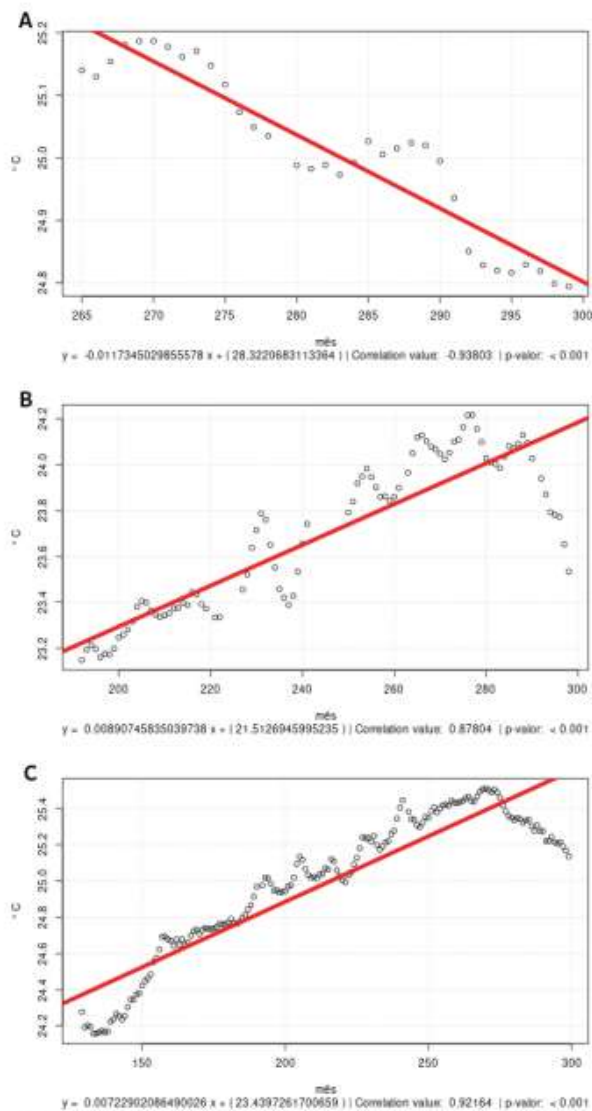
Tabela 3. Correlações lineares de Temperatura (MedMov60M mínimas) em função dos meses.

Estação	Correlação linear	p-valor
Irajá	-0.8823	< 0.001
Santa Cruz	-0.80581	< 0.001
Alto da Boa Vista	-0.56637	< 0.001
Jardim Botânico	-0.56541	< 0.01
Guaratiba	0.56598	< 0.001
São Cristóvão	0.9382	< 0.001

Fonte: Autores.

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

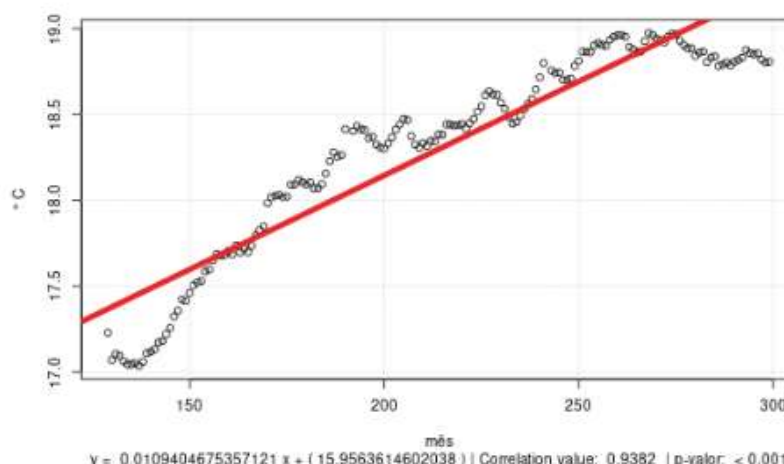
Figura 5. Santa Cruz, Guaratiba e São Cristóvão – Temperatura (MedMov60M médias)



Das estações que apresentam redução na MedMov60M das temperaturas médias, Santa Cruz é a mais significativa quanto à intensidade do

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio decaimento (Figura 5A) e as que apresentam tendência de alta são Guaratiba (Figura 5B) e São Cristóvão (Figura 5C).
Fonte: Autores.

