Avaliação da Validade da Lei de Benford aos dados do Hidroweb, CAMELS-BR e CABra

João Marcos Carvalho[1] ; Henrique Degraf[2]; Daniel Henrique Marco Detzel[3]& Eloy Kaviski[4]

Abstract: Benford's Law, also known as the Law of Anomalous Numbers, describes a counterintuitive yet consistent pattern in the distribution of leading digits in numerical datasets derived from natural phenomena. First formalized in 1938, it states that in many naturally occurring datasets, lower digits appear more frequently as leading digits than higher ones, following a logarithmic distribution—where the digit 1 appears about 30% of the time, while the digit 9 occurs in only about 5% of cases. This principle has been successfully applied to various datasets, including city populations, atomic weights, death rates, water consumption, sediment transport, income metrics, and hydrological measurements. Given that hydrological time series can contain biases or inconsistencies, Benford's Law offers a statistical tool for assessing data quality. This study evaluates the adherence of three Brazilian hydrological datasets to Benford's Law, aiming to identify potential irregularities. The datasets include daily discharge and precipitation records from 715 gauging stations from Hidroweb, CAMELS-BR, and CABra. When analyzed as a whole, both discharge and precipitation data closely followed the expected Benford distribution. At the individual series level, most precipitation records also showed strong adherence. While discharge data generally followed the expected distribution, greater variability was observed among individual time series. These results suggest a general absence of distortions or bias and highlight the potential of Benford's Law as a diagnostic tool to identifying measurement issues, refining data filtering processes, and guide the prioritization of gauging stations for hydrological analysis.

Resumo: A Lei de Benford, também conhecida como Lei dos Números Anômalos, descreve um padrão contraintuitivo, porém recorrente, na distribuição dos dígitos iniciais de conjuntos numéricos derivados de fenômenos naturais. Formalizada em 1938, a lei estabelece que, em muitos conjuntos de dados naturais, dígitos menores tendem a aparecer com maior frequência na posição inicial do que dígitos maiores, seguindo uma distribuição logarítmica — na qual o dígito 1 ocorre em cerca de 30% dos casos, enquanto o dígito 9 aparece em apenas 5%. Essa regularidade estatística foi aplicada com sucesso a diversos dados, como populações de cidades, pesos atômicos, consumo de água, indicadores de renda e séries hidrológicas. Considerando que séries hidrológicas podem apresentar inconsistências, a Lei de Benford se mostra uma ferramenta útil como indicador de qualidade. Neste estudo, analisa-se a aderência de três bases hidrológicas brasileiras à Lei de Benford, com o objetivo de identificar divergências. Os dados avaliados incluem séries diárias de vazão e precipitação provenientes de 715 estações fluviométricas extraídas do Hidroweb, CAMELS-BR e CABra. Quando analisados em conjunto, tanto os dados de vazão quanto de precipitação apresentaram aderência à distribuição de Benford. Individualmente, a maioria das séries de precipitação também apresentou forte conformidade. Embora as vazões, em geral, sigam a distribuição teórica, observou-se maior variabilidade entre as séries individuais. Esses resultados indicam que certas estações podem ter

1) UFPR. Av. Cel. Francisco H. dos Santos nº 100, joao.huf.carvalho@gmail.com
2) UFPR. Av. Cel. Francisco H. dos Santos nº 100, degraf.henrique@gmail.com
3) UFPR. Av. Cel. Francisco H. dos Santos nº 100, detzel@ufpr.br
3) UFPR. Av. Cel. Francisco H. dos Santos nº 100, eloy.dhs@gmail.com

padrões artificiais ou vieses, e reforçam o potencial da Lei de Benford como ferramenta diagnóstica para identificação de problemas de medição, filtragem e priorização de estações em análises hidrológicas.

Palavras-Chave – Lei de Benford, Consistência de Bases de Dados, Séries Hidrológicas.

INTRODUÇÃO

A qualidade e confiabilidade dos dados hidrológicos são fundamentais para a análise e gestão dos recursos hídricos. No Brasil, diversas bases de dados como o Hidroweb (ANA), o CAMELS-BR (Chagas et al., 2020, dados disponíveis em https://zenodo.org/records/15025488) e o CABra (Almagro et al., 2021, dados disponíveis em https://zenodo.org/records/4070147) reúnem séries temporais de variáveis hidrometeorológicas, amplamente utilizadas por pesquisadores, gestores e tomadores de decisão. No entanto, a origem, os métodos de processamento e a consistência dessas séries podem variar significativamente entre as fontes e podem afetar a qualidade da informação disponibilizada. Assim sendo, a utilização de meios de aferição da qualidade de dados tem grande valia.

