

DespesasGovPB: um conjunto de dados estruturado sobre despesas públicas do Governo da Paraíba

Ricardo Anastácio de Souza¹, Damires Yluska de Souza Fernandes¹,
Alex Sandro da Cunha Rego¹

¹ Instituto Federal da Paraíba (IFPB) – João Pessoa – PB – Brasil

anastacio.ricardo@academico.ifpb.edu.br

damires@ifpb.edu.br, alex@ifpb.edu.br

Abstract. *This paper presents DespesasGovPB, an open dataset on the budget execution of the Paraíba State Government from 2015 to 2025. It consolidates 2,700,835 records on commitment, liquidation, payment, and outstanding payables in a reusable analytical structure. The dataset includes derived variables for analyzing time intervals between expenditure stages, fiscal seasonality, supplier concentration, electoral context, and atypical records. Built through extraction, integration, standardization, enrichment, and validation procedures, it follows FAIR and open-data principles. Published in CSV and Parquet, with notebooks and documentation, it can support auditing, social control, fiscal efficiency studies, and data science applications in public management.*

Resumo. *Este artigo apresenta o DespesasGovPB, um dataset aberto sobre a execução orçamentária do Governo da Paraíba entre 2015 e 2025. A base consolida 2.700.835 registros de empenho, liquidação, pagamento e restos a pagar em uma estrutura analítica reutilizável. O conjunto inclui variáveis derivadas para analisar prazos entre fases da despesa, sazonalidade fiscal, concentração de fornecedores, contexto eleitoral e registros atípicos. Construído por meio de atividades de extração, integração, padronização, enriquecimento e validação, segue princípios FAIR e de dados abertos. Disponibilizado em CSV e Parquet, com notebooks e documentação, pode apoiar auditorias, controle social, estudos de eficiência fiscal e aplicações de ciência de dados na gestão pública.*

1. Introdução

A transparência pública depende de dados acessíveis e bem documentados para sustentar a *accountability*, entendida como a obrigação de agentes públicos prestarem contas e justificarem suas decisões perante instâncias de controle e a sociedade [Bovens 2007, Brelàz et al. 2025]. No Brasil, a Lei de Acesso à Informação – LAI (Lei 12.527/2011) [Brasil 2011] ampliou a disponibilização de informações governamentais, especialmente no campo orçamentário e financeiro. Ainda assim, persistem lacunas de estruturação, documentação, interoperabilidade e qualidade dos dados, o que dificulta análises avançadas e reproduzíveis [Brelàz et al. 2025, Godoy et al. 2025, Pipino et al. 2002].

O Portal da Transparência do Governo da Paraíba disponibiliza dados sobre as fases da execução da despesa pública, como empenho, liquidação, pagamento e restos a

pagar [[Governo do Estado da Paraíba 2026](#)]. Essas informações, porém, são publicadas em arquivos separados e usam estruturas e nomes de campos nem sempre equivalentes. Por isso, é necessário integrar os arquivos de cada fase da despesa e padronizar seus campos antes de calcular prazos, comparar valores e reconstruir a trajetória da execução orçamentária.

Nesse contexto, o *DespesasGovPB* foi desenvolvido para transformar registros administrativos dispersos em uma base analítica integrada da execução orçamentária estadual. Com ela, é possível investigar a concentração de pagamentos por órgão ou fornecedor, os intervalos entre empenho, liquidação e pagamento, padrões sazonais do exercício fiscal ou do calendário eleitoral e comportamentos financeiros ou temporais atípicos. O *dataset* incorpora variáveis derivadas para essas análises, seguindo normativos como o Manual Técnico do Orçamento (MTO), o Manual de Contabilidade Aplicada ao Setor Público (MCASP) e a Classificação das Funções do Governo (CO-FOG) [[Brasil. Secretaria de Orçamento Federal \(SOF\) 2025](#), [Tesouro Nacional 2025](#), [United Nations Statistics Division 2016](#)]. Sua modelagem adota os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) [[Wilkinson et al. 2016](#)], interoperabilidade semântica, granularidade longitudinal entre empenhos e pagamentos, e práticas de dados abertos, com documentação pública e formatos reutilizáveis, para favorecer integração e uso analítico.