Figura 6. São Cristóvão – Temperatura (MedMov60M mínimas)



MedMov60M das temperaturas mínimas em São Cristóvão apresentam acréscimo de aproximadamente 2 °C.

Fonte: Autores.

A tendência de alta das temperaturas verificadas nas estações São Cristóvão e Guaratiba (que apresentam maiores séries temporais) correspondem às estimativas apontadas por Silva (2014) para todo o estado do RJ. Utilizando dados anuais (máximas, médias e mínimas) das estações do INMET (Instituto Nacional de Meteorologia), Silva (2014) identificou aumento da temperatura no período de 1921 a 1990 na Praça XV, com ascendência mais significativa a partir dos anos 1950, na ordem de 1 °C. Silva (2014) também resgata trabalhos de Brandão (1987), que constatou no período de 1851 a 1900, na mesma estação, dois ciclos quentes (± 100 anos) intercalado por um frio

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

(± 40 anos). No mesmo estudo, Silva (2014) aponta a mesma tendência em outras estações do INMET localizadas no Rio de Janeiro.

Além das já citadas, vasta literatura, utilizando diferentes fontes de dados, indicam, aumento da temperatura no Rio de Janeiro por meio observacional (SILVA e DEREZYNSKI, 2014; DEREZYNSKI, CALADO e BARROS, 2017; FARIAS e BRANDÃO, 2010) e aumento para os próximos anos por meio de modelagem matemática preditiva (SILVA, 2014).

O estudo de Brandão (1987), indicando ciclos quente/frio na Praça VX com duração aproximada de 140 anos (o ciclo frio foi compreendido num período de 40 anos, entre 1891 e 1931, aproximadamente) sugere que o atual período quente termine por volta do ano 2031. Por ausência de dados antigos, o Sistema Alerta Rio é incapaz de evidenciar esses ciclos quente/frio propostos por Brandão (1987). Entretanto, os estudos ambientais para além do mero acompanhamento observacional sugerem contínuo aumento da temperatura global, explicitadas nos relatórios do IPCC (Painel Intergovernamental sobre Mudanças Climáticas), vinculado à ONU, indicando a exploração, degradação e poluição ambiental como determinantes para as Mudanças Climáticas.

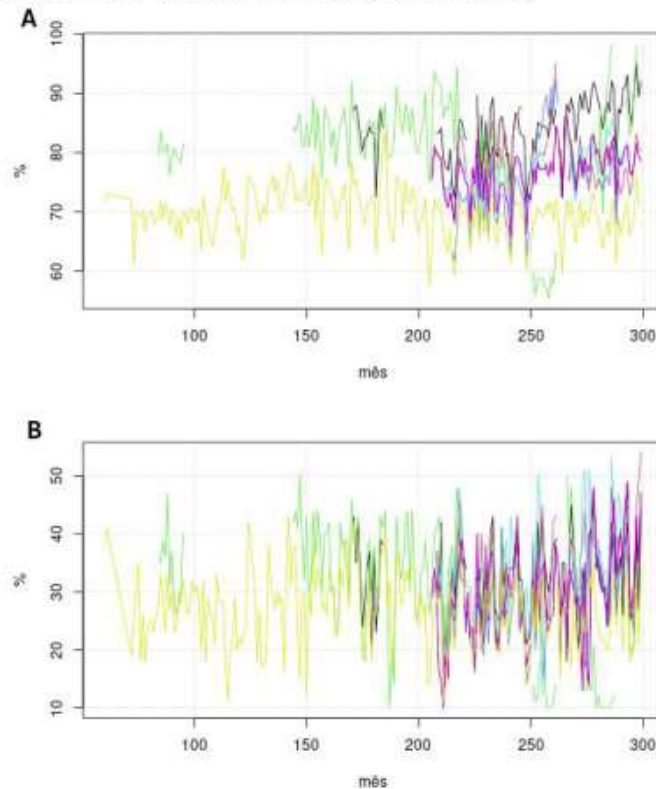
Silva (2014) e Lucena (2010) relatam sobre a influência da urbanização na temperatura local. Os autores evidenciam significativos aumentos de temperaturas em momentos marcados por forte urbanização, como recordes máximos de temperatura ocorridos em Bangu em alguns anos das décadas de 1940-60.

