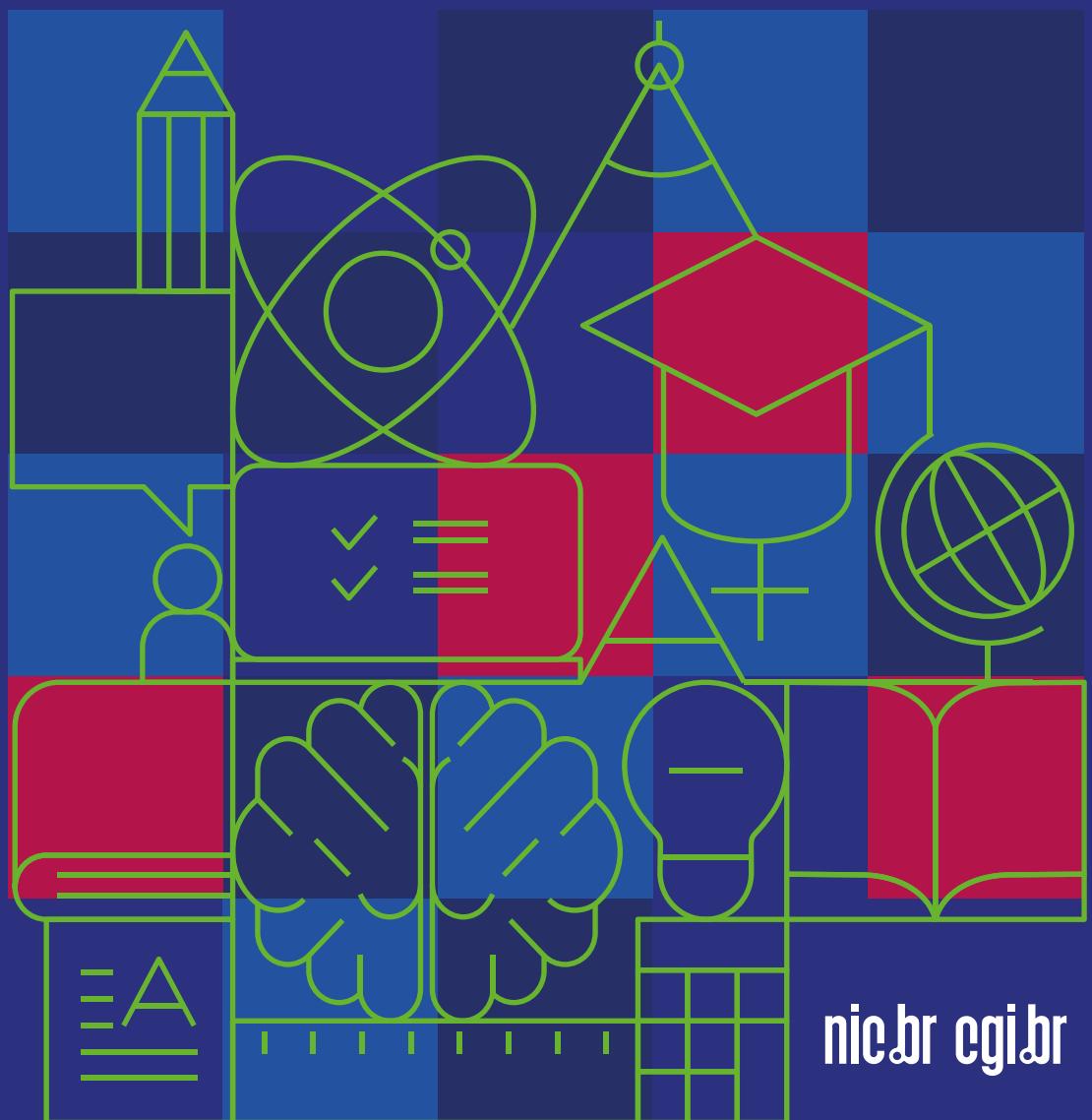


Medidor Educação Conectada (tecnologia SIMET)

DESENVOLVIMENTO DE CRITÉRIO DE AVALIAÇÃO DA VELOCIDADE DE INTERNET MEDIDA NAS ESCOLAS



nic.br egi.br

Medidor Educação Conectada (tecnologia SIMET)

DESENVOLVIMENTO DE CRITÉRIO DE AVALIAÇÃO DA VELOCIDADE DE INTERNET MEDIDA NAS ESCOLAS

2025

Núcleo de Informação e Coordenação do Ponto BR – NIC.br

Diretor Presidente: Demi Getschko

Diretor Administrativo: Ricardo Narchi

Diretor de Serviços e Tecnologia: Frederico Neves

Diretor de Projetos Especiais e de Desenvolvimento: Milton Kaoru Kashiwakura

Diretor de Assessoria às Atividades do CGI.br: Hartmut Richard Glaser

Créditos

Autoria: Cristiane Honora Millan, Gabriela de Lima Marin, Paulo Kuester Neto, Philipp Kleer

Agradecimentos

À equipe de Medições do Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações (Ceptro.br): Holger Araújo Wiehen, Erik Jhordan, Felipe Amorim de Melo, Henrique de Moraes Holschuh, João Paulo Vieira de Carvalho, Khloe Camargo, Leandro Dybal Bertoni, Leandro Ferrari, Paulo Kuester Neto, Cristiane Honora Millan, Gabriela de Lima Marin.

Sobre o Comitê Gestor da Internet no Brasil (CGI.br)

O Comitê Gestor da Internet no Brasil é responsável por estabelecer diretrizes estratégicas relacionadas ao uso e ao desenvolvimento da Internet no país. Coordena e integra todas as iniciativas de serviços de Internet, promovendo a qualidade técnica, a inovação e a disseminação dos serviços ofertados.

Sobre o Núcleo de Informação e Coordenação do Ponto BR (NIC.br)

O Núcleo de Informação e Coordenação do Ponto BR foi criado para implementar as decisões e os projetos do Comitê Gestor da Internet no Brasil, sendo responsável por coordenar e integrar as iniciativas e serviços da Internet no país.

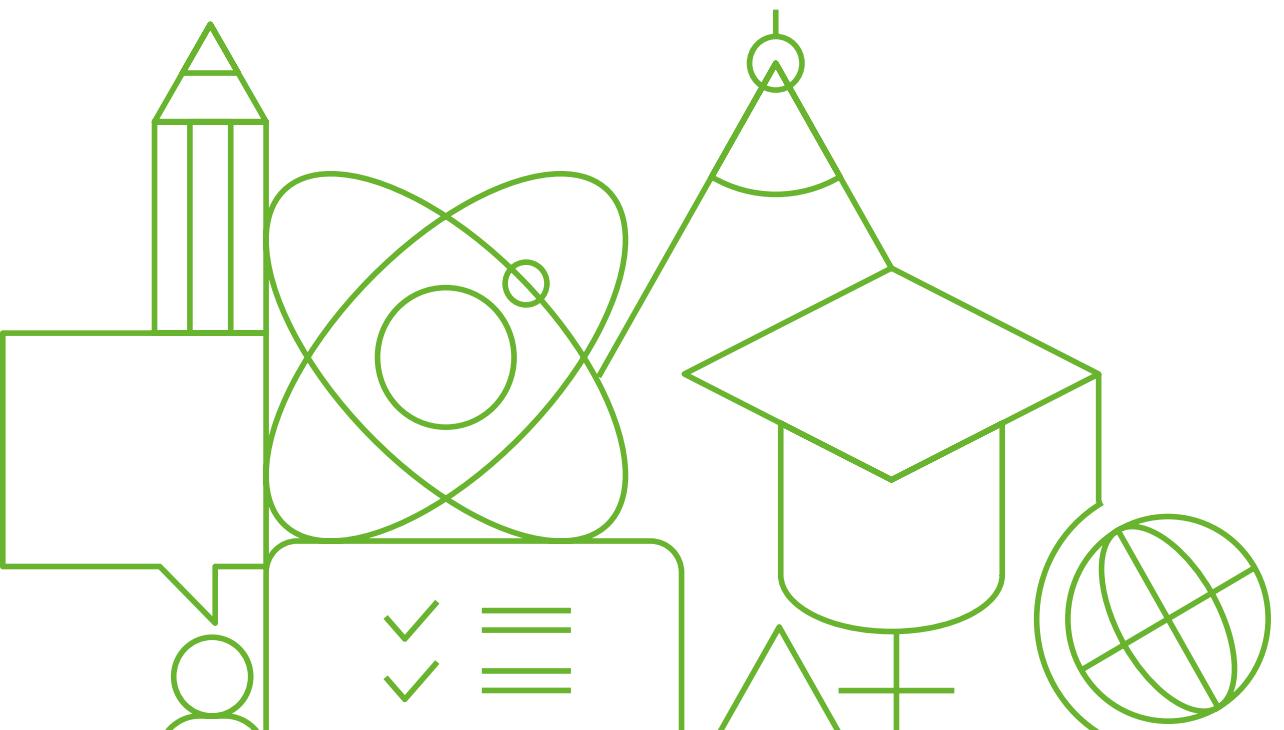
Sobre o Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações do NIC.br (Ceptro.br)

O Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações do NIC.br tem por objetivo desenvolver projetos voltados à melhoria da qualidade da Internet e à disseminação de seu uso, com especial atenção aos aspectos técnicos e de infraestrutura.

