UNIVERSIDADE FEDERAL FLUMINENSE

RODRIGO CHIMELLI SILVA

CUMULATIVE EVALUATION AND OUTLIER DETECTION IN COMPUTER NETWORKS

NITERÓI

2026

RODRIGO CHIMELLI SILVA

CUMULATIVE EVALUATION AND OUTLIER DETECTION IN COMPUTER NETWORKS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
ANTONIO AUGUSTO DE ARAGÃO ROCHA

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S586c Silva, Rodrigo Chimelli
CUMULATIVE EVALUATION AND OUTLIER DETECTION IN COMPUTER NETWORKS / Rodrigo Chimelli Silva. - 2026.
109 f.

Orientador: Antonio Augusto de Aragão Rocha.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2026.

1. Redes de Computadores. 2. Detecção de Outliers. 3. Indicador de Desempenho. 4. Produção intelectual. I. Rocha, Antonio Augusto de Aragão, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Rodrigo Chimelli Silva

CUMULATIVE EVALUATION AND OUTLIER DETECTION IN COMPUTER NETWORKS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

BANCA EXAMINADORA

Prof. Antonio Augusto de Aragão Rocha - Orientador, UFF

Prof. Célio Albuquerque, UFF

Prof. José Ferreira de Rezende, UFRJ

Niterói

2026

Por me ensinar a cada dia a ser forte e corajoso,
e me contagiar com sua crença sincera em mim,
Dedico este trabalho a Victor Menezes Araújo,
com fé e confiança,
e a mais profunda gratidão.

Agradecimentos

Agradeço primeiramente a Deus, suplicando que por meio da vida profissional, me faça canal de Seu amor no mundo.

Agradeço a meus pais, Marco Aurélio e Salma, que me ensinaram o valor da educação e do trabalho, sempre me provendo o trigo para que eventualmente eu fosse capaz de produzir meu próprio pão. E também à minha irmã Karine, exemplo de vida pelo empenho no trabalho e espírito aventureiro.

Agradeço a minha avó Natália e minha tia Mannoun, pelos conselhos, histórias e cafés da tarde. E a toda a minha família Nabti, Chimelli e Pereira da Silva.

Agradeço a meus irmãos de vida, André Abicaram, Gabriel Azevedo, José Estevão Diniz, Leonardo Macedo, Lucas Souza de Oliveira, João Simões, Rafael Vidal, Marcus Roale, Marcos Magalhães e Camilla Brandão. Quão sortudo é o homem que tem amigos.

Agradeço a meu orientador, Guto, pelas oportunidades concedidas, orientação de excelência e profunda humanidade. E também a todos os professores que passaram pela minha trajetória.

Agradeço a meus colegas de pós graduação, especialmente André, Alex e Victor, companheiros de almoço e risadas.

Agradeço aos funcionários da pós graduação na UFF, e a todos do laboratório Midiacom, em particular Marister cuja energia sempre alegrou as manhãs e tardes de estudo.

Agradeço a Luís Raposo e Alessandra Assaf, por me assistirem há tanto tempo na caminhada da vida.

Agradeço profundamente a todos os que oraram, intercederam, ou à sua própria maneira torceram para a finalização e o sucesso deste trabalho.

Por fim, agradeço novamente a Deus, por ter posto cada uma das pessoas mencionadas em minha vida.

"Ora, predigo-lhe que no momento mesmo em que a senhora verificar, com terror, que, malgrado todos os seus esforços, não somente não se aproximou do alvo, mas até mesmo dele se afastou — nesse momento, predigo-lhe —, a senhora atingirá o alvo e verá acima de si a força misteriosa do Senhor, que a terá guiado com amor, sem que a senhora o soubesse."

Fiódor Dostoiévski
Os Irmãos Karamázov

Resumo

À medida que as redes de computadores se tornam mais complexas, observa-se um aumento contínuo de exigências relacionadas a aplicações de alta demanda, como, por exemplo, o streaming de vídeo. Do mesmo modo crescem as expectativas dos usuários finais em relação à qualidade da rede. A avaliação do desempenho de redes deve, portanto, acompanhar esses desafios, e, mais especificamente, o monitoramento de backbones precisa estar equipado com ferramentas robustas para essa tarefa. Outro desafio importante reside na grande quantidade de dados gerados pelas redes modernas, provenientes de múltiplos enlaces, nós e dimensões de monitoramento. Essa abundância torna a inspeção manual impraticável e reforça a necessidade de métodos capazes de resumir, processar e interpretar o comportamento da rede de forma significativa e acionável. Esta dissertação busca contribuir para o desenvolvimento de mecanismos transparentes de avaliação da qualidade de redes de computadores, bem como para a identificação de outliers e dos contextos espaço-temporais de monitoramento associados à degradação do desempenho da rede. Em direção a esse objetivo, duas frentes são desenvolvidas: o desenvolvimento de um novo indicador de qualidade de redes, e a proposição de uma técnica de detecção de outliers em conjuntos de dados tabulares. A primeira parte do trabalho introduz ANPI (Aggregated Network Performance Index), um novo índice de desempenho de redes de computadores baseado em Funções de Distribuição Acumuladas (CDFs) de métricas de fluxos. A métrica proposta agrega informações acerca das distribuições de diferentes dimensões de monitoramento em diferentes pontos da rede, tanto para pontos de interesse internos quanto externos ao backbone analisado. Diferentes níveis de agregação são testados e comparados para se entender de que forma as diferentes dimensões e localizações devem ser combinadas para gerar uma pontuação final. Um dos principais benefícios da métrica proposta é a não diluição de desempenho local insatisfatório, mesmo quando grande parte da rede permanece saudável. Dessa forma, a estratégia apresentada é sensível não apenas à degradação generalizada do desempenho, mas também a desempenhos insatisfatórios restritos a contextos específicos, permitindo priorizar situações relevantes para investigação. Além disso, as métricas brutas são processadas de modo a permitir

agregação e comparação não enviesadas, uma vez que diferenças como a distância geográfica tornam inadequada a comparação direta entre medições em contextos distintos. Continuando os esforços de diagnóstico transparente, a segunda parte da dissertação apresenta CCBOS (Complementary Cumulative-Based Outlier Score), um novo método de detecção de outliers que se baseia no tradicional método HBOS (Histogram-Based Outlier Score), ampliando-o com a inclusão de informação ordinal para a detecção de outliers que se encontram nas caudas das distribuições. Assim, é esperado que ambas as frentes sejam utilizadas em conjunto para a detecção de baixo desempenho e posterior identificação dos contextos de monitoramento mais fortemente associados à degradação do indicador. Assim, as duas frentes poderão colaborar para o funcionamento saudável das redes de computadores em diferentes escalas, além de fornecerem mais informações para apoiar intervenções em políticas, algoritmos e arquiteturas.

Palavras-chave: Redes de Computadores. Detecção de Outliers. Indicador de Desempenho.

Abstract

As computer networks become increasingly complex, there is a continuous growth in the demand for high-performance applications, such as video streaming. At the same time, end users' expectations regarding network quality continue to rise. Consequently, network performance evaluation must evolve to meet these challenges, and, more specifically, backbone monitoring must be supported by robust tools capable of performing this task effectively. Another significant challenge lies in the massive volume of data generated by modern networks, originating from multiple links, nodes, and monitoring dimensions. This abundance makes manual inspection impractical and reinforces the need for methods capable of summarizing, processing, and interpreting network behavior in a meaningful and actionable manner. This dissertation aims to contribute to the development of transparent mechanisms for computer network quality assessment, as well as to the identification of outliers and the spatiotemporal monitoring contexts associated with network performance degradation. To achieve this goal, the work is organized into two complementary directions: the development of a novel network quality indicator and the proposal of a new outlier detection technique for tabular datasets. The first part of the dissertation introduces ANPI (Aggregated Network Performance Index), a novel computer network performance index based on the Cumulative Distribution Functions (CDFs) of flow-level metrics. The proposed metric aggregates information from the distributions of multiple monitoring dimensions across different points in the network, considering both internal and external points of interest relative to the analyzed backbone. Different aggregation strategies are evaluated and compared to determine how distinct monitoring dimensions and network locations should be combined to produce a final performance score. One of the main advantages of the proposed metric is its ability to preserve the impact of localized performance degradation, even when the majority of the network remains healthy. Thus, the proposed strategy is sensitive not only to widespread performance degradation, but also to poor performance restricted to specific contexts, allowing relevant situations to be prioritized for further investigation. Furthermore, the raw metrics are processed to enable unbiased aggregation and comparison, since factors such as geographical distance make direct comparisons between measurements collected under different conditions in-

appropriate. Extending the goal of transparent network diagnosis, the second part of the dissertation presents CCBOS (Complementary Cumulative-Based Outlier Score), a novel outlier detection method based on the traditional HBOS (Histogram-Based Outlier Score). The proposed method extends HBOS by incorporating ordinal information, enabling the detection of outliers located in the tails of feature distributions. Thus, both contributions are expected to be used together: first, to detect network performance degradation through the proposed quality indicator and, subsequently, to identify the spatiotemporal monitoring contexts most strongly associated with the observed degradation. Together, these two approaches have the potential to support the healthy operation of computer networks at different scales while providing valuable insights to guide improvements in network policies, algorithms, and architectures.