Neste contexto, a Lei de Benford ou Lei dos Números Anômalos, formalizada por Benford (1938) com base em observações anteriores de Newcomb (1881), descreve um padrão estatístico recorrente em dados provenientes de fenômenos naturais. Embora muitas vezes vista como uma curiosidade matemática, e recebendo pouca atenção nas ciências da natureza, a lei apresenta aplicações como ferramenta para auditoria, controle de qualidade e validação de dados em diversas áreas, incluindo contabilidade, ciências da Terra e hidrologia.

Entretanto, mesmo sendo pouco utilizada nos recursos hídricos, estudos como os de Nigrini e Miller (2007), Sambridge *et al.* (2010) e Sowby (2018), demonstram a sua capacidade de servir como indicador de rápido e eficaz de: Anomalias, Dados Incompleta; Erros de Medição; Controle de Qualidade; Validação de Modelos Númericos.

Nigrini e Miller (2007) utilizaram dois grandes conjuntos de dados hidrológicos: as vazões médias anuais registradas em estações fluviométricas nos Estados Unidos (de 1874 a 2004) e o banco de dados global de lagos e áreas úmidas (GLWD). Enquanto os dados de vazão apresentaram uma conformidade quase perfeita com a Lei de Benford, os dados referentes a perímetros e áreas de corpos hídricos mostraram desvios significativos, com indícios de distribuição segundo uma lei de potência.

Sambridge et al. (2010) demonstraram que a Lei de Benford se aplica amplamente a conjuntos de dados das ciências naturais, incluindo observações em geofísica, astronomia e matemática. Os autores analisam 15 conjuntos de dados e identificam que a conformidade com a Lei está associada à presença de uma ampla faixa dinâmica e à ausência de restrições artificiais. Por fim, Sowby (2018) demonstrou que 3 bases de dados públicos de uso de água potável nos EUA seguem a Lei de Benford, indicando seu potencial como ferramenta simples e eficaz para verificar a integridade e a plausibilidade de dados hidrológicos e modelagens associadas. Em testes comparativos entre dois conjuntos de dados, o conjunto conhecido por ser mais preciso também apresentou maior conformidade com a Lei de Benford, reforçando sua utilidade na avaliação da qualidade dos dados.

Embora a comunidade matemática ainda busque uma justificativa teórica sólida para a Lei de Benford, existem evidências crescentes de que suas propriedades podem ser uma característica comum nas ciências físicas (Sambridge (2011)). Nas últimas décadas a comunidade científica tem aumentado seu interesse nas propriedades desta lei, a ponto de existir um repositório centralizado e dedicado a estudos envolvendo esta lei (http://www.benfordonline.net/).

Diante das questões apresentadas, este estudo propõe a aplicação da Lei de Benford como ferramenta diagnóstica para investigar a consistência das séries de vazão e precipitação das três bases de dados brasileiras Hidroweb, CAMELS-BR e CABra. Por meio da comparação entre a distribuição teórica dos primeiros dígitos e a frequência observada nas séries temporais, busca-se identificar possíveis distorções ou padrões artificiais que possam comprometer análises de recursos hídricos baseadas nestas informações.

MATERIAIS E MÉTODOS

Dados Utilizadas

Para esta análise foram utilizadas três bases de dados, sendo elas o Hidroweb da Agência Nacional de Águas e Saneamento Básico, o CABra (Almagro *et al.*, 2021) e CAMELS-BR (Chagas *et al.*, 2020). O Hidroweb se trata da principal base de dados hidrometeorológicos observados do Brasil, e reúne mais de 3.500 estações cadastradas em sua base. Enquanto isso o CAMELS-BR e CABra são conjuntos de dados padronizados e consistidos de parte do banco de dados da ANA, sendo um dos seus objetivos, serem utilizados como testes de bancada para estudo dos recursos hídricos brasileiros. O CAMELS-BR conta com aproximadamente 897 estações pré-selecionadas, enquanto o CABra possui cerca de 735.