O trabalho apresenta três contribuições: (i) um *dataset* aberto e documentado da execução orçamentária da Paraíba entre 2015 e 2025; (ii) uma metodologia de curadoria, integração, padronização, enriquecimento e validação do conjunto de dados; e (iii) artefatos de uso, incluindo arquivos em CSV e *Parquet*, *notebooks* e dicionário de dados.

O artigo está organizado da seguinte forma: a Seção 2 apresenta fundamentos e trabalhos correlatos; a Seção 3 descreve a construção do *dataset* e as condições de reprodutibilidade do *pipeline* e dos artefatos publicados; a Seção 4 discute resultados analíticos; e a Seção 5 reúne considerações finais, limitações e trabalhos futuros.

2. Referencial Teórico

Esta seção reúne os fundamentos teóricos e o contexto dos dados que sustentam o *DespesasGovPB*. Para isso, aborda princípios de dados abertos, FAIR e *accountability*, apresenta o ciclo de execução da despesa pública e discute trabalhos correlatos na área.

2.1. Dados abertos, FAIR e accountability

A transparência fiscal exige compreensibilidade, documentação e interoperabilidade para que registros administrativos se convertam em informação reutilizável [[Brelàz et al. 2025](#)]. Nesse sentido, publicar dados não basta: é preciso que eles permitam conferir decisões, auditar registros e reutilizar informações em novas análises. Princípios e diretrizes como FAIR, as boas práticas de dados na Web propostas pelo World Wide Web Consortium (W3C) e os princípios da *Open Data Charter* convergem ao recomendar metadados estruturados, identificadores persistentes, formatos processáveis por máquina, abertura por padrão, temporalidade, interoperabilidade, documentação adequada e licenças explícitas [[Wilkinson et al. 2016](#), [World Wide Web Consortium \(W3C\) 2017](#), [Open Data Charter 2015](#)]. A ciência de dados se beneficia desse processo ao extrair padrões e identificar anomalias em grandes volumes financeiros [[Godoy et al. 2025](#)].

2.2. Execução da despesa pública

A execução da despesa pública corresponde à materialização dos compromissos previstos na Lei Orçamentária Anual. De acordo com o MTO e o MCASP, esse ciclo compreende três fases centrais, podendo ainda gerar restos a pagar quando a obrigação financeira ultrapassa o exercício fiscal [Brasil. Secretaria de Orçamento Federal (SOF) 2025, Tesouro Nacional 2025]. O fluxo ilustrado na Figura 1 evidencia a passagem do compromisso orçamentário ao pagamento ou à inscrição em restos a pagar.



Figura 1. Processo da execução da despesa pública.

O processo inicia-se com o empenho, ato administrativo que reserva o recurso orçamentário para um fim específico e formaliza o compromisso de gasto do ente público. A liquidação consiste na verificação do direito do credor, confirmando a entrega do bem ou a prestação do serviço. O pagamento, por sua vez, efetiva a transferência financeira. Caso a obrigação de pagamento persista para o exercício seguinte, o registro é tratado como resto a pagar.

2.3. Trabalhos correlatos

Na literatura nacional, ainda são pouco frequentes *datasets* abertos voltados à execução da despesa pública em nível transacional. Os trabalhos mais próximos do *DespesasGovPB* podem ser organizados em dois grupos. O primeiro reúne bases relacionadas a etapas anteriores da despesa, especialmente contratações e compras públicas. Nesse grupo, o *LiP-Set* [Silva et al. 2022] organiza documentos de licitações públicas, favorecendo estudos sobre conteúdo textual, classificação documental e recuperação de informação, enquanto o *ICPSet* [Oliveira et al. 2024] estrutura itens de compras públicas com foco semântico, isto é, representa e classifica produtos e serviços para facilitar a comparação e a padronização dos itens adquiridos. Embora relevantes, essas iniciativas se concentram em fases anteriores à execução financeira da despesa.

O segundo grupo reúne bases de dados públicos em nível agregado. Nessa linha, um banco de dados municipal [Davis 2022] consolida despesas por função e indicadores de educação, saúde, segurança pública, desenvolvimento, saneamento e situação fiscal, enquanto o *BrStats* [Toledo et al. 2023] reúne estatísticas municipais sobre população, economia, trabalho, educação e saúde. Essas bases ampliam o acesso a informações públicas, mas operam com indicadores agregados por município ou área temática. Sendo assim, diferem do propósito deste trabalho, que é preservar a trajetória da despesa pública em registros transacionais, do empenho ao pagamento ou à permanência como obrigação pendente.