No Ceptro.br, o setor Medições é responsável pelo desenvolvimento da família de medidores do Sistema de Medição de Tráfego Internet (SIMET), que inclui soluções destinadas ao setor público, como o Medidor Educação Conectada e o SIMET Saúde. Essas medições constituem subsídio essencial para fomentar estudos, gerar análises e propor ações que resultem em políticas de conectividade mais eficazes.

ÍNDICE

Resumo Executivo	5
Introdução	6
Métodos e Resultados	8
Conclusão	23
Referências	24
Anexo	25



RESUMO EXECUTIVO

Para apoiar o Ministério da Educação (MEC) nas políticas de melhoria da conectividade na educação pública, o NIC.br fornece dados de conectividade escolar obtidos por meio do Medidor Educação Conectada. As medições revelam que a velocidade de *download* de Internet experimentada nas escolas pode, com frequência, ficar acima ou abaixo do plano contratado, variando consideravelmente entre escolas e faixas de planos. Essa variabilidade pode refletir tanto oscilações no desempenho da rede quanto mudanças nas condições de uso dentro da própria escola, como horários de maior demanda. Esses fatores reforçam a necessidade de um método capaz de representar, de forma consistente, a velocidade de *download* tipicamente disponível para os alunos e professores. O objetivo deste estudo foi desenvolver um indicador robusto, comparável e de fácil comunicação, baseado na métrica de velocidade de *download*, que resuma adequadamente as conexões utilizadas no ambiente escolar.

METODOLOGIA

Para avaliar qual indicador representa melhor a velocidade de *download* tipicamente disponível para alunos e professores, foram simuladas séries de medições com padrões semelhantes aos observados nas escolas, porém com a inserção proposital de flutuações e valores extremos (*outliers*). Essa abordagem permitiu comparar, de forma controlada, a precisão e a consistência dos diferentes métodos propostos para identificar medições atípicas em cada escola.

PRINCIPAIS ACHADOS

- **Percentil 95 (P95): o melhor equilíbrio.** Mostrou-se o indicador mais estável e representativo: reduz o efeito dos picos muito altos, mantém boa sensibilidade para captar mudanças reais na velocidade e é fácil de explicar e interpretar.
- **Percentil 99 (P99): tende a inflar os resultados.** Por ignorar apenas 1% dos maiores valores, costuma produzir velocidades acima do que a escola realmente experimenta no dia a dia. Além disso, é menos eficiente em sinalizar comportamentos realmente extremos.
- **Intervalo interquartil (IQR): bom, mas limitado.** Funciona bem quando a série de medições é estável. Porém, em escolas com grande oscilação de velocidade, torna-se conservador demais, acabando por descartar muitas medições reduzindo a representatividade do indicador.
- **Método adaptativo: útil, porém pouco padronizável.** Em alguns casos gera resultados próximos ao P95, mas sua lógica muda conforme o padrão da escola. Isso dificulta a comparação entre escolas e atrapalha a criação de um critério uniforme para políticas públicas.

CONCLUSÃO E RECOMENDAÇÃO

O percentil 95 (P95) mostrou-se a métrica mais adequada para representar a velocidade efetivamente disponível para alunos e professores nas escolas: captura o desempenho típico observado no uso cotidiano, minimiza o impacto de picos momentâneos e permite comparações justas entre escolas, redes e faixas de plano. Por esse conjunto de propriedades, o P95 se destaca como referência prática para definição de metas e para o monitoramento da conectividade escolar.

INTRODUÇÃO

O acesso à Internet de qualidade é condição essencial para que as Tecnologias da Informação e Comunicação (TIC) sejam plenamente incorporadas ao ambiente escolar, ampliando oportunidades de aprendizagem e viabilizando usos que exigem maior velocidade de *download* e baixa latência (plataformas educacionais *on-line*, videoaulas e aplicações colaborativas).

Para responder a essa necessidade, o governo brasileiro vem implementando políticas de universalização da conectividade nas escolas públicas e, em 2023, lançou a Estratégia Nacional de Educação Conectada (ENEC) (BRASIL, 2023), que define parâmetros de qualidade para garantir conectividade adequada ao uso pedagógico. Esses parâmetros foram formalizados pela Resolução CNEC nº 2 (DOU, 2024) e incluem, entre outros, a definição de velocidades ideais de conexão por aluno de forma a servir como orientação objetiva sobre qual plano de conexão à Internet a escola deveria contratar.

Para verificar a aderência das escolas a esses parâmetros, é indispensável medir a conexão de forma contínua e padronizada. Desde 2018, o NIC.br, em parceria com o Ministério da Educação (MEC), executa o projeto Conectividade na Educação, que realiza o monitoramento por meio do Medidor Educação Conectada (Tecnologia SIMET/CEPTRO.br, box 1 e figura 1). Instalado nas escolas, o medidor coleta automaticamente, em intervalos regulares, indicadores como velocidade de *download* e *upload*, latência, *jitter* e perda de pacotes, permitindo constatar objetivamente o atendimento (ou não) aos critérios da ENEC.

Análises dos dados coletados pelo Medidor Educação Conectada e dados declarados de plano contratado mostraram que é bastante comum ocorrerem tanto medições acima quanto abaixo da velocidade contratada. A frequência e a intensidade desses desvios variam conforme a faixa de plano, mas em todos os grupos observa-se uma ampla dispersão: escolas com o mesmo plano podem apresentar desempenhos muito distintos.

Esses resultados indicam que o valor declarado do plano nem

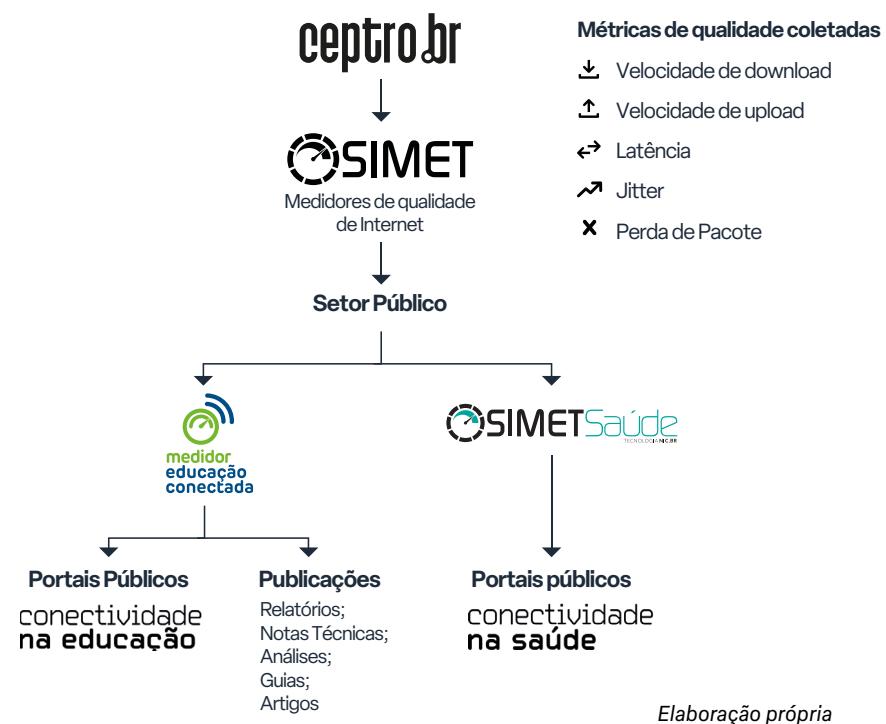
¹ Além do monitoramento, o NIC.br apoia a formulação de políticas públicas com guias técnicos e publicações como o Guia Conectividade na Educação (NIC.br, 2021), o Panorama da qualidade da Internet nas escolas públicas brasileiras (NIC.br, 2024) e outros (acesse em <https://medicoes.nic.br/publicacoes/>).

sempre reflete a velocidade experimentada pelos alunos. Fatores como renegociações contratuais, divergências entre o plano nominal e o serviço entregue, variações temporais de desempenho ou desconhecimento do gestor sobre o pacote vigente podem distorcer essa informação. Nesse contexto, as medições objetivas e automatizadas do Medidor Educação Conectada constituem uma fonte mais fidedigna e comparável para representar a experiência real de conectividade nas escolas.