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Umidade

As estações São Cristóvão e Guaratiba também possuem as maiores séries temporais (Figura 7). Nota-se também pela Umidade média que Alto da Boa Vista e São Cristóvão apresentam microclima bem distintos, com predomínio de mais umidade no Alto da Boa Vista, região de floresta, enquanto em São Cristóvão, mais urbano, é menos úmido (Figura 7A).

Figura 7. Umidade (médias e mínimas mês)

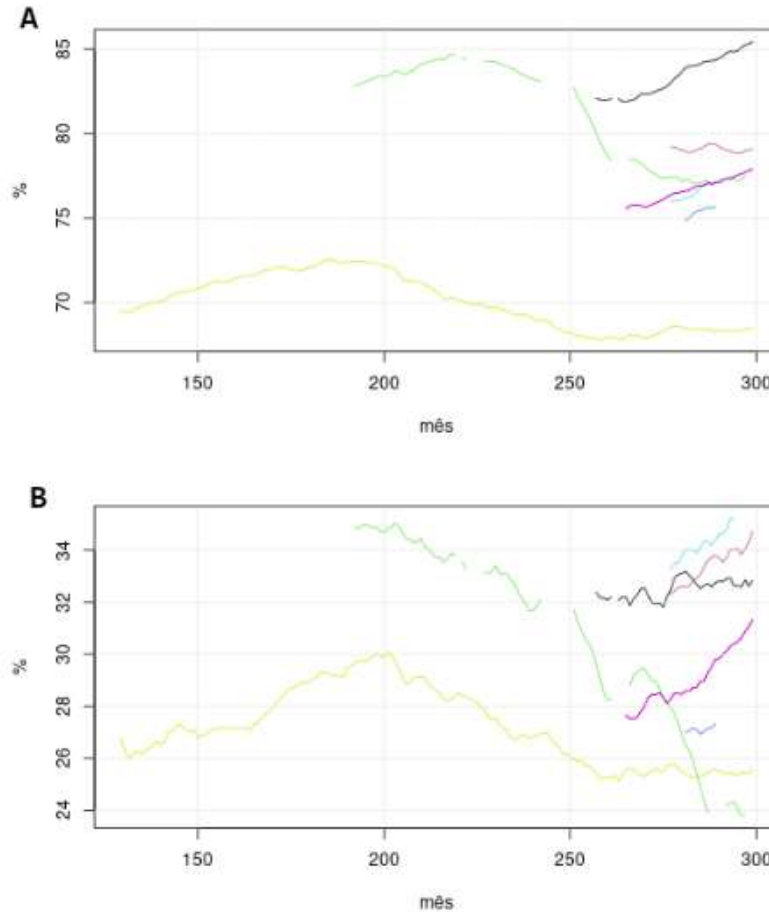


Umidade média mensal das estações oscila entre 55 e 98 % (Figura 7 A) e as mínimas entre 10 e 54 % (Figura 7B).

Fonte: Autores

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Figura 8. Umidade (MedMov60M médias e mínimas)



As MedMov60M da umidade média notabiliza São Cristóvão como a estação menos úmida do Sistema Alerta Rio, com larga distância das demais. Alto da Boa Vista possui altos níveis de umidade e Guaratiba reduz bruscamente seu nível, saindo do intervalo 80-85 % para 75-80 % (Figura 8A).

Já na MedMov60M da umidade mínima, Guaratiba apresenta sucessivos recordes de mínimas, variando em torno de 10 pontos percentuais para menos, ao passo que Alto da

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Boa Vista e Santa Cruz superam os recordes máximos (Figura 8B).

