COS868 - Projeto do curso

Fernando Dias

5 de Janeiro de 2025

Este é o relatório do projeto do curso de probabilidade e estatística do Programa de Engenharia em Sistemas de Computação (PESC) da COPPE/UFRJ. Esse curso foi ministrado no trimestre 2024/3 pela professora Rosa Leão.

O código utilizado para a obtenção dos gráficos e tabelas está disponível em https://github.com/Fdms-3741/trabalhoprobabilidade.

1 Introdução

O objetivo desse trabalho é fazer uma avaliação estatística das taxas de download e upload de dois tipos de dispositivos: Smart TVs e Chromecasts. Nessa análise, investiga-se a distribuição das taxas enviadas e recebidas de modo a avaliar se essas distribuições seguem alguma distribuição conhecida da literatura. Além disso, é feito um teste comparativo entre as diferentes taxas para verificar se ambas seguem uma mesma distribuição.

Para esse trabalho, utilizou-se a linguagem `Python3` para construção de scripts que processam os datasets e geram as tabelas e gráficos deste relatório. As bibliotecas `numpy` e `scipy` são utilizadas para operações numéricas em arrays e calcular algumas das estatísticas vistas nesse relatório. A biblioteca `pandas` é responsável por manipular o dataset através de tabelas. As bibliotecas `matplotlib`, `seaborn` e `statsmodels` foram utilizadas para criar os gráficos.

A metodologia para cálculo dos dados é a mesma para todas as questões: Primeiramente, ambos os arquivos CSV são importados e unidos em um único dataset. Após isso, o dataset é transformado em um `pandas.DataFrame` que contém as colunas descritas na tabela 1. Com o dataset nesse formato, os gráficos, tabelas e boxplots são gerados diretamente através da biblioteca `seaborn`, que automaticamente filtra e separa os dados dessa tabela para a criação de multiplos gráficos.

Nome	Tipo	Descrição
Data e hora	`datetime[64]ns`	Data e hora da observação
Tipo de dispositivo	Smart TV|Chromecast	Separa dispositivos entre Chromecast ou Smart TV
Direção do fluxo	Upload|Download	Define se o bps é de upload ou de download
bps	`float`	Valor de bps naquele minuto

Tabela 1: Colunas dos datasets gerais

1.1 Tratamento dos valores nulos

Os valores de taxas consistem em valores que variam em diversas ordens de grandeza. Portanto, é necessário um tratamento dos dados para que as estatísticas possam ser calculadas sem o risco de perda de precisão. A sugestão dada pelo roteiro do projeto consiste em calcular a estatística do log dos dados ao invés dos próprios valores. Isso porém

A primeira alternativa para lidar com o problema da existencia de zeros nos dados consiste em calcular o logaritmo dos valores somados a 1. Assim, todas as entradas nulas serão consideradas e terão valor 0 e o erro para entradas grandes é negligenciável. Uma segunda consequência é o limitar o conjunto de dados na região $[0, \infty)$, que impede que valores bem pequenos de bps se tornem outliers com ordem de grandeza muito grande. A operação de conversão seguiu a seguinte fórmula:

$$b'_n = \log_{10}(b_n + 1) \; \forall \; n$$

A segunda alternativa consiste em tirar os valores 0 das entradas antes de fazer o logaritmo. Para manter consistência entre datasets e preservar a região do conjunto de dados, a mesma função de conversão é utilizada em ambos os casos.

A escolha entre os datasets vai depender da interpretação desejada dos resultados obtidos e dos objetivos com a análise, já que ambos os datasets fornecem resultados diferentes. Ao longo desse trabalho, é possível ver situações onde um dataset é mais apropriado para explicar um fenômeno do que outro, e essas situações serão elencadas ao longo desse trabalho.

2 Estatísticas gerais

Para as estatísticas gerais, os dados foram divididos em quatro conjuntos distintos: Upload e download da Smart TV e upload e download do Chromecast. Os resultados para as estatísticas gerais podem ser vistas na tabela 2. Já os histogramas, com e sem exclusão de zeros, e as CDFs empíricas podem ser vistas nas figuras 1, 2 e 3 respectivamente. Os boxplots representando os dois datasets está na figura 4.

Tipo de dispositivo	Direção do fluxo	n	Média	Desvio padrão	Variância
Chromecast	Download	1620529	3.800	1.663	1.289
	Upload	1620529	3.350	0.459	0.678
Smart TV	Download	4417903	2.351	6.721	2.592
	Upload	4417903	2.158	4.110	2.027

Tabela 2: Estatísticas globais

Na tabela 2, vemos as estatísticas globais de cada conjunto de dados analisado, além da quantidade de entradas em cada. Vemos que, para cada tipo de dispositivo, as médias das taxas são próximas, porém o download possui um desvio padrão significativamente maior que o upload. Isso indica que alguma característica do download torna a taxa mais variável do que a taxa do upload, o que pode ser explicado pelos diferentes tipos de taxas para diferentes serviços de streaming. Já o desvio padrão é da mesma ordem de grandeza, e as vezes até maior, do que a média em todos os casos. Isso é um indício de um grande espalhamento dos dados, o que pode ser esperado de dispositivos que podem alternar entre períodos de alta demanda (streaming de vídeo em 4K) ou baixa demanda (música ou desligado).

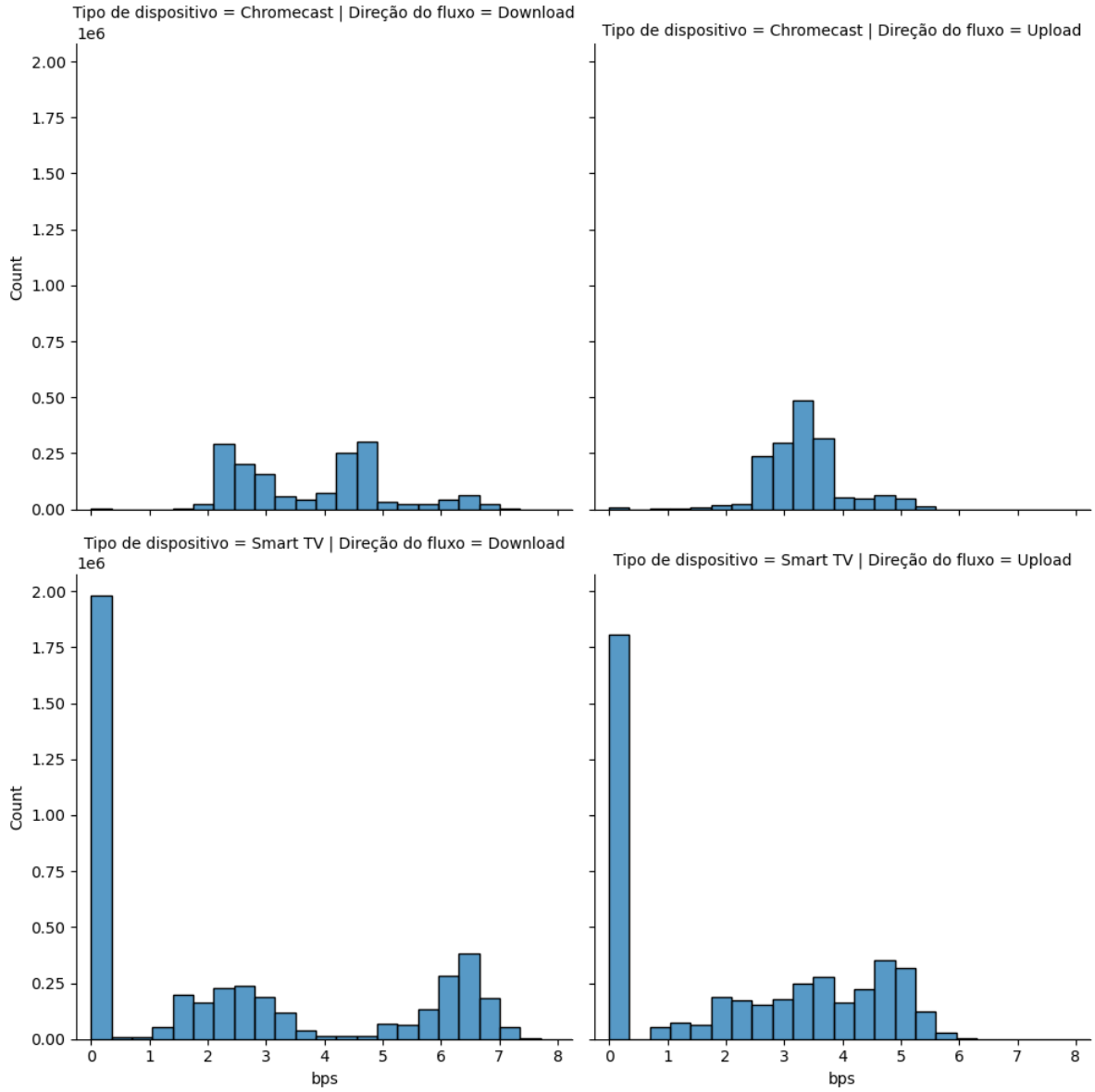
Vemos os datasets sendo representados integralmente pelos histogramas da figura 1. Em todos os histogramas *do relatório*, os bins dos histogramas foram calculados utilizando a regra de sturges onde $bins = \lceil 1 + \log_2(n) \rceil$. Uma primeira diferença notável está no impacto que a presença de zeros tem no formato do histograma dos conjuntos de dados das Smart TVs. Esse comportamento não foi observado no Chromecast e, sem mais detalhes sobre como o experimento foi conduzido, não é possível apontar o motivo dessa diferença.

Com a remoção dos valores nulos, é possível observar com mais detalhes o formato do histogramas na figura 2. Nessa figura, podemos distinguir a presença de uma distribuição multimodal nos histogramas de download comparado aos histogramas de upload. Isso pode ser explicado devido a presença de variáveis independentes que afetam o padrão de transmissão de dados, mas sem mais informações sobre os dispositivos não é possível definir a causa desse comportamento. As distribuições de upload não parecem possuir esse comportamento multimodal, e a distribuição da Smart TV possui um espalhamento maior do que a distribuição do :/Chromecast.