Um dos principais desafios para estimar a Internet efetivamente disponível nas escolas está na diversidade de fatores que influenciam as medições. Elementos como a tecnologia de acesso utilizada (fibra, rádio, satélite, 4G etc.), oscilações no desempenho da rede e variações nas condições de uso ao longo do dia podem afetar momentaneamente a velocidade observada (Guia Conectividade na Educação, 2021). Por isso, é necessário adotar uma metodologia capaz de captar o comportamento típico da conexão, evitando que o indicador seja distorcido por oscilações pontuais ou medições extremas.

O objetivo deste estudo foi desenvolver um indicador robusto, comparável e de fácil comunicação, baseado na métrica de velocidade de *download*, que resuma adequadamente o desempenho das conexões no ambiente escolar.

Figura 1 – Estrutura das iniciativas do Ceptro.br/NIC.br voltadas ao setor público, desenvolvidas a partir de medições de qualidade da Internet feitas por meio do Sistema de Medição de Tráfego Internet (SIMET). As soluções Medidor Educação Conectada e SIMET Saúde realizam medições contínuas da qualidade da conexão em escolas e unidades de saúde, gerando dados, portais e publicações que subsidiam políticas de conectividade baseadas em evidências.



O SISTEMA DE MEDIÇÃO (SIMET)

O que é

O **SIMET (Sistema de Medição de Tráfego Internet)** é uma plataforma desenvolvida pelo **CEPTRO.br**, núcleo do **NIC.br**, para avaliar a qualidade da conexão de Internet no Brasil.

O que mede

Durante cada teste, o sistema coleta automaticamente:

- Velocidade de **download** e **upload** (protocolo TCP);
- Latência, *jitter* e perda de pacotes (protocolo UDP)
- Versão do IP (IPv4/IPv6), ASN e geolocalização (quando autorizada).

Versões do sistema

- **SIMETBox**: versão embarcada, realiza medições contínuas e automatizadas que não são afetadas por uso simultâneo do equipamento em atividade pedagógica ou administrativa.
- **Medidor Educação Conectada**: versão software utilizada no projeto Conectividade na Educação, que executa medições diretas no computador da escola, indicando a velocidade efetivamente disponível para os alunos.

MÉTODOS E RESULTADOS

A metodologia adotada neste estudo foi estruturada em três etapas complementares. Na primeira, buscou-se caracterizar a relação entre a velocidade de *download* contratada² e a velocidade medida nas escolas, avaliando tanto a ocorrência de medições acima ou abaixo do plano quanto a magnitude da variação entre escolas com faixas de contratação semelhantes.

Na segunda etapa, foram testados diferentes métodos de exclusão de medições atípicas (daqui em diante também chamadas de *outliers*) para cada escola. Para avaliar o desempenho desses métodos, foram simuladas séries de medições com características semelhantes às observadas nas escolas monitoradas, às quais foram adicionados valores extremos artificiais. Com base nessas simulações, os métodos foram avaliados quanto à sua capacidade de identificar corretamente os *outliers* sem gerar excesso de falsos positivos, permitindo comparar o equilíbrio entre sensibilidade e robustez de cada abordagem.

Na terceira etapa, o método escolhido foi aplicado às medições provenientes

² Informação sobre o plano contratados foi obtida por meio da parceria com a EACE (Entidade Administradora da Conectividade de Escolas) que, no contexto do projeto Aprender Conectado, promoveu a conexão e a instalação da SIMETBox (gratuita) em milhares de escolas.

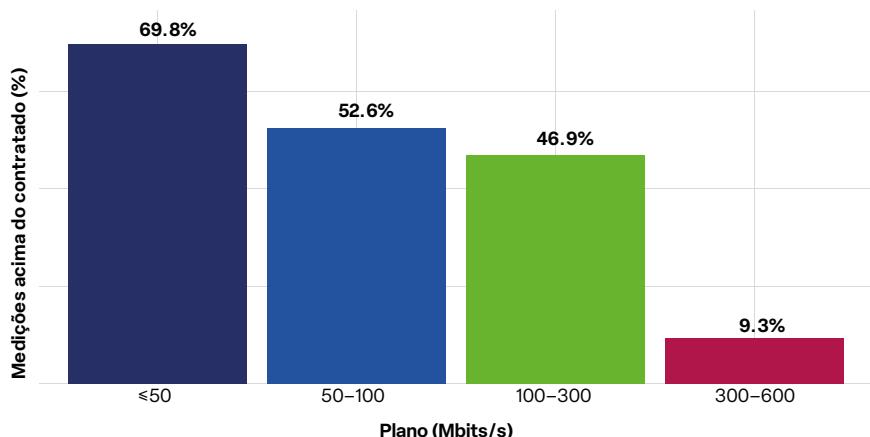
das escolas com SIMETBox para avaliar seu efeito sobre a estimativa de velocidade disponível, por escola, em cada faixa de plano contratado.

ETAPA 1: RELAÇÃO ENTRE A VELOCIDADE CONTRATADA E A VELOCIDADE DISPONÍVEL PARA ALUNOS E PROFESSORES

Na primeira etapa foi avaliado como a velocidade contratada se relaciona com a velocidade realmente disponível para alunos e professores, verificando se as medições refletem o desempenho esperado e identificando casos de entrega acima ou abaixo do plano. Também foi examinada a variação observada entre escolas com planos semelhantes.

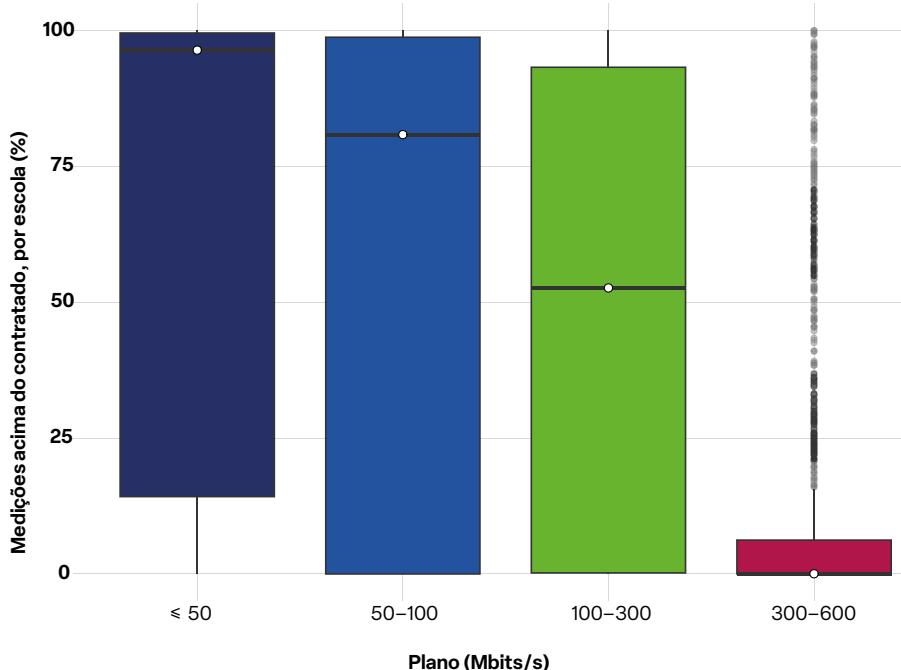
Para isso, foram utilizadas 6.816.861 medições provenientes de 7.423 escolas, realizadas entre 12/07/2022 e 28/10/2025 por meio das SIMETBox, dispositivos que permaneceram ligados de forma contínua — 24 horas por dia, 7 dias por semana — executando medições automáticas a cada quatro horas.

Figura 2 – Frequência de medições acima da velocidade contratada por faixa de plano.



Observou-se que a proporção de medições acima da velocidade contratada diminui conforme aumenta a faixa de velocidade do plano (figura 2). Nos planos de até 50 Mbps, 69,8% das medições superaram o valor contratado, indicando ampla margem de entrega superior. Essa proporção cai para 52,6% nos planos de 50 a 100 Mbps e para 46,9% nos planos de 100 a 300 Mbps. Já entre os planos de 300 a 600 Mbps, apenas 9,3% das medições ficaram acima do contratado. Esses resultados sugerem que, quanto menor o plano, maior a probabilidade de o usuário receber velocidades superiores às contratadas, possivelmente porque nas faixas mais baixas há maior folga técnica e de provisionamento das redes, enquanto nos planos mais altos o desempenho tende a se aproximar do limite nominal ofertado.

Figura 3 – Proporção de medições acima da velocidade contratada por escola e faixa de plano.



Nos planos de até 50 Mbps e de 50 a 100 Mbps, a maioria das escolas apresenta proporções de medições acima da contratada muito elevadas. A mediana próxima de 100%, o que significa que, em pelo menos metade dessas escolas, praticamente todas as medições registraram velocidades superiores ao plano contratado (figura 3).