O *DespesasGovPB* reúne, em uma única base analítica, diferentes fases da execução orçamentária do Governo da Paraíba entre 2015 e 2025. Diferentemente das bases

voltadas à contratação ou a estatísticas agregadas, o *dataset* preserva a trajetória da despesa desde o empenho até o pagamento ou sua permanência como obrigação pendente. Com isso, viabiliza análises sobre tempo de execução, concentração de fornecedores, sazonalidade fiscal e registros atípicos.

3. Método da Pesquisa

O presente trabalho caracteriza-se como pesquisa aplicada, descritiva e quantitativa, voltada à construção, curadoria e disponibilização de um conjunto de dados abertos no âmbito da Companhia de Processamento de Dados da Paraíba (CODATA), empresa vinculada ao Governo da Paraíba. A construção do *DespesasGovPB* envolveu etapas de extração, integração, padronização, enriquecimento e validação de dados de execução orçamentária, com foco em uso analítico, rastreabilidade e conformidade legal.

Como a base inclui identificação de credores, seu processamento exigiu tratamento distinto para pessoas físicas e jurídicas. Na versão pública, os registros de credores pessoa física (PF) foram tratados por minimização, com retirada de CPF e nome, e por pseudonimização, com substituição do CPF por um identificador artificial estável. Essa estratégia reduz a exposição de identificadores diretos e preserva o vínculo analítico entre registros do mesmo credor, conforme a Lei Geral de Proteção de Dados Pessoais – LGPD (Lei nº 13.709/2018) e a literatura sobre privacidade e pseudonimização [Brasil 2018, Sweeney 2002, Pfitzmann and Hansen 2010]. Para pessoas jurídicas (PJ), o CNPJ é mantido nos casos compatíveis com a Lei de Acesso à Informação – LAI (Lei 12.527/2011) [Brasil 2011]. Os detalhes desse tratamento são apresentados na Subseção 3.2.

3.1. Fontes e escopo dos dados

O *DespesasGovPB* cobre o período de 2015 a 2025 e utiliza como fontes primárias os registros de empenho, liquidação, pagamento e restos a pagar do Portal da Transparência do Governo da Paraíba [Governo do Estado da Paraíba 2026]. Esses registros constituem os insumos brutos e abrangem as despesas orçamentárias registradas nessas fases, incluindo custeio, capital, pessoal, encargos sociais, previdência e transferências pagas pelo Estado. A base não contempla receita orçamentária. Como fonte complementar, utiliza informações cadastrais de fornecedores pessoa jurídica obtidas por consulta institucional de CNPJ (SERPRO/Receita Federal) [Receita Federal do Brasil / SERPRO 2026].

3.2. Metodologia de construção do *dataset*

O *DespesasGovPB* foi construído por meio de um *pipeline* de ELT (*Extract, Load, Transform*) em cinco etapas sequenciais, apresentadas na Figura 2. A classificação como ELT é adotada porque os insumos brutos são primeiro extraídos e carregados em um Sistema de Gerenciamento de Banco de Dados Relacional (DuckDB), preservando o *snapshot* inicial, e as transformações de integração, limpeza e criação de variáveis derivadas ocorrem após essa carga.

A **etapa 1 (extração)** coleta os insumos brutos: arquivos do Portal da Transparência [Governo do Estado da Paraíba 2026] e dados cadastrais de fornecedores obtidos por consulta institucional de CNPJ (SERPRO/Receita Federal)



Figura 2. Pipeline de construção do *DespesasGovPB*.

[[Receita Federal do Brasil / SERPRO 2026](#)]. Nessa etapa, eles ainda não são integrados nem enriquecidos; apenas são carregados e organizados em um banco de dados DuckDB. A **etapa 2 (integração e enriquecimento)** relaciona empenhos, liquidações, pagamentos e restos a pagar por chaves dos arquivos-fonte (exercício, número de empenho, unidade orçamentária, credor e data de pagamento), associando credores PJ aos dados cadastrais de CNPJ. Quando há múltiplas liquidações ou pagamentos para um mesmo empenho, os valores são acumulados, e as datas registram o evento mais recente de cada fase, evitando duplicidades. Por exemplo, um empenho associado a um fornecedor PJ pelo campo `cnpj_credor` passa a incorporar atributos cadastrais como `razao_social`, `cnae_principal`, `porte_empresa`, `municipio_empresa` e `estado_empresa`, permitindo analisar o gasto público também por perfil econômico e localização do fornecedor.