Keywords: Computer Networks. Outlier Detection. Performance Index.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AA Average Availability

AD Average Delay

ANPI Aggregated Network Performance Index

CCBOS Complementary Cumulative-Based Outlier Score

CCDF Complementary Cumulative Distribution Function

CDF Cumulative Distribution Function

CND National Data Center (Centro Nacional de Dados)

DNS Domain Name System

ECCDF Empirical Complementary Cumulative Distribution Function

ECDF Empirical Cumulative Distribution Function

ECOD Empirical-Cumulative-distribution-based Outlier Detection

GM Geometric Mean

HBOS Histogram-Based Outlier Score

HM Harmonic Mean

HTTP Hypertext Transfer Protocol

IP Internet Protocol

KPI Key Performance Indicator

KQI Key Quality Indicator

P2E PoP-to-External endpoint

P2I PoP-to-Internal endpoint

P2P PoP-to-PoP

PDF Probability Density Function

PL Packet Loss

PoP Point of Presence

QoE Quality of Experience

QoS Quality of Service

RNP Rede Nacional de Ensino e Pesquisa

RTA Round-Trip Average

RTT Round-Trip Time

TCP Transmission Control Protocol

UF Federative Unit (Unidade Federativa)

Contents

1 Introduction

In the present context, Network Performance faces several challenges. As networks become larger and more complex, there is a continuous increase in new demands related to developments such as cloud and edge computing, the emergence of IoT (Internet of Things), the development of 5G/6G networks, video streaming and other high performance services. At the same rate, end-user expectations regarding service quality continue to increase. The evaluation of network performance must therefore keep pace with this growing complexity and demand. More specifically backbone monitoring needs to be equipped with robust tools for the assessment of performance.

Evaluating networks, however, is far from trivial, and several frameworks have been proposed to address these intrinsic obstacles, each with its own strengths and weaknesses.

Firstly, network assessment depends on several factors such as delay, packet loss, and availability, each with its own particularities, challenges, and consequences. These factors can be orthogonal or interdependent to each other, and the process of choosing which ones to consider can significantly impact the results, since each one of them reveals a different aspect of the overall situation. At the same time, there is no complete consensus on the set of dimensions that should be considered in a general evaluation framework.

Another challenge lies in the sheer quantity of data generated by modern networks. Network monitoring infrastructures continuously produce large volumes of measurements across time, links, nodes, and multiple performance dimensions. This abundance of data makes manual inspection increasingly impractical and reinforces the need for methods capable of summarizing, processing, and interpreting network behavior in a meaningful and actionable way.

Simply presenting the performance of each metric separately fails to capture the full complexity of network performance. How should two networks be compared when each one performs well on a different set of metrics? One possible way to address this semantic challenge is to carefully aggregate the data into a score that remains comparable across

time, scope, and geographical context.

Aggregation necessarily entails some loss of information; nonetheless, the benefit of enabling a broader and more coherent understanding of the network often outweighs this cost. In this sense, the loss of detail may be accompanied by a substantial gain in knowledge and practical utility.

As stated previously, the aggregation process must be designed with great care. Important information must not be lost in the process; otherwise, problems that should be reflected through score degradation may remain hidden, leading to the mistaken conclusion that the network is performing well when the opposite is actually the case.

A performance index must therefore be sensitive to local bottlenecks, not allowing good global behavior to mask local problems. Even when considering a single metric, different scenarios may imply different baselines. For example, the raw delay values measured between distant machines cannot be directly compared to those observed between pairs that are much closer to each other. Thus, even within the same performance dimension, the aggregation process requires careful treatment.

Moreover, an ideal framework for network evaluation should not only describe the quality of the network, but also indicate directions for improvement. Evaluation is not an end in itself, but rather a means of supporting understanding, which, in turn, motivates interventions and improvements.

Operational actions should be aided by the evaluation framework, and if important information is lost or masked in the aggregation process, additional mechanisms are needed to indicate measurement contexts associated with performance degradation so that they may guide corrective action. From this perspective, it is important not only to evaluate the network, but also to identify source, destination, and time contexts associated with irregular performance.

One possible approach for this goal is the use of anomaly/outlier detection techniques. Of particular interest are those that assign outlier scores to each data point and can therefore indicate where multidimensional data become sparse or deviate from expected behavior. A network evaluation framework should therefore be consistent with the detection of such outliers, and the cooperative relationship between both components should be validated so that their joint use can effectively contribute to network improvement.

Despite the existence of several approaches for network evaluation and outlier detection, important gaps remain. Many methods focus on isolated metrics, while others

aggregate dimensions in ways that hinder explainability or fail to preserve sensitivity to local degradation. Likewise, outlier detection methods are not always incorporated into a broader framework that relates detected outliers to the actual degradation of network performance. As a result, there remains a need for approaches capable of combining holistic evaluation, explainability, and the identification of measurement contexts associated with performance degradation in a coherent manner.

In order to better understand the state of a network, this work proposes the Aggregated Network Performance Index (ANPI), a framework that leverages the Cumulative Distribution Function (CDF) and the Harmonic Mean (HM) to process and aggregate large quantities of data into a single score, merging different aspects of performance monitoring.

In addition, the Complementary Cumulative-Based Outlier Score (CCBOS) is introduced as an outlier detection method that takes advantage of the popular HBOS method, with the ability of CDFs to identify outliers in the tails of the distributions. The CCBOS is employed with the goal of identifying data points that worsen the score, thereby supporting a better understanding of how performance degradation can be remediated.

Finally, both components are jointly validated, aiming not only to show the individual capabilities of the proposed artifacts, but primarily their shared value, when collaborating in the complementary tasks of network performance assessment and improvement.

1.1 Contributions

The main contributions of this work are as follows:

1) The proposal of ANPI, a novel network performance index that leverages CDFs for single-dimension scoring, Harmonic Means for locally-sensitive aggregation, dynamic thresholds and excesses, rather than raw values, for the unbiased assessment of network quality.

2) The introduction of CCBOS, an outlier-detection scoring method which adapts the traditional histogram-based method HBOS by employing CCDFs (Complementary Cumulative Distribution Functions) for the identification of spatiotemporal contexts which are most associated with ANPI's degradation. This method leverages HBOS's density-information technique within a context where outliers are found on the tail end of data distributions.

3) The validation of the effectiveness of ANPI and CCBOS not only individually but as collaborators in the main goal of network performance evaluation.

1.2 Organization

The remainder of this work is organized as follows. Chapter 2 provides background on network metrics, statistical definitions and outlier detection methods, as well as details the utilized dataset. Chapter 3 gives a review of works related to performance evaluation and anomaly/outlier detection in the field of Computer Networks. Then, Chapter 4 presents and details the performance index which motivated the present work and serves as its baseline. Following, Chapter 5 introduces, motivates, and thoroughly details our proposed index. After that, Chapter 6 provides numerical comparisons between the proposed index and the baseline. Moreover, it explores the new index parameters and design choices, along with the motivation for the recommended configuration. Chapter 7, then, presents the CCBOS and its motivations, followed by Chapter 8 which jointly validates the effectiveness of both ANPI and CCBOS, showing that their combined use can provide beneficial information for network operators. Furthermore, this section compares CCBOS with other outlier-detection scoring techniques. Finally, Chapter 9 concludes the work and outlines possible directions for future research.