Nos planos de 100 a 300 Mbps, a mediana cai para cerca de metade das medições acima do contratado, indicando maior equilíbrio entre entregas superiores e inferiores ao valor contratado. Já nos planos mais altos, entre 300 e 600 Mbps, a proporção de medições acima do contratado é baixa e bastante dispersa entre escolas, sugerindo que, nessa faixa, o desempenho tende a se aproximar ou ficar ligeiramente abaixo da velocidade contratada. Esse comportamento reforça que as distribuições de velocidade variam de forma relevante dentro e entre escolas, o que torna essencial compreender como essas distribuições se estruturam antes de definir o método mais adequado para exclusão de valores atípicos.

ETAPA 2: COMPARAÇÃO DE DIFERENTES MÉTODOS DE EXCLUSÃO DE MEDIÇÕES ATÍPICAS (*OUTLIERS*)

A segunda etapa concentrou-se na caracterização estatística das medições e na comparação entre métodos de identificação e exclusão de *outliers*, com o objetivo de encontrar o procedimento mais adequado para resumir de forma robusta as distribuições observadas. A análise incluiu uma amostra de 3.200 escolas públicas equipadas com o Medidor Educação Conectada (SIMET, versão *software*), cada uma com pelo menos 1.000 medições coletadas entre janeiro de 2024 e julho de 2025, garantindo representatividade.

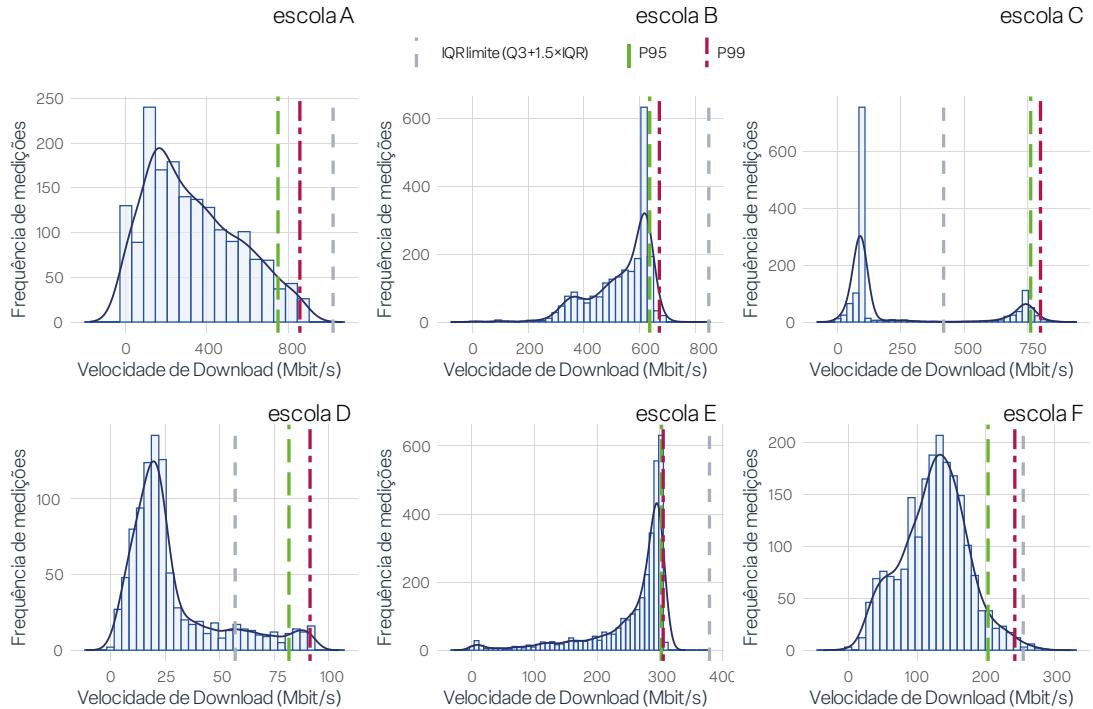
Para cada escola da amostra foram calculadas estatísticas descritivas das medições de velocidade de *download* (tabela 1), permitindo caracterizar a forma da distribuição, e identificar padrões de assimetria ou dispersão. Com essa análise, verificou-se que as escolas apresentam diferentes formatos de distribuição de velocidades de *download* (figura 4).

Tabela 1 – Definição das métricas estatísticas utilizadas na caracterização das distribuições de velocidade de download.

Métrica	Definição
Média	Valor obtido pela soma de todas as medições dividido pelo número total de observações. ³
Mediana	Valor central da distribuição, que separa a metade superior da metade inferior das observações.
Desvio padrão	Medida de dispersão que expressa o quanto as medições se afastam, em média, da média aritmética.
Intervalo interquartil (IQR)	Diferença entre o terceiro quartil (Q3, 75%) e o primeiro quartil (Q1, 25%), representando a variação da porção central dos dados.
Percentil 70 (P70)	Valor abaixo do qual estão 70% das observações.
Percentil 90 (P90)	Valor abaixo do qual estão 90% das observações.
Percentil 95 (P95)	Valor abaixo do qual estão 95% das observações.
Percentil 99 (P99)	Valor abaixo do qual estão 99% das observações.

³ Cada observação equivale à uma medição.

Figura 4 – Velocidades de download medidas em seis escolas monitoradas pelo medidor Educação Conectada. O eixo horizontal representa a velocidade de download (Mbit/s) e o eixo vertical, a frequência de medições em cada faixa de velocidade. As barras cinzentas indicam quantas medições ocorreram dentro de cada intervalo de velocidade, enquanto a curva azul suavizada mostra o formato geral da distribuição.



Os resultados revelam ampla diversidade de formas de distribuição de velocidade entre as escolas. Observam-se casos de alta assimetria, com pico concentrado perto do limite superior e cauda curta, o que torna P95 e P99 muito próximos (por exemplo, na escola E). Há perfis unimodais com cauda longa à direita, nos quais a densidade diminui gradualmente e os percentis superiores são mais distantes entre si (escolas A e D); distribuições mais simétricas e próximas da normal (escola F); e situações claramente bimodais, com dois picos distintos separados por um vale (escola C). Também aparecem distribuições fortemente concentradas em valores baixos, com queda rápida da frequência conforme a velocidade.

Os marcadores verticais na figura 4 ajudam a ilustrar essa variedade de formas de distribuição das medições de velocidade das escolas: quando P95 e P99 são semelhantes, há pouca dispersão em torno do patamar dominante; quando P95 fica muito abaixo do P99, a cauda é longa e os percentis capturam essa extensão; o limite IQR ($Q3+1,5\times IQR$) tende a aparecer ainda mais à direita.

Em cenários bimodais, os percentis se alinham para o pico de maior velocidade, enquanto uma parcela relevante das medições permanece concentrada no outro pico, dificultando a síntese por um único número.

Essas evidências mostram que a diversidade de formatos exige um método para lidar com valores extremos e resumir de forma confiável o comportamento típico da conexão.

AVALIAÇÃO DE MÉTODOS DE DETECÇÃO DE MEDIÇÕES ATÍPICAS

Para comparar diferentes métodos de detecção de *outliers*, adotou-se uma abordagem baseada em simulações que permitiu testar o desempenho dos métodos propostos em cenários controlados e com padrões semelhantes aos observados nas medições reais. O procedimento envolveu: (i) geração de séries de medições simuladas e com características semelhantes às das escolas monitoradas, (ii) injeção controlada de *outliers* e (iii) avaliação do desempenho de cada método proposto para a detecção desses *outliers*.

Simulação de Dados

A escolha das distribuições de dados para a simulação se baseou na sua capacidade de representar a diversidade de padrões observados nas medições reais, incluindo cenários com comportamento previsível, assimetria e variação acentuada. São elas: Normal (Gauss), Bimodal, Lognormal e Beta.

Para cada distribuição, foram geradas amostras de tamanho fixo (1000 observações). A descrição detalhada de cada distribuição, bem como os parâmetros utilizados nas simulações, encontra-se apresentada no anexo.

