

Especificação de um Experimento de Ciência de Dados para Análise de Áreas Contaminadas

Rosa Virginia Encinas Quille^{1,2,3}, Gabriela dos Santos Luchetti Vieira⁴, Leandro Gomes de Freitas², Pedro Luiz Pizzigatti Corrêa^{1,3}, Solange Nice Alves de Souza^{1,3}, Alexandre Muselli Barbosa², Felipe Valencia de Almeida³

¹Escola de Artes, Ciências e Humanidades – Universidade de São Paulo

²Instituto de Pesquisas Tecnológicas

³Escola Politécnica - Universidade de São Paulo

⁴Instituto de Ciência e Tecnologia - Campus de Sorocaba – Universidade Estadual Paulista “Júlio de Mesquita Filho”

{encinas, pedro.correa, ssouza, fvalencia}@usp.br

{lfreitas, muselli}@ipt.br

{gabriela.luchetti}@unesp.br

Abstract. *This work presents the methodological aspects of an experiment with Data Science to optimize the management of environmental data from Contaminated Sites management systems. Recent studies show that there is a lack of maturity models in the management of environmental data, such as the monitoring of contaminated sites. This scenario presents opportunities for innovation, through the application of computational tools associated with the digital transformation, intelligence, and data analytics. Therefore, a real case study is presented, involving the preliminary qualitative analysis of an environmental monitoring dataset from the EACH - USP University Campus. It could be observed that, despite the quality of the database, there are management and availability problems, confirming the opportunity to carry out a proof of concept experiment with that methodology.*

Resumo. *Este trabalho apresenta um experimento de aplicação de Ciência de Dados para otimizar a gestão de dados ambientais do Gerenciamento de Áreas Contaminadas. Estudos recentes indicam que existe uma carência de modelos de maturidade na gestão de dados ambientais, como é o caso do monitoramento das áreas contaminadas. Este cenário apresenta oportunidades de inovação, pela aplicação de ferramentas computacionais associadas à transformação digital, inteligência e capacidade analítica. Sendo assim, são apresentados aspectos metodológicos e um estudo de caso real, envolvendo a análise preliminar qualitativa dos dados de monitoramento ambiental do Campus da USP Leste. Observou-se que, apesar da riqueza da base de dados, existem problemas de gestão e de disponibilidade, confirmando a oportunidade para a realização de provas de conceito com a metodologia proposta.*

1. Introdução

A Ciência de Dados tornou-se uma abordagem transformadora para a análise de dados nas diferentes áreas, utilizando metodologias e tecnologias atuais para obter *insights* a partir de *datasets* complexos [Cao 2016, Iskamto 2023, Kumar et al. 2023, Ismail et al. 2024]. Na área ambiental, essa abordagem ampliou a compreensão dos fenômenos naturais e antropogênicos, contribuindo para o monitoramento, a modelagem e a previsão de impactos relacionados à qualidade do solo, do ar e da água. Em particular, o Gerenciamento de Áreas Contaminadas (GAC) se beneficia do uso de técnicas analíticas modernas para explorar grandes volumes de dados gerados por sensores, campanhas de campo e sistemas de monitoramento contínuo. Por exemplo, modelos de aprendizado de máquina têm sido empregados para prever os níveis de contaminação do solo e identificar os principais fatores de influência [Anifowose and Anifowose 2024, Bi et al. 2024].

As áreas contaminadas trazem implicações diretas para a saúde pública, o meio ambiente e a economia local, provocando a desvalorização de imóveis e o abandono de terrenos *brownfields*. A contaminação pode alcançar recursos hídricos, comprometendo mananciais e ecossistemas. Segundo a Companhia Ambiental do Estado de São Paulo (CETESB), essas áreas têm origem acidental ou decorrente de atividades como postos de combustíveis, indústrias e disposição inadequada de resíduos [CETESB 2022]. O estado de São Paulo é referência nacional na estruturação de políticas públicas e instrumentos legais voltados ao GAC, com destaque para a Lei Estadual nº 13.577/2009 e, em nível federal, a Resolução CONAMA nº 420/2009. Atualmente, o estado registra mais de 6.984 áreas contaminadas e reabilitadas [CETESB 2020], com predominância dos hidrocarbonetos derivados de petróleo, o que reforça a necessidade de medidas específicas para prevenção e remediação.