Através da distribuição cumulativa empírica da figura 3, podemos fazer as mesmas observações feitas através do histograma, o que mostra que a escolha dos bins foi bem representativa. As observações sobre o comportamento multimodal e o espalhamento do dataset são feitos através da observação da variação da taxa de crescimento da curva. Para observar o efeito multimodal nos gráficos de download, é possível observar que há duas partes do gráfico onde o crescimento da curva aparenta "acelerar" com mais rapidez do que as demais regiões. Já para o espalhamento, é possível observar que o crescimento da curva se encontra sobre uma região do eixo X menor no caso do Chromecast do que no caso da Smart TV, o que indica uma maior concentração de dados em uma região de suporte menor, ou seja, um menor espalhamento dos dados.

Os boxplots da figura 4 servem tanto para fazer uma comparação direta entre os quatro datasets quanto para observar o efeito da exclusão dos valores nulos. Na figura 4b o efeito dos valores nulos só

Figura 1: Histogramas dos dados com a **inclusão** de valores nulos



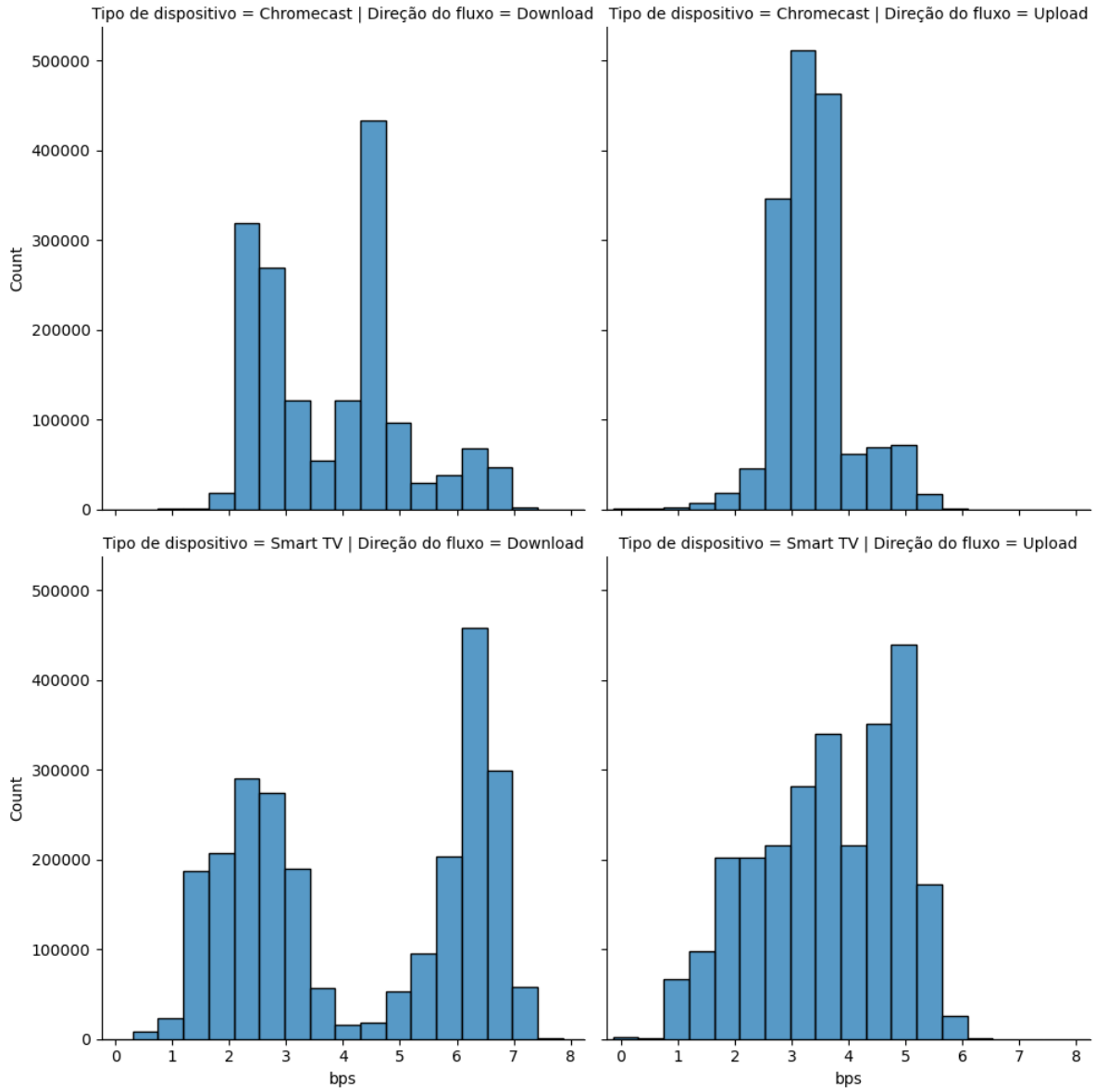
está presente nos dados da Smart TV, como observado anteriormente nos histogramas, e o efeito disso é o maior tamanho da distância inter-quartis (IQR) do que nos dados do chromecast. Ademais, como o fundo do box está em 0 em ambos os casos, sabemos que ao menos 25% dos dados nesses datasets é de valores nulos.

Com a exclusão dos zeros, como visto na figura 4a, os dados de um dispositivo já apresentam um comportamento similar aos dados do segundo dispositivo. Pode-se observar que os boxes da smarttv possuem uma distância interquartil maior que os boxes do Chromecast, tanto no caso de download quanto do upload. Vemos que o efeito do espalhamento afeta o formato dos whiskers e a aparição de outliers entre os dados de download e upload de ambos os dispositivos, mas a informação de multimodalidade não é observável. Com esse gráfico, podemos comparar apenas as estatísticas gerais entre os datasets de uma forma mais visual do que a observação em tabelas, porém nem todas as características do dataset puderam ser exploradas. Uma alternativa para observar a multimodalidade nesse tipo de visualização é utilizar um gráfico do tipo `violinplot`.

Nessa seção, as observações mais relevantes são elencadas a seguir:

- A exclusão dos valores nulos tem um impacto significativo na distribuição da Smart TV

Figura 2: Histogramas dos dados com a **exclusão** de valores nulos

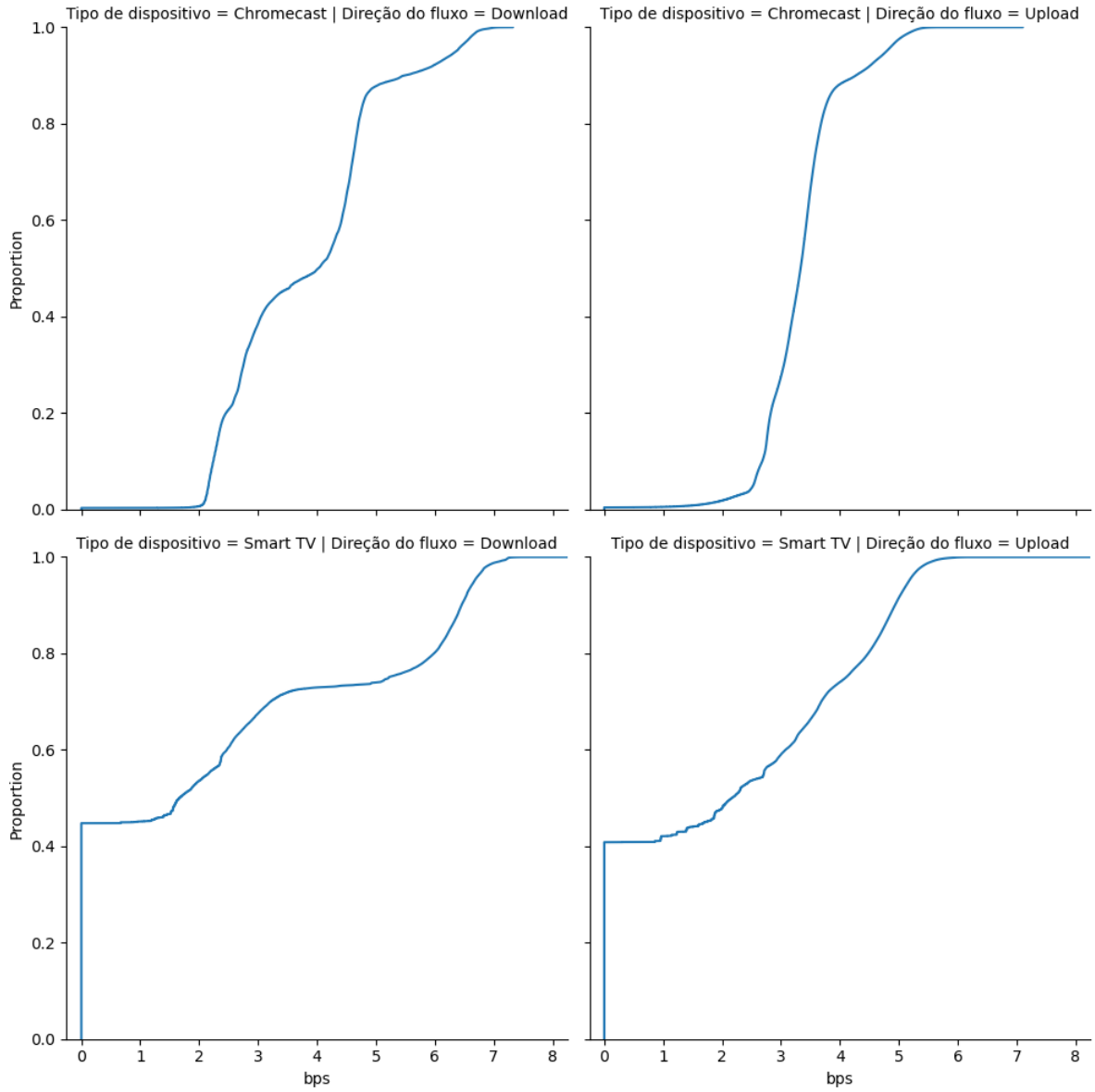


- Todos os dados possuem uma variância bem alta
- A distribuição para o Download possui dois picos característicos de uma multimodal, enquanto o upload não apresenta essa característica.
- Visualmente, é possível ver que a Smart TV apresenta distribuições que são mais espalhadas do que o Chromecast.
- As diferenças observadas podem ser resultado de diferentes tipos de serviços em um tipo de dispositivo que exigem taxas diferentes de consumo do que no outro, porém não é possível definir se esse é o caso sem mais informações sobre o experimento.