Com o objetivo de realizar uma análise comparativa dos dados através da Lei de Benford, foram selecionadas as 715 estações comuns às três bases de dados. As séries de vazão foram avaliadas nas três fontes, enquanto as séries de precipitação foram consideradas apenas para o CAMELS-BR e o CABra. Isso foi feito, pois os dados destas duas bases de dados não se tratam de medições pontuais, como as presentes no Hidroweb, mas sim uma precipitação média sobre cada uma das bacias hidrográficas traçadas a partir da coordenada da estação de referência. A Tabela 1 apresenta um resumo das informações utilizadas neste estudo.

Tabela 1 – Dados Utilizados

Fonte	Tipo	Período
CAMELS-BR v2	Precipitação* e Vazão	1980 - 2024
CABra	Precipitação* e Vazão	1980 - 2010
Hidroweb	Vazão	> 10 anos de dados

*Valor médio sobre a bacia hidrográfica

Lei de Benford e Métricas de Avaliação

A Lei dos Números Anômalos, também conhecida como Lei de Benford, surge a partir da observação feita por Simon Newcomb, por volta de 1881, que nota que as primeiras páginas de livros de tabelas logarítmicas eram muito mais utilizadas que as últimas. Esta observação o levou a intuir que números que começam com o dígito 1 são mais utilizados que números que começam com o dígito 9.

A partir desta ideia de Newcomb, Benford (1938) compila os primeiros dígitos de aproximadamente 20.229 observações de diferentes fontes como - população de cidades, dados financeiros, pesos atômicos, comprimentos de rios etc - e demonstra que o primeiro dígito d das diferentes amostras segue uma distribuição de probabilidades dada pela equação 1.

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(1 + \frac{1}{d}\right) \quad (1)$$

Posteriormente a Lei dos Números Anômalos foi generalizada, de modo que é possível encontrar a distribuição de probabilidades, considerando não só o primeiro dígito, mas também os dois primeiros, ..., e também probabilidades condicionais (por exemplo: Dado que o primeiro dígito

é 2, qual deve ser a probabilidade do quarto dígito ser igual a 4). Além desta generalização, vale comentar que a Lei de Benford é independente da base númerica utilizada, ou seja, ela não só é válida para a base decimal, mas também bases como binária, octal, hexadecimal etc.

De modo geral, a lei tende a se aplicar bem a conjuntos numéricos que representam ocorrências naturais (SOWBY, 2018), e é justamente essa propriedade que motiva sua aplicação neste estudo - Verificar a aderência dos três bancos de dados hidrometeorológicos supracitados à Lei dos Números Anômalos.

Para quantificar a aderência entre as distribuições empíricas e teórico dos primeiros dígitos, optou-se por utilizar o Erro Máximo Absoluto (EMA) - definido como a maior diferença absoluta entre as frequências observadas e aquelas previstas pela distribuição de Benford. O EMA foi calculado individualmente para cada estação analisada.

Além disso, aproveitando que a base de dados CAMELS-BR possui uma lista de atributos relativos a cada uma das bacias hidrográficas presente no seu banco, foi verificada a correlação de Spearman entre estas variáveis e o EMA das estações. Ao todo foram utilizados 59 atributos, sendo eles relativos a topografia, uso do solo, hidrologia, clima, geologia, localização, checagem de qualidade e solo. Vale comentar que a base CABra também possui este tipo de informação, porém julgou-se suficiente até o momento a utilização de apenas uma fonte de dados de atributo das bacias.

RESULTADOS

A avaliação dos resultados se deu em três passos:

1. Comparação direta de todas as estações com a distribuição teórica da Lei de Benford;
2. Visualização espacial do EMA;
3. Comparação entre os valores de EMA dos diferentes bancos de dados;

A Figura 1 apresenta o conjunto de gráficos referentes à primeira etapa de análise de resultados. Através dela é possível observar que, em média, todos os bancos e tipos de dados avaliados são aderentes à Lei de Benford. Para os dados de precipitação, o comportamento médio apresenta menor variância entre as séries. Em particular, a frequência do dígito 1 é de aproximadamente 37%, enquanto a distribuição teórica prevê cerca de 30%. Para os dados de vazão, o comportamento médio apresenta maior variabilidade. Apesar disso, os valores médios estão mais próximos da distribuição teórica do que os observados nas séries de precipitação.