2 Background

This chapter sets out an overview of several concepts, notations, objects and methods, all of which are fundamental for the present investigation. Notably, the real dataset adopted in this work is presented and detailed, together with statistical and mathematical definitions, outlier-detection methods and network related concepts.

The chapter is divided in the following way: 2.1 establishes common terms and conventions utilized throughout the work. Section 2.2 then introduces and explains several classical network measurements which are employed in this study. Next, Section 2.3 introduces and describes the datasets used in the work, both in structure and composition. Finally, Sections 2.4 and 2.5 provide conceptual background regarding mathematical concepts and outlier detection methods, respectively.

2.1 Notation and Definitions

Firstly, it is important to present some terminology and conventions adopted and frequently mentioned throughout this and the following chapters. These terms are not taken from previous literature, but were chosen to improve the clarity of the work by providing a precise language for discussing specific concepts.

Scenario: The communication setting under analysis. All scenarios use PoPs (Points of Presence) as sources and differ in the type of destination. In this work, there are 3 possible scenarios, all P2P (PoP to PoP), P2E (PoP to External point of interest) and P2I (PoP to Internal point of interest). Here, internal and external are defined relative to the network under analysis.

Measure: The specific metric or service being measured. The measurement types considered in this work are RTA (Round-Trip Average), PL (Packet Loss), HTTP, and DNS Resolve.

Context: A tuple composed of a scenario, an origin PoP, a destination, and a measurement type.

Measurement: A tuple composed of a measurement type and a destination.

2.2 Network Metrics

As previously mentioned, several dimensions can be taken into account when evaluating network quality and performance. In this work, five were selected:

`HTTP`: Represents the average HTTP page load time, in milliseconds, between PoPs and monitored web pages.

`DNS Resolution`: Represents the average DNS resolution time, in milliseconds, between PoPs and monitored DNS servers.

`RTA`: Represents the average round-trip time, in milliseconds, measured using the `ping` tool.

`Packet Loss`: Represents packet loss (PL), measured as the percentage of sent packets that did not receive a response.

`Availability`: Represents the fraction of time during which a machine, such as a PoP, was operational and responsive.

Both `ping_rta` and `ping_pl` are obtained from the `ping` measurement tool. This tool sends network probes from a source to a destination and waits for the corresponding replies.

The round-trip time (RTT) of each probe is defined as the time elapsed between the transmission of the packet and the arrival of the corresponding response. The RTA is then computed as the average RTT across all probes.

Based on these replies, it is also possible to determine the fraction of packets lost during the measurement process, which constitutes the Packet Loss measurement type.

It is also worth noting that, while the other four metrics are associated with network paths and therefore have both source and destination attributes, Availability is associated with a single machine and therefore does not have a destination attribute.

2.3 Datasets Description

The data used in the ANPI experiments and tests consist of KPI (Key Performance Indicators) datasets collected from the RNP (Rede Nacional de Ensino e Pesquisa - National Teaching and Research Network) backbone. The RNP backbone is composed of several Points of Presence (PoPs), one in each of the 27 Brazilian federative units, and links connecting pairs of PoPs at the logical layer of the network. The RNP backbone also provides connectivity to external networks, such as Monet and RedCLARA, enabling campus connectivity on a global scale.

Figure 1 shows the RNP backbone topology and its heterogeneity. While some PoPs, such as DF (Brasília), have as many as nine direct neighbors, others, such as Rio Branco (AC), have only one. This non-uniformity is also reflected in the link bandwidths, which range from 1 to 300 Gb/s.

Due to this asymmetry in network connectivity, different PoPs are expected to produce distinct measurement distributions. For example, the latencies between RR (Boa Vista) and CE (Fortaleza), which are connected by a 1 Gb/s link, are not expected to be comparable to those between SP (São Paulo) and PR (Curitiba), whose link has a capacity of 300 Gb/s and whose endpoints are geographically much closer to each other.

The data adopted for the ANPI derivation are divided into three axes: Latency, Packet loss, and Availability. Although Latency is more heterogeneous in terms of the number of measurement types (as shown in 2.3.2,2.3.4, and 2.3.3), it shares the same tabular structure as the Packet Loss measurements.

Table 1 describes the semantics of each of the columns from Packet Loss and Latency Datasets. It should be noted that the information concerning the destination of the Packet Loss and Latency measurements is included in the "_measurement" column of the respective datasets.

Table 1: Latency and Packet Loss datasets' columns

Columns	Description
`_time`	Time of measurement
`_value`	Value of measurement
`_measurement`	Performed measurement (including destination)
`host_id`	id of host
`p_int`	Type of Destination ('p2p', 'p2e', 'p2i')
`pop`	Source Pop
`type`	Type of Measurement

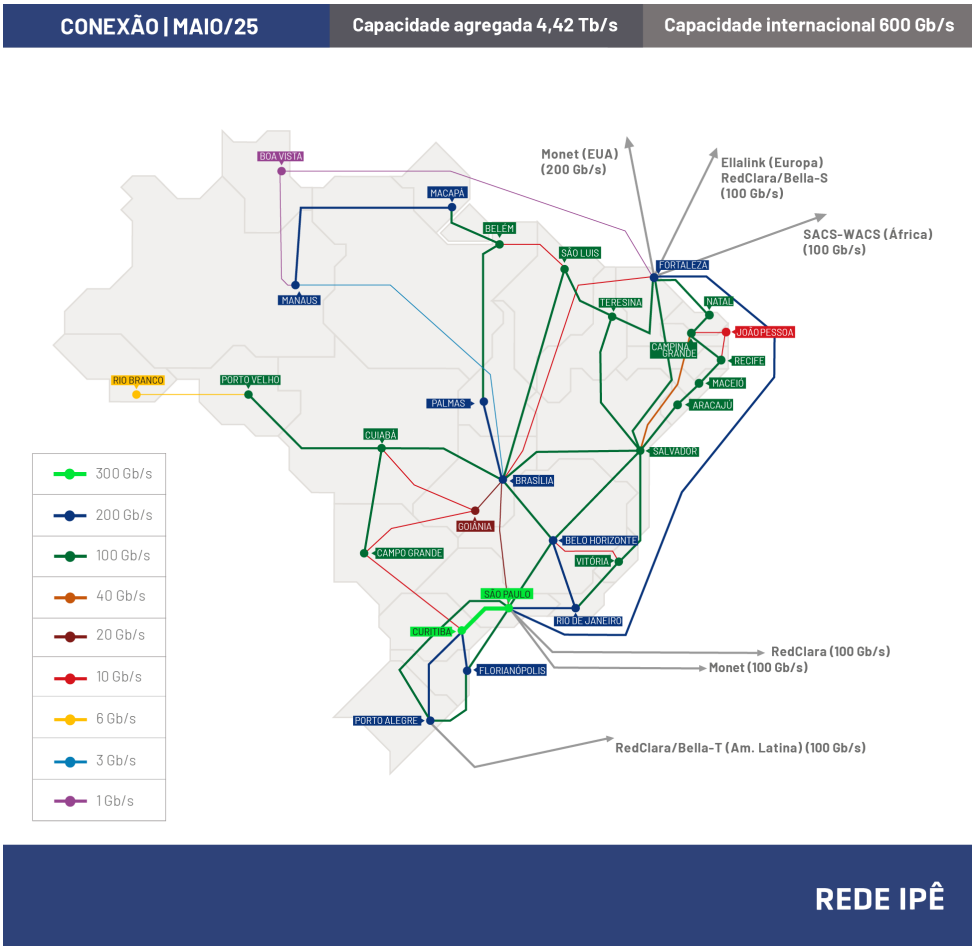

Figure 1: RNP backbone topology

Similarly, Table 2 presents the semantics of the columns in the Daily Availability datasets, which was derived from a dataset of unavailability events logs in the analyzed semester.

The choice for daily aggregation for ANPI was made for outlier detection purposes, which are further explained in 7. The main motivation was to achieve the highest temporal granularity possible, in order not to lose important information, brought about by coarser time frame aggregations.