3 Estatísticas por horário

Nessa análise, será feita a observação de estatísticas gerais por horário. O objetivo é observar diferenças no padrão de distribuição de dados devido a hora de observação. Para isso cada um dos quatro datasets

Figura 3: CDFs empíricas dos dados com a inclusão de valores nulos

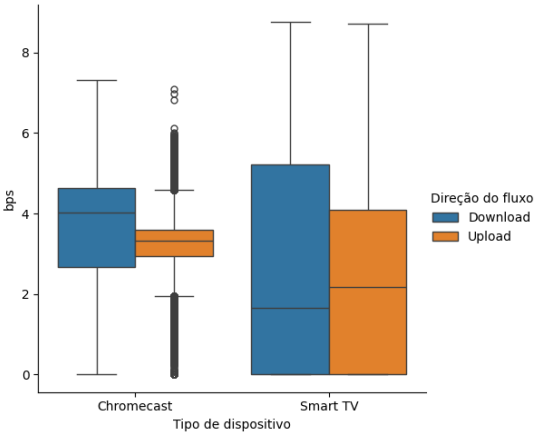


observados anteriormente são divididos em 24 subconjuntos de dados, um para cada hora do dia, onde todos os dias e todos os dispositivos de cada conjunto são utilizados para a formação dos dados. A tabela 4 contém todas as estatísticas separadas por horário, e os gráficos das figuras 5, 6 e 7 mostram as variações da média, desvio padrão e variância nos quatro datasets por horário, respectivamente. Finalmente, a figura 8 mostra os boxplots dos quatro conjuntos de dados, com cada box referente a uma hora do dia.

Um dos objetivos dessa análise é encontrar qual o horário de pico, ou seja, o horário com o maior tráfego médio observado entre os quatro datasets. Para encontrar esse dado, é necessário contabilizar o tráfego de todos os dispositivos observados, mesmo que o dispositivo não esteja transmitindo nem recebendo dados. Assim, será utilizado nessa análise os dados sem a exclusão das entradas nulas. Para comparar os efeitos da exclusão dos zeros, temos a tabela 3. Vemos que, com a exclusão dos zeros, a diferença entre a menor média e a maior média é de no máximo 0.9 log(bps), com todos os resultados entre 3 e 4.5. Já com o acréscimo dos zeros, os resultados de média variam significativamente (entre 0.7 e 3.4) para o caso da Smart TV, o que mostra que a disparidade entre os resultados observados dos dois dispositivos se dá pela adição dos zeros.

Na tabela 4, vemos as estatísticas gerais separadas por horário. Temos o desvio padrão, a média e a variância, para download e depois upload, dos dispositivos chromecast e Smart TVs. Através dessa

(a) Sem exclusão de zeros



(b) Com exclusão de zeros

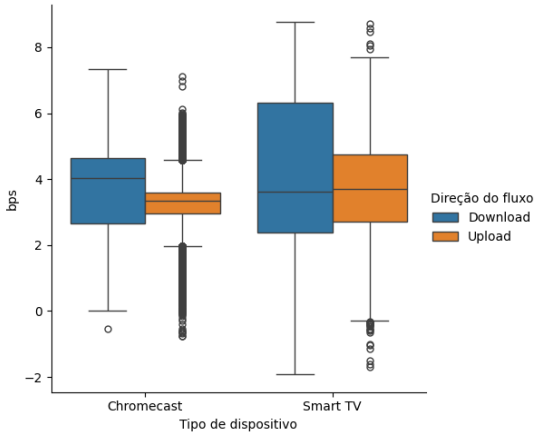


Figura 4: Boxplot dos dados gerais

		Sem zeros		Com zeros	
		Mínimo	Máximo	Mínimo	Máximo
Tipo de dispositivo	Direção do fluxo				
Chromecast	Download	3.572052	4.077841	3.565706	4.052698
	Upload	3.161117	3.542044	3.156929	3.521546
Smart TV	Download	3.697172	4.519928	0.735541	3.396095
	Upload	2.938141	3.831448	0.768513	3.124258

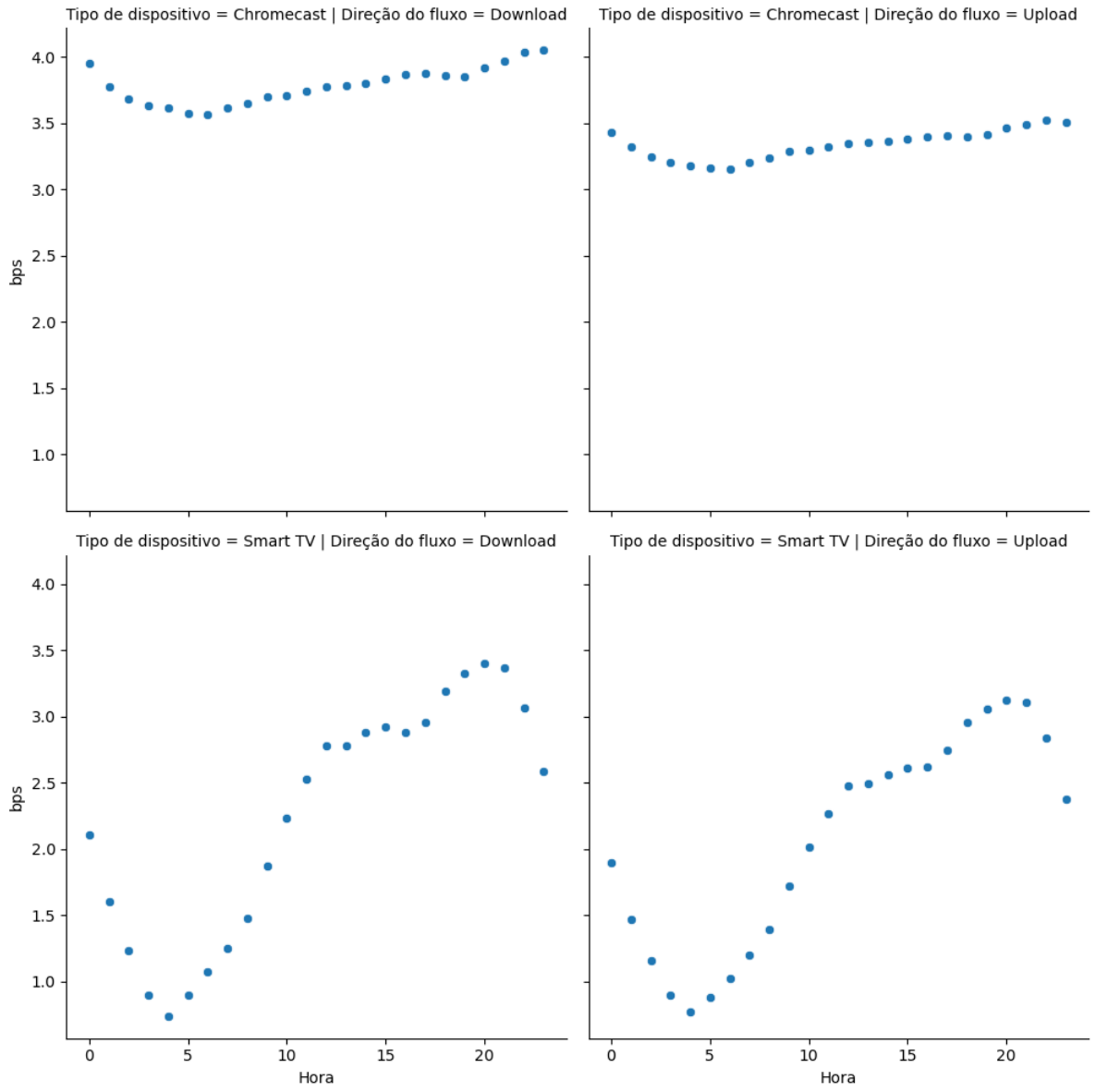
Tabela 3: Comparação da média com a exclusão dos zeros