Investigando a distribuição espacial dos resultados (Figuras 2), é possível observar que para os dados de precipitação a região mais próxima do litoral geralmente apresentam diferenças maiores que a região central do país. As maiores divergências concentram-se nas macro regiões hidrográficas do Atlântico Leste e Atlântico Nordeste Oriental. No caso dos dados de vazão, a macro região do São Francisco apresenta uma concentração de estações com maior EMA do que as demais regiões, sendo que essa concentração ocorre especialmente na porção central e norte. A macro região do Uruguaia, por outro lado, apresenta forte aderência com à Lei dos Números Anômalos.

Ao comparar os resultados dos dados de precipitação do CABra com os do CAMELS-BR (Figura 3), é possível observar que no geral existe uma boa correlação, porém algumas séries temporais do CAMELS-BR possuem um EMA maior que aos do CABra. Essas estações com maior desvio também aparecem no mapeamento da Figura 2, sendo que a sua maioria se encontra na parte mais ao norte da macro região do Atlântico Leste, e mais a oeste na região do Atlântico Nordeste Oriental.

Figura 1 – Distribuição de Frequências (Esquerda) e EMA (Direita) dos Diferentes Dados Analisados.

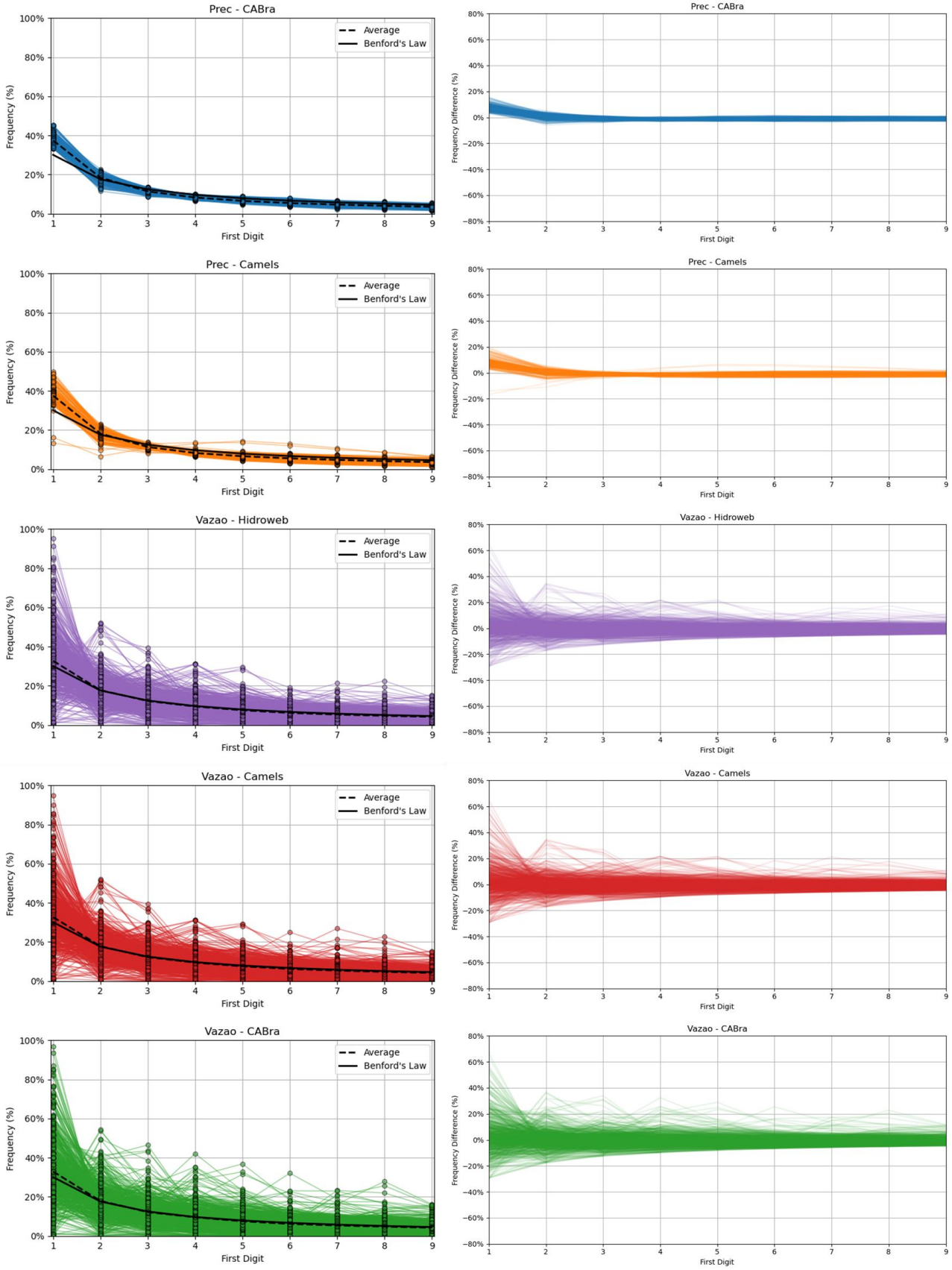


Figura 2 – Dados de Precipitação (Acima) e Vazão (Abaixo) – Erro Máximo Absoluto

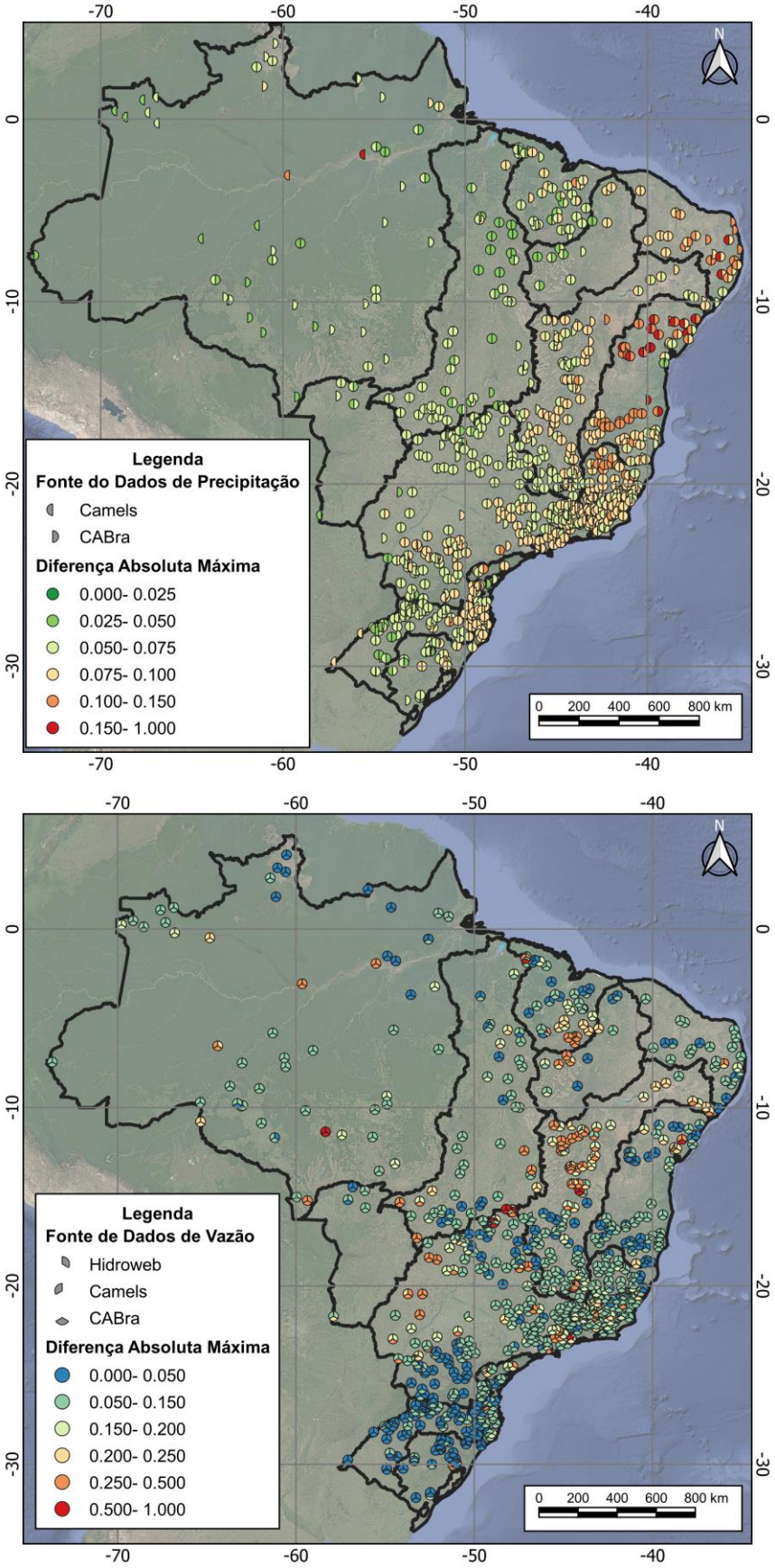
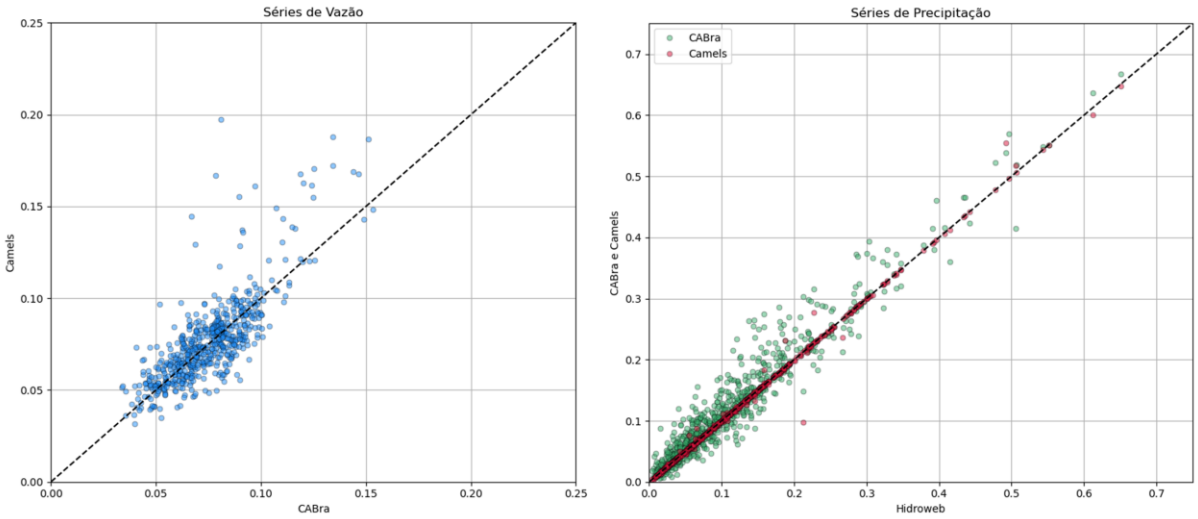