MedMov60M da umidade média mensal ficam entre 68 e 85 % (Figura 8A); enquanto as MedMov60M da umidade mínimas mensais registram entre 24 e 35 % (Figura 8B).

Fonte: Autores

Submetidas a correlação linear, as MedMov60M das médias mensais de 2 estações apresentam tendência de redução (Guaratiba e São Cristóvão) e 4 de alta (Tabela 4).

Quanto à Umidade, as estações com maior extensão da série histórica também apresentam tendência de redução, evidenciando que os dias dessas regiões tem se tornado menos úmidos.

Tabela 4. Correlações lineares da Umidade (MedMov60M médias) em função dos meses.

Estação	Correlação linear	p-valor
Guaratiba	-0.89838	< 0.001
São Cristóvão	-0.72536	< 0.001
Irajá	0.9428	< 0.001
Alto da Boa Vista	0.97512	< 0.001
Jardim Botânico	0.97955	< 0.001
Santa Cruz	0.99192	< 0.001

Fonte: Autores

Em Guaratiba, a brusca queda nos níveis de umidade (Figura 9B) e temperatura (Figura 5B) pode ter relação com a modificação do uso do solo, pois a região tem passado, nos últimos anos, por expansão da área urbana, demandando

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

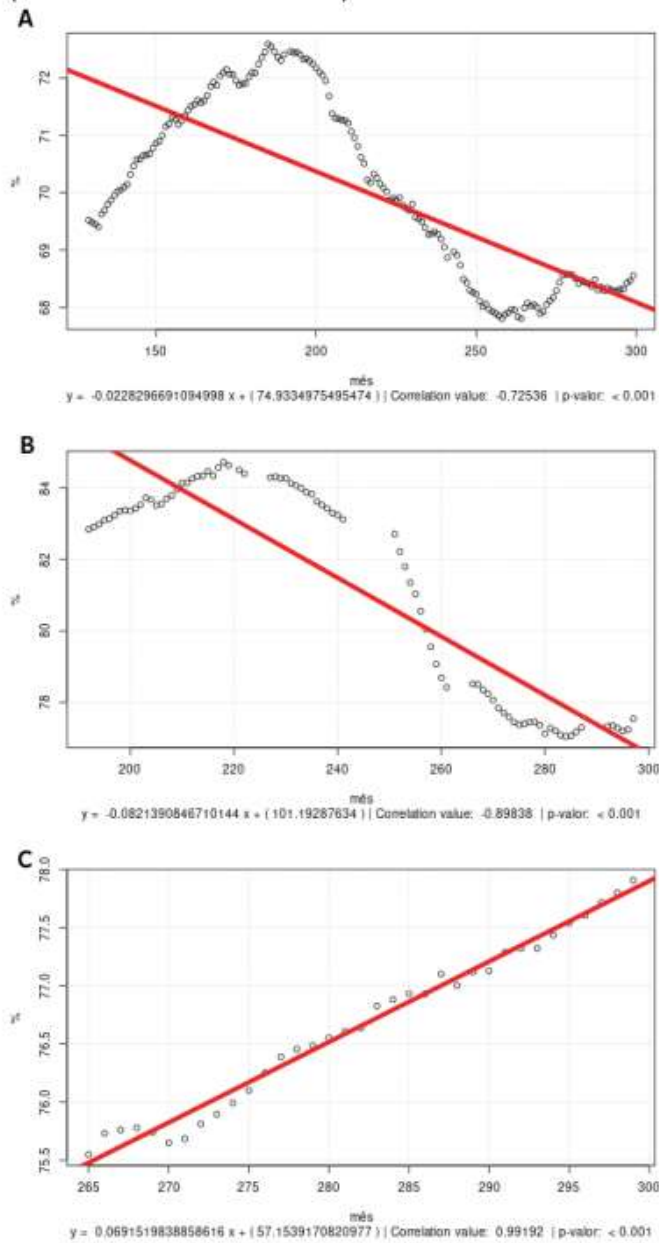
desmatamento de área de manguezal. Barros, Deus e Costa (2020) investigaram a expansão urbana na Bacia Hidrográfica do Rio Cabuçu-Piraquê, onde encontra-se o bairro de Guaratiba, e apontaram crescimento da região a partir da década de 2000, sobretudo após o período chamado “Era dos Megaeventos Esportivos” inaugurado pelos XV Jogos Pan-Americanos em 2007, marcado por melhorias viárias em Guaratiba, favorecendo a especulação imobiliária. Segundo Mendonça (2018), no período de 1965 a 2016 (dois anos após a Copa do Mundo 2014 ocorrida no Brasil, tendo o Rio de Janeiro sido cidade sede de jogos), a área de Floresta de Manguezal encolheu 7 Km² (- 30%), a Planície Hipersalina perdeu 2 Km² (- 23%) e a Área Urbana cresceu 43 Km² (+ 519%). Guaratiba também tem registrado acentuada redução da MedMov60M da umidade mínima, evidenciando recordes mínimos sucessivos na umidade (Figura 10B).