Tipo de dispositivo						Chromecast					Smart TV
Direção do fluxo			Download			Upload			Download			Upload
	Desvio padrão	Média	Variância	Desvio padrão	Média	Variância	Desvio padrão	Média	Variância	Desvio padrão	Média	Variância
Hora												
0	1.436	3.952	2.064	0.794	3.432	0.631	2.624	2.104	6.887	2.038	1.894	4.157
1	1.319	3.775	1.741	0.694	3.322	0.482	2.460	1.601	6.051	1.938	1.467	3.758
2	1.207	3.685	1.457	0.584	3.243	0.341	2.226	1.228	4.956	1.775	1.153	3.151
3	1.186	3.636	1.407	0.554	3.202	0.307	1.904	0.897	3.627	1.569	0.893	2.464
4	1.189	3.618	1.414	0.560	3.178	0.314	1.699	0.735	2.887	1.433	0.768	2.055
5	1.173	3.571	1.376	0.539	3.159	0.290	1.876	0.891	3.522	1.531	0.875	2.345
6	1.170	3.565	1.371	0.551	3.156	0.303	1.999	1.072	3.999	1.624	1.024	2.640
7	1.194	3.616	1.426	0.580	3.200	0.337	2.097	1.244	4.400	1.732	1.197	3.000
8	1.219	3.652	1.486	0.624	3.241	0.389	2.307	1.477	5.323	1.878	1.391	3.528
9	1.228	3.696	1.509	0.630	3.286	0.397	2.500	1.868	6.251	1.992	1.717	3.971
10	1.232	3.707	1.518	0.638	3.297	0.407	2.623	2.229	6.884	2.058	2.016	4.236
11	1.230	3.741	1.514	0.638	3.321	0.407	2.656	2.526	7.055	2.065	2.265	4.268
12	1.239	3.778	1.537	0.636	3.348	0.404	2.654	2.775	7.044	2.039	2.473	4.158
13	1.260	3.785	1.588	0.654	3.354	0.428	2.646	2.778	7.001	2.034	2.488	4.137
14	1.257	3.797	1.581	0.652	3.362	0.425	2.690	2.875	7.237	2.053	2.557	4.218
15	1.274	3.832	1.623	0.659	3.380	0.435	2.676	2.919	7.163	2.028	2.606	4.116
16	1.307	3.865	1.709	0.691	3.399	0.478	2.599	2.875	6.758	1.966	2.620	3.866
17	1.315	3.879	1.729	0.704	3.407	0.496	2.532	2.958	6.415	1.895	2.744	3.591
18	1.288	3.857	1.661	0.689	3.401	0.474	2.498	3.191	6.241	1.830	2.951	3.350
19	1.288	3.852	1.659	0.695	3.417	0.484	2.508	3.321	6.292	1.809	3.053	3.275
20	1.323	3.922	1.750	0.702	3.468	0.493	2.490	3.396	6.201	1.780	3.124	3.168
21	1.364	3.967	1.861	0.737	3.493	0.543	2.474	3.366	6.125	1.769	3.103	3.131
22	1.403	4.036	1.969	0.771	3.521	0.595	2.507	3.060	6.289	1.859	2.837	3.458
23	1.469	4.052	2.159	0.832	3.507	0.693	2.578	2.585	6.648	1.984	2.374	3.939

Tabela 4: Estatísticas separadas por horário

tabela, é possível determinar os horários de maior demanda, que serão selecionados utilizando o valor médio observado em cada hora. Desta tabela, vemos que há uma diferença significativa no padrão de média entre Smart TVs e Chromecast, e essa diferença se dá pelo horário.

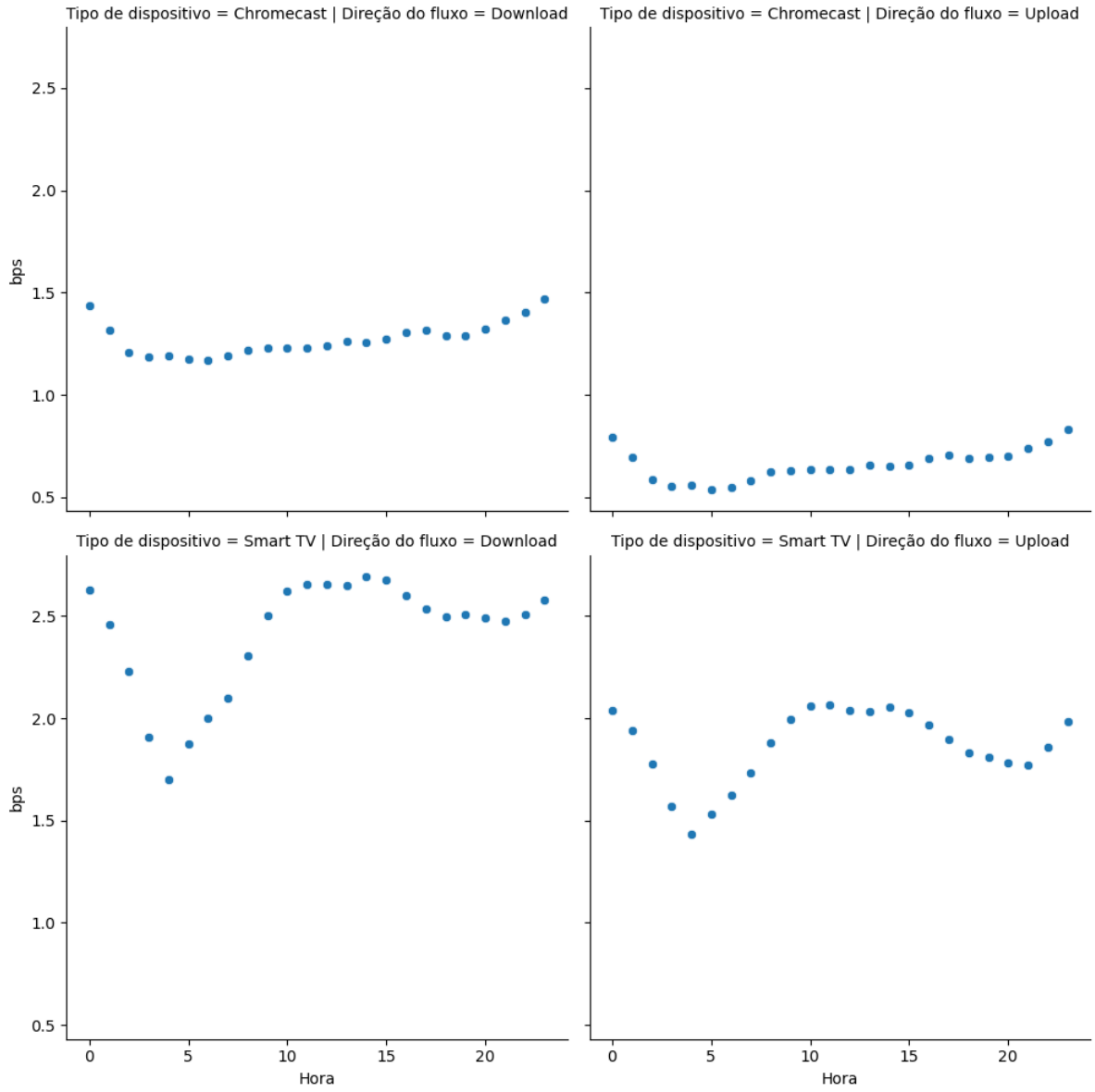
Figura 5: Variação da média por horário



Podemos ver o comportamento da média de forma mais clara ao analisar os gráficos da figura 5. A variação da média é altamente impactada pelo horário do dia observado no caso da Smart TV, o que é esperado dado que o uso desses dispositivos pode variar ao longo do dia, e tem maior uso durante a noite. Os dados do Chromecast porém seguem esse padrão com uma intensidade bem menor, se mostrando bem mais estáveis ao longo de 3.5 e 4 log(bps). Esse gráfico mostra um padrão de uso que pode ser dividido em três etapas: Um vale de baixo uso entre 0 e 10 horas, um consumo de banda constante entre 10 e 16 horas, um pico entre 18 e 22. Essas regiões são compatíveis com o padrão de uso das pessoas que assistem televisão.

Tanto a variância na figura 7 quanto o desvio padrão na figura 6 possuem comportamentos que são compatíveis com as três regiões observadas para a média no parágrafo anterior. Nesse caso, o período de estabilidade da média na parte da tarde é o período de maior variância dos dados no caso da Smart TV. O maior uso em média definido como um pico entre 18 e 22 hrs possui um decréscimo em relação ao desvio padrão se comparado com a região anterior. Se for feita a suposição que esses dados estão