Table 2: Daily Availability Data

Column	Description
`day`	Day of measurement
`month`	Month of measurement
`pop`	PoP associated with the measurement
`av_percent`	Availability percentage
`minutes_down`	Total number of minutes down
`unav_frac`	Fraction of unavailability computed from the availability data

2.3.1 Scenarios

All source points for performed measurements were PoPs from the RNP backbone. Regarding destination, however, three different scenarios were taken into account:

P2P: the destination is another PoP in the backbone

P2E: the destination is a point of interest external to the RNP Network

P2I: the destination is a point of interest internal to the RNP Network

By measuring data from these 3 use cases, the health of the network can be assessed in a more general approach. It is worth noting that these 3 scenarios make sense only in measurements of Packet Loss and Latency, since Availability data does not involve sources and destinations.

In the P2P scenario, only two measurement types are considered: ping_rta and ping_pl. For P2E and P2I, however, in addition to these 2, dns_resolve and http measurements are also considered, due to the nature of the endpoints.

Subsections 2.3.2, 2.3.3 and 2.3.4 provide additional information from P2P, P2E and P2I data, respectively.

2.3.2 P2P Measurements

As previously stated, datasets concerning the P2P scenario were comprised only of ping_rta and ping_pl measurements.

The value of Packet Loss measurements ranged from 0 to 100, representing the percentage of packets that were lost during probing. As for Latency related measurements, only the rta was considered, which represents the average rtt of probes that traveled from the source to the destination PoPs.

Table 3 contains row examples from both measurement types. It should be observed that the destination PoP is referenced inside the '_measurement' columns where the 'ac', for example, represents that the measurement destination is the 'AC' (Acre) PoP.

It is also worth noting that in the analyzed semester, the number of packet loss measurements in the P2P scenario was 180,135,185 while the number of rta measurements were 179,575,727.

Table 3: P2P measurement datasets row examples

_time	_value	_measurement	host_id	p_int	pop	type
2025-01-01T00:16:18.323693174Z	9.0	ping_ndac_pl	6156	p2p	RJ	ping_pl
2025-01-01T00:00:22.754205198Z	55.32	ping_ndac_rta	6156	p2p	RJ	ping_rta

2.3.3 P2E Measurements

The datasets concerning the P2E scenario were comprised of all 4 different measurement types with Table 4 presenting row examples from all four of them. It is worth noting that the destination is still present in the '_measurement' column but now as an IP address (e.g. 8.8.8.8.) or the name of a webpage (e.g. apple) instead of an UF code.

Considering the semester of interest, the number of Packet Loss, Rta, Dns Resolve and HTTP measurements in the P2E scenario are respectively 20,788,300, 20,782,836, 110,670,359 and 68,286,148.

Table 4: P2E measurement datasets row examples

_time	_value	_measurement	host_id	p_int	pop	type
2025-01-01T00:00:22.737073414Z	0.0	ping_dns_8.8.8.8_pl	6156	p2e	RJ	ping_pl
2025-05-12T05:30:31.7486193Z	30.44	ping_dns_1.1.1.1_rta	6296	p2e	SE	ping_rta
2025-02-12T21:09:08.742635202Z	143.0	dns_resolve_google_200.137.53.53	6218	p2e	AM	dns_resolve
2025-01-01T00:00:23.155265682Z	70.0	http_page_apple	6156	p2e	RJ	http

2.3.4 P2I Measurements

The P2I dataset is analogous to that related to P2E, however, instead of reaching external points of interest, measurement's destinations are internal to the RNP network.

Table 5 contains row examples from all measurement types of the P2I scenario.

Considering the analyzed semester, the number of Packet Loss, Rta, Dns Resolve and HTTP measurements in the P2I scenario are respectively 13,858,889, 13,856,318, 13,796,681 and 20,911,595.

Table 5: P2I measurement datasets row examples

_time	_value	_measurement	host_id	p_int	pop	type
2025-01-01T00:00:22.728616421Z	0.0	ping_dns_edudns_200.137.53.53_pl	6156	P2I	RJ	ping_pl
2025-01-01T00:00:22.724073305Z	0.50	ping_dns_edudns_200.137.53.53_rta	6156	P2I	RJ	ping_rta
2025-06-30T23:55:29.610966806Z	160.0	dns_resolve_amazon_200.19.16.53	6308	P2I	TO	dns_resolve
2025-05-17T07:36:12.624038525Z	697.0	http_page_rnp.br	6297	P2I	CE	http

2.3.5 Availability Data

Data regarding Availability is the most unique out of the 3 dimensions of evaluation. It is not divided in the 3 aforementioned scenarios since Availability is a characteristic of a PoP as unit in a time frame, and not of packet transmissions. Thus data concerning availability does not contain source and destination attributes.

Table 6 provides examples of rows of the daily availability tables.

Table 6: Daily availability row examples

day	month	pop	`availability_percent`	`minutes_down`
14	1	PA	92.71	105.0
26	1	PB	14.51	1231.0

2.4 Mathematical Background

We now turn our attention to statistical and mathematical concepts at the center of this work.

2.4.1 Probability and Cumulative Functions

One of the conceptual foundations of both pillars of this work is the Cumulative Distribution Function(CDF), which is defined in the following way:

Definition 1 (CDF). *Let X be a random variable, the CDF of X ($CDF(x)$) is defined as:*

$$F_X(x) = \mathbb{P}(X \leq x)$$

In natural language, the CDF yields exactly the percentage of values that lies beneath a desired threshold, given a probability distribution.

The CCDF (Complementary Cumulative Distribution Function) of a random variable X, as the name indicates, has a dual role to the CDF:

Definition 2 (CCDF). *Let X be a random variable, the CCDF of X ($CCDF(x)$) is defined as:*

$$\bar{F}_X(x) = \mathbb{P}(X > x)$$

Definition 3 (PDF)**.** *Let X be a continuous random variable. The probability density function (PDF) of X, denoted by $f_X(x)$, is a function that describes the relative likelihood of X assuming values around x. It satisfies:*

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(t)\,dt$$

for any $a < b$.

The PDF can also be thought of as the derivative (when it is defined) of the CDF. Conversely, the CDF can be defined as the primitive of the pdf.

Since the probability and cumulative distribution functions must be estimated from the available datasets, empirical counterparts to the PDF, CDF, and CCDF are required. The following definitions provide these roles.

Definition 4 (ECDF)**.** *Let $X_1, X_2, \ldots, X_n$ be a sample of observations. The empirical cumulative distribution function (ECDF), denoted by $\hat{F}_n(x)$, is defined as:*

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

where $\mathbf{1}_{\{X_i \leq x\}}$ is an indicator function that equals 1 if $X_i \leq x$ and 0 otherwise.

Definition 5 (ECCDF)**.** *Let $X_1, X_2, \ldots, X_n$ be a sample of observations. The empirical complementary cumulative distribution function (ECCDF), denoted by $\widehat{\overline{F}}_n(x)$, is defined as:*

$$\widehat{\overline{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i > x\}}$$

where $\mathbf{1}_{\{X_i > x\}}$ is an indicator function that equals 1 if $X_i > x$ and 0 otherwise.

Definition 6 (Histogram)**.** *Let $X_1, X_2, \ldots, X_n$ be a sample of observations divided into a set of disjoint bins $B_1, B_2, \ldots, B_m$. A histogram represents the distribution of the sample by counting the number of observations that fall within each bin. For each bin B_j, the histogram count is defined as:*

$$h_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in B_j\}}$$

where h_j is the number of observations assigned to bin B_j.

Figure 2 provides an example for each of these concepts.

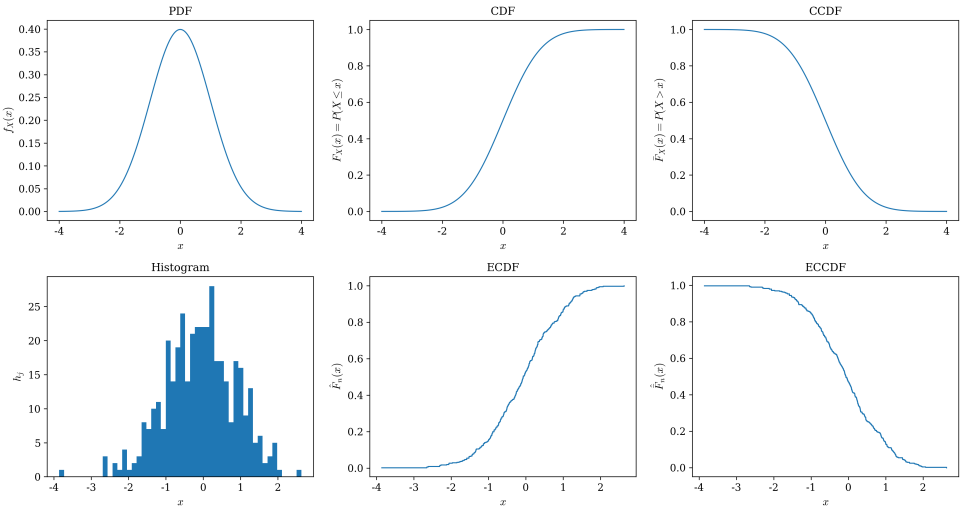

Figure 2: Examples of theoretical and empirical distribution functions. The top row shows the PDF, CDF, and CCDF of a standard normal random variable. The bottom row shows their empirical counterparts obtained from a finite sample: histogram, ECDF, and ECCDF.