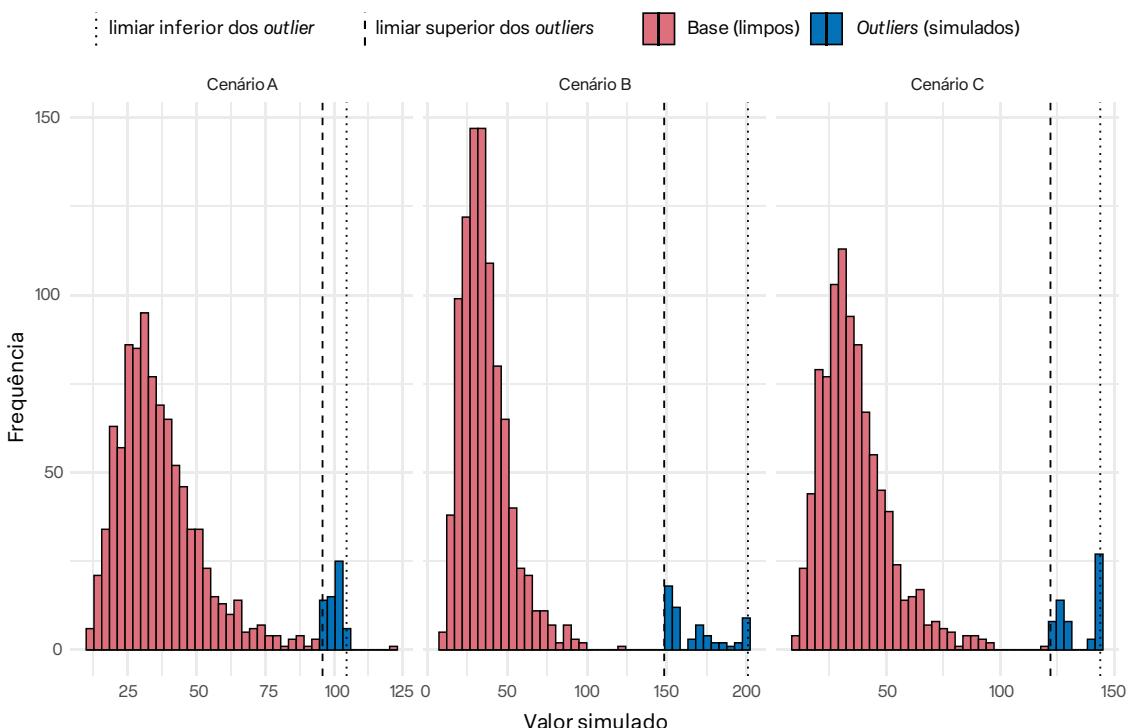
Injeção de *outliers*

Após a geração de cada conjunto de dados base, uma pequena porcentagem (entre 1 e 5%) de valores discrepantes foi injetada de forma controlada. Para garantir que esses valores fossem genuínos *outliers* e não apenas parte da cauda natural da distribuição, sua geração foi vinculada às propriedades estatísticas do próprio conjunto de dados “limpo”. Para isso calculou-se os valores de Q1, Q3 e o IQR (Q3-Q1). Em seguida definiu-se um limiar inicial (L) na cauda superior definido por: $L = Q3 + k \times IQR$, em que o valor de k é sorteado dentro de um intervalo pré-definido. A partir desse limiar, estabeleceu-se ainda um limite superior $L_{sup} = L + m \times IQR$ em que m controla a largura da faixa onde os *outliers* serão sorteados. Os valores discrepantes foram então gerados dentro do intervalo entre L e L_{sup} . Essa abordagem assegura que os *outliers* sejam gerados a uma distância estatisticamente significativa da maioria dos dados, posicionando-os de forma consistente na cauda da distribuição, independentemente do formato do conjunto de dados. Os *outliers* foram posicionados seguindo dois tipos de padrões:

- **Outliers próximos ao corte:** valores que aparecem logo acima do limiar estatístico, formando um pequeno agrupamento próximo à borda da distribuição original (figura 5, Cenário A). Esses pontos representam medições que estão um pouco acima do comportamento habitual, mas ainda relativamente próximas da distribuição típica.
- **Outliers distantes do corte:** valores muito afastados da distribuição principal, situados em regiões onde praticamente não há dados originais (figura 5, Cenários B e C). Esses pontos simulam medições claramente anômalas, que se destacam por estarem muito acima do padrão observado.

Assim, foram geradas diferentes intensidades de anomalias (distribuições com medições atípicas), desde desvios leves até valores extremamente discrepantes, refletindo melhor a variedade de comportamentos observados em medições reais e possibilitando uma avaliação mais abrangente dos métodos de detecção.

Figura 5 – Distribuição de valores simulados em três cenários, destacando a inclusão de outliers. As barras em cor rosa representam os dados originais (base com dados ‘limpos’), enquanto as barras em azul correspondem aos outliers injetados. A linha tracejada indica o ponto a partir do qual os valores passam a ser considerados fora do padrão, e a linha pontilhada marca o limite máximo estabelecido para esses valores em cada cenário.



Métodos de detecção de outliers avaliados

Quatro estratégias de definição de limites superiores foram aplicadas às distribuições simuladas:

1. **P95** – valores acima do percentil 95.
2. **P99** – valores acima do percentil 99.
3. **IQR** – limite clássico definido como $Q3 + 1,5 \times IQR$.
4. **Adaptativo⁴** – regra dinâmica baseada na assimetria da distribuição.

Nesta etapa de simulação, os métodos foram empregados apenas para sinalização: os *outliers* injetados não foram removidos, apenas rotulados como *outliers*. Essa escolha metodológica permitiu isolar o desempenho de cada método, analisando sua capacidade de detectar valores extremos sem alterar a estrutura dos dados.

Avaliação de Desempenho

Em cada réplica simulada, os valores identificados como *outliers* por cada um dos métodos foram comparados aos rótulos verdadeiros — isto é, aos *outliers* incluídos no processo de simulação. O desempenho foi avaliado a partir dos seguintes indicadores:

- **Recall/sensibilidade (Taxa de Verdadeiros Positivos):** indica a capacidade do método de identificar os *outliers* reais — quanto maior, melhor a sensibilidade.
- **FPR (Taxa de Falsos Positivos):** mostra a proporção de medições típicas incorretamente classificadas como *outliers*.
- **Precisão:** reflete a confiabilidade das sinalizações, isto é, quantas das observações marcadas como *outliers* realmente o são.
- **F1-score:** combina precisão e recall em uma única métrica, expressando o equilíbrio entre detectar muitos *outliers* e evitar falsos alarmes

Esse procedimento foi repetido em múltiplas reamostragens (*bootstrap*), permitindo comparações mais robustas entre os métodos, em diferentes distribuições e níveis de contaminação por *outliers*. O uso de reamostragem é consolidada na prática de avaliação de classificadores (Fawcet, 2006).

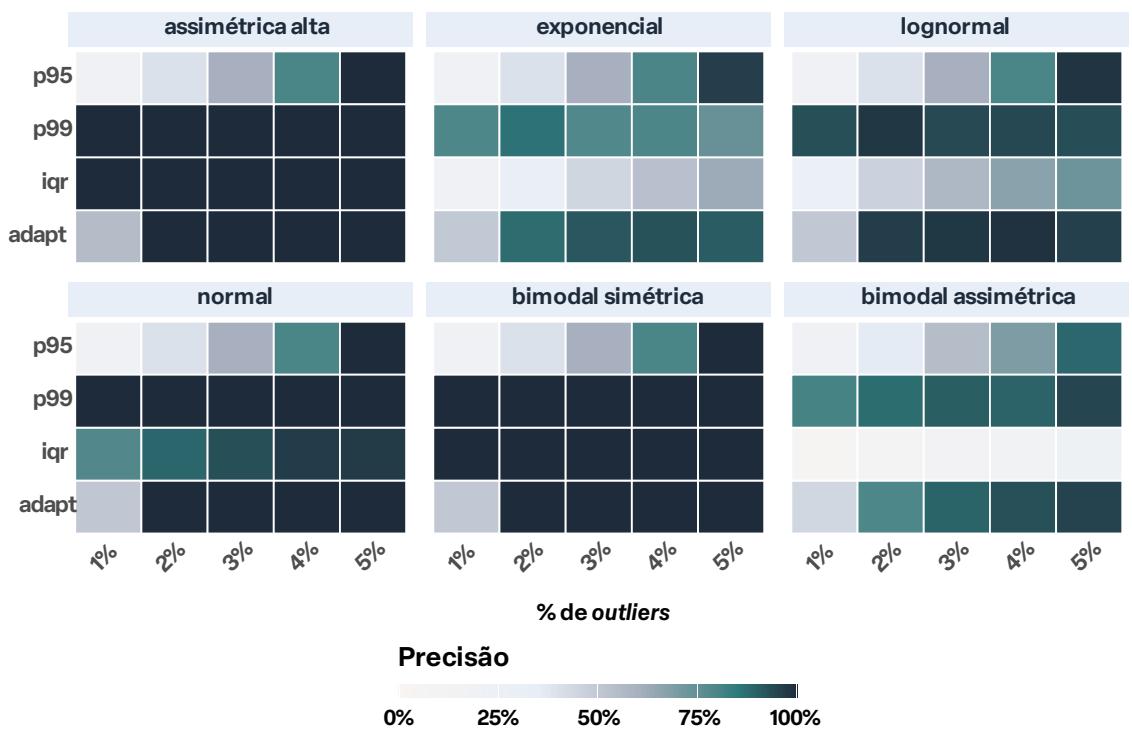
⁴ O método adaptativo ajusta automaticamente o critério de corte conforme a forma da distribuição e o tamanho da amostra. Em vez de usar um único limiar fixo (ex.: P95 ou P99), ele seleciona — ou combina — regras robustas (baseadas em IQR e faz tratamentos específicos quando há multimodalidade (duas ou mais modas).