Figura 3 – Comparação entre o EMA das séries do Hidroweb e demais fontes de dados.



Em relação às vazões, a comparação utilizou como base os dados do Hidroweb. A partir da Figura 4 é possível observar que os dados do CAMELS-BR possuem uma boa aderência aos dados do Hidroweb, enquanto os dados do CABra apresentam uma dispersão maior, assim como EMA's mais elevados.

A Tabela 2 apresenta os coeficientes de correlação de Spearman entre os valores de EMA e os atributos físicos das bacias. Os principais destaques são que para precipitação, as maiores correlações estão associadas à Longitude da estação (*gauge_lon*, ρ = 0,64), indicando que regiões mais a leste do país (litorâneas) tendem a apresentar maior desvio da distribuição teórica. Da mesma forma, a duração de eventos de precipitação extrema possui um valor importante (*high_prec_dur*, ρ = 0,53), o que pode estar relacionado à maior dificuldade de medição durante chuvas intensas, resultando em maior distorção.

Para a vazão, os atributos com maior correlação com o EMA incluem a declividade da curva de permanência de vazões (entre os percentis 33% e 66%, *slope_fdc*), a frequência e duração de vazões baixas (*low_q_freq* e *low_q_dur*) e o escoamento de base (*baseflow_index*) - que foi o único atributo a apresentar correlação positiva com o EMA.

Tabela 2 – Coeficientes de correlação de Spearman entre os atributos físicos das bacias e os valores de EMA.