2.4.2 Averages

Central tendencies measures such as the median and the arithmetic mean are frequently chosen as value aggregators. Although they may seem good candidates for our purposes, close inspection reveals that the median is very insensitive to outliers. And while the arithmetic mean may not suffer so strongly from this effect, it can still mask very bad results if the majority of other values behaves normally.

In our context this is important since we want the performance index to be affected by low performance in specific contexts such as rta measurements between AC and AM. Since there are many of such specific contexts, it is important to encounter more suitable alternatives.

Luckily, there are plenty of other definitions of average which are much more responsive to outliers. The most famous of which are the Geometric and Harmonic Means.

Definition 7 (Geometric Mean)**.** *Let $X = (x_1, x_2, \ldots, x_n)$ be a collection of numbers. The Geometric Mean (GM) of this collection, denoted by $GM(X)$, is defined as:*

$$GM(X) = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}.$$

Definition 8 (Harmonic Mean)**.** *Let $X = (x_1, x_2, \ldots, x_n)$ be a collection of numbers. The Harmonic Mean (HM) of this collection, denoted by $HM(X)$, is defined as:*

$$HM(X) = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

This behavior is related to the classical inequality of means, which states that, for positive values,

$$AM(X) \geq GM(X) \geq HM(X),$$

where $AM(X)$ denotes the Arithmetic Mean of the collection X.

As a toy example of this difference in sensitivity to outliers, consider Table 7, in which nine values are fixed at 10, while the last value varies across the rows. As the last value approaches zero, the arithmetic mean remains close to 9, since the nine high values compensate for the single low one. In contrast, both the geometric and harmonic means decrease more sharply, with the harmonic mean approaching zero considerably faster.

Table 7: Comparison between arithmetic, geometric, and harmonic means

Dataset	Arithmetic Mean	Geometric Mean	Harmonic Mean
[10,10,10,10,10,10,10,10,10,10]	10.00	10.00	10.00
[10,10,10,10,10,10,10,10,10,4]	9.40	9.12	8.70
[10,10,10,10,10,10,10,10,10,2]	9.20	8.51	7.14
[10,10,10,10,10,10,10,10,10,1]	9.10	7.94	5.26

2.5 Outlier Scores

In this dissertation, an outlier is understood as a sample whose values deviate substantially from the reference distribution of comparable observations. In our setting, this deviation is characterized not only by local density, but also by the relative magnitude and ordering of the observed values.

Therefore, our interest lies not merely in detecting observations located in low-density regions, but specifically in those which lie in the right tail of the corresponding distribu-

tions. This follows from the preprocessing choices adopted in this work, which associate larger numerical values to worse performance.

One of the most common outlier detection approaches consists of assigning each sample a non-negative score, which aims to quantify how anomalous the sample is. Scores closer to zero indicate that the sample is closer to the expected behavior, whereas higher scores indicate that the sample is considered to be more anomalous by the method.

(Samariya; Thakkar, 2023) provides an extensive review of several outlier scoring methods, covering different mechanisms, computational complexities, advantages, and disadvantages. The authors divide these methods into seven categories based on their underlying principles: statistical, density-based, distance-based, clustering-based, isolation-based, ensemble-based, and subspace-based methods.

When choosing one method among the many existing alternatives, it is essential to balance simplicity and robustness. While the selected method should rarely produce false negatives, thereby identifying most or all existing outliers, it must also be able to process incoming data within a feasible time frame.

2.5.1 HBOS

One of the most popular outlier scoring methods is HBOS (Histogram-Based Outlier Score) (Goldstein; Dengel, 2012), which assigns a score to each entry in the tabular dataset based on the marginal distributions of its attributes.

The main idea is that, for each attribute, belonging to a sparse bin should increase the corresponding sample's probability of being an outlier. Thus, a column-wise outlier contribution is calculated for each attribute, and these contributions are then aggregated to obtain the sample's final outlier score.

For each attribute i, a histogram is computed and normalized so that its maximum height is equal to 1. Then, for each sample p, the HBOS score is computed as:

$$\mathrm{HBOS}(p) = -\sum_{i=1}^{d} \log(\mathrm{hist}_i(p))$$

HBOS estimation by marginal distributions is particularly appropriate for the dataset considered in this work, which contains a large number of measurements described by attributes with different ranges and scales. Since the score is computed based on the em-

pirical distribution of each feature, HBOS allows attributes with different magnitudes to contribute to the outlier score without requiring complex modelling assumptions. Moreover, its low computational complexity makes it well suited for large datasets, such as the one analyzed in this work.

Another important aspect is that the proposed scenario does not require online or streaming detection. Therefore, more complex real-time detection mechanisms are not necessary. Instead, the focus is on obtaining an interpretable and computationally efficient method capable of highlighting observations that deviate from the expected behavior of the network. HBOS satisfies these requirements while remaining simple to configure and easy to interpret, since anomalous samples can be associated with low-density regions of the corresponding feature distributions.

It is also worth noting that the distribution-based nature of HBOS contributes to the interpretability of the framework. Since the final outlier score is obtained from feature-wise contributions, it is possible to inspect the extent to which each evaluation dimension contributes to the abnormality of a given sample. Therefore, the method not only provides a detection mechanism aligned with the statistical behavior of the data, but also supports an explanation of the detected outliers in terms of the individual attributes considered in the evaluation.

Furthermore, HBOS has been widely used and positively evaluated in outlier detection tasks, especially as a strong and efficient baseline for unsupervised outlier detection. Its combination of scalability, simplicity, and competitive performance makes it a suitable method for the proposed network performance analysis framework.

HBOS, however, does not utilize magnitude or order information regarding the values of the samples, leveraging only density information. Since in our context outliers are present in the tails, our work proposes a new method, inspired by the many qualities of HBOS and adapted for scenarios where the tail hypothesis holds.

2.5.2 ECOD

Motivated by the simplicity of HBOS and the large number of real world scenarios where outliers of interest lie on the tails of the column distributions (extreme values), (Li; Zhao, *et al.*, 2020) proposed the ECOD (Empirical Cumulative Outlier Detection) which leverages column-wise ECDFs to create outlier subscores, which are then aggregated in a similar fashion of HBOS.

ECOD is derived in the following steps:

1. Compute the left-tail and right-tail ECDFs for each feature.

$$\widehat{F}^{(j)}_{\text{left}}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i^{(j)} \leq z\}, \qquad \widehat{F}^{(j)}_{\text{right}}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i^{(j)} \geq z\}, \qquad z \in \mathbb{R}$$

2. Compute the sample skewness coefficient for each feature.

$$\gamma_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(j)} - \overline{X}^{(j)}\right)^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^{(j)} - \overline{X}^{(j)}\right)^2\right]^{3/2}},$$

   where

$$\overline{X}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)}$$

3. Aggregate tail probabilities to obtain the outlier score O_i for each sample X_i.

$$O_{\text{right-only}}(X_i) = -\sum_{j=1}^{d} \log\left(\widehat{F}^{(j)}_{\text{right}}\left(X_i^{(j)}\right)\right), \quad O_{\text{left-only}}(X_i) = -\sum_{j=1}^{d} \log\left(\widehat{F}^{(j)}_{\text{left}}\left(X_i^{(j)}\right)\right)$$

$$O_{\text{auto}}(X_i) = -\sum_{j=1}^{d}\left[\mathbf{1}\{\gamma_j < 0\}\log\left(\widehat{F}^{(j)}_{\text{left}}\left(X_i^{(j)}\right)\right) + \mathbf{1}\{\gamma_j \geq 0\}\log\left(\widehat{F}^{(j)}_{\text{right}}\left(X_i^{(j)}\right)\right)\right]$$

$$O_i = \max\{O_{\text{left-only}}(X_i), O_{\text{right-only}}(X_i), O_{\text{auto}}(X_i)\}$$

For the last step, three aggregation strategies are performed: considering only right tails, considering only left tails, or choosing either the left or right tail depending on each attribute's skewness.