Em São Cristóvão (Figura 9A) a redução de MedMov60M umidade média se inicia aproximadamente 20 meses antes de Guaratiba (Figura 9B). Embora o tamanho da série histórica seja reduzida, em Santa Cruz a MedMov60M de umidade média apresenta cada vez valores mais altas, indicando dias mais úmidos (Figura 9C).

Guaratiba e São Cristóvão também apresentam tendência de redução nas MedMov60M da umidade mínima, indicando renovação de recordes mínimos enquanto as demais possuem tendência de aumento (Tabela 5).

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Figura 9. São Cristóvão, Guaratiba e Santa Cruz – Umidade (MedMov60M médias)



Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio MedMov60M da umidade médias mensais de São Cristóvão (Figura 9A) e Guaratiba (Figura 9B) em forte tendência de queda enquanto Santa Cruz aumenta de forma consistente (Figura 9C).

Fonte: Autores

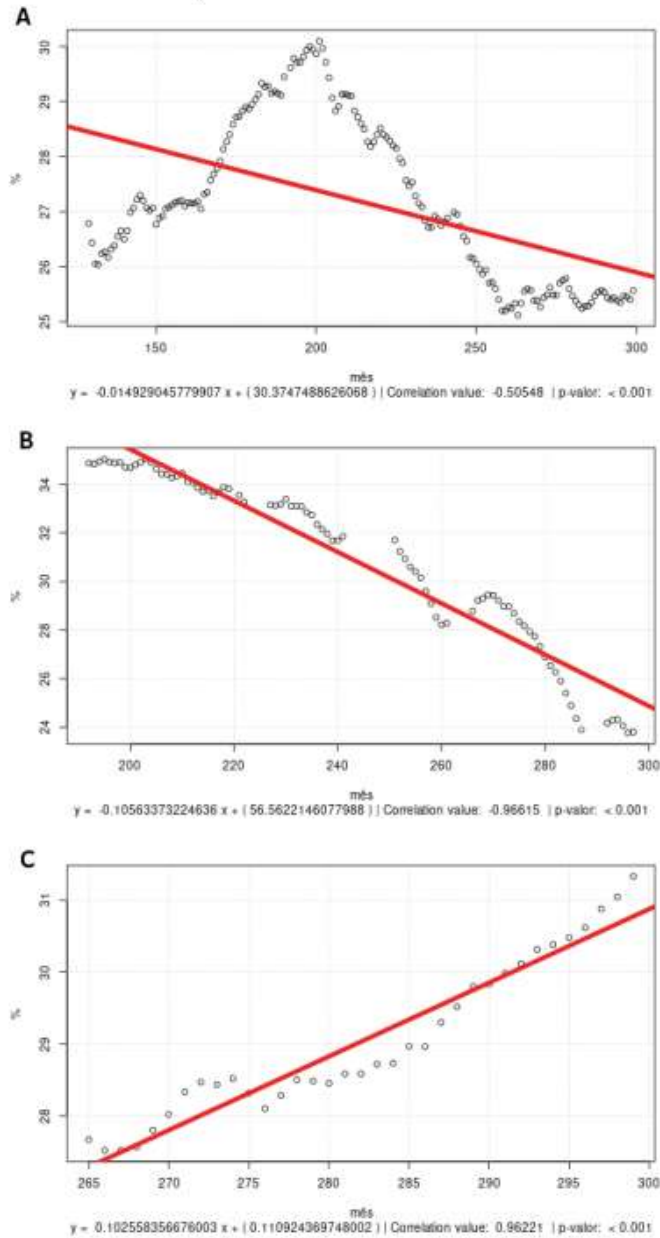
Tabela 5. Correlações lineares da Umidade (MedMov60M mínimas) em função dos meses.