Um dos fenômenos mais críticos no GAC é a Intrusão de Vapores, a migração de gases tóxicos ou inflamáveis do solo para o interior de edificações, onde podem representar sérios riscos à saúde humana. O Brasil, no entanto, ainda carece de normativas específicas e bancos de dados estruturados para lidar com essa complexidade [Gouvêa Jr. et al. 2018]. As informações ambientais são produzidas em diversas fases do processo de gerenciamento, desde o diagnóstico inicial até a remediação e o monitoramento contínuo, envolvendo uma ampla gama de métodos de coleta e análise de dados, como amostragens em campo, modelagem de transporte de contaminantes e controle de equipamentos [Moraes et al. 2014]. Apesar da riqueza e diversidade de dados gerados, o GAC ainda enfrenta barreiras significativas em termos de integração, interoperabilidade e uso eficiente da informação [Horst et al. 2017, Barbosa et al. 2017].

Iniciativas como o projeto MOIRA [Appelgren et al. 1996], estudos de integração geoespacial em São Paulo [Barbosa et al. 2017], e as abordagens de Asante-Duah [Asante-Duah 2019] e Palanisamy et al. [Palanisamy 2016] evidenciam a importância de sistemas de informação robustos e bem estruturados na gestão de áreas contaminadas. Com a crescente complexidade e volume dos dados ambientais, práticas da Ciência de Dados tornam-se essenciais para potencializar a análise, gerar novos conhecimentos e apoiar a transformação digital da gestão ambiental.

Este trabalho propõe justamente investigar essas possibilidades por meio da aplicação de metodologias de Ciência de Dados a um estudo de caso relacionado ao GAC em áreas urbanas brasileiras.

2. Estudo de caso USP - Leste

Os estudos ambientais preliminares do campus USP Leste, iniciados em 2004 e impulsionados por um forte apelo político para sua construção, estabeleceram medidas mitigadoras para possíveis alterações no meio físico, biótico e antrópico [USP 2004]. Esses estudos revelaram a presença de gás metano no subsolo, um gás inflamável originado da degradação de matéria orgânica presente nos materiais de dragagem depositados no terreno entre as décadas de 1980 e 1990, provenientes das obras de retificação do Rio Tietê [SERVMAR 2014]. Diante desse cenário, foram conduzidos diversos estudos de diagnóstico e monitoramento ambiental visando garantir a segurança das obras e dos futuros usuários do campus. Durante a instalação dos edifícios, sistemas de engenharia foram implementados para evitar o acúmulo de gases e vapores sob as construções, além de uma extensa rede de monitoramento de gases no solo [IPT 2011].

Em 2014 foi feita a primeira campanha para a instalação dos Poços de Monitoramento de Gases do solo (PMGs), resultando em uma série histórica regular desde então. Além do monitoramento regular, o histórico de estudos ambientais da USP Leste conta com uma vasta quantidade de relatórios emitidos pelas empresas de consultoria e órgãos públicos, o que permite o acesso a um amplo acervo de metadados importantes para a compreensão dos fenômenos e da dinâmica da migração dos gases no terreno. Como exemplo destes dados podem ser relacionados: resultados de análises químicas de solo e água subterrânea, os perfis litológicos das sondagens realizadas, modelos matemáticos e pareceres técnicos. Além desses relatórios, considera-se importante também a complementação das informações por outras fontes de dados, tais como os dados climáticos disponibilizados por empresas de meteorologia.

Para a qualidade dos dados durante a etapa de levantamento e montagem dos conjuntos de dados, são verificados fatores, como a calibração dos equipamentos de medição, a metodologia dos trabalhos de campo e a qualificação técnica das equipes executoras. Nas informações apresentadas nos relatórios consultados, todos esses quesitos foram devidamente cumpridos e comprovados pelas consultorias responsáveis [WEBER 2019].

3. Metodologia

O GAC pode ser abordado sob duas frentes metodológicas complementares: a gestão e a análise dos dados ambientais. A primeira está relacionada à organização, curadoria e interoperabilidade dos dados ao longo de seu ciclo de vida; a segunda envolve a aplicação de técnicas analíticas para gerar conhecimento a partir dos dados coletados. Neste trabalho, o foco é direcionado exclusivamente à frente analítica, considerando que a base de dados já se encontra estruturada e acessível. Essa delimitação é alinhada às recomendações de trabalhos atuais, que destacam a importância de distinguir entre gestão e análise na arquitetura de sistemas de dados ambientais, em contextos com grande volume e diversidade de dados [Quille et al. 2023].