A **etapa 3 (padronização e limpeza)** trata nulos, ajusta tipos de dados e converte datas para o formato ISO 8601 (AAAA-MM-DD). Classificações orçamentárias são padronizadas conforme MTO e MCASP [[Brasil. Secretaria de Orçamento Federal \(SOF\) 2025](#), [Tesouro Nacional 2025](#)]; campos textuais recebem normalização de escrita e variações de grafia da mesma entidade são corrigidas. Na **etapa 4 (geração de variáveis derivadas)**, são produzidos indicadores por empenho, como intervalos entre fases, mês e trimestre do pagamento, marcadores eleitorais, marcadores de inconsistência e variáveis de agregação por fornecedor usadas para medir a concentração de valores. Entre elas estão o índice de Gini, usado aqui para medir a concentração dos valores liquidados entre credores, e a distribuição acumulada por credor.

Por fim, a **etapa 5 (validação da estrutura e da consistência)** verifica a integridade estrutural da base e sua aderência aos arquivos de origem, considerando chaves, completude de campos e correspondência entre registros. Foram adotados quatro procedimentos: unicidade pela chave composta (exercício, número de empenho, unidade orçamentária, credor e data de pagamento), mensuração da completude de campos pela razão entre valores preenchidos e total de registros (razão 1,0 indica ausência de nulos), conferência manual de uma amostra com os arquivos-fonte e triagem de anomalias.

Para identificação de valores anômalos, aplicou-se o *Isolation Forest* [[Liu et al. 2008](#)], método não supervisionado para identificar observações incomuns em bases sem rótulo prévio. O modelo usou quatro atributos financeiros

(valor_empenhado, valor_liquidado, valor_pago e saldo_apagar) e dois intervalos temporais em dias, padronizados com *StandardScaler*. A execução adotou `n_estimators=200`, `contamination=0.01` e `random_state=42`; o parâmetro de contaminação define uma triagem conservadora de 1% da base, não uma taxa observada de erro ou irregularidade. Na sensibilidade, `contamination=0.005`, `0.01` e `0.02` selecionaram 13.504, 27.009 e 54.017 registros, com inclusão aninhada entre os conjuntos.

Ao final do processo de extração, integração, limpeza, enriquecimento e validação, o conjunto consolidado passou a compreender 2.700.835 registros e 61 variáveis. A versão pública oficial foi depositada no Zenodo, sob DOI <https://doi.org/10.5281/zenodo.19502820> [Anastácio de Souza et al. 2026], com o arquivo `despesasgovpb.parquet` disponível diretamente, o arquivo `despesasgovpb.csv` empacotado em `despesasgovpb.zip`, o dicionário completo de dados (`dicionario_de_dados.md`), seis *notebooks* Jupyter, `requirements.txt` e licença CC BY 4.0 [Creative Commons 2013]. As análises deste artigo referem-se exclusivamente a esse *snapshot*.

3.3. Dicionário de dados

Ao término da etapa de pré-processamento e limpeza, o *DespesasGovPB* consolidado é composto por 61 variáveis: 42 provenientes do Portal da Transparência, 8 originárias do enriquecimento cadastral via API de CNPJ e 11 geradas por criação de variáveis derivadas. A Figura 3 apresenta um fragmento real do arquivo CSV, evidenciando a granularidade da base e a coexistência de campos orçamentários, temporais e financeiros. A Tabela 1 exibe um subconjunto de variáveis selecionadas para ilustrar auditorias, análises temporais e estudos de eficiência fiscal. O dicionário completo das 61 variáveis descreve nome técnico, tipo, percentual de preenchimento, descrição, valores possíveis, exemplos e proveniência, quando aplicável; sua disponibilidade pública é abordada na Subseção 3.4.