Atributo*	Precipitação CABra	Precipitação CAMELS	Vazão CABra	Vazão CAMELS	Vazão Hidroweb
gauge_lat	-0.04	-0.01	0.24	0.23	0.23
gauge_lon	***0.53***	***0.64***	0.19	0.19	0.19
area_ana	0.12	0.05	0.02	-0.03	-0.02
area_gsim	0.12	0.06	0.01	-0.03	-0.03
p_mean	-0.41	-0.46	-0.25	-0.22	-0.22
pet_mean	-0.10	-0.08	0.30	0.29	0.28
et_mean	-0.19	-0.26	0.00	0.04	0.04
aridity	0.25	0.30	0.28	0.25	0.25
p_seasonality	0.25	0.28	0.26	0.22	0.23
asynchronicity	-0.27	-0.29	0.14	0.12	0.12

high_prec_freq	0.25	0.33	-0.18	-0.16	-0.16
high_prec_dur	0.43	***0.53***	0.07	0.04	0.04
low_prec_freq	0.30	0.41	-0.06	-0.03	-0.03
low_prec_dur	0.17	0.25	0.32	0.29	0.29
q_mean	-0.30	-0.33	-0.23	-0.20	-0.19
runoff_ratio	-0.24	-0.26	-0.22	-0.18	-0.18
stream_elas	0.02	0.00	-0.40	-0.39	-0.39
slope_fdc	-0.05	-0.07	***-0.69***	***-0.67***	***-0.66***
baseflow_index	-0.20	-0.21	***0.64***	***0.62***	***0.61***
hfd_mean	-0.36	-0.35	0.13	0.13	0.12
Q5	-0.15	-0.16	0.33	0.34	0.35
Q95	-0.26	-0.31	-0.39	-0.36	-0.35
high_q_freq	0.22	0.22	-0.49	-0.48	-0.48
high_q_dur	0.32	0.30	-0.41	-0.43	-0.43
low_q_freq	0.03	0.00	***-0.63***	***-0.61***	***-0.61***
low_q_dur	0.05	0.02	***-0.55***	***-0.54***	***-0.54***
zero_q_freq	0.22	0.29	0.00	0.02	0.02
elev_gauge	-0.04	-0.01	0.11	0.10	0.10
elev_mean	0.10	0.16	0.06	0.06	0.06
slope_mean	0.31	0.33	-0.15	-0.12	-0.12
area	0.14	0.06	-0.01	-0.05	-0.05
crop_perc	-0.07	-0.06	-0.01	-0.05	-0.05
crop_mosaic_perc	-0.01	0.04	-0.01	-0.07	-0.07
forest_perc	0.07	-0.01	-0.13	-0.07	-0.07
shrub_perc	0.12	0.15	0.26	0.24	0.24
grass_perc	0.09	0.11	0.14	0.12	0.12
barren_perc	0.05	0.11	0.22	0.20	0.19
imperv_perc	0.10	0.13	0.14	0.12	0.12
wet_perc	0.06	0.01	0.06	0.02	0.02
snow_perc	0.14	0.10	-0.03	-0.03	-0.04
dom_land_cover_perc	-0.08	-0.02	-0.08	-0.04	-0.05
geol_class_1st_perc	-0.23	-0.23	-0.07	-0.03	-0.03
geol_class_2nd_perc	0.20	0.21	0.05	0.03	0.03
carb_rocks_perc	0.16	0.11	0.02	0.01	0.01
geol_porosity	-0.16	-0.25	0.01	-0.01	-0.02
geol_permeability	-0.29	-0.25	0.00	-0.04	-0.05
sand_perc	-0.02	0.02	0.36	0.35	0.35
silt_perc	0.01	-0.05	-0.44	-0.42	-0.42
clay_perc	0.01	0.00	-0.30	-0.30	-0.30
org_carbon_content	0.08	0.09	-0.15	-0.13	-0.13
bedrock_depth	-0.10	-0.06	0.33	0.29	0.30

water_table_depth	0.34	0.40	-0.07	-0.05	-0.04
q_quality_control_perc	0.06	0.05	-0.05	-0.06	-0.05
q_stream_stage_perc	0.02	0.08	0.19	0.19	0.19
consumptive_use	0.07	0.20	0.06	0.02	0.01
consumptive_use_perc	0.27	0.38	0.15	0.11	0.11
reservoirs_vol	0.19	0.17	-0.01	-0.05	-0.05
regulation_degree	0.22	0.21	-0.03	-0.06	-0.06