A metodologia especificada de análise de dados segue uma estrutura adaptada do ciclo de análise em Ciência de Dados, composta por oito passos: (1) Estabelecer Questionamentos, (2) Aquisição de Dados, (3) Limpeza de Dados, (4) Modelagem, (5) Análise de Dados, (6) Apresentação, (7) Análise de Resultados, e (8) Refinar o problema. Os passos da metodologia descritos a seguir descrevem o que pode ser feito em cada etapa no contexto para a análise de áreas contaminadas.

Estabelecer Questionamentos: Nesta etapa, devem ser definidos com clareza os objetivos do problema a ser analisado. No contexto deste estudo, formulamos os seguintes questionamentos orientadores: Quais são os compostos químicos responsáveis pela contaminação dos solos? Quais são os principais parâmetros indicadores a serem monitorados? Como a distribuição dos poluentes pode ser interpretada com base nas características de cada área? Quais compostos estão se acumulando ao longo do tempo? Quais estão se atenuando? Os sistemas de remediação estão sendo eficazes na redução das concentrações? Existem eventos com comportamentos anômalos? Em quanto tempo as metas de remediação poderão ser atingidas?

Aquisição de Dados: A aquisição de dados consiste em coletar dados conforme as necessidades do projeto, com base nos questionamentos previamente estabelecidos. Um exemplo de aquisição de dados para o monitoramento da intrusão de gases e vapores é o caso do campus USP Leste, onde são realizadas medições periódicas. Essas medições foram realizadas em diferentes dispositivos de monitoramento, tais como poços (PMGs), caixas de passagem, drenos e sistemas de ventilação. Para este experimento, centra-se especificamente no tipo PMGs. Assim, as variáveis coletadas são CH₄ (%), CO₂ (%), O₂ (%), H₂S (ppm), CO (ppm), VOC (ppm) e LEL (%) utilizando instrumentos portáteis como o GEM 5000, o MX6 iBrid e BW GasAlert Micro 5. O conjunto de dados cobre um período superior a oito anos, compreendido entre 2.014-04-08 e 2.022-06-30, com 100.570 observações, para maior detalhamento do dataset analisado na seguinte referencia [Quille et al. 2025].

Limpeza dos Dados: Esta etapa compreende a aplicação de procedimentos para a garantia da integridade, consistência e qualidade analítica do conjunto de dados. Inicialmente, foram excluídos atributos e registros irrelevantes para os objetivos da análise, tais como colunas com rótulos genéricos (e.g., *unknown*, que não agregavam valor informacional. Em seguida, foram tratados valores ausentes e identificadas inconsistências semânticas, como categorias duplicadas ou mal formatadas (por exemplo, 'E' versus 'E' com espaço). Também foi realizada a detecção e o tratamento de valores extremos (*outliers*), bem como a correção de erros de codificação e de formatação — como o uso incorreto de separadores decimais (e.g., 79_1% ou 87..6%) — além da padronização de formatos de dados (e.g., *float*, *datetime* e outras representações numéricas).

Modelagem: Esta etapa envolve a construção e aplicação de modelos estatísticos ou de aprendizado de máquina para explorar relações entre variáveis e realizar previsões. Em áreas contaminadas, os dados estão sujeitos a incertezas e ruídos, e os sistemas físicos geralmente apresentam comportamento não linear, possibilitando múltiplos modelos ajustáveis. Nesse contexto, podem ser aplicadas técnicas de regressão supervisionada, como regressão linear, *Random Forest* e *Decision Tree*, entre outras. Essas abordagens demonstraram alto desempenho preditivo em estudo recentemente aceito no LatinX ICML 2025 [Quille and de Almeida 2025], ao modelar séries temporais multivariadas de gases monitorados no solo.

Análise de Dados: Nesta etapa, aplicam-se métodos estatísticos e computacionais para avaliar os dados processados, com foco no tipo de análise que deseja-se realizar, por exemplo, análise exploratória, análise preditiva, entre outros. São utilizadas técnicas como estatísticas descritivas, identificação de *outliers*, correlações e análise multivariada. A interpretação é potencializada por visualizações, como gráficos, mapas temáticos,

tabelas dinâmicas e matrizes de dispersão, conforme proposto por Tukey [Tukey 1977]. Representações visuais eficazes mostram padrões, tendências e anomalias que orientam hipóteses e decisões.

Apresentação: A apresentação dos resultados deve priorizar a clareza, acessibilidade e profundidade analítica, podendo incorporar ferramentas de *Analytics* e *Business Intelligence* para representar os achados de forma interativa e compreensível. O uso de painéis interativos, visualizações geoespaciais e gráficos comparativos permite a comunicação eficiente dos resultados do monitoramento, apoiando o processo de avaliação de desempenho ambiental e eficiência operacional. Segundo [Few 2009], a apresentação visual de dados facilita a identificação de padrões, exceções e pontos críticos, sendo essencial para comunicar descobertas complexas a públicos técnicos e não técnicos.

Análise de Resultados: A interpretação dos resultados deve ser conduzida de forma crítica, considerando as limitações dos dados, a sensibilidade dos modelos utilizados e a relevância das conclusões obtidas. Ferramentas de *Visual Analytics* (que combinam análise automatizada e visualização interativa) são úteis nesta fase para integrar múltiplas dimensões dos dados e facilitar o engajamento dos diversos *stakeholders* envolvidos. Como destacado por [Thomas and Cook 2005], a abordagem de *Visual Analytics* permite não apenas visualizar os dados, mas também promover a descoberta guiada e a geração de insights em tempo real, o que é muito requerido em contextos de gestão ambiental e tomada de decisão.

Refinar o Problema: A etapa de refinamento representa uma retroalimentação fundamental dentro do ciclo de ciência de dados aplicada. Trata-se da reavaliação crítica das questões iniciais à luz dos resultados obtidos, com o objetivo de ajustar hipóteses, redefinir variáveis de interesse, aprimorar o desenho experimental e coletar novos dados direcionados. Este processo iterativo enfatiza a importância da revisão contínua das perguntas analíticas para alcançar resultados mais relevantes. Ao refinar o problema, aumenta-se o potencial de gerar conhecimento significativo e aplicável às práticas no GAC.

4. Resultados e discussões

4.1. Análise exploratória dos dados

Na Tabela 1, observa-se que as medições se concentram majoritariamente nos edifícios E1 (14,50 %), E2 (11,93 %), E3 (11,92 %) e E4 (10,21 %) que representam mais de 48 % do total. Por outro lado, E14, E13 e E12 somam menos de 10 %, evidenciando a priorização de áreas críticas no plano de monitoramento. A quantidade de conjuntos de PMGs no *dataset* é de 128 conjuntos. Um conjunto são dois poços no mesmo local: exemplo, PMG-A (raso) + PMG-B (profundo); às vezes, três poços: PMG-A, PMG-B e PMG-C.

A limpeza de dados foi realizada conforme descrito na seção de metodologia. A Figura 1 apresenta a matriz de dados ausentes (sendo a observação 1 a inicial e 100570 e 90921 as últimas antes e após o tratamento), seguida da distribuição de frequências do novo *dataset* e da matriz de correlação entre variáveis. O tratamento considerou as variáveis com maior completude e relevância analítica (CH₄, O₂, CO₂, H₂S), resultando na preservação de 90.921 das 100.570 observações, o que corresponde à exclusão de aproximadamente 9 % dos dados. As variáveis VOC e LEL, por apresentarem elevada proporção de valores ausentes, foram descartadas nesta etapa. A distribuição das

Tabela 1. Distribuição das frequências de medições por edificação na USP Leste

ID	Edifício	AF	RF	RF %	CF	CF %
E1	Conjunto Laboratorial	14.586	0,145	14,50 %	14.586	14,50 %
E2	Bloco Inicial (conjunto didático)	11.996	0,119	11,93 %	26.582	26,43 %
E3	I-3 Biblioteca	11.986	0,119	11,92 %	38.568	38,35 %
E4	Edifício I-4	10.272	0,102	10,21 %	48.840	48,56 %
E5	Edifício I-1 Parte 1	7.706	0,077	7,66 %	56.546	56,23 %
E6	Edifício I-1 Parte 2	6.850	0,068	6,81 %	63.396	63,04 %
E7	Enfermaria	5.982	0,059	5,95 %	69.378	68,98 %
E8	I-3 Auditórios	5.978	0,059	5,94 %	75.356	74,93 %
E9	CAT	5.978	0,059	5,94 %	81.334	80,87 %
E10	Ginásio Poliesportivo	5.978	0,059	5,94 %	87.312	86,82 %
E11	Incubadora	5.124	0,051	5,09 %	92.436	91,91 %
E12	Guarda Universitária	3.096	0,031	3,08 %	95.532	94,99 %
E13	Transportes	2.880	0,029	2,86 %	98.412	97,85 %
E14	Portaria P3	2.158	0,022	2,15 %	100.570	100,00 %

AF = Absolute Frequency; RF = Relative Frequency; CF = Cumulative Frequency.

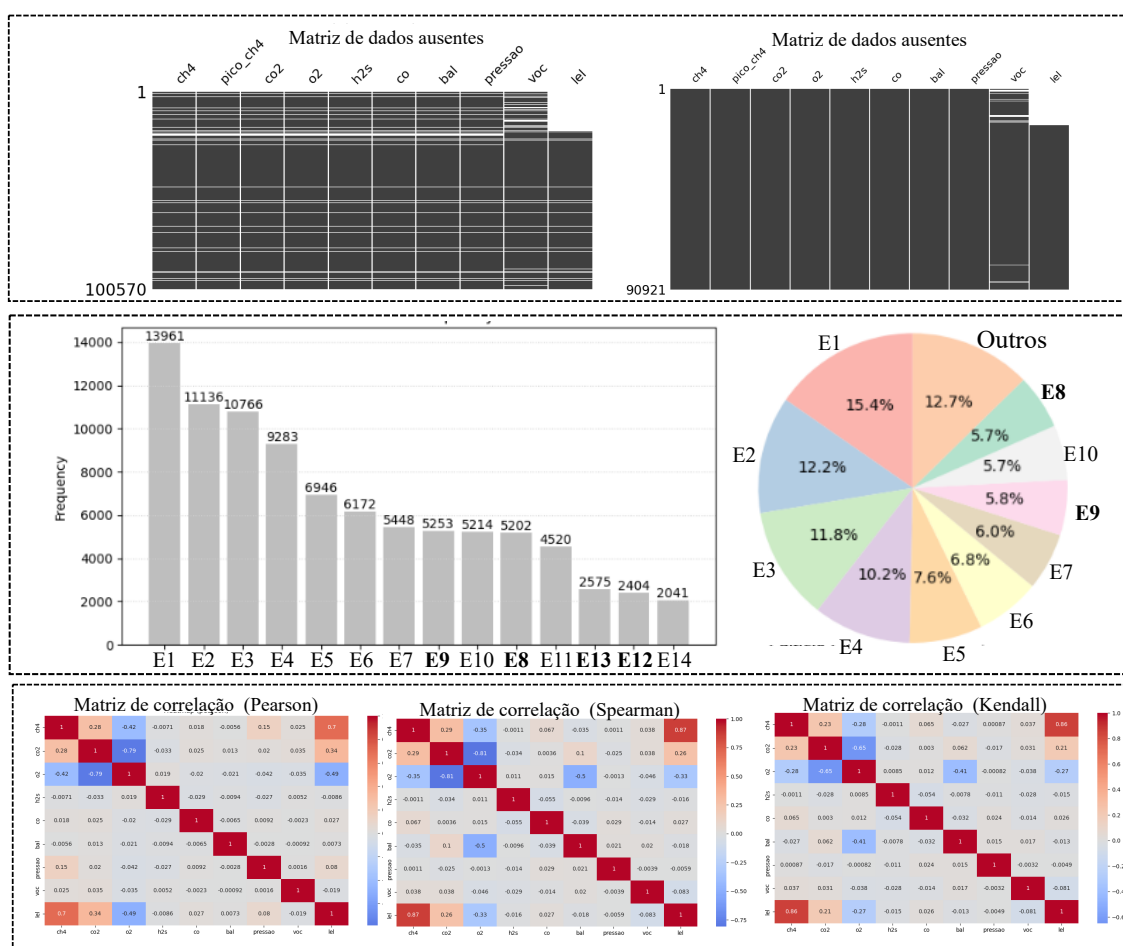


Figura 1. Matriz de correlação de variáveis e dados vazios.

frequências manteve a representatividade das principais edificações (E1 a E4), concentrando cerca de 50 % das observações. Observou-se uma forte correlação entre CH₄ e LEL, dado que a explosividade da mistura gasosa está diretamente relacionada à presença do CH₄, que atua como combustível. Também foram identificadas correlações moderadas entre CH₄, O₂ e CO₂, refletindo interações físico-químicas características do ambiente subterrâneo. Nesse meio, o biogás (mistura do CH₄ e CO₂) pode deslocar o oxigênio,

enquanto reações microbiológicas (em profundidades mais rasas), consomem CH_4 e O_2 , resultando no aumento da concentração de CO_2 .

4.2. Análise de Resultados

Os conjuntos com maior frequência observacional destacam-se o PMG-119 e PMG-120 como os mais monitorados ($\text{AF} = 1.022$ e 1.019 respectivamente). Essa alta granularidade espacial permite acompanhar a variabilidade da emissão de gases ao longo do tempo e em diferentes pontos de risco. A tipologia dos PMGs inclui poços rasos 'A' ($\sim 0,30$ m), profundos 'B' ($\sim 1,00$ m) e em múltiplas profundidades. A Tabela 2 apresenta uma resumo

Tabela 2. Distribuição das medições por tipo de PMG

Poço	AF	RF	%RF	CF	%CF
B	43.688	0,496.6	49,66 %	43.688	49,66 %
A	43.601	0,495.6	49,56 %	87.289	99,23 %
C	679	0,007.7	0,77 %	87.968	100,00 %

AF = Absolute Frequency; RF = Relative Frequency; CF = Cumulative Frequency.

A Tabela 2 apresenta a distribuição das 87.968 medições realizadas conforme a classificação dos PMGs em três categorias distintas: B, A e C. Os poços do tipo B correspondem a 49,66 % do total e são instalados em profundidades intermediárias, geralmente em torno de $\sim 1,00$ m abaixo da camada de brita. Esses poços são utilizados para monitorar a concentração de gases na zona não saturada do solo, possibilitando o acompanhamento da migração gasosa em camadas intermediárias. Os poços do tipo A, que representam 49,56 % das medições, são estruturas mais rasas, instaladas sob as lajes ou no tapete de brita ($\sim 0,30$ m de profundidade), e são empregados para avaliar a presença de gases próximos à interface com o ambiente interno das edificações, sendo essenciais para identificar riscos imediatos de intrusão dos gases. Já os poços do tipo C, responsáveis por 0,77 % das medições, são os mais profundos ($\sim 1,30$ m) e utilizados em locais com estrutura construtiva mais complexa, como a Portaria P3. Esses têm como principal finalidade a investigação de perfis verticais detalhados do solo, permitindo detectar acúmulos de gases em profundidade e antecipar seu potencial de ascensão. Essa diferenciação funcional entre os tipos de poço é fundamental para o diagnóstico ambiental e a prevenção de riscos à saúde e à infraestrutura.

Observa-se que os poços do tipo A e B concentram, juntos, mais de 99 % das medições (49,56 % e 49,66 %, respectivamente), refletindo o foco do plano de monitoramento em estruturas principais e edificações prioritárias. A baixa frequência relativa dos poços do tipo C (0,77 %) indica seu uso pontual e estratégico. Essa classificação permite avaliar o gradiente de gases ao longo da profundidade e auxilia na detecção de riscos de intrusão em diferentes camadas do solo. As concentrações médias de CH_4 , CO_2 , VOC e CO são baixas, com valores de mediana iguais a zero, indicando que a maior parte das medições não detectou presença significativa desses compostos. No entanto, os valores máximos revelam picos expressivos, como 78,8 ppm de CO e 72 ppm de VOC, sugerindo eventos pontuais que exigem atenção. O coeficiente de variação (CV) elevado para pressão (105,5), VOC (65,0) e CO (7,7) indica alta dispersão nos dados, o que pode estar relacionado a variações sazonais, falhas locais de ventilação ou presença intermitente de fontes emissoras. A presença de valores negativos em O_2 e CO_2 sugere interferência instrumental ou ruídos de leitura, que devem ser tratados na etapa de limpeza de dados.

A Figura 2 apresenta a evolução temporal de CH₄ nos poços das edificações com maior impacto. Observa-se que, embora em alguns pontos tenha ocorrido uma redução recente nos níveis do gás, persistem registros significativamente mais elevados em comparação com os demais edifícios. Nos gráficos (a), (c), (e) e (g), destacam-se picos históricos entre 4,5 % e 8,0 % de CH₄, valores que ultrapassam o limite inferior de inflamabilidade (5 %), conforme definido pela CETESB. Apesar da queda nas concentrações após a implementação dos sistemas de ventilação, essas áreas continuam apresentando níveis, considerando o valor de 5% como a concentração mínima de CH₄ no subsolo capaz de gerar alguma possibilidade de risco em casos de intrusão em construções, conforme o *framework* de Avaliação de Risco por Metano da ASTM E2993-16 (2016) [ASTM 2016].

Esse comportamento é coerente com os dados dos relatórios técnicos, que apontam a persistência de concentrações com potencial de risco em poços localizados nos conjuntos laboratoriais, com destaque para os pontos PMG-49B, PMG-53A/B e PMG-55B, nos quais foram registrados episódios com concentrações de até 39,6 % de CH₄. A recorrência de valores acima do referencial de 5%, mesmo após anos de operação dos sistemas de exaustão, reforça a necessidade de atenção prioritária a essas edificações, tanto na sequência dos monitoramentos periódicos, como na manutenção da operação dos sistemas de ventilação para mitigar potenciais cenários de intrusão e riscos aos usuários do campus. Essa constatação corrobora com as recomendações de continuidade do gerenciamento ambiental da área descritas nos relatórios técnicos das consultorias [SERVMAR 2014].

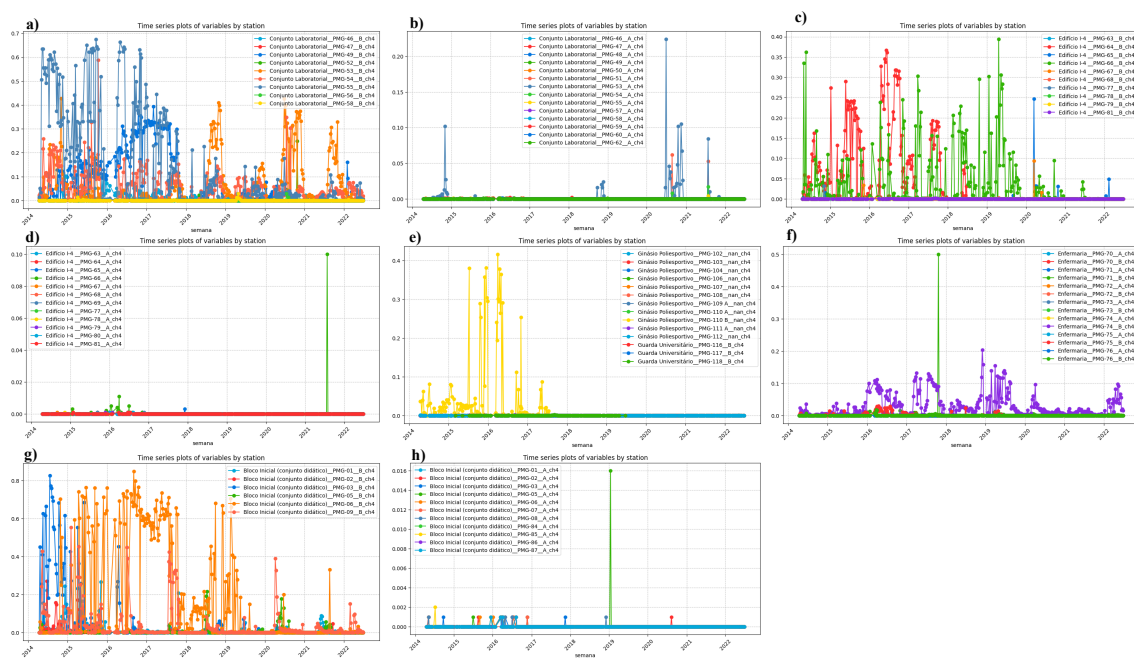


Figura 2. Evolução temporal da concentração de metano (CH₄)

5. Conclusões

As variáveis analisadas no *dataset* de monitoramento incluem gases com potencial risco à saúde humana e ao meio ambiente, destacam-se CH₄, H₂S e VOC como os principais contaminantes, devido ao seu potencial de inflamabilidade, toxicidade e intrusão de vapores. Observou-se alta frequência de medições nas edificações prioritárias (E1 a E4), com

concentrações que, em alguns pontos, ultrapassaram o limite inferior de explosividade do CH₄ em PMGs específicos. Apesar da atuação dos sistemas de ventilação, persistem episódios de risco potencial (concentrações de CH₄ > 5%), mas percebe-se claramente que os sistemas de engenharia instalados estão atenuando o problema ao longo do tempo nas edificações. A análise exploratória revelou correlações consistentes entre variáveis (CH₄ e LEL) e dispersão elevada em pressão, VOC e CO.

Agradecimentos

Os autores expressam seus agradecimentos ao apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processos 19/21693-0, 17/50343-2, e 17/17047-0.

Referências

- Anifowose, B. and Anifowose, F. (2024). Artificial intelligence and machine learning in environmental impact prediction for soil pollution management – case for eia process. *Environmental Advances*, 17:100554.
- Appelgren, A., Bergstrom, U., Brittain, J., Gallego, D., Hakanson, L., Monte, L., et al. (1996). An outline of a model-based expert system to identify optimal remedial strategies for restoring contaminated aquatic ecosystems: the project moira. Technical report, ENEA.
- Asante-Duah, K. (2019). *Management of contaminated site problems*. CRC Press.
- ASTM (2016). Standard guide for evaluating potential hazard as a result of methane in the vadose zone. ASTM Standard E2993–16. ASTM Committee E50 on Environmental Assessment, Risk Management and Corrective Action.
- Barbosa, M., Bertolo, R. A., and Hirata, R. (2017). A method for environmental data management applied to megasites in the state of sao paulo, brazil. *Journal of Water Resource and Protection*, 09(3):322–338.
- Bi, Z., Sun, J., Xie, Y., Gu, Y., Zhang, H., Zheng, B., Ou, R., Liu, G., Li, L., Peng, X., Gao, X., and Wei, N. (2024). Machine learning-driven source identification and ecological risk prediction of heavy metal pollution in cultivated soils. *Journal of Hazardous Materials*, 476:135109.
- Cao, L. (2016). Data science and analytics: a new era. *International Journal of Data Science and Analytics*, 1(1):1–2.
- CETESB (2020). Relação de Áreas contaminadas e reabilitadas no estado de são paulo. *Relatório Oficial - CETESB*. Dados conforme CETESB, acessado em novembro de 2024.
- CETESB (2022). Manual de gerenciamento de Áreas contaminadas. Technical report, Companhia Ambiental do Estado de São Paulo.
- Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press.
- Gouvêa Jr., J. C., Bertolo, R. A., and Hart, S. T. (2018). Intrusão de vapores do solo: breve histórico sobre desenvolvimento da tecnologia, cenário brasileiro e avanços recentes. *Holos Environment*, 18(2):240–270.

- Horst, J., Welty, N., Schnobrich, M., Sinha, P., and Kulkarni, P. (2017). Digital innovation: The next disruptive but transformative remediation frontier. *Groundwater Monitoring and Remediation*, 37(3):19–27.
- IPT (2011). Relatório técnico 123582-205/11 – final. avaliação de risco à saúde humana. gleba i da each/usp. Technical report, IPT.
- Iskamto, D. (2023). Data science: Trends and its role in various fields. *Adpebi International Journal of Multidisciplinary Sciences*, 2(2):165–172.
- Ismail, F. B., Xuan, A. T. Z., Rusilowati, U., and Williams, J. (2024). Exploring the frontier of data science: Innovations, challenges, and future directions. *International Transactions on Education Technology (ITEE)*, 2(2):163–172.
- Kumar, A., Kapil, A., and Ahlawat, D. (2023). Exploring the data science. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–7.
- Moraes, S., Teixeira, C., and Maximiano, A. (2014). *Guia de elaboração de planos de intervenção para o gerenciamento de áreas contaminadas*. IPT - Instituto de Pesquisas Tecnológicas do Estado de São Paulo.
- Palanisamy, G. (2016). Arm data file standards version 1.2. Technical report, DOE Office of Science Atmospheric Radiation Measurement (ARM) Program
- Quille, R. V. E. and de Almeida, F. V. (2025). Evaluation of machine learning regression techniques for analyzing contaminated soils. In *LatinX in AI Workshop at the International Conference on Machine Learning (ICML)*, Vancouver, Canada. Accepted for presentation.
- Quille, R. V. E., de Almeida, F. V., Ohara, M. Y., Corrêa, P. L. P., de Freitas, L. G., Alves-Souza, S. N., de Almeida, J. R., Davis, M., and Prakash, G. (2023). Architecture of a data portal for publishing and delivering open data for atmospheric measurement. *International Journal of Environmental Research and Public Health*, 20(7).
- Quille, R. V. E., dos Santos Luchetti Vieira, G., de Freitas, L. G., Corrêa, P. L. P., de Souza, S. N. A., Barbosa, A. M., and de Almeida, F. V. (2025). Dataset de poços de monitoramento de gases da usp leste. In *Anais do Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA 2025)*, Maceió/AL, Brasil. Sociedade Brasileira de Computação.
- SERVMAR (2014). Relatório de investigação detalhada, avaliação de risco à saúde humana e plano de intervenção na ai-01 e investigação detalhada de gases – ma/12936/14/bls. Technical report, SERVMAR.
- Thomas, J. J. and Cook, K. A. (2005). *Visual Analytics: The Scope and Challenges*, volume 3787 of *Lecture Notes in Computer Science*. Springer.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- USP (2004). Relatório ambiental preliminar: Campus usp leste. Technical report, USP.
- WEBER (2019). Relatório técnico: Evolução do monitoramento de intrusão de gases e da operação do sistema de ventilação – 2º trimestre/2019. projeto: 311.1264.14/e21vmgs-vs.02 – usp leste. Technical report, WEBER AMBIENTAL.