Exercício	Unidade	Programa	Credor	Data empenho	Valor empenhado	Data pagamento	Valor pago
2025	FUNDO ESPECIAL DO PODER J	PROGRAMA DE GEST	ROBERTO CLAUDINO DA SILVA	2025-03-25	R\$ 234.09	2025-04-07	R\$ 234.09
2025	SECRETARIA DE ESTADO DO D	ASSISTÊNCIA SOCI	LUANDERSON VAGNER PEREIRA	2025-08-15	R\$ 499.20	2025-08-18	R\$ 499.20
2025	SECRETARIA DE ESTADO DA E	EDUCAÇÃO PARA CR	CONSELHO ESCOLAR EEPG DOM	2025-03-17	R\$ 1.417.00	2025-03-17	R\$ 1.417.00
2025	SECRETARIA DE ESTADO DA E	EDUCAÇÃO PARA CR	AFONSO HENRIQUE PATRICIO	2025-07-23	R\$ 48.000.00	2025-07-25	R\$ 48.000.00

Figura 3. Fragmento do *DespesasGovPB*.

Tabela 1. Subconjunto de variáveis do *DespesasGovPB*.

Variável	Descrição	Tipo	Exemplo
<code>exercicio</code>	Ano de exercício orçamentário.	int	2023
<code>nome_unidade_orcamentaria</code>	Unidade orçamentária responsável pelo gasto.	string	SEC. DE ESTADO DA SAÚDE
<code>nome_programa</code>	Programa orçamentário associado à despesa.	string	SAÚDE INTEGRAL
<code>nome_acao</code>	Ação orçamentária vinculada à despesa.	string	MANUTENÇÃO HOSPITAL REGIONAL
<code>nome_natureza_despesa</code>	Classificação econômica da despesa.	string	MATERIAL DE CONSUMO
<code>nome_fonte_recurso</code>	Fonte financeira utilizada.	string	REC. VINCULADOS AO FES

Tabela 1 – continuação

Variável	Descrição	Tipo	Exemplo
nome_credor	Nome do credor PJ; substituído pelo id_credor_pf para PF na versão pública.	string	DISTRIBUIDORA ABC LTDA
tipo_credor	Tipo do credor: pessoa jurídica (PJ), pessoa física (PF) ou outros.	string	PESSOA JURIDICA
cnpj_credor	CNPJ do credor PJ; nulo para PF e outros na versão pública.	string	12345678000190
id_credor_pf	Identificador pseudônimo do credor PF (HMAC-SHA256 com sal, prefixo PF_); nulo para PJ. A identificação integral de PF encontra-se disponível apenas na fonte governamental originária.	string	PF_a3f2c1b8d5e7f
data_empenho	Data de emissão do empenho.	date	2024-12-10
valor_empenhado	Valor inicialmente comprometido.	float	100,00
data_liquidacao	Data de liquidação da despesa.	date	2024-12-12
valor_liquidado	Valor confirmado após a execução do serviço ou entrega do bem.	float	50,00
data_pagamento	Data do pagamento efetivo.	date	2024-12-20
valor_pago	Valor efetivamente pago.	float	40,00
saldo_apagar	Diferença entre o valor empenhado e o valor pago ainda registrada como saldo a pagar.	float	60,00
municipio_empresa	Município sede do fornecedor.	string	JOÃO PESSOA
tempo_liquidacao	Faixa de tempo entre empenho e liquidação.	string	1–30 DIAS
tempo_pagamento	Faixa de tempo entre liquidação e pagamento.	string	1–30 DIAS
flag_datas_inconsistentes	Indica datas fora da ordem cronológica esperada.	binary	0

3.4. Ferramentas e reprodutibilidade

As etapas de construção e análise do *DespesasGovPB* foram executadas em Python 3.11, com apoio das bibliotecas *Pandas*, *NumPy*, *DuckDB*, *scikit-learn* e *Requests*. Esta subseção delimita a reprodução possível a partir dos artefatos depositados no Zenodo (Subseção 3.2).

Os seis *notebooks* registram extração, enriquecimento cadastral via CNPJ, integração das fases da despesa, limpeza, análise exploratória e detecção de registros atípicos. Desse modo, documentam decisões de curadoria e os procedimentos analíticos retomados na Seção 4.

A reprodutibilidade do processo é parcial. Os *notebooks* permitem reproduzir as análises do artigo a partir do *snapshot* público, usando o *Parquet* e o CSV empacotado. Já a extração bruta depende do *lakehouse* institucional da CODATA, e o enriquecimento cadastral requer acesso ao SERPRO/Receita Federal. Por isso, a versão pública já incorpora esses dados, permitindo reanálise sem nova coleta. A pseudonimização de credores pessoa física também não é integralmente reproduzível por razões de privacidade.