Figura 6 – Matriz de confusão para avaliação da performance dos métodos na detecção de *outliers*. Os quadrantes em azul reúnem os casos cujo rótulo verdadeiro é outlier – VP (outlier sinalizado) e FN (outlier não sinalizado). Os quadrantes em rosa representam os casos cujo rótulo verdadeiro é não-outlier – FP (não-outlier sinalizado) e VN (não-outlier não sinalizado). As caixas em azul escuro indicam as métricas: precisão = $VP / (VP + FP)$, que expressa a confiabilidade da sinalização, e recall = $VP / (VP + FN)$, que mede a cobertura.

		Rótulo do dado simulado		
		outlier	não-outlier	
Sinalizador	sinalizado	Verdadeiro Positivo VP (outlier sinalizado)	Falso Positivo FP (não-outlier sinalizado)	Precisão = $VP / VP + FP$ (confiabilidade da sinalização)
	não sinalizado	Falso Negativo FN (outlier não sinalizado)	Verdadeiro Negativo VN (não-outlier não sinalizado)	
Recall = $VP / VP + FN$ (cobertura da sinalização)				

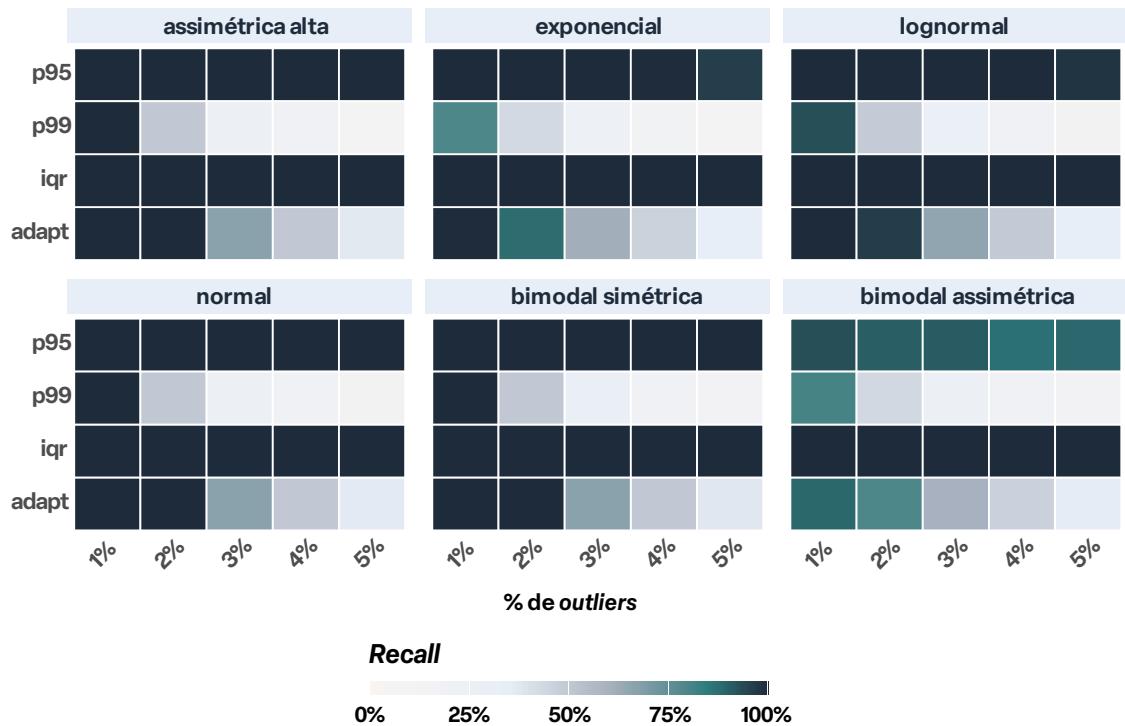
Considerando-se a precisão média (figura 7), em distribuições simétricas, como a normal e a bimodal simétrica, todos os métodos convergem para valores altos, pois há pouca ambiguidade sobre o que constitui um valor extremo. À medida que se introduzem caudas longas, como nas distribuições log-normal e exponencial, emergem diferenças relevantes: o P95 preserva precisão elevada e relativamente estável, detectando os *outliers* reais (verdadeiros positivos) na parte superior da distribuição sem inflacionar falsos positivos; o P99 torna-se extremamente seletivo, o que lhe garante, em vários cenários, uma precisão ainda maior, porém às custas da não sinalização de muitos *outliers*, o que aumenta os falsos negativos; o IQR assume postura conservadora, sinaliza poucos pontos e, por consequência, tende a manter alta a razão de acertos entre os sinalizados, mas ao mesmo tempo falha em marcar uma parcela relevante de *outliers* em distribuições com cauda longa. No entanto uma avaliação centrada apenas na métrica de precisão é insuficiente, pois um método pode parecer preciso simplesmente por sinalizar poucas observações com sendo outliers. Portanto, para uma avaliação mais robusta há necessidade de observar também o recall.

Figura 7 – Precisão média dos métodos de detecção de outliers em função da proporção de contaminação simulada (1% a 5%).



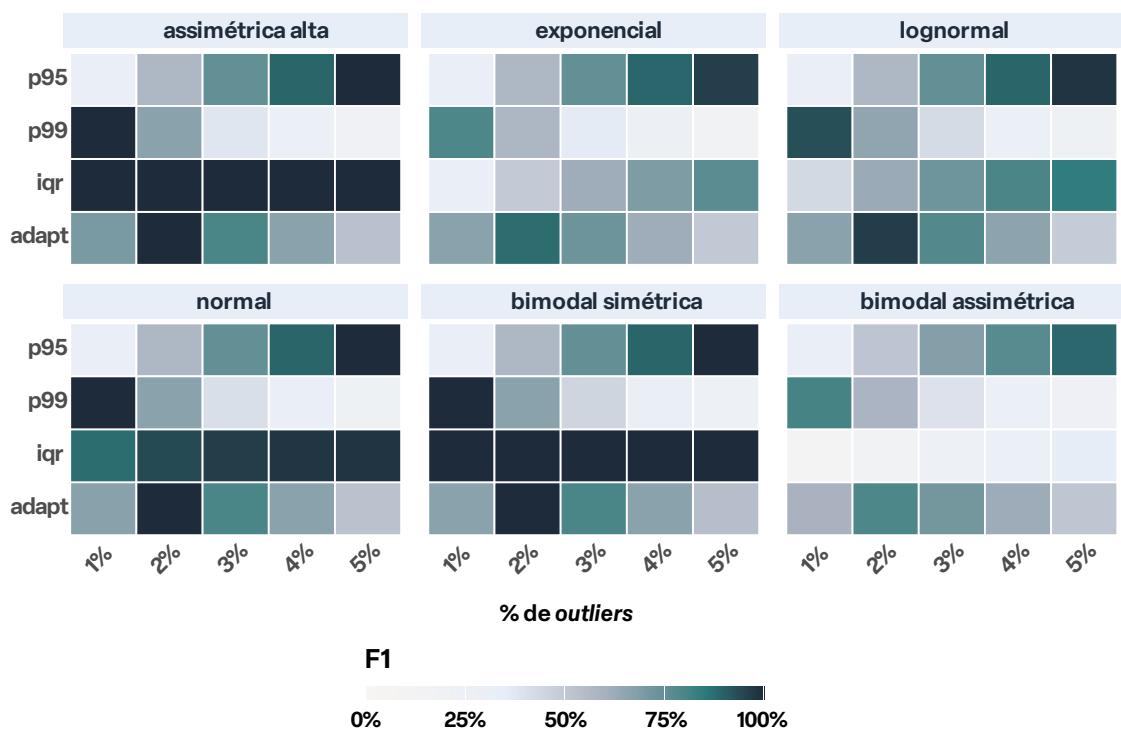
O indicador do *recall* (figura 8) revela a capacidade de um método capturar efetivamente os outliers quando eles existem. Nessa dimensão, o IQR tende a apresentar coberturas muito altas, sobretudo em distribuições assimétricas, pois sinaliza uma faixa maior da cauda de valores extremos; entretanto, esse ganho de sensibilidade costuma ser acompanhado de perdas em precisão, uma vez que aumentam os falsos positivos. O P99, por cortar apenas o topo da cauda, apresenta *recall* sistematicamente baixo, o que o torna pouco adequado quando se deseja um valor de referência que reconheça a fração superior praticável do desempenho. O P95 posiciona-se em um ponto intermediário: captura a maior parte dos outliers relevantes sem alargar demais a zona de sinalização, preservando, portanto, um patamar de confiabilidade compatível com o uso como indicador.