The idea behind this is that some distributions may have only a right tail, only a left tail, or both. In this way, each sample receives the attribute-wise aggregation that maximizes its outlier score. In our context, outliers are always expected to occur on the right side. Therefore, for baselining purposes, we use a modified version of ECOD in which the only possible aggregation method is the right-tail aggregation.

ECOD, however, does not depend on binning, but only on the ECDF and ECCDF. And although this can generally be interpreted as a strength, by removing work related to parameter selection, binning discretization will prove to be very beneficial for our purposes.

Therefore, one of our goals is to merge the benefits present in both HBOS and ECOD.

3 Related Work

This chapter reviews the literature related to the present work, mirroring the dual structure of the dissertation, divided in two sections concerning Performance Evaluation in Computer Networks and Outlier/Anomaly detection both in and outside of the field.

3.1 Performance Evaluation

Traditional network evaluation relies on KPIs, such as latency, packet loss, and bitrate, to assess networks' QoS (Quality of Service).

In the context of campus network, for instance, (Ming; Man, 2024) collects several network performance indicators and aggregates them using the entropy weight method, quantifying the relative importance of each indicator in the context of higher education. User surveys were also conducted for robustness.

Similarly, (Sun; Zhang; Zhao, 2022) employs a SAE (Sparse AutoEncoder) to reduce data dimensionality and extract relevant features for network evaluation. These features are then combined with an improved GRA-TOPSIS method and a Cloud Model in order to classify network QoS into five categories, ranging from unqualified to excellent.

As in the present work, both studies propose strategies for aggregating different performance metrics. However, neither explicitly aims to preserve sensitivity to localized performance degradation during aggregation, nor do they couple the resulting assessment with an associated outlier detection mechanism.

Beyond static evaluation, other works focus on QoS prediction, motivated by the need to anticipate performance degradation and support proactive network management.

For example, (Pasco; Bernardini; Antônio, 2023) uses data-stream learning algorithms for Round-Trip Time (RTT) and One-Way Delay (OWD) prediction. Likewise, (Bicski; Pekar, 2024) aims to predict network service degradation, characterized by increased

latency, buffering, and outages. The authors show that early flow characteristics can be used to predict the state of non-observable flow segments. This makes it possible to better understand current network conditions and anticipate future degradation.

In a different direction, (Gomes *et al.*, 2025) proposes a methodology based on changes in KPI statistics to answer general questions regarding client and server performance. In addition, the work leverages LLMs to present the conclusions of the framework in a way that is more accessible to non-specialists.

The growing use of smartphones and other devices for intensive and time-sensitive applications, such as video streaming, has also motivated the development of a more user-centered approach to network evaluation. Rather than focusing exclusively on objective network metrics, this perspective seeks to account for the experience perceived by end users. In this context, the QoE (Quality of Experience) paradigm has become increasingly prominent (Kougioumtzidis *et al.*, 2022), (Yang *et al.*, 2018).

Traditional QoE formulations, often depend on subjective user feedback, which can make their implementation costly and resource-intensive. For this reason, several works attempt to bridge the gap between QoE and QoS by inferring KQIs (Key Quality Indicators) from KPIs.

In particular, (Panahi; Jalilvand; Najafi, 2025) uses KPIs such as jitter, latency, and packet loss to train a model that predicts MOS (Mean Opinion Score). The work also filters video comments in order to extract performance-related information, thereby facilitating the collection of user opinions. By combining both strategies, the proposed framework provides a more robust and resource-efficient approach to user experience assessment.

A similar objective is pursued by (Zhang *et al.*, 2022), which leverages a GAT (Graph Attention Network) to model and learn the structural causal dependencies between KPIs and KQIs. In doing so, the work uses the already well-understood behavior of KPIs to predict future values of KQIs.

In contrast to these predictive approaches, ANPI is not intended as a forecasting model; rather, it evaluates observed monitoring data and aggregates them into a context-sensitive measure of network performance. Nevertheless, the information produced by ANPI and CCBOS may complement network-quality prediction methods by providing interpretability to the assessment of observed performance degradation and the spatiotemporal monitoring contexts associated with it.

From a broader user-oriented perspective, (Ohlsen; Sermpezis; Newcomb, 2025) proposes the Internet Barometer, an Internet evaluation framework based on a three-layer understanding of quality: Use Cases, Network Requirements, and Measurement Tables. Since each use case has distinct demands, quality should be assessed through measurements that are appropriate to the corresponding context. One of the main contributions of this barometer is precisely its user-centric view of quality, highlighting that the same network may present different perceived quality depending on the purpose for which it is being used.

The present work shares the context-sensitive perspective of these approaches, although it does not directly address QoE, focusing instead on KPIs obtained from network monitoring. ANPI uses historical minima to adapt the reference against which measurements from each context are compared, thus taking into account the specific conditions under which each measurement was collected.

Finally, (Rede Nacional de Ensino e Pesquisa, 2025), described in detail in Chapter 4, proposes an index that aggregates packet loss, latency, and PoP availability through arithmetic means to derive a single score representing the performance of the RNP backbone. However, the use of raw monitoring values and global aggregation through simple averages may introduce contextual biases and allow well-performing contexts to compensate for local irregularities. These limitations constitute one of the main motivations for the present work, which seeks to provide a more context-sensitive and locally responsive alternative for network performance assessment.

Although these studies illustrate the diversity of approaches to network performance evaluation, a distinctive feature of the performance index proposed in this dissertation is its integration with a purpose-built outlier detection method. This coupling provides an additional layer of explainability by helping identify the spatiotemporal contexts and measurements associated with score degradation, thereby supporting both diagnosis and remedial action. In this way, the proposed framework aims to contribute to a more efficient, context-aware, and explainable assessment of network quality.

3.2 Irregularity Identification

Due to their distributed and complex nature, networks are not only frequent targets of malicious actions but often suffer from local technical malfunctions and bottlenecks. Network operators must therefore be able to identify both types of events in order to

take the necessary measures of prevention and protection regarding infrastructure and performance.

One particularly popular strategy for achieving this goal is anomaly/outlier detection, which has given rise to a wide range of studies based on different underlying mechanisms. Anomaly/outlier detection aims at identifying behavior in data which deviates from the expected or desired. However, there is an important distinction when approaching this task in computer networks:

(Thottan; Ji, 2003) divides network anomaly detection into two groups: performance-related and security-related anomalies. Security-related anomalies involve deliberate malicious activities and security breaches, such as intrusions and attacks, whereas performance-related anomalies are associated with failures that affect the functionality and efficiency of different network components and dimensions.

These definitions yield different approaches in the field. Notably security-related detection tend to rely more on supervised machine learning, since labels are a direct way of utilising the necessary information regarding the malicious nature of the event. When security is not a concern however, purely statistical methods can provide a rich set of tools inherited from the traditional study of outlier data.

When considering specific works, the distinction between both definitions and paradigms can be very subtle. For example, papers which are motivated by security concerns, sometimes take into consideration the performance-related definition when deriving detection strategies.

For the purpose of disambiguation in this dissertation we propose the utilization of the words 'anomaly' and 'outlier' in security and performance-related scenarios, respectively. This is not uniform in literature, however, where both are used interchangeably.

3.2.1 Anomaly Detection in Computer Networks

Due to the spatial and graph-structured nature of computer network data, several works address anomaly detection through graph-based techniques, such as GNNs (Graph Neural Networks) and GAT (Graph Attention Network), either in the modelling phase or in the detection phase.

(Kaya; Ozdem; Das, 2025), for instance, combines GCN (Graph Convolutional Network) and LSTM (Long-Short Term Memory) to model data while taking into account

both the spatial graph structure and temporal dynamics. This work performs supervised packet-level multiclass classification and is particularly suitable for integration with Wireshark.

(Yılmaz; Das, 2025) focuses on aggregating both local and global information through a combination of GCN and GAT, to identify both kinds of anomalies on real campus network data. Although the model is able to handle the complex structure of large networks, it incurs high computational overhead due to the transformation of log data into graph structures. Moreover, the model has low interpretability, providing limited insight into why a given sample was deemed anomalous.