4. Resultados e Discussão

Esta seção apresenta evidências do que o *DespesasGovPB* permite observar sobre a execução orçamentária estadual. Primeiro, descreve atributos relevantes para uso analítico, como completude de campos, cobertura do enriquecimento cadastral, consistência estrutural e sinalizações para revisão. Em seguida, discute padrões temporais e concentração

de valores por fornecedor.

4.1. Condições da base para uso analítico

Antes da análise substantiva, é necessário descrever características do *DespesasGovPB* que afetam seu uso analítico. Consideram-se quatro dimensões: completude de campos, cobertura do enriquecimento cadastral, consistência estrutural e sinalização para revisão, pois elas condicionam análises que dependem de datas de liquidação, datas de pagamento e informações cadastrais de fornecedores. A Tabela 2 sintetiza esses indicadores.

Tabela 2. Condições do *DespesasGovPB* para uso analítico.

Grupo de variáveis	Dimensão	Valor	Observação
<i>Completude de campos</i>			
Chaves orçamentárias	Completude	100%	Exercício, unidade, programa, ação, natureza e fonte
Valores financeiros	Completude	100%	Zeros indicam ausência de fase registrada, não valor ausente
Data de empenho	Completude	100%	Campo obrigatório na fonte de origem
Data de liquidação	Completude	89,27%	10,73% de registros sem liquidação até o recorte analisado
Data de pagamento	Completude	87,77%	12,23% de registros sem pagamento até o recorte analisado
Identificadores de credor	Completude	99,92%	CNPJ para PJ e identificador pseudonimizado para PF
<i>Enriquecimento cadastral</i>			
Enriquecimento via CNPJ	Cobertura cadastral	>99%	Entre credores PJ, a cobertura cadastral supera 99%; cerca de 40% dos registros da base correspondem a pessoas jurídicas
<i>Consistência estrutural</i>			
Duplicidades	Unicidade	0	Verificadas pela chave composta definida na integração
Inconsistências temporais	Consistência temporal	<0,01%	2 registros com datas fora da ordem esperada
<i>Sinalização para revisão</i>			
Valores negativos	Sinalização estrutural	7,06%	190.715 registros associados a estornos ou anulações
Registros atípicos	Triagem automatizada	1,00%	27.009 registros do <i>dataset</i> consolidado sinalizados pelo <i>Isolation Forest</i> , sem exclusão automática

O *DespesasGovPB* apresenta alta completude nos campos centrais da execução orçamentária, como chaves, valores e data de empenho. Ausências em datas de liquidação e pagamento não representam, necessariamente, inconsistências, pois podem indicar despesas ainda não liquidadas ou não pagas no recorte analisado. A cobertura cadastral aplica-se apenas a credores pessoa jurídica, o que deve ser considerado em análises sobre localização e situação cadastral de fornecedores.

4.2. Padrões temporais e concentração de valores

A análise dos intervalos entre fases indica tramitação em prazos curtos na maior parte dos registros. O intervalo médio entre empenho e liquidação foi de 5 dias, com mediana de 1 dia; entre liquidação e pagamento, foi de 4 dias, com mediana de 0 dias. Como o *dataset* consolida múltiplos eventos de um mesmo empenho em uma única linha, esses valores medem o tempo entre as datas finais registradas para cada fase, não todos os eventos transacionais.

Também se observa concentração temporal dos pagamentos: cerca de 30% ocorrem no último trimestre, e novembro e dezembro respondem por 22% do total liquidado. Esse padrão é compatível com maior intensidade de execução no encerramento do exercício fiscal, mas sua explicação causal demanda investigação específica. A Figura 4 apresenta a evolução temporal da despesa empenhada entre 2015 e 2025.