Estação	Correlação linear	p-valor
Guaratiba	-0.96615	< 0.001
São Cristóvão	-0.50548	< 0.001
Alto da Boa Vista	0.66172	< 0.001
Jardim Botânico	0.96094	< 0.001
Santa Cruz	0.96221	< 0.001
Barra_Riocentro	0.96599	< 0.001

Fonte: Autores

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

Figura 10. São Cristóvão, Guaratiba e Santa Cruz – Umidade (MedMov60M mínimas)



MedMov60M da umidade mínimas mensais de São Cristóvão (Figura 10A) e com mais intensidade em Guaratiba (Figura 10B)

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio apresentam tendência de queda, enquanto Santa Cruz de alta (Figura 10C).
Fonte: Autores

CONCLUSÕES

O Sistema Alerta Rio é uma importante fonte de dados para monitoramento climatológico mas possui séries históricas ainda incipientes em termos climáticos. Algumas estações possuem períodos de dados faltantes, que no caso de Guaratiba chegam a dois anos. A complexidade com que os dados são disponibilizados foi contornada com uso do Knime, que se mostrou muito eficiente também no tratamento dos dados e aliado ao R, na geração dos gráficos. Cabe salientar que os autores recomendam fortemente que se o uso de Knime se der em ambiente Linux Ubuntu, que se observe a versão utilizada neste trabalho (20.04), pois a versão mais recente de Ubuntu (22.04) apresenta uma incompatibilidade com Knime, que mesmo sendo contornada através de mudanças de configuração, não foi satisfatória e inviabilizou a execução do trabalho.

São Cristóvão e Guaratiba são as estações com maiores séries temporais. Analisando as médias móveis de 60 meses, a temperatura das duas estações apresentam tendência de alta nas médias e mínimas, e de redução da umidade, também nas médias e mínimas. Especialmente em Guaratiba, outros trabalhos relacionam a expansão urbana e degradação ambiental com a alteração do microclima, acarretando em aumento de temperatura.

As médias móveis com janelas de 60 meses se mostraram muito eficientes em evidenciar movimentos de

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio
tendências, porém exige longas séries históricas para se obter resultados satisfatórios.