(Latif-Martínez *et al.*, 2025) leverages GAT to group flows with similar histories in order to detect contextual anomalies. After defining the contexts, anomaly scores are computed from the absolute error between observed and predicted OD flows. The use of GAT provides some interpretability, since its coefficients can be understood as the probability of pairs of flows belonging to the same context.

Also concerned with explainability, (Liu *et al.*, 2022) proposes MSCA, an anomaly detection framework that not only leverages multiple sketches, clustering, and voting strategies for detection, but also finds association rules within anomalous sets. A broad review and taxonomy of explainability in anomaly detection can be found in (Li; Zhu; Van Leeuwen, 2023).

Nevertheless, not all techniques are based on graph methods. (Wakui; Kondo; Teraoka, 2022), for example, leverages LSTM-RNN to capture the periodicity of Internet backbone traffic. By predicting flow sizes, it associates the presence of anomalies with higher prediction errors. Thus, the methodology can be understood as self-supervised forecasting used for unsupervised anomaly detection. It also detects anomalies in real time and thus is expected to prompt interventions.

(Rezaei; Taheri; Shojafar, 2025) focuses on private and resource-efficient point anomaly detection in 5G networks through the use of Federated Learning, which enables deployment at the network edge, reducing both the possibility of threats and latency. Its use of pre-trained LLMs for explainability, however, can hinder performance in highly specialized contexts or in scenarios involving encrypted traffic.

(Baldoni; Battisti, 2025) proposes a new engineered representation of network traffic conditions which, when coupled with a simple PCA (Principal Component Analysis)-based anomaly detection method, can identify anomalies through flow reconstruction

errors. The representation is constructed using histograms and provides a compact and up-to-date view of the current network conditions. Due to its low complexity, real-time diagnosis can be performed. It is also suitable for simple detection methods, since it changes significantly in the presence of attacks.

Finally, in the context of SDN (Software Defined Network), where routing management is centralized, networks become particularly prone to attacks. (Khanal; Kumar; Ansari, 2025) addresses this through the training of a lightweight IDS (Intrusion Detection System) that compares different machine learning detectors, and deploying the best one on the controller for synchronous cyberattack detection. The IDS also reduces detection time by feature selection.

Unlike these approaches, the present work adopts an outlier-detection perspective, aiming to identify spatiotemporal monitoring contexts associated with performance degradation, regardless of whether the degradation results from intentional or unintentional events.

3.2.2 Outlier Detection in Computer Networks

Following the outlier perspective, (Michalak *et al.*, 2021) applies classic detection methods, such as Local Outlier Factor (LOF), Robust Kernel-Based Outlier Factor (RKOF), and Generalized Extreme Studentized Deviate (GESD), combined with moving-window strategies to find outliers characterized by slow changes over time in packet-characteristic time series.

(Ali *et al.*, 2017) applies LOF and Cluster-Based Local Outlier Factor (CBLOF) to detect performance spikes and drops in the Ping End-to-end Reporting (PingER) Internet performance dataset. Similarly to the present work, the paper aims to localize these outliers both spatially and temporally, in order to better understand what is affecting performance.

(Arifuzzaman; Islam; Arslan, 2021) employs traditional machine learning techniques, such as Random Forest (RF), Neural Network (NN), and Support Vector Machine (SVM), for Transmission Control Protocol (TCP) socket-level outlier detection and root-cause identification. Moreover, it generalizes the dataset, which is network-specific, to make it possible to train detection and identification models for other networks.

(Rao *et al.*, 2018) applies Principal Component Analysis (PCA) to TCP and performance Service-Oriented Network monitoring ARchitecture (perfSONAR) datasets in the

context of file transfers. The authors conclude that simple feature extraction techniques can be of great aid to outlier detection in network monitoring, especially when labeled normal-behavior data is available.

(Kiran *et al.*, 2020) uses a similar approach, comparing three different feature extraction methods: PCA, Isolation Forest, and autoencoder. (Marnerides; Schaeffer-Filho; Mauthe, 2014), although not recent, provides a survey on backbone network outlier detection and is important for a historical perspective of the field, which can be fruitful for the future of backbone monitoring.

(Silveira *et al.*, 2010) uses statistical testing to detect instances where flows on the same link are strongly correlated. This technique is interesting because it can detect changes that are not noticeable in single flows, since they may not produce high spikes, but that still interfere with the network due to the number of flows affected by those changes.

Together, these works show that network outlier detection can be approached from different perspectives, including statistical deviation, spatial and temporal localization, feature extraction, root-cause identification, and correlation between flows. This diversity motivates the adoption of methods capable of identifying measurement contexts associated with performance degradation without relying exclusively on attack labels or predefined failure classes.

The network-oriented works reviewed above do not specifically investigate the use of Cumulative Distribution Functions (CDFs) for detecting outliers located in the tails of feature distributions. Since this is the case of the context of the present work, we next discuss some works that try to combine CDFs and outlier detection strategies, though not specifically in Computer Networks.

3.2.2.1 Cumulative Techniques in Outlier Detection

Both in and outside Networks applications, outlier detection is very broad, with different techniques such as clustering, ML, and statistical tests being employed for such purposes. The following field-agnostic papers employ cumulative-distribution-based techniques in outlier detection, similarly to the technique proposed in this work.

(Li; Zhao, *et al.*, 2020), for example, introduces COPOD, which leverages a combination of CDFs and copulas to find outliers in the tails of feature distributions, while also considering interdependence between features.

(Zhu *et al.*, 2021) takes another approach. Concerned with the problem of inhomogeneous cluster densities, which produce bias toward higher-density clusters, the authors use a multidimensional CDF in a transform-and-shift method, balancing cluster densities and transforming local low-density regions into global low-density regions.

Moreover, (Li; Wang; Li, 2025) computes the ECDF of each feature, approximates it using a spline, and then takes the derivative to approximate the probability distribution function (PDF). After that, IDOS uses the PDF for giving outlier scores to each sample. This last step is very similar to the Histogram Outlier-Based Score (HBOS) which uses a histogram instead of a parametric PDF. The HBOS will later be introduced, being one of the cornerstones of the present work.

Following the previous paper, (Li; Li, 2025) uses the same CDF-spline technique. However, inter-feature dependencies are calculated through copulas, and features are not treated as independent when searching for outliers. In a sense the work combines both IDOS and COPOD, leveraging both's strengths for modeling and predictions.

(Angiulli, 2020), on the other hand, defines the problem of concentration of outliers, where high dimensionality degrades notions of distance and density, thus deteriorating several detection mechanisms based on such notions. The paper then uses the CDF to propose a new outlier definition, one that is not affected by distance degradation as dimensionality increases. The authors then propose the CFOF method which they theoretically prove to be immune to the concentration effect.

Finally, (Jäntschi, 2019) takes the data distribution and calculates the CDF values for each point. It then creates intervals for acceptable CDF values and assigns points that fall outside those intervals as outliers. This method stems from the idea that values very close to 0 or 1 in the CDF space correspond to observations located in the distribution tails and, therefore, may represent extreme behavior. (Duraj; Szczepaniak, 2021), (Han *et al.*, 2022), and (Bouman; Bukhsh; Heskes, 2024) provide examples of systematic and extensive benchmarking approaches for comparing outlier detection methods.

Overall, these works support the relevance of cumulative-based approaches for outlier detection, especially in scenarios where the data may present skewed distributions, heterogeneous densities, or tail-driven anomalous behavior. This is particularly relevant for network performance monitoring, where degradation events may appear as extreme values or rare combinations of measurements rather than as explicitly labeled attacks.

Building on these properties, CCBOS was designed to complement ANPI. In particu-

lar, its bin-based discretization is consistent with the threshold-based evaluation adopted by ANPI, while its cumulative component incorporates magnitude and ordinal information. Together, these properties help relate high outlier scores to the types of performance degradation to which ANPI is designed to be sensitive, in a simple and scalable fashion.

4 Performance Index Use Case

The development of the ANPI was motivated by institutional concerns regarding the foundations of the QIM 3 index currently being employed in the RNP. This index, which will henceforth be called 'baseline index', is thoroughly described in (Rede Nacional de Ensino e Pesquisa, 2025) and will be given a general overview in the present chapter, both to motivate the new score and its changes, and to better understand it by means of comparison.