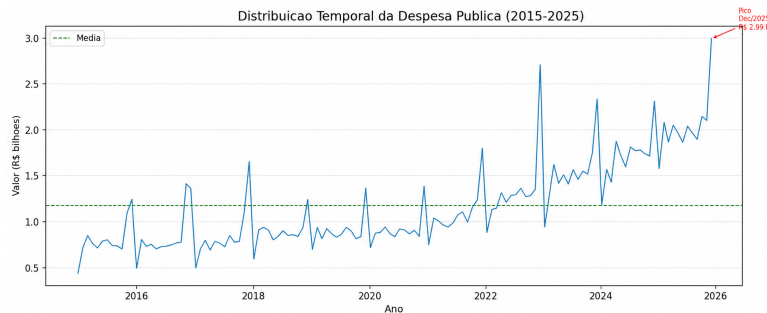


Figura 4. Distribuição temporal da despesa pública empenhada (2015–2025).

A série mensal apresenta sazonalidade recorrente e variações expressivas. O pico mais alto ocorre em dezembro de 2025, quando a despesa empenhada supera R\$ 2,9 bilhões, resultado compatível com a concentração de registros orçamentários no fim do exercício.

A concentração de valores por fornecedor também é elevada. Os 10% maiores fornecedores concentram 98,89% do valor liquidado, e o índice de Gini de 0,99, calculado sobre o valor liquidado por credor, reforça esse padrão. Os dez maiores fornecedores respondem ainda por 54,46% do volume total empenhado. Essa concentração deve ser lida à luz do escopo da base (Subseção 3.1), que inclui despesas de pessoal, encargos e transferências concentradas em credores institucionais.

4.3. Triagem de registros atípicos

Além dos indicadores descritivos, o *DespesasGovPB* permite sinalizar registros que merecem inspeção posterior. Os marcadores estruturais identificaram 190.715 registros com valores negativos, associados a estornos ou anulações, e 2 registros com datas fora da ordem esperada entre fases da despesa.

A aplicação do *Isolation Forest* selecionou os 27.009 registros com maiores escores de atipicidade, equivalentes ao percentil de 1,00% definido na parametrização do modelo (Subseção 3.2). Esses casos combinam valores financeiros ou intervalos temporais destoantes do padrão geral, mas não devem ser interpretados automaticamente como erro ou irregularidade; servem como triagem inicial para auditorias e análises especializadas.

5. Considerações Finais

O *DespesasGovPB* reduz a distância entre a disponibilização formal de dados públicos e seu uso em auditoria, transparência e gestão fiscal. Ao integrar empenho, liquidação, pagamento e restos a pagar, o *dataset* permite observar a execução da despesa pública em granularidade pouco explorada em bases abertas nacionais. O conjunto consolida

2.700.835 registros da Paraíba entre 2015 e 2025, oferece variáveis derivadas para análise fiscal e disponibiliza artefatos no Zenodo para reanálise do *snapshot* público.

As análises exploratórias evidenciam concentração temporal e financeira: cerca de 30% dos pagamentos ocorrem no último trimestre, novembro e dezembro respondem por 22% do total liquidado, os 10% maiores credores concentram 98,89% do valor liquidado e o índice de Gini é 0,99. O *Isolation Forest*, parametrizado para selecionar o percentil mais discrepante da base, isolou 27.009 registros como candidatos à revisão. Esses resultados não indicam irregularidades por si só, mas mostram que a base apoia perguntas sobre concentração, sazonalidade e triagem da execução da despesa.

Entre as limitações, a base depende da qualidade dos registros originais, da cobertura do enriquecimento cadastral por CNPJ e da consolidação de múltiplos eventos por empenho; essa escolha favorece a análise longitudinal, mas simplifica eventos intermediários. Registros sinalizados pelo *Isolation Forest* são triagem, não evidência de irregularidade.

Como trabalhos futuros, pretende-se incorporar receitas, convênios e transferências recebidas pelo Estado, construir indicadores comparativos e avaliar a portabilidade para outros estados ou recorte nacional.

Financiamento e conflito de interesses

Esta pesquisa não recebeu financiamento, e os autores declaram não haver conflito de interesses.

Uso de Inteligência Artificial

Foram utilizadas ferramentas de IA generativa (OpenAI GPT-5.4 e Claude Sonnet 4.6, Anthropic) apenas para revisão textual, organização redacional, diagnóstico de erros e apoio à programação. A responsabilidade final por resultados, tabelas, referências e validação permanece integralmente com os autores.