*Atributo segue a nomenclatura do CAMELS-BR, disponível em https://zenodo.org/records/15025488

CONCLUSÕES

A aplicação da Lei de Benford às séries hidrometeorológicas brasileiras das bases Hidroweb, CAMELS-BR e CABra revelou-se uma ferramenta eficaz para avaliação preliminar da consistência e integridade dos dados. De modo geral, os resultados demonstram que tanto os dados de precipitação quanto os de vazão apresentaram boa aderência à distribuição teórica da Lei de Benford, quando considerados em conjunto, porém estações individuais podem apresentar desvios consideráveis.

Não foi feito uma investigação a fundo de cada uma das estações, porém vale ressaltar que não necessariamente uma discordância com a Lei dos Números Anômalos indica um problema nos dados em si, mas sim que existe um indicativo de que eles não seguem um comportamento natural. Por exemplo, no caso de medições de vazão a jusante de um barramento, é esperado que ocorra uma diferença.

As análises revelaram que:

- As séries de precipitação foram mais estáveis e menos dispersas, embora apresentem uma leve super-representação do dígito 1;
- As séries de vazão apresentaram maior variabilidade entre estações, mas com média global mais próxima da distribuição teórica;
- Regiões litorâneas, especialmente no Atlântico Leste e Atlântico Nordeste Oriental, concentraram estações com maiores desvios, tanto para precipitação quanto para vazão;
- A bacia do São Francisco apresentou agrupamentos notáveis de estações com altos valores de EMA, enquanto a bacia do Uruguai destacou-se pela elevada conformidade com a Lei.

A comparação entre bases mostrou que:

- Os dados de precipitação do CAMELS-BR e do CABra são coerentes entre si, embora o CAMELS-BR apresente EMA ligeiramente superior em algumas regiões;
- Os dados de vazão do CAMELS-BR se alinham mais fortemente com o Hidroweb, enquanto o CABra mostrou maior dispersão.

A análise de correlação de Spearman com os atributos físicos das bacias revelou que:

- Para precipitação, a longitude e a duração de eventos extremos estão positivamente associadas aos desvios, sugerindo que condições extremas e posicionamento geográfico influenciam a aderência à Lei de Benford;
- Para vazão, atributos como declividade da curva de permanência, escoamento de base e duração/frequência de vazões baixas apresentaram correlações relevantes com o EMA,

o que pode indicar influência de características hidrológicas estruturais na conformidade com a distribuição esperada.

Esses resultados reforçam o potencial da Lei de Benford como ferramenta complementar em processos de validação e diagnóstico de bases de dados hidrológicos. Sua aplicação permite a detecção de padrões artificiais, priorização de estações para revisão ou filtragem e aperfeiçoamento de modelagens e análises de larga escala no contexto da hidrologia brasileira.

REFERÊNCIAS

ALMAGRO, André et al. CABra: a novel large-sample dataset for Brazilian catchments. Hydrology and Earth System Sciences, v. 25, n. 6, p. 3105-3135, 2021.

BENFORD, Frank. The law of anomalous numbers. Proceedings of the American philosophical society, p. 551-572, 1938.

CHAGAS, Vinícius BP et al. CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. Earth System Science Data, v. 12, n. 3, p. 2075-2096, 2020.

NEWCOMB, Simon. Note on the frequency of use of the different digits in natural numbers. American Journal of mathematics, v. 4, n. 1, p. 39-40, 1881.

NIGRINI, Mark J.; MILLER, Steven J. Benford's law applied to hydrology data—results and relevance to other geophysical data. Mathematical Geology, v. 39, p. 469-490, 2007.

SAMBRIDGE, Malcolm; TKALČIĆ, Hrvoje; JACKSON, A. Benford's law in the natural sciences. Geophysical research letters, v. 37, n. 22, 2010.

SAMBRIDGE, Malcolm et al. Benford's law of first digits: from mathematical curiosity to change detector. Asia Pacific Mathematics Newsletter, v. 1, n. 4, p. 1-6, 2011.

SOWBY, Robert B. Conformance of Public Water Use Data to Benford's Law. Journal-American Water Works Association, v. 110, n. 12, p. E52-E59, 2018.