Figura 6: Variação do desvio padrão por horário

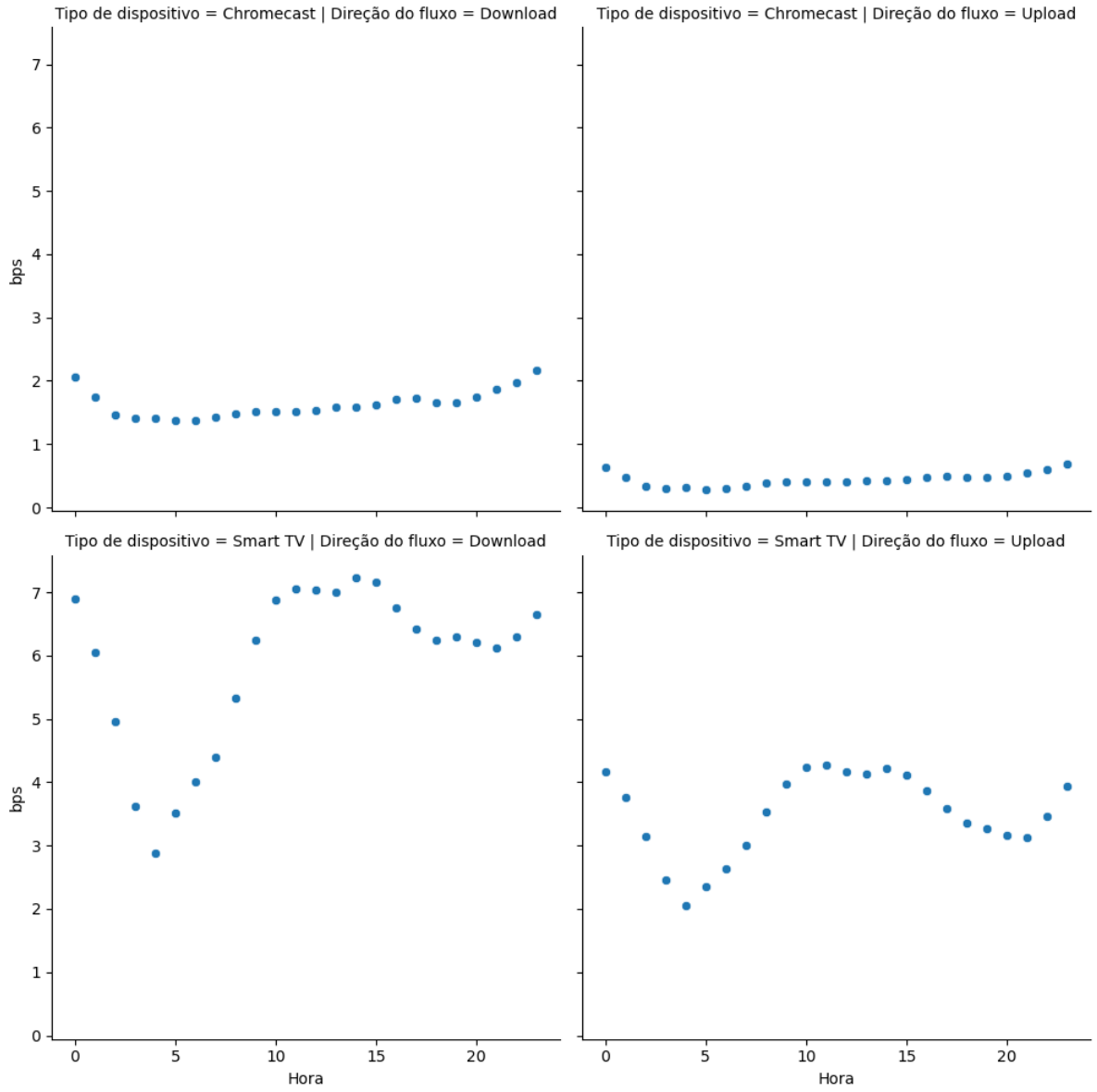


relacionados com o consumo de serviços de streaming em geral, seja canais ao vivo ou conteúdo gravado, uma possível explicação para isso é que com o maior número de aparelhos simultaneamente recebendo vídeo, mais dispositivos estão em um estado estável onde a taxa de bps não varia e é a mesma. Com isso, o desvio padrão diminui já que mais dispositivos apresentam o mesmo valor de bps. Analogamente, a região do vale também tem um decréscimo do desvio padrão, já que é o equivalente a dizer que mais dispositivos estão em um mesmo estado "estável" (desligado nesse caso), que possui uma taxa de bits por segundo (0) bem mais estável que no estado ligado.

Finalmente, os boxplots da figura 8 comparam as distribuições para diferentes horas do dia. Vemos que o chromecast possui um comportamento similar e estável ao longo do dia, seguindo o mesmo formato que os boxplots analisados na seção 2. Já para a Smart TV, vemos o impacto dos zeros afetando drasticamente o resultado dos boxplots ao longo do dia. As mesmas regiões descritas anteirormente se destacam nessa imagem, e nesse gráfico vemos que o vale contém basicamente todos os dados em um valor muito próximo de zero, com apenas outliers que chegam a taxas próximas às observadas nos horários comuns e de pico.

Nessa seção, podemos observar as seguintes características dos dados:

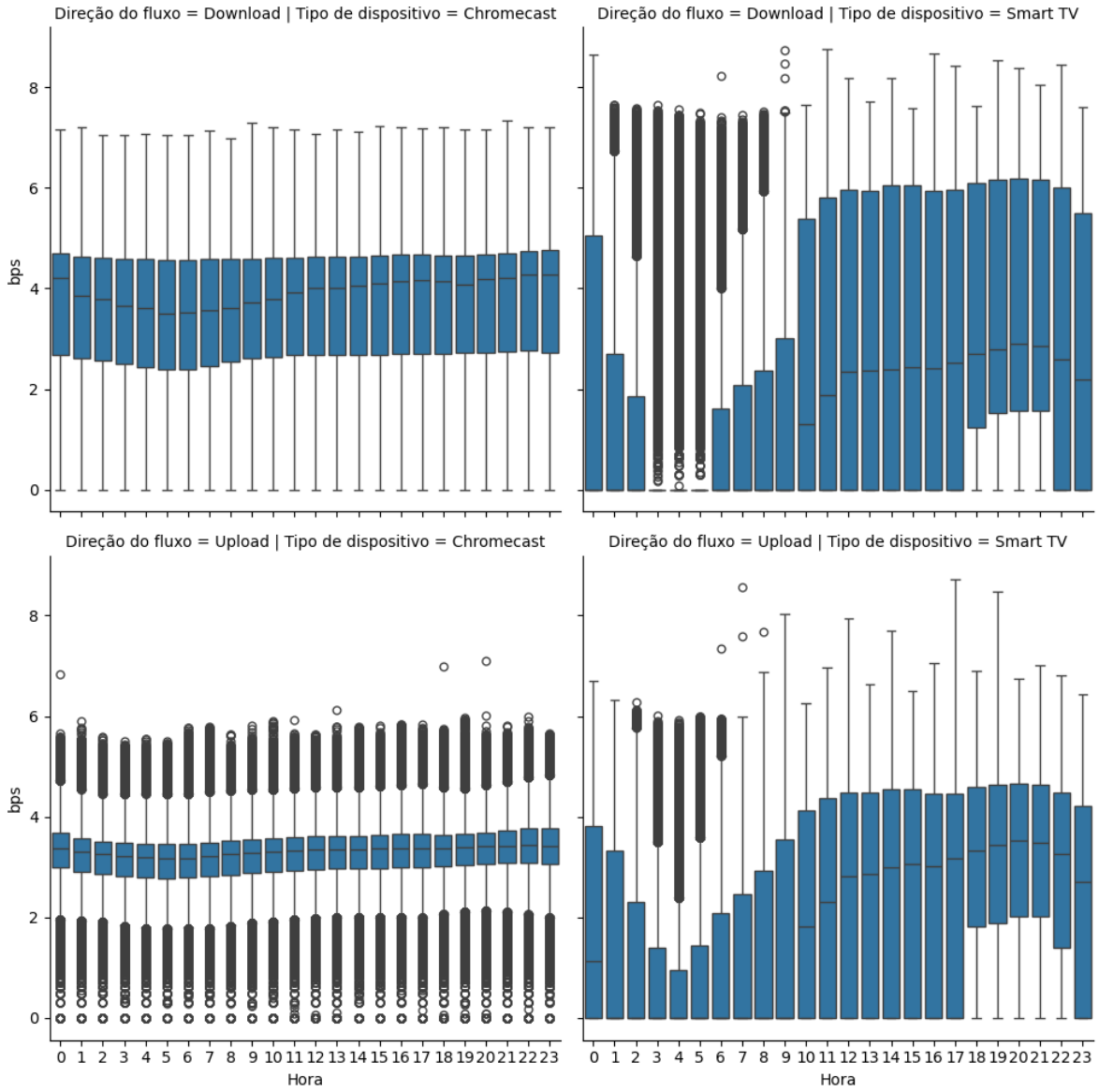
Figura 7: Variação da variância por horário



- A presença de zeros no dataset impacta consideravelmente a maior e menor médias observadas, o que facilita a observação dos diferentes períodos de uso dos dispositivos.
- A Smart TV apresenta uma variação de uso ao longo do dia bem maior que o Chromecast, que é bem mais estável. Não é possível explicar o porquê desse comportamento.
- Através da Smart TV, é possível observar três padrões de comportamento distintos ao longo do dia, que são condizentes com o uso de televisão por uma população.
- A variância apresenta um decréscimo tanto no momento de menor uso quanto no momento de maior uso na Smart TV. Isso pode ser um resultado de mais dispositivos tendendo para um mesmo estado "em uso" ou "desligado" na qual a taxa consumida é constante e similar nesses dois cenários.

Por último, com base nos resultados da tabela 4, pode-se encontrar os horários de maior média para serem utilizados na próxima seção. Esses horários estão representados na tabela 5.

Figura 8: Boxplot por horário



4 Análise dos horários de pico

Nessa seção, quer-se avaliar o comportamento dos dados nos horários de pico. Para isso, utilizamos a informação da tabela 5 para dividir os dados em quatro datasets distintos. Nesse caso, a tabela de dataset original é filtrada para conter apenas as entradas cujo horário na coluna "Data e hora" é o mesmo do que foi encontrado na seção anterior, para os quatro conjuntos de dados.

Nesse caso, o objetivo é caracterizar o comportamento dos dispositivos que estão transmitindo dados. Assim, não faz parte da análise predizer se um dispositivo vai transmitir ou não. Com isso, essa seção vai analisar os resultados para o dataset excluindo os valores nulos. Para fins de comparação, o histograma com a inclusão dos zeros também será apresentado.

4.1 Metodologia

A análise dessa seção consiste em fazer a estimação por máxima verossimilhança dos parâmetros das distribuições normal e gamma em cada um dos datasets. A estimação por MLE é desenvolvida no apêncide desse relatório, e nela é possível ver que não há uma forma fechada para a estimação dos parâmetros da função gamma. Portanto, essa estimação é feita utilizando soluções numéricas providenciadas pela

Tipo de dispositivo	Direção do fluxo		0
Chromecast	Download	Média	23
	Upload	Média	22
Smart TV	Download	Média	20
	Upload	Média	20

Tabela 5: Horários de pico para cada dataset

biblioteca `scipy`. Uma vez que os parâmetros são encontrados, eles são utilizados para desenhar as curvas das PDFs teóricas juntos aos histogramas.

Após essa análise inicial, é feito o *Probability plot* de cada um dos datasets, comparando-os com as distribuições normal e gamma. Finalmente, um QQPlot é feito para comparar os datasets do Chromecast e da Smart TV e identificar se o Download ou o upload é similar entre dispositivos.

4.2 resultados

As tabelas com os parâmetros encontrados são as tabelas 6 e 7, para a normal e a gamma respectivamente. O histograma que compara as PDFs pode ser visto na figura 9. O *Probability Plot* para a normal e a gamma podem ser vistos nas figuras 11 e 12. Finalmente, a comparação entre os datasets através do QQPlot pode ser vista na figura 13.

Tipo de dispositivo	Direção do fluxo	μ	σ^2
Chromecast	Download	4.077841	1.438828
	Upload	3.539628	0.731279
Smart TV	Download	4.286281	2.002812
	Upload	3.780892	1.162781

Tabela 6: Estimação de parâmetros via MLE para a distribuição normal

Tipo de dispositivo	Direção do fluxo	α	λ
Chromecast	Download	7.870926	1.930165
	Upload	23.287792	6.579145
Smart TV	Download	3.947653	0.920995
	Upload	8.547361	2.260667

Tabela 7: Estimação de parâmetros via MLE para a distribuição gamma

Na figura 9, podemos comparar os histogramas as distribuições cujos parâmetros foram estimados pela estimação de máxima verossimilhança. Esses parâmetros estão descritos nas tabelas 6 e 7. Vemos que ambos os ajustes feitos seguem razoavelmente bem os histogramas. As piores representações ficam para os histogramas multimodais observados para os dados de Download de ambos os dispositivos.