Figura 8 – *Recall* médio dos métodos de detecção de *outliers* em função da proporção de *outliers* injetados nas distribuições simuladas (1% a 5%).



Em distribuições com cauda longa e com percentual de *outliers* mais altos (o padrão de distribuição de medições mais frequente nos dados empíricos das escolas) o P95 mantém valores de F1 consistentemente altos e, na maioria dos cenários, supera tanto o IQR (que sacrifica precisão para ganhar sensibilidade) quanto o P99 (que perde demasiada sensibilidade). Em distribuições simétricas, os métodos praticamente empatam, o que confirma que o ganho do P95 se manifesta exatamente onde é necessária uma métrica robusta à assimetria. Em distribuições bimodais, especialmente quando existe um segundo pico bem mais alto, o P95 continua apresentando um desempenho sólido, conseguindo equilibrar a capacidade de detectar *outliers* com a cautela para não exagerar nos falsos alarmes. Nesses mesmos cenários, o P99 acaba deixando muitos *outliers* reais passarem despercebidos (perde cobertura), enquanto o IQR tende a marcar pontos demais como anômalos, aumentando desnecessariamente os falsos positivos.

Figura 9 – F1 médio dos métodos de detecção de *outliers* em função da proporção de contaminação simulada (1% a 5%).



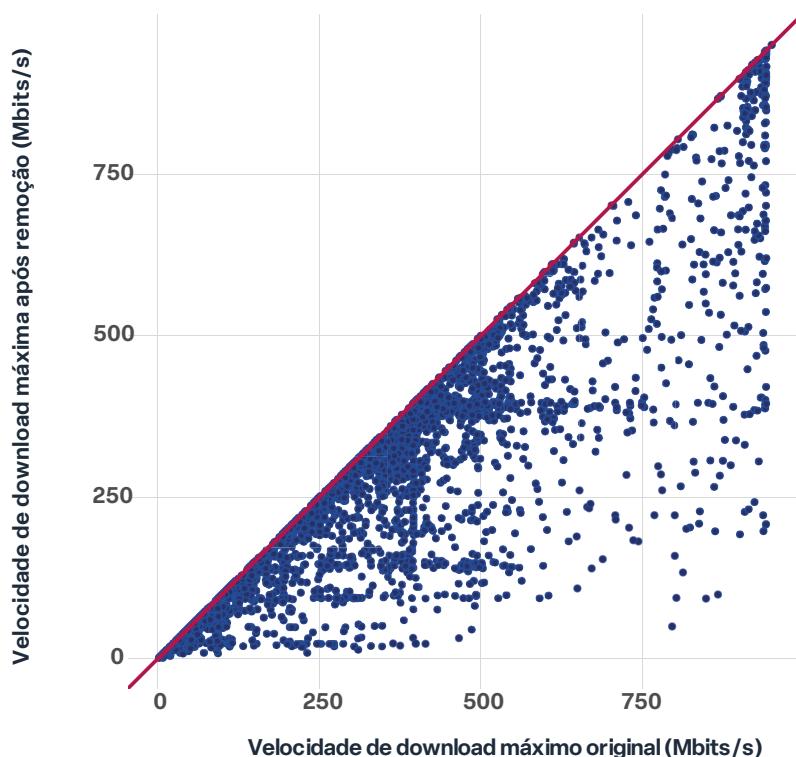
Ao utilizar o percentil 95 como critério de exclusão, algumas situações podem levar à retirada de valores legítimos. Em distribuições muito simétricas e concentradas, o P95 tende a ficar próximo do máximo normal, de modo que cortar acima dele pode eliminar observações que representam apenas variações esperadas. Em distribuições bimodais, com dois picos bem definidos, o P95 pode cair dentro do segundo pico e acabar descartando parte de um grupo que não é, na prática, *outlier*. Em ambos os casos, ainda que alguns valores legítimos sejam perdidos, essa escolha é coerente com a estratégia de priorizar robustez e evitar superestimações: é preferível subestimar ligeiramente o patamar de referência do que correr o risco de inflá-lo por valores ocasionais ou concentrações atípicas.

Esses resultados sustentam um encadeamento lógico. Ao estimar a velocidade de referência a ser associada à experiência dos alunos, pretende-se um valor que represente a porção superior praticável da distribuição, sem ser dominado por observações raras ou circunstanciais. A métrica precisa ser estável diante da heterogeneidade entre escolas, comparável ao longo do território e de fácil comunicação para fins de monitoramento de políticas públicas. O P95 satisfaz simultaneamente esses requisitos: responde adequadamente a caudas longas, evita superestimações em cenários com poucos valores muito altos e mantém desempenho equilibrado quando se considera, ao mesmo tempo, o acerto das sinalizações e a capacidade de cobertura.

ETAPA 3: APLICAÇÃO DO P95 NAS MEDIÇÕES SIMETBOX

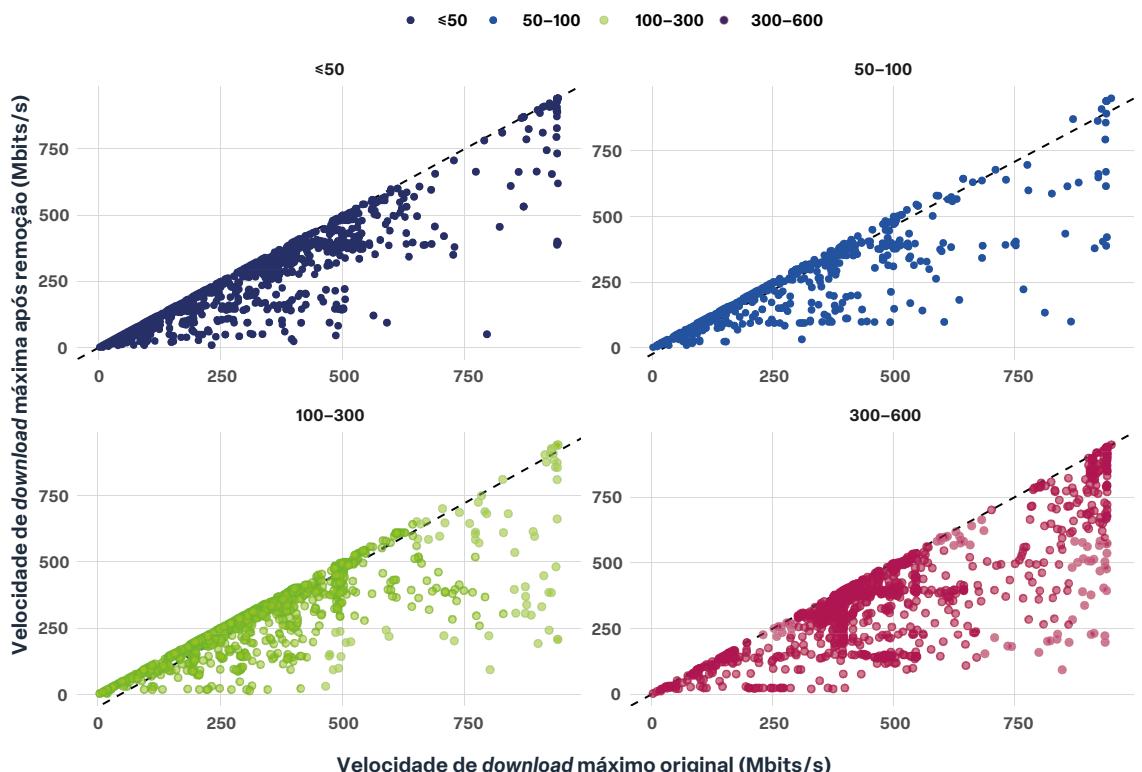
Nesta etapa, método do percentil 95 (P95) foi aplicado às medições originais do SIMETBox para avaliar seu efeito sobre a estimativa de velocidade disponível, por escola. Ao substituir o máximo simples pelo P95 por escola, buscou-se uma estimativa mais robusta e representativa do desempenho habitual da rede, menos sujeita a distorções causadas por medições atípicas e, portanto, mais adequada para diagnósticos operacionais e comparações entre escolas.

Figura 10 – Efeito da aplicação do P95 por escola sobre a estimativa de velocidade disponível. Relação entre o valor máximo original e o valor obtido após aplicar o corte no percentil 95 (P95) de cada escola. Pontos na reta rosa diagonal indicam medições estáveis, em que o P95 coincide com o máximo original. Pontos abaixo da linha representam escolas cujo valor máximo diminuiu após a remoção dos valores acima do P95— quanto mais distante da linha, maior a influência de medições excepcionalmente altas sobre a estimativa inicial.