It is important to notice, that even though this work was born from this specific use case, ANPI is general for usage in other Computer Networks, and its core components, such as CDF, HM, dynamic thresholds and unbiasing through historical minima are powerful techniques that can be generalized even beyond the field.

The baseline index is made of 3 components: PL (Packet Loss - originally PP from "Perda de Pacotes"), AD (Average Delay - originally RM from "Retardo Médio") and AA (Average Availability - originally DM from "Disponibilidade Média"), which will be outlined in the following subsections.

4.1 Baseline Index

4.1.1 Packet Loss (PL)

The PL component is responsible for indicating the quality of performance of the backbone with regard to packet loss, which is a critical indicative of reliability in data delivery.

Furthermore, being packet loss a good proxy for congestion in networks, this component of the score can indicate problems in load distribution, or even high stress in the network.

Mathematically, the Packet Loss term of the baseline index is given by the following formula:

$$PL = 10 * (3.83 - P\%)$$

where

$$P\% = \frac{44\,P\%_{P2P} + 28\,P\%_{P2I} + 28\,P\%_{P2E}}{100}$$

$$BaselineIndex = AD + PL + AA = 3.32 + 3.33 + 3.26 = 9.91$$

$$ANPI = AD + PL + AA = 2.10 + 3.30 + 3.31 = 8.71$$

and $P\%_{P2P}$, $P\%_{P2I}$ and $P\%_{P2E}$ are the average Packet Loss in each of the scenarios.

It is worth highlighting that the higher weight given to $P\%_{P2P}$ reflects that the interconnectivity between PoPs possesses higher importance to the health of the network than connectivity of PoPs with other points in the Network.

Since this component has an upper limit of 33.33, corresponding to one third of the total score, packet loss rates up to 0.5% do not penalize the component. In contrast, rates at or above 3.83% reduce its contribution to zero.

4.1.2 Average Delay (AD)

The AD component, which is responsible for scoring latency in the network is further divided in 3 subcomponents: AD1, AD2 and AD3. Each of them related to one of the 3 aformentioned scenarios, respectively, P2P, P2I and P2E.

This component is intended to aid the guarantee that data is travelling with low delay, a condition that has become even more critical due to high demanding applications such as video streaming and online gaming.

4.1.2.1 AD1

The AD1 component is responsible for encapsulating the assessment of latency in the P2P scenario.

The only latency-related measurement type in this scenario is that of ping_rta, and

since high latency should decrease this component, the AD1 is given by an inverse scaling of the AD_{P2P} (simple ping_rta average) with an experimentally defined weight, based on the target value of $27ms$.

$$AD_1 = \frac{660.15}{AD_{P2P}}$$

4.1.2.2 AD2

The AD2 is responsible for the P2I scenario, which in addition to the ping_rta, has latency information in the form of dns_resolve and http measurements as previously stated.

As with AD1, AD2 is computed through an inverse scaling with empirically defined weights, followed by summation.

$$AD2 = \frac{84}{AD_{P2I\_PING}} + \frac{1904}{AD_{P2I\_HTTP}} + \frac{289.23}{AD_{P2I\_DNS}}$$

4.1.2.3 AD3

The AD3 is responsible for the P2E scenario, and works similarly to the AD2 component, the only difference being the empirical weights, since the latter deals with external while the former with internal destinations.

$$AD_3 = \frac{4435}{AD_{P2E\_PING}} + \frac{386}{AD_{P2E\_HTTP}} + \frac{102}{AD_{P2E\_DNS}}$$

4.1.3 Average Availability (AA)

The AA component assesses the availability term of the final score and it is given by the following formula:

$$AA = \frac{AA_{\text{average}} \times 33.33}{99.91}$$

where

$$AA_{\text{average}} = \frac{\sum_{UF} A_{UF}\, w_{UF}}{\sum_{UF} w_{UF}}$$

and

$$AA_{\text{average}} = \frac{A_{AC} \cdot 8 + A_{AL} \cdot 10 + \cdots + A_{SP} \cdot 39 + A_{MIA} \cdot 36}{465}$$

- A_{UF}: Availability (%) of the PoP in each Federative Unit (UF).
- w_{UF}: Weight assigned to the corresponding Federative Unit.

The choice of weights is due to expert-defined parameters, which are listed in 6.3.

4.1.4 Final Score

The Final score is given by:

$$AD + PL + AA$$

$$AD = AD_1 + AD_2 + AD_3$$

Each component has a target score of 33.33, so that the final score is designed around an ideal value of 100.

4.1.5 Drawbacks

Since aggregation is done by arithmetic means, specific contexts (e.g. rta from RJ to AC in P2P) where the network may be behaving badly are easily compensated by those where everything is working well. Thus mistakes and inefficiencies that should be caught by the score to motivate intervention by network management stay under the radar and thus are not remediated.

Another drawback lies in the fact that the same thresholds (most notably in latency measurements) are being used in every context to define which measurements comply with the network requirements. This fails to take into account differences such as geographical distance that should make some contexts have higher thresholds than others.

If a measurement is assigned a threshold that is too high for its context, it will always seem (falsely) to perform optimally, and problems will not be remediated, since they are not caught by the score. By contrast, If a measurement is associated to a threshold that

is too low for its context, it will seem that it almost always fails, since it needs more leniency due to performance-unrelated characteristics.

5 Proposal Methodology

In this chapter, the Aggregate Network Performance Index (ANPI) will be introduced, and all aspects involving its motivations, development and conceptual background will be examined thoroughly.

5.1 ANPI: Overview and Calculation

Since ANPI was designed as an improvement over the baseline index, its high-level structure and selected design choices were preserved to maintain comparability between the two approaches. Methodological choices and parameter values were retained whenever no clear limitation or motivation for change was identified. This allows ANPI to be evaluated not as an entirely independent metric, but as a refinement of the existing methodology.

5.1.1 Historical Minimum

For each combination of scenario, measure, PoP and destination, the ideal value varies. If we take for example the P2P scenario, RTA measurements between (PR,SP) and (AM, RS) are not directly comparable since it is already expected that PoPs further apart from each other will have longer Latency values due to the longer propagation period.

Moreover, Figure 1 shows that to go from the AM PoP (Manaus) to the PR PoP (Curitiba) the minimum number of hops between PoPs is 3, while PR (Curitiba) and São Paulo (SP) are directly connected by a link.

Furthermore, when we focus on bandwidth, differences in constraints also arise. For instance, the link that connects the SP and PR PoPs has a bandwidth of 300 Gb/s, and for the AM-RS path with the shortest number of hops between PoPs to be traversed, it must pass through the link that connects AM to DF (Brasília) which has only 3 Gb/s.

Therefore, setting a measure-fixed threshold as a baseline to which every measurement

must adhere is far from appropriate, since it may be impossible for some contexts to maintain it, whilst being very easy for others, even when they are performing badly. Therefore, a score that takes into consideration the particularities of different contexts (geographical distances, connectivity due to the topology, etc.) must be properly defined for means of inter-context comparability.

Figure 3 shows the plotting of geographical distance versus historical minimum value of ping_rta for each of the (27*26)/2 = 351 unordered pair of PoPs in the P2P scenario. The red line is the regression line of the 2 variables and shows that the 2 concepts possess a strong linear correlation, with an R^2 score of 0.7932. This instigates the idea that geographical distance is the major factor that accounts for differences in Latency that are not related to congestion.

The green line represents the relationship between the Distance Between pairs of PoPs and the Latency data would achieve if there was a straight line that connected both PoPs and the only delay was due to propagation latency. For the construction of the green line, data velocity through the straight link was approximated to 2/3 of the velocity of light, which is commonly considered as a good approximation of propagation velocity based on links' materials.

The difference between the green and the red lines accounts for Latency brought about by transmission, processing, queuing delay and also by the fact that the fastest path between two PoPs normally has to pass through at least a third one.

The figure also brings to light the fact that the majority of outliers above the red line is composed of PoPs from the Northern region of Brasil. This merits attention, since it possibly signifies low connectivity in that region.

The historical minimum thus, can contain a lot of information regarding its respective context. When taken from a dataset constructed in a long time frame, it is expected to represent closely the best possible performance we can achieve in that particular context, both regarding congestion, chosen path, and other variables which can affect the