A figura 10 mostra qual seria o resultado dos ajustes caso os zeros não fossem desconsiderados. Temos que o caso do Chromecast é pouco alterado, com a distribuição gamma apresentando um *skew* maior para a esquerda, em direção aos valores nulos. Já o dataset da Smart TV é amplamente prejudicado, com a distribuição normal obtendo valores de σ^2 maiores do que o caso anterior e a distribuição gamma se concentrando completamente nos valores nulos. Vemos que as distribuições nesse caso seriam uma péssima representação dos dados observados. Isso se dá porque esses datasets refletem a união de dois comportamentos distintos: Dispositivos desligados (taxa nula) e dispositivos ligados (taxa não nula). Uma distribuição ajustada nesse dataset não seria capaz de representar esse comportamento de estados internos.

Os *Probability plots* das figuras 11 e 12 mostram o quão bem os datasets seguem as distribuições normal e gamma, respectivamente. De todos os 8 gráficos, o gráfico que mais se aproxima visualmente da linha de 45° é a curva de upload do Chromecast para a distribuição normal. Esse resultado pode ser comparado com o histograma: A região entre o percentil 80% e 95% da distribuição, a curva em formato de "S" observada no gráfico bate com a "menor coluna do meio", a coluna do 4 log(bps), no histograma. Exceto também pela coluna mais alta, quase todas as colunas do histograma ficam próximas à curva da normal. Portanto, isso indica que o Upload do Chromecast pode ser substituído por uma variável aleatória normal com os parâmetros iguais aos da tabela 6.

O QQPlot da figura 13 tenta determinar se o fluxo de download e upload entre os dispositivos segue a mesma distribuição. Nenhum dos gráficos é suficientemente próximo a curva de 45°, mas é possível observar que os datasets de upload tem uma similaridade maior do que os datasets de upload.

Portanto, as principais conclusões dessa seção são:

- Os horários de pico escolhidos estão na tabela 5
- A exclusão dos zeros é necessária nessa seção, já que se quer avaliar a distribuição resultante dos dispositivos que estão transmitindo dados. A não exclusão dos zeros resulta em um ajuste ruim de curvas.
- Os melhores ajustes de curva com base nos histogramas estão para os dados de Upload, já que os dados de download apresentam um comportamento multimodal.
- A melhor curva de *Probability Plot* é para o upload do Chromecast, e dado que a curva mantém-se na mesma direção ao longo de todo o gráfico, pode-se caracterizar esse dataset como uma variável aleatória normal.
- Com base no QQPlot, não é possível dizer que os datasets do Chromecast é similar ao da Smart TV, para nenhum dos dois casos.

5 Análise de correlação

Nessa seção, será feita uma análise de correlação entre as taxas de download e upload dos dispositivos. Nesse caso, a estrututra do dataset é diferente da utilizada até então: Ao invés da coluna categórica "Direção do fluxo" para dizer se o valor é de upload ou de download, há duas colunas: Uma para o valor da taxa de upload e outro para a taxa de download.

Nesse caso, é desejado avaliar a correlação incluindo todos os dados. Se a hipótese de que alguns dispositivos possuem taxa nula porque estão desligados, essas taxas devem ser nulas tanto para upload quanto para download, o que não vai prejudicar o resultado da correlação tanto quanto uma variável livre e outra nula fariam. Porém, múltiplas entradas do vetor $(0, 0)$ podem introduzir um viés que incrementa artificialmente o resultado de correlação de pearson. Portanto, para avaliar a estatística, ambas as alternativas (com e sem os zeros) serão avaliadas.

Nessa seção, decidiu-se por avaliar a correlação entre taxas de download e upload que ocorrem na mesma hora, já que não é possível associar taxas de upload e download que aconteceram em horários diferentes. Assim, já que o upload e o download do Chromecast acontecem em horários distintos, decidiu-se por avaliar a correlação em **três** datasets diferentes:

- Chromecast as 23hrs
- Chromecast as 22hrs
- Smart TV as 20hrs

5.1 Metodologia

Para remover os zeros do dataset, o dataset é filtrado de modo que, se pelo menos uma entrada de bps (download ou upload) é zero, toda a entrada é removida. Isso vai remover tanto os pontos $(0, 0)$ repetidos como qualquer ponto $(0, x)$ e $(0, y)$.

A correlação de pearson é calculada utilizando a biblioteca `scipy`, que retorna a estatística e o valor-p associado.

5.2 Resultados

Os resultados das estatísticas podem ser vistos nas tabelas 8 e 9, para resultados com valores nulos e sem valores nulos respectivamente. Já o scatter plot pode ser visto na figura 14.

Podemos ver através das tabelas 8 e 9 que a presença ou não das entradas nulas pouco influenciou no resultado da correlação, que indica uma correlação forte ($\rho > 0.7$) em todos os casos, com um valor-p menor que 7 casas decimais. O resultado positivo de todos os casos indica que o valor de upload tende a crescer quando o valor de download cresce, e vice-versa. Porém, esse resultado não nos dá mais informações a respeito do tipo de relação e nem a taxa de crescimento de uma variável em função da outra.

Com base no gráfico da figura 14, podemos observar que os dados apresentam uma correlação positiva com o tipo de correlação difícil de distinguir devido a quantidade excessiva de pontos. Pode-se dizer que o segundo e o terceiro gráfico aparentam possuir correlações lineares, mas o primeiro apresenta uma tendência de seguir um "S", que claramente é não linear. Vemos que o impacto das entradas nulas é observado devido aos dois conjunto de pontos que se apresentam em $(x, 0)$ e $(0, y)$, muito mais predominantes no caso do Smart TV do que no caso do Chromecast.

Hora	Tipo de dispositivo	ρ	p-valor
20	Smart TV	0.915609	0.000000
22	Chromecast	0.776742	0.000000
23	Chromecast	0.792504	0.000000

Tabela 8: Resultados da correlação de pearson sem exclusão de zeros

Hora	Tipo de dispositivo	ρ	p-valor
20	Smart TV	0.900466	0.000000
22	Chromecast	0.776669	0.000000
23	Chromecast	0.792728	0.000000

Tabela 9: Resultados da correlação de pearson com exclusão de zeros

Com isso, pudemos concluir nessa seção que

- Três datasets com horários de pico distintos devem ser analizados para fazer uma correlação justa (upload e download no mesmo horário).
- As taxas de upload e download possuem uma forte correlação positiva entre si durante horários de pico, independente se o dataset contém as entradas nulas ou não.
- Os gráficos scatter sugerem uma relação linear entre as entradas do Chromecast, mas uma relação não-linear (formato de "s") no caso da Smart TV.

6 G-Test

O G-Test é um teste de hipótese que avalia se dois datasets, com lista de frequências de observação, possuem distribuições similares. Nessa seção, queremos observar se as distribuições para as taxas de upload e download são similares entre um mesmo dispositivo nos horários de pico. Para isso, serão comparados os mesmos datasets da seção anterior.

6.1 metodologia

Para o cálculo do G-Test, é necessário criar duas listas de frequências, uma para cada distribuição, e compará-las no teste. Essa lista de frequências é equivalente ao resultado numérico de um histograma,

onde cada bin representa o intervalo e a contagem é a frequência de observações desse intervalo. Com isso, utilizou-se funções da biblioteca `numpy` para encontrar os intervalos dos bins apropriados segundo a regra de sturges para as taxas de download. Após isso, os mesmos intervalos dos bins são utilizados para fazer a contagem da taxa de upload e da taxa de download, e ambas as contagens são passadas para a função que calcula a estatística.

6.2 resultados

Os resultados da estatística e do valor-p podem ser vistos na tabela 10.

Hora	Tipo de dispositivo	Estatística	p-valor
20	Smart TV	464123.754381	0.000000
22	Chromecast	147134.065797	0.000000
23	Chromecast	139680.940307	0.000000

Tabela 10: Resultados G-Test

Como podemos ver, os resultados são bem diferentes entre si, e o valor-p é menor que 7 casas decimais. Portanto, todas as hipóteses que as distribuições são iguais foram rejeitadas. Esse resultado era esperado dado as observações das seções anteriores: Os histogramas entre as taxas de download e upload possuem formatos diferentes, onde uma é multimodal e a outra não.

Com isso, nessa seção podemos simplesmente concluir que todos os testes de hipótese foram rejeitados, ou seja, nenhuma distribuição de upload é similar a de download.

7 Considerações finais

Nesse trabalho, foram observados os dados de taxa de upload e download de dois tipos de dispositivos: Smart TVs e Chromecasts, que foram coletados por uma ISP de médio porte.

Através desses dados, foi possível observar diferenças no comportamento de upload e de download dos dispositivos, que resultam em distribuições diferentes; observar padrões de uso desses dispositivos que variam em função da hora do dia, observar correlações entre taxas e comparar o comportamento desse dataset com o comportamento de distribuições teóricas.

A análise sobre a exclusão ou não dos valores nulos trouxe uma discussão a respeito da interpretação dos dados e dos diferentes casos onde essa informação é relevante ou não é relevante. Algumas comparações foram feitas entre os datasets com e sem a presença de valores nulos e foi possível observar o impacto dessas amostras nos resultados. A presença ou não dos zeros só será definida ao definir um objetivo específico para a análise.

Figura 9: Histogramas **com** a exclusão de entradas nulas. A curva azul é a normal e a curva laranja é a gamma.