Os resultados indicados na figura 10 mostram que a maioria das escolas apresenta redução do valor máximo após a aplicação do corte no percentil 95 (P95), evidenciada pela concentração de pontos abaixo da linha diagonal. Esse padrão indica que, em grande parte dos casos, as medições originais continham picos isolados de velocidade que inflavam artificialmente a estimativa de desempenho máximo. Em algumas escolas, a diferença entre o valor original e o P95 é expressiva, chegando a reduções da ordem de centenas de Mbits/s, o que reforça a ocorrência de *outliers* de alta magnitude. Por outro lado, há um subconjunto de escolas cujos pontos permanecem próximos à diagonal, sugerindo medições estáveis e coerentes ao longo do tempo, nas quais o P95 coincide com o máximo.

Figura 11 – Efeito da aplicação do P95 por escola sobre a estimativa de velocidade máxima, por faixa de plano contratado.



A redução do valor máximo após a aplicação do P95 ocorre em praticamente todas as faixas de plano, mas com intensidades diferentes (figura 11). Nos planos de até 50 Mbps, a variação é pequena, indicando medições mais consistentes e com baixa incidência de picos extremos. Já nas faixas intermediárias, entre 50 e 300 Mbps, observa-se maior dispersão e quedas mais acentuadas, o

que sugere a presença de medições isoladas significativamente superiores ao padrão habitual. Nos planos mais altos (300–600 Mbps), a redução tende novamente a ser menor, refletindo um comportamento mais estável — possivelmente por limitações técnicas da rede ou menor ocorrência de medições infladas.

Em síntese, os resultados indicam que a aplicação do P95 remove valores extremos sem alterar o comportamento típico das escolas com medições estáveis, resultando em uma estimativa mais fiel e representativa da velocidade efetivamente experimentada pela maioria dos alunos.

CONCLUSÃO

Uma parcela significativa das escolas apresenta medições acima da velocidade contratada, sobretudo nos planos de menor faixa, evidenciando que o valor declarado do plano nem sempre reflete a velocidade realmente experimentada pelos alunos. Nesse contexto, torna-se ainda mais importante utilizar métricas robustas, capazes de representar o desempenho típico sem distorções causadas por valores extremos, e recorrer às medições objetivas e automatizadas do Medidor Educação Conectada — tanto na versão software quanto na SIMETBox — como fonte mais fidedigna e comparável da experiência real de conectividade nas escolas.

As simulações e análises realizadas confirmam que o percentil 95 (P95) é a métrica que melhor equilibra robustez estatística, representatividade da experiência escolar e facilidade de comunicação. Ao descartar apenas os 5% de valores mais extremos, o P95 reflete com fidelidade a velocidade efetivamente disponível para os alunos, reduz a influência de picos ocasionais e mantém consistência em diferentes tipos de distribuição observados nas escolas brasileiras.

Embora situações específicas — como distribuições bimodais — possam introduzir desafios adicionais, os resultados mostram que, em comparação a alternativas como o P99, o IQR e o método adaptativo, o P95 oferece o compromisso mais sólido entre sensibilidade e precisão. Assim, recomenda-se sua adoção como indicador da qualidade de Internet disponível para alunos em diagnósticos de conectividade escolar conduzidos pelo SIMET. Essa escolha garante um indicador robusto e comparável entre escolas.

REFERÊNCIAS

BRASIL. **Decreto nº 11.713, de 26 set. 2023.** Institui a Estratégia Nacional de Escolas Conectadas. Brasília: Presidência da República, 26 set. 2023. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/D11713.htm. Acesso em: 29 abr. 2024.

BRASIL. **Resolução CENEC nº 2, de 22 fev. 2024.** Estabelece os parâmetros de conectividade para fins pedagógicos nos estabelecimentos de ensino da rede pública de educação básica. *Diário Oficial da União*, 4 mar. 2024. Disponível em: <https://www.in.gov.br/web/dou/-/resolucaocenec-n-2-de-22-de-fevereiro-de-2024-546279176>. Acesso em: 29 abr. 2024.

CENTRO DE ESTUDOS E PESQUISAS EM TECNOLOGIA DE REDES E OPERAÇÕES (CEPTRO). **Relatórios, artigos e guias para melhorar a qualidade da Internet no Brasil.** São Paulo: NIC.br, 2023. Disponível em: <https://medicoes.nic.br/publicacoes/>. Acesso em: 29 abr. 2025.

FAWCETT, T. **An introduction to ROC analysis.** *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006.

GRUPO INTERINSTITUCIONAL DE CONECTIVIDADE NA EDUCAÇÃO (GICE). **Guia Conectividade na Educação.** São Paulo: GICE, 2021. E-book (PDF). Disponível em: <https://medicoes.nic.br/publicacoes/>. Acesso em: 29 abr. 2025.

ANEXO-DETALHAMENTO DAS SIMULAÇÕES DOS DADOS

O anexo apresenta os parâmetros e configurações utilizados em cada cenário de simulação, descrevendo como as diferentes distribuições foram ajustadas para reproduzir padrões de medição típicos observados nas escolas, incluindo casos de estabilidade, bimodalidade e alta assimetria.

- **Distribuição de Gauss (normal):** Esta distribuição foi utilizada para simular medições de velocidade que se comportam de maneira previsível e simétrica, como aquelas realizadas em redes de alta qualidade e com pouca variação. A inclusão da distribuição Normal garante que as métricas de detecção de *outliers* sejam testadas em um cenário “ideal”, servindo como base para comparação com as distribuições mais complexas e assimétricas. A combinação de duas distribuições de Gauss pode gerar os padrões que resultam em um histograma com dois picos distintos, que é a definição de uma distribuição bimodal e que também são encontradas no histórico de medições das escolas. A depender da escolha de parâmetros é possível gerar uma bimodal simétrica ou assimétrica (ver escolha de parâmetros no material anexo).
- **Distribuições Lognormal:** ideal para modelar dados com assimetria à direita. Isso significa que a maioria dos valores se concentra em um ponto mais baixo, com uma cauda longa se estendendo para valores mais altos. Esse é um padrão muito comum em medições de velocidade de internet, onde a maioria das conexões tem velocidades moderadas, mas algumas poucas podem ser excepcionalmente altas. Ambas possuem uma flexibilidade permite simular cenários de alta variabilidade, como aqueles encontrados em redes com desempenho inconsistente ou em locais com infraestrutura precária, que representam um desafio real para a detecção de *outliers*.
- **Distribuição Beta:** cria uma distribuição com assimetria à esquerda. O pico de frequência se encontra em valores altos, com uma cauda longa se estendendo para valores mais baixos. No contexto da sua análise, isso poderia simular uma rede de internet extremamente robusta, onde a maioria das medições atinge a velocidade máxima contratada, mas ocasionalmente ocorrem quedas de desempenho raras e significativas. Ela é bastante versátil e, por isso, a depender da escolha de parâmetros é possível gerar dados com característica **uniforme**.

Resumo dos Cenários (k, m, limites)

	Q3	IQR	k	m	limiar = Q3 + k*IQR	lim_sup= limiar + m*IQR	Total pontos	Outliers
A: k=3, m=0.5 (corte perto, faixa estreita)	43.4	17.48	3.0	0.50	95.82	104.56	1000	60
B: k=6, m=3.0 (corte longe, faixa larga)	43.4	17.48	6.0	3.00	148.25	200.68	1000	60
C: k=4.5, m=1.25 (intermediário)	43.4	17.48	4.5	1.25	122.04	143.88	1000	60

Distribuição	Descrição do Formato
Normal	Forma de sino, simétrica. Os dados se concentram em torno da média de 50, com a frequência diminuindo uniformemente em ambas as direções. Ideal para dados com comportamento previsível e sem assimetria.
Lognormal	Assimétrica à direita. A maioria dos dados se agrupa em valores mais baixos, com uma cauda longa e esticada em direção a valores mais altos. Representa bem dados que não podem ser negativos e que têm um crescimento exponencial.
Weibull	Assimétrica à direita. Os dados começam a se agrupar, atingem um pico e depois decaem gradualmente. É mais flexível que a Lognormal, podendo modelar caudas com diferentes características.
Beta (Assimétrica Alta)	Assimétrica à esquerda. A maioria dos dados se concentra em valores altos (próximos de 100), com uma cauda longa se estendendo para valores mais baixos. Simula cenários onde o desempenho é quase sempre excelente, com poucas e significativas quedas.
Bimodal Simétrica	Dois picos, simétricos. A distribuição tem dois picos de frequência distintos, um em torno de 30 e outro em 70. Não possui uma cauda dominante para a direita ou esquerda.
Bimodal Assimétrica	Dois picos, assimétricos. A distribuição tem dois picos, mas um é significativamente mais alto que o outro. O pico em 30 é mais proeminente, enquanto o pico em 70 é menor.
Uniforme	Plana e retangular. Todos os valores entre 0 e 100 têm a mesma probabilidade de ocorrência. Não tem picos ou caudas.



nic.br egi.br