REFERÊNCIAS BIBLIOGRÁFICAS

- AMORIM, M. C. de C. T. Climatologia e gestão do espaço urbano. **Mercator - Revista de Geografia da UFC**, vol. 9, núm. 1, diciembre, 2010, pp. 71-90.
- BARROS, F. da C.; DEUS, L. A. B. de; COSTA, A. J. S. T. da. Análise sobre tendências de expansão urbana na Bacia Hidrográfica do Rio Cabuçu-Piraquê, Zona Oeste da cidade do Rio de Janeiro. **Revista de Geografia do Programa de Pós-Graduação em Geografia da UFJF**. V. 10, Nº 1, 2020.
- BRANDÃO, A.M.P.M. **Tendências e Oscilações Climáticas na Área Metropolitana do Rio de Janeiro**. Dissertação (Mestrado em Geografia). Faculdade de Letras, Filosofia e Ciências Humanas, Universidade de São Paulo, São Paulo, 1987.
- COSTA, D. R. da. **Linguagem R: Conceitos e Aplicações**. Ed. Aya. Ponta Grossa, 2022.
- DERECZYNSKI, C. P.; CALADO, R. N.; BARROS, A. B. de. Chuvas Extremas no Município do Rio de Janeiro: Histórico a partir do Século XIX. **Anuário do Instituto de Geociências – UFRJ**. Vol. 40 - 2 / 2017 p. 17-30.
- FARIAS, H. S. de. BRANDÃO, A. M. de P. M. A influência antrópica na evolução semanal da temperatura do ar no bairro Maracanã/RJ. **Revista Brasileira de Climatologia**. Ano 6, Volume 6, junho/2010.
- FIALHO, E. S.; FERNANDES, L. A.; CORREA, W. de S. C. Climatologia Urbana: Conceitos, Metodologias e Técnicas. **Revista Brasileira de Climatologia**, Ano 15 – Edição Especial – XIII Simpósio Brasileiro de Climatologia Geográfica – JUN 2019.
- HORA, G. S.; SANTOS JÚNIOR, G. P. dos S.; MENEZES, J. S. S.; REHEM NETO, A. N. Avaliação de ferramentas de mineração de dados: uma abordagem com o Modelo TAM. **Interfaces Científicas - Exatas e Tecnológicas**. Aracaju V.2 N.3 p. 109 - 121 Fev. 2018.
- MENDONÇA, J. M. **Expansão urbana na bacia de drenagem e variações dos conteúdos de nutrientes no sedimento de um manguezal em eutrofização (Manguezal de Guaratiba/RJ)**. Monografia (Graduação em Geografia). Instituto de Geociências. Universidade Federal Fluminense/RJ. 2018.
- METELO, M.; BERNARDINO, J.; PEDROSA, I. Avaliação de Ferramentas Open Source para Data Science usando a Metodologia OSSpal. RISTI - **Revista Ibérica de Sistemas e Tecnologias de Informação**.

Caracterização Climatológica no Município do Rio de Janeiro/RJ utilizando dados do Sistema Alerta Rio

- MOLINARO, C. A.; LEAL, A. A. F. Big data, machine learning e a preservação ambiental: instrumentos tecnológicos em defesa do meio ambiente. **Veredas do Direito**, Belo Horizonte, v.15 n.31 p.201-224 Janeiro/Abril de 2018. Disponível em: <<http://www.domhelder.edu.br/revista/index.php/veredas/article/view/1142>>. Acesso em: 05/11/2022.
- OLIVEIRA-JÚNIOR, J. F.; LYRA, G. B; MENDES, L. S. Modelos de Zolnier e Campbell-Norman para estimativa da temperatura do ar no Rio de Janeiro. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.19, n.8, p.727-733, 2015.
- SILVA, W. L da. **Caracterização Climatológica e Mudanças Climáticas no Estado do Rio de Janeiro**. Dissertação (Mestrado). Universidade Federal do Rio de Janeiro. Centro de Ciências Matemáticas e da Natureza – CCMN. Instituto de Geociências – IGEO. Programa de Pós-Graduação em Meteorologia – PPGM. Rio de Janeiro – RJ, 2014.
- SILVA, W. L da. DEREZYNSKI, C. P. Caracterização Climatológica e Tendências Observadas em Extremos Climáticos no estado do Rio de Janeiro. **Anuário do Instituto de Geociências – UFRJ**. Vol. 37 - 2 / 2014 p. 123-138.
- SILVEIRA, J. G. da. A Rio-92, os movimentos ecologistas e a Política Nacional do Meio Ambiente: uma reflexão sobre a construção das políticas ambientais brasileiras na década de 1990. **Revista Hydra**. Volume 5, número 9. abril de 2021.
- SOFFIATI, A. A Conferência de Estocolmo e as posturas políticas diante da questão ambiental. IN: SILVA, M. W. da; RAMOS, T. T.; RIBEIRO, D. de A. (org.). **Pesquisas socioespaciais e ambientais** [livro eletrônico]: II Seminário de pesquisa do programa de pós graduação em geografia da UFF: Campos dos Goytacazes. 1. ed. -- São Carlos, SP: Editora Cubo, 2021.