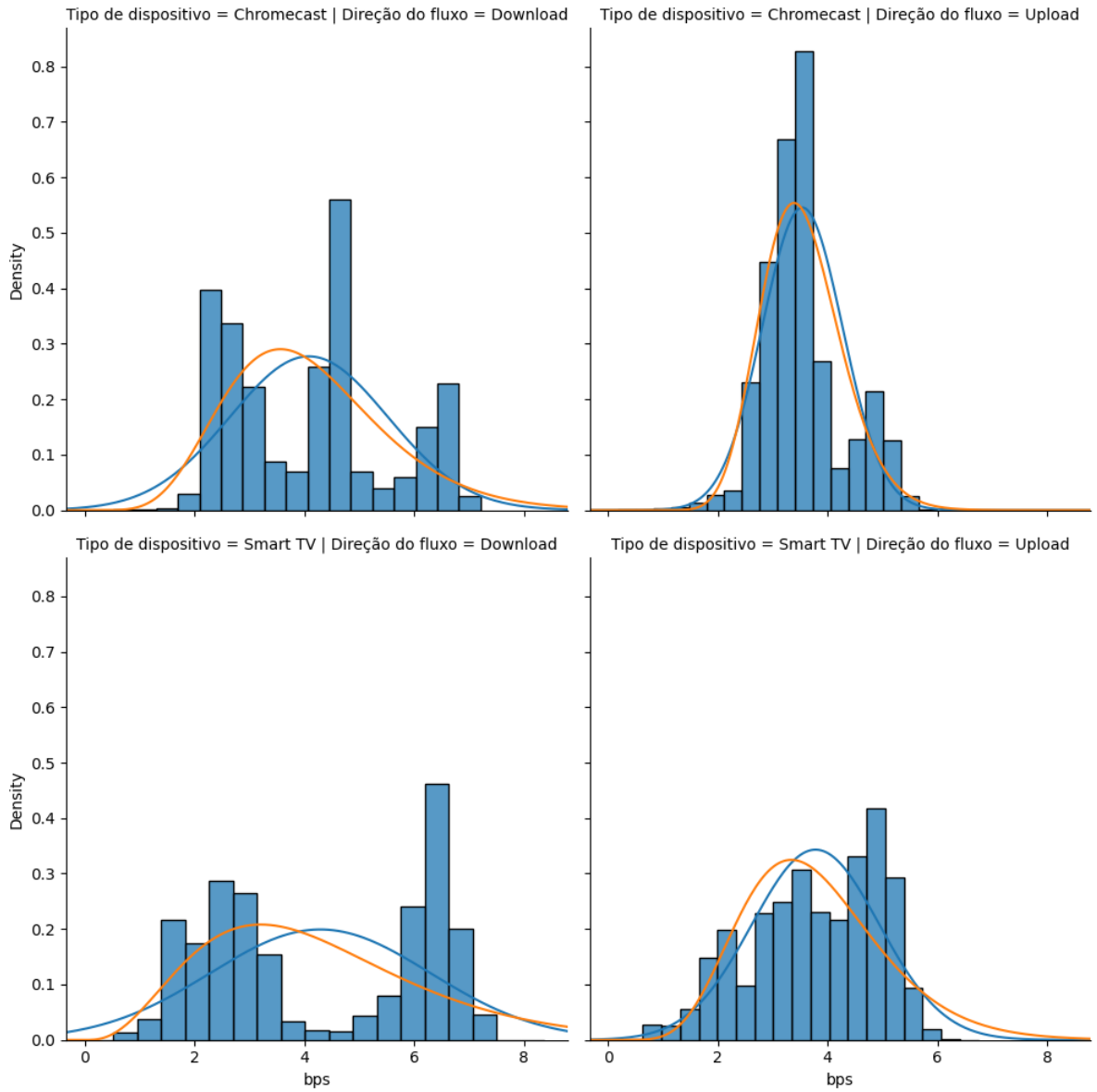


Figura 10: Histogramas **sem** a exclusão de entradas nulas. A curva azul é a normal e a curva laranja é a gamma.

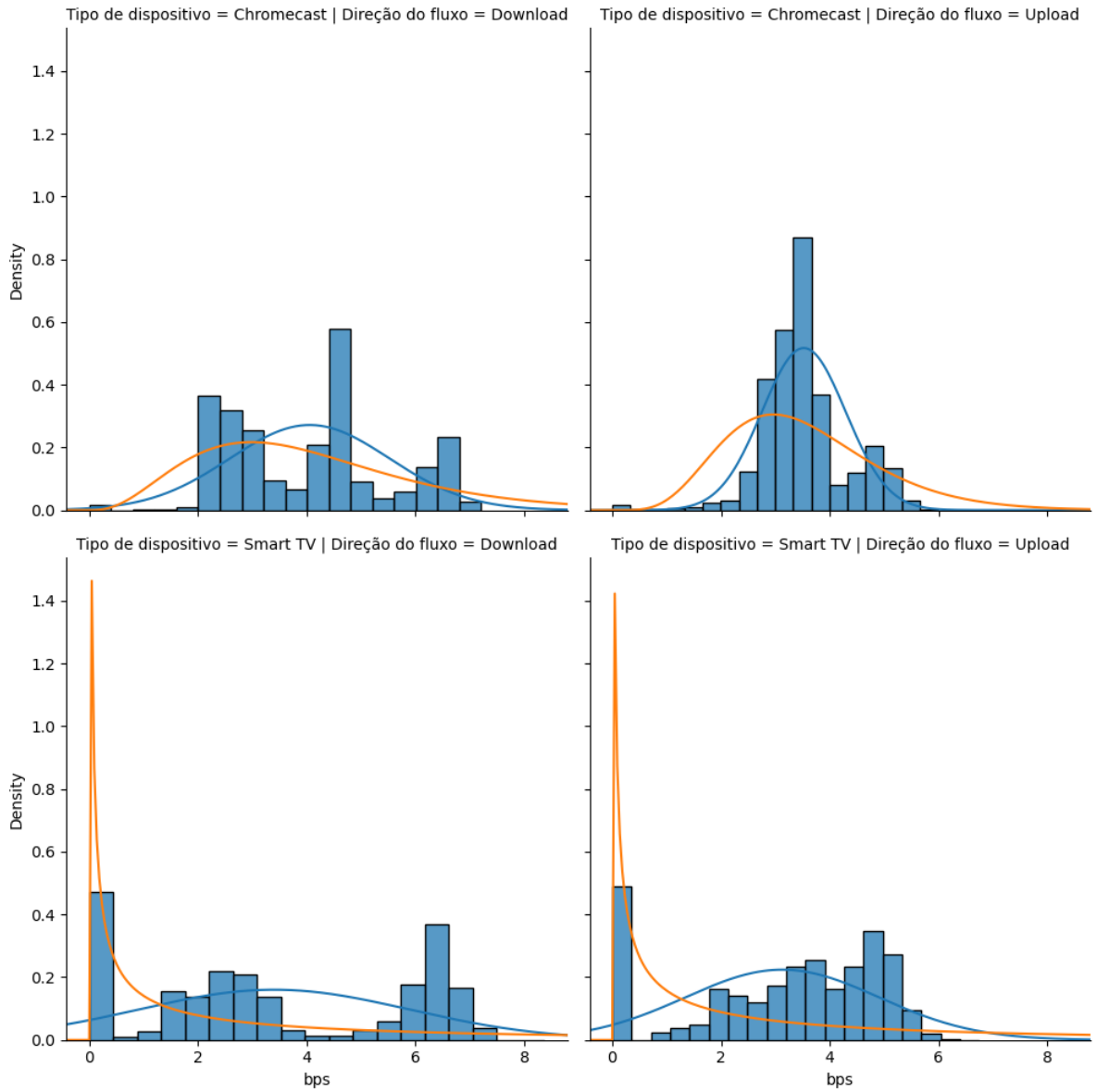


Figura 11: *Probability Plot* para a distribuição normal

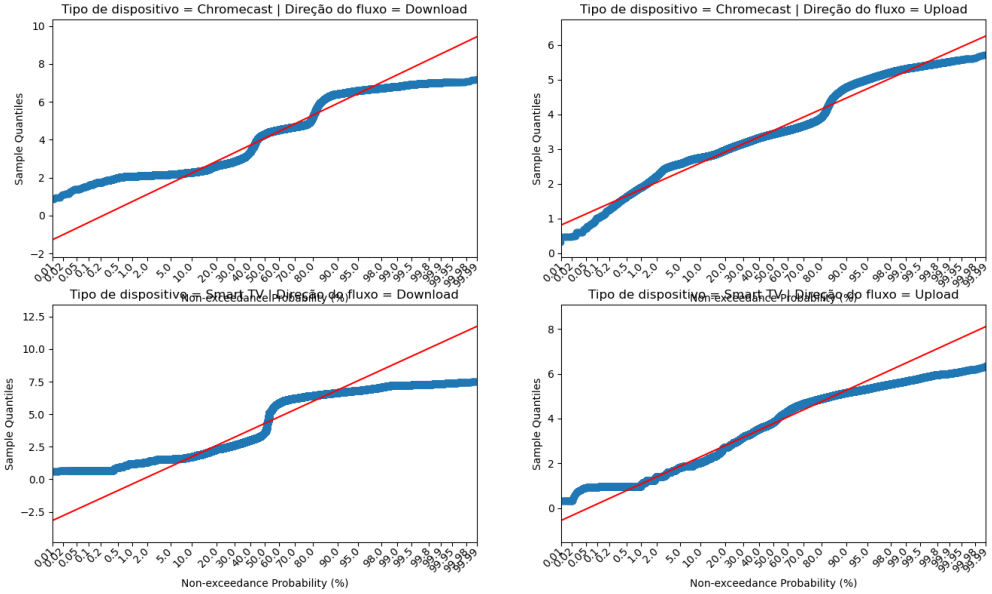


Figura 12: *Probability Plot* para a distribuição gamma

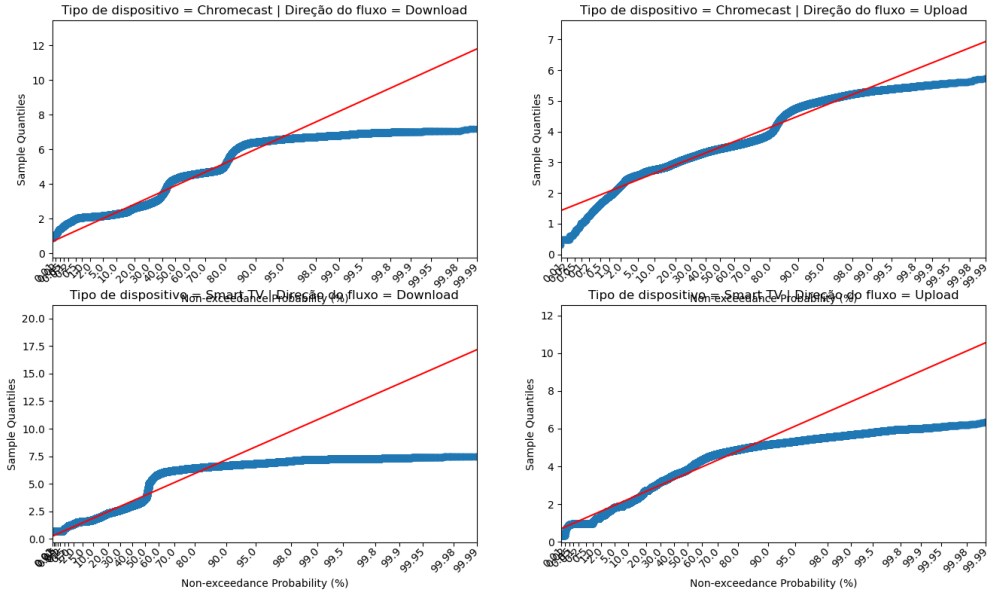


Figura 13: QQPlots

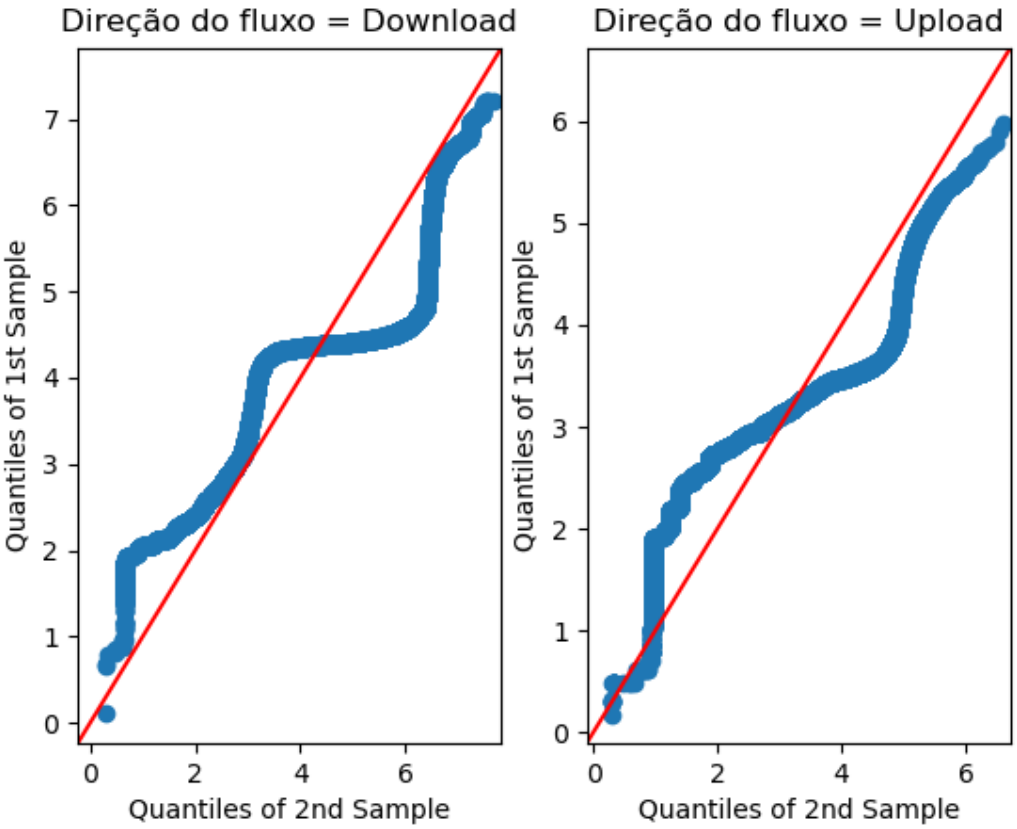
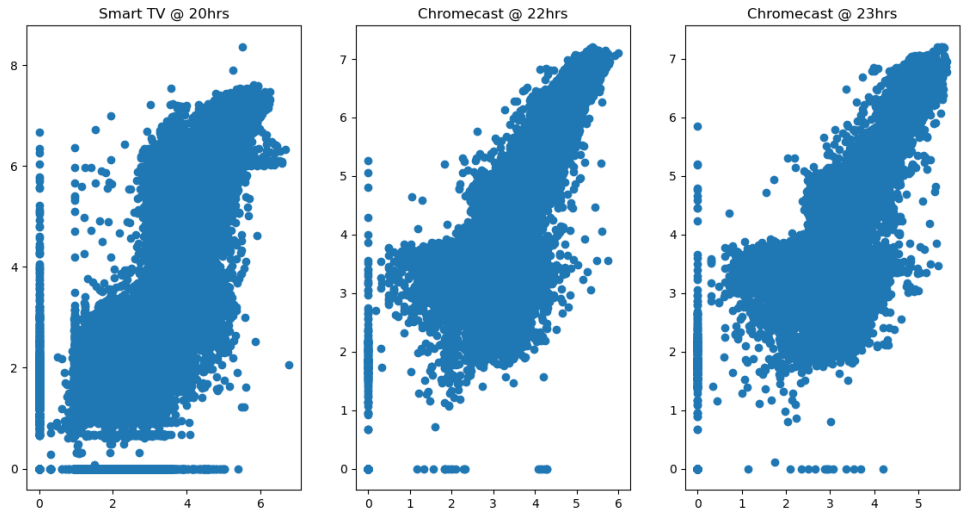


Figura 14: Gráficos de correlação entre os datasets sem exclusão de zeros



Apêndice 1 - Cálculo dos estimadores de máxima verossimilhança

Para uma sequência de V.A.s $X = [X_1, X_2, X_3]$ que representam amostras i.i.d de uma distribuição $f(x|\theta)$, temos que a função de verossimilhança pode ser escrita como

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

O estimador MLE é encontrado ao maximizar $\mathcal{L}(\theta)$:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \mathcal{L}(\theta)$$

1 - Normal com μ e σ desconhecidos

Temos a pdf da normal como

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Assim, $\mathcal{L}(\theta)$ é

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

Usamos o logaritmo de $\mathcal{L}(\theta)$ para calcular o valor máximo e obtemos

$$l(\theta) = \ln(\mathcal{L}(\theta)) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

Podemos agora calcular as derivados parciais $x_i^2 - 2x_i\mu + \mu_i^2$

$$\frac{\partial l}{\partial \hat{\mu}} = -\sum_{i=1}^{n} \frac{(2\hat{\mu} - 2x_i)}{2\sigma^2} = 0$$

$$\therefore \sum_{i=1}^{n} 2\hat{\mu} = \sum_{i=1}^{n} 2x_i \quad \therefore n\hat{\mu} = \sum_{i=1}^{n} x_i \quad \therefore \hat{\mu} = \frac{\sum x_i}{n}$$

Para σ, temos:

$$l(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{2}\frac{4\pi\sigma}{2\pi\sigma^2} - \frac{\sum}{2}\left(\frac{-2}{\sigma^3}\right) = 0$$

Assim, temos:

$$\frac{n}{\sigma} = \frac{\sum}{\sigma^3} \quad \therefore \sigma^2 = \frac{\sum}{n} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Onde trocamos μ por $\hat{\mu}_{MLE}$

2 - Destribução gamma

Temos a pdf da destribução gama como:

$$f(x|\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

Calculamos a verossimilhança como:

$$L(\theta) = \prod_{i=1}^{n} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} = \left(\frac{\lambda^{\alpha}}{\Gamma(\alpha)}\right)^{n} e^{-\lambda \sum_{n} x_i} \prod_{i=1}^{n} x_i^{\alpha-1}$$

O logaritmo da verossimilhança:

$$\ell(\theta) = n\alpha \ln(\lambda) - n \ln\left(\Gamma(\alpha)\right) - \lambda \underbrace{\sum_{i=1}^{n} x_i}_{S} + (\alpha-1) \sum_{i=1}^{n} \ln(x_i)$$

Calculamos as derivadas e igualamos a zero para obter os parâmetros:

$$\frac{\partial \ell}{\partial \lambda} = \frac{n\alpha}{\lambda} - S = 0$$

$$\frac{n\alpha}{\lambda} = S \therefore \hat{\lambda}_{MLG} = \frac{\hat{\alpha}}{\bar{x}}$$

Agora para α

$$\frac{\partial \ell}{\partial \alpha} = n \ln(\hat{\lambda}) - \frac{n}{\Gamma(\alpha)} \Gamma'(\alpha) + \sum_{i=1}^{n} \ln(x_i) = 0$$

Vamos chamar $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ de $\phi(\alpha)$, que é conhecida como a função digamma.

Assim, substituindo λ por $\frac{\alpha}{\bar{x}}$, temos:

$$n \ln\left(\frac{\hat{\alpha}}{\bar{x}}\right) - n\phi(\hat{\alpha}) + \sum_{i=1}^{n} \ln(x_i) = 0$$

$$n \ln(\hat{\alpha}) - n \ln(\bar{x}) - n \phi(\alpha) + \sum_{i=1}^{n} \ln(x_i) = 0$$

fatorando n, temos

$$\ln(\hat{\alpha}) - \ln(\bar{x}) - \phi(\alpha) + \overline{\ln(x_i)} = 0$$

Observação: Temos que não existe forma analítica para encontrar os valores dos parâmetros da distribuição gamma. Sendo assim, o código encontrará os valores dos parâmetros via MLE através de otimização numérica.