Referências

- Anastácio de Souza, R., Fernandes, D. Y. d. S., and Rego, A. S. d. C. (2026). DespesasGovPB: um conjunto de dados estruturado sobre despesas públicas do governo da paraíba. Zenodo. Disponível em: <https://doi.org/10.5281/zenodo.19502820>. Acesso em: 10 abr. 2026.
- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4):447–468.
- Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011 (Lei de acesso à informação). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 11 abr. 2026.
- Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018 (Lei geral de proteção de dados pessoais). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 11 abr. 2026.
- Brasil. Secretaria de Orçamento Federal (SOF) (2025). Manual técnico do orçamento – MTO 2025. Disponível em: <https://www1.siof.planejamento.gov.br/mto/lib/exe/fetch.php/mto2025:mto2025.pdf>. Acesso em: 17 mar. 2026.
- Brelàz, G. d., Dias, T. F., Reinecke, L. F. G., Nascimento, A. B. F. M. d., and Rodrigues, D. C. (2025). Governo aberto: Caminhos para transparência, dados abertos, participação, colaboração e accountability. *Cadernos Gestão Pública e Cidadania*, 30:e92960.
- Creative Commons (2013). Attribution 4.0 international (CC BY 4.0). Disponível em: <https://creativecommons.org/licenses/by/4.0/>. Acesso em: 28 mar. 2026.
- Davis, P. G. (2022). Indicadores e dados municipais: um banco de dados para avaliar a eficiência das despesas públicas. In *Anais do IV Dataset Showcase Workshop (DSW), SBBD 2022*, pages 23–34, Porto Alegre. Sociedade Brasileira de Computação.
- Godoy, R. M. d., Menezes, J. P. C. B., Ferreira, J. C. d. S., and Souza, M. C. d. (2025). Panorama das pesquisas sobre dados governamentais abertos no período 2014 a 2024: uma análise de conteúdo. *Revista do Serviço Público*, 76(4):681–705.
- Governo do Estado da Paraíba (2026). Portal da transparência da paraíba: Seção de execução orçamentária. Disponível em: <https://transparencia.pb.gov.br/>. Acesso em: 02 abr. 2026.
- Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.
- Oliveira, G. P. et al. (2024). ICPSet: um conjunto de dados estruturado de itens de compras públicas. In *Anais do VI Dataset Showcase Workshop (DSW), SBBD 2024*, pages 103–113, Fortaleza. Sociedade Brasileira de Computação.
- Open Data Charter (2015). International open data charter principles. Disponível em: <https://opendatacharter.org/principles/>. Acesso em: 30 mar. 2026.
- Pfitzmann, A. and Hansen, M. (2010). A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. Version 0.34. Disponível em: <http://dud.inf>.

- tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf. Acesso em: 30 jun. 2026.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.
- Receita Federal do Brasil / SERPRO (2026). Consulta CNPJ: serviço de dados cadastrais de pessoas jurídicas. Serviço de consulta cadastral via API REST institucional (SERPRO/Receita Federal), acesso por convênio com a CODATA-PB; sem documentação pública disponível. Campos coletados por CNPJ: razão social, porte, natureza jurídica, atividade econômica principal (CNAE) e município; coleta realizada de jan. a abr. 2026; acesso restrito a conveniados.
- Silva, M. O. et al. (2022). LiPSet: um conjunto de dados com documentos rotulados de licitações públicas. In *Anais do IV Dataset Showcase Workshop (DSW), SBBD 2022*, pages 45–56, Porto Alegre. Sociedade Brasileira de Computação.
- Sweeney, L. (2002). k -anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.
- Tesouro Nacional (2025). Manual de contabilidade aplicada ao setor público – 11ª edição. Disponível em: <https://www.tesourotransparente.gov.br/publicacoes/manual-de-contabilidade-aplicada-ao-setor-publico-mcasp/2025/26>. Acesso em: 17 mar. 2026.
- Toledo, J. M., Moura, T. J. M., and Timoteo, R. D. A. (2023). BrStats: a socioeconomic statistics dataset of the Brazilian cities. In *Anais do V Dataset Showcase Workshop (DSW), SBBD 2023*, pages 67–78, Belo Horizonte. Sociedade Brasileira de Computação.
- United Nations Statistics Division (2016). Classification of the functions of government (COFOG). Disponível em: <https://unstats.un.org/unsd/classifications/Family/Detail/4>. Acesso em: 30 mar. 2026.
- Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.
- World Wide Web Consortium (W3C) (2017). Data on the web best practices. W3C Recommendation. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 28 mar. 2026.