

Análise da malha aérea brasileira (2000–2025): evolução, recuperação pós-pandemia e conectividade com o turismo internacional

Ronald Rabelo dos Santos Filho
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
ronald.santos@dcc.ufmg.br

Orientador: João Guilherme Maia de Menezes
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
jgmm@dcc.ufmg.br

Resumo—This paper presents a longitudinal analysis of the Brazilian commercial aviation network from 2000 to 2025. Using ANAC microdata and foreign entry records from the Ministry of Tourism, the study investigates the relationship between flight supply and international visitor flows, testing the hypothesis that certain airports function as logistical hubs without meaningfully contributing to leisure tourism. The analysis also examines whether post-pandemic recovery was distributed across the network or concentrated in major hubs. To handle approximately 25 million raw records, a cloud-based ETL pipeline was built on Google BigQuery. Two original metrics are proposed: the Tourist Efficiency Index (ETE) and the Regional Recovery Index (IRR), both of which reveal structural imbalances that aggregate statistics tend to conceal.

Index Terms—Air Transportation; International Tourism; Big Data; Post-Pandemic Recovery; Connectivity.

Resumo—Este artigo apresenta uma análise longitudinal da aviação comercial brasileira entre 2000 e 2025. A partir de microdados da ANAC e registros de entrada de estrangeiros do Ministério do Turismo, investiga-se a relação entre oferta de voos e fluxo de visitantes internacionais, testando a hipótese de que certos aeroportos funcionam como polos logísticos sem contribuição significativa ao turismo de lazer. O estudo também avalia se a recuperação pós-pandemia foi distribuída na malha ou concentrada em grandes hubs. Para processar aproximadamente 25 milhões de registros brutos, foi construído um pipeline ETL em nuvem com Google BigQuery. São propostas duas métricas originais: o Índice de Eficiência Turística (ETE) e o Índice de Recuperação Regional (IRR), que revelam desequilíbrios estruturais que as estatísticas agregadas tendem a ocultar.

Palavras-chave—Transporte Aéreo; Turismo Internacional; Big Data; Recuperação Pós-Pandemia; Conectividade.

I. INTRODUÇÃO

Este trabalho parte da suspeita de que uma possível percepção comum de que mais voos equivale a mais turismo não se sustenta empiricamente no Brasil, e os dados dos últimos 25 anos permitem testá-la.

A questão tem relevância prática clara. O Brasil é o maior país da América do Sul, com infraestrutura aérea concentrada no eixo Rio-São Paulo, enquanto destinos com vocação turística consolidada, como destinos na região Nordeste, dependem de conexões indiretas para receber visitantes internacionais. A pergunta que orienta esta pesquisa é: a malha aérea brasileira serve ao turismo, ou serve principalmente à logística de cargas e ao trânsito corporativo?

Para investigar isso, cruzou-se a base de Voo Regular Ativo (VRA) da ANAC, com cerca de 25 milhões de registros operacionais entre 2000 e 2025, com os dados de chegadas de estrangeiros do Ministério do Turismo. O volume de dados exigiu uma infraestrutura robusta, foi feita a construção de um pipeline ETL em nuvem antes de qualquer análise estatística. Foram propostas duas métricas originais, ETE e IRR, para capturar dimensões que correlações globais não revelam.

Este artigo está estruturado da seguinte forma: a Seção II apresenta os fundamentos técnicos e econômicos que embasam as escolhas metodológicas; a Seção III descreve o processo de coleta, limpeza e integração dos dados; a Seção IV discute os resultados; e a Seção V apresenta as conclusões e limitações do estudo.

II. FUNDAMENTAÇÃO TEÓRICA E TECNOLÓGICA

A. ETL em Ambientes de Nuvem

O tratamento de grandes volumes de dados governamentais exige pipelines de Extração, Transformação e Carga (ETL) que sejam auditáveis e escaláveis. A migração desses processos para ambientes *cloud-native* consolidou-se na última década como prática padrão em projetos analíticos [3]. Uma aplicação recente e metodologicamente próxima deste trabalho é o pipeline construído para dados do SUS, que demonstrou desempenho estável no processamento massivo de registros de saúde pública no BigQuery [4], contexto comparável ao dos dados da ANAC em termos de volume e fragmentação.

B. Qualidade de Dados em Bases Abertas

Trabalhar com dados governamentais abertos impõe problemas que muitas vezes só aparecem durante a análise

exploratória. Quebras metodológicas em séries históricas, inconsistências de *encoding* e ausência de dicionários de dados atualizados são recorrentes. A adoção de rotinas sistemáticas de validação programática reduz esse tipo de problema antes que ele comprometa os modelos [5]. No caso deste trabalho, essa etapa revelou uma anomalia relevante nos dados de 2010, descrita na Seção IV.

C. Infraestrutura Aeroportuária e Turismo

A relação entre capacidade aeroportuária e turismo não é direta. A expansão física de terminais viabiliza o tráfego, mas não o explica por si só [6]. Análises focadas no Brasil mostram que o país apresenta dificuldade estrutural em converter vantagens logísticas em desempenho turístico [7], o que motivou a criação de uma métrica de conversão específica para este estudo.

III. METODOLOGIA

A. Coleta e Ingestão dos Dados

As duas fontes primárias foram: os registros de operações aéreas da ANAC (base VRA, cerca de 25 milhões de linhas) e os dados de chegadas de turistas internacionais do Ministério do Turismo [1], [2]. Ambas estão disponíveis como dados abertos, mas distribuídas em centenas de arquivos separados por período.

Para automatizar o download, foi desenvolvido um script em Python com `concurrent.futures.ThreadPoolExecutor`, permitindo a extração paralela dos arquivos e reduzindo significativamente o tempo de coleta em comparação com uma abordagem sequencial.

B. Tratamento e Padronização

A etapa de transformação foi a mais trabalhosa e revelou problemas que não estavam documentados nas fontes:

- **Dados do MTur:** A unificação exigiu normalização de cabeçalhos (variações ortográficas entre anos distintos), remoção de acentuação via `unicodedata` e eliminação de duplicatas originadas por arquivos sobrepostos.
- **Dados da ANAC (VRA):** O principal desafio foi o tamanho dos arquivos: carregá-los diretamente causava esgotamento de memória. A solução foi ler os arquivos em *chunks* iterativos. Durante esse processo, foram encontrados arquivos com caracteres `latin1` encapsulados em `utf-8`, além de valores textuais representando nulos (como a string literal "null"), que precisaram ser detectados e corrigidos antes da carga.
- **Dados:** Os registros usavam o formato brasileiro (DD/MM/YYYY HH:MM), que foi convertido para ISO 8601 para compatibilidade com o BigQuery. Uma etapa de *data profiling* verificou a integridade tipológica das colunas antes da carga final.

C. Armazenamento e Análise

Após a limpeza, os dados foram consolidados em um bucket S3 (staging) e transferidos ao BigQuery via Airbyte. As análises estatísticas foram feitas em Python (bibliotecas Pandas e SciPy) integrado ao Google Colab, com visualizações em Matplotlib e Seaborn.

D. Métricas Propostas

Para responder à pergunta central do trabalho, foram definidas duas métricas originais:

Eficiência Turística por Estado (ETE): razão entre o total de turistas internacionais recebidos por um estado e o total de voos comerciais realizados no mesmo período. A ideia é isolar quanto de tráfego aéreo se converte efetivamente em visitantes.

$$ETE_{uf} = \frac{\text{Total de Turistas Internacionais}_{uf}}{\text{Total de Voos Comerciais Realizados}_{uf}}$$

Índice de Recuperação Regional (IRR): razão entre o valor de uma variável em 2022 e seu valor em 2019 (ano de referência pré-pandemia), expressa em percentual. Calculado separadamente para voos e para turistas, permite comparar a velocidade de retomada de cada dimensão.

IV. RESULTADOS E DISCUSSÕES

A. Correlação Global e seus Limites

A correlação de Pearson entre volume de voos e chegadas de turistas internacionais foi de **0,77** (p-valor: $1,25 \times 10^{-3}$), com Spearman em **0,61**. A relação existe, mas essa visão agregada esconde o que é mais interessante nos dados.

B. A Anomalia de 2010 e a Mitigação Estatística

Durante a análise exploratória, um salto abrupto no volume absoluto de voos a partir de 2010 chamou atenção. O crescimento era vertical demais para refletir qualquer fenômeno econômico real. Investigando as notas metodológicas da ANAC, constatou-se que a base VRA passou a incorporar operações de aviação fretada e de carga naquele ano, uma mudança de escopo que não estava destacada na documentação principal.

Em vez de manipular ou excluir arbitrariamente os registros brutos da fonte governamental para forçar um alinhamento, a estratégia para mitigar esse viés foi puramente analítica. Para evitar que os modelos fossem distorcidos pela inclusão massiva de voos de carga, o trabalho apoiou-se na correlação de postos de Spearman (que avalia o ordenamento das variáveis e é imune a saltos absolutos anômalos) e na formulação da métrica ETE, que atua como um normalizador matemático intrínseco. Esse passo metodológico garantiu que a infraestrutura logística fosse avaliada sem comprometer a integridade da série histórica original.

C. ETE: Logística vs. Lazer por Estado

A métrica ETE separou com clareza dois perfis de estado. O **Amazonas** apresentou ETE de **1.43**: alto volume de operações aéreas, sustentado pela Zona Franca de Manaus e pelo abastecimento de cargas na região Norte, mas conversão mínima em turismo internacional. A **Bahia**, com ETE de **2.89**, representa o oposto: um estado que aproveita melhor sua infraestrutura aérea para fins turísticos.

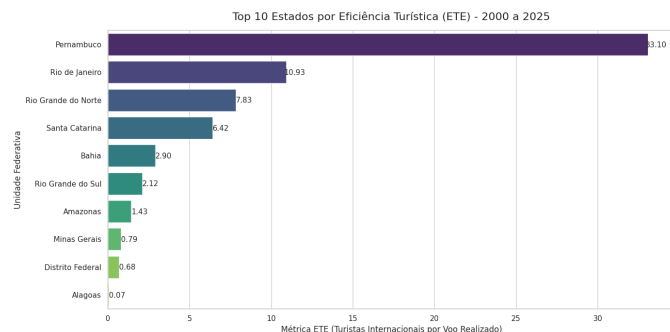


Figura 1. Top 10 Estados Brasileiros por Eficiência Turística (2000-2025).

O resultado mais expressivo foi o de **Pernambuco**, com ETE de **33.10**. Um índice desse nível indica que o estado recebe um volume de turistas internacionais desproporcional à sua malha de voos diretos. A interpretação é de demanda reprimida: visitantes chegam a Recife por rotas com escalas em outros hubs, o que encarece e alonga a viagem. A conectividade direta de longo curso simplesmente não existe na proporção que a demanda indicaria.

D. Pandemia e Velocidade de Recuperação

O impacto da COVID-19 nos dados foi assimétrico. O total de voos recuou **52%** entre 2019 e 2020 sustentado, em parte, pela continuidade das operações de carga. O turismo internacional, sem essa âncora, caiu **66%**.

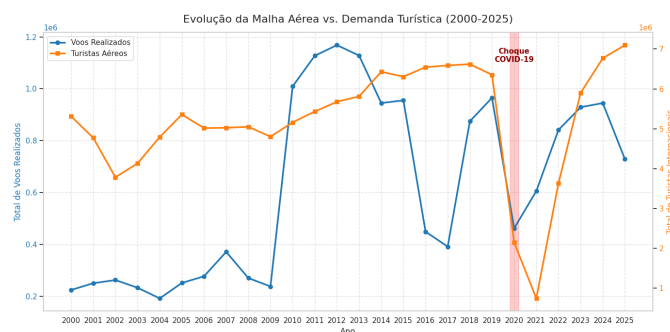


Figura 2. Evolução Histórica de Voos Realizados vs. Turistas Internacionais (Eixos Duplos).

A recuperação em 2022 também foi assimétrica. O IRR dos voos chegou a **91,45%**, sinalizando retomada quase completa da infraestrutura operacional. O IRR do turismo ficou em **55,96%**, a demanda voltou muito mais devagar. O CAGR do turismo entre 2021 e o período mais recente foi de **75,64%**

ao ano, contra **4,73%** da infraestrutura de voos. A oferta está crescendo em ritmo muito inferior ao da demanda, o que pressiona a capacidade instalada principalmente nos estados com ETE alto.

V. CONCLUSÕES

Este trabalho partiu da hipótese de que volume de voos e turismo não são equivalentes, e os dados confirmaram essa premissa. A métrica ETE conseguiu separar estados que funcionam como polos logísticos daqueles com vocação turística real. O caso de Pernambuco ilustra bem o problema prático: há demanda consolidada, mas a malha aérea direta não acompanha.

A combinação entre a engenharia de dados e a modelagem analítica foi a base do trabalho, não apenas um meio para um fim. Sem a correção estrutural das inconsistências de *encoding* durante o pipeline ETL e sem a mitigação estatística da quebra metodológica de 2010 (através da métrica ETE e da correlação de Spearman), a avaliação do cenário brasileiro teria sido severamente enviesada.

Algumas limitações merecem menção. Os dados do MTur são estimativas, e há razões para suspeitar de subnotificação em estados com menor infraestrutura de controle de fronteiras. Além disso, a análise é agregada por estado, não por aeroporto, uma granularidade maior permitiria identificar terminais específicos que concentram o gargalo, sendo esta a extensão natural mais imediata do estudo.

Por fim, os números do IRR sugerem que o descompasso entre oferta e demanda deve se agravar antes de melhorar. Rotas internacionais diretas para o Nordeste não aparecem nos dados como uma conveniência mercadológica, mas despontam como uma lacuna estrutural com impacto econômico mensurável.

REFERÊNCIAS

- [1] Agência Nacional de Aviação Civil (ANAC). "Index of /Dadosabertos/Voos E Operações Aéreas/Voo Regular Ativo (VRA)/". 2021.
- [2] Ministério do Turismo. "Estimativas de Chegadas de Turistas Internacionais ao Brasil". Portal de Dados Abertos, 2024. [Online].
- [3] "The Role of ETL (Extract-Transform-Load) Pipelines in Scalable Business Intelligence: A Comparative Study of Data Integration Tools," *Global ASRC Conference*. [Online].
- [4] A. M. Shimaoka et al., "Scalable ETL Pipeline for Health Data Ingestion: Application to the Brazilian Unified Health System (SUS)," *Asklepion Informação em Saúde*, vol. 5, no. 1, Apr. 2026.
- [5] V. Jain, "Scalable Microservice-Based Data Quality Framework Using Great Expectations and BigQuery on Google Kubernetes Engine," *Int. Journal of Science and Research (IJSR)*, vol. 14, no. 5, May 2025.
- [6] M. A. R. Demant, "Infraestrutura aeroportuária e o desenvolvimento do tráfego aéreo regional no Brasil," *ResearchGate*, Jan. 2011.
- [7] V. A. Fernandes, R. R. Pacheco, and E. Fernandes, "A dynamic analysis of air transport and tourism in Brazil," *Journal of Air Transport Management*, vol. 105, p. 102297, Sep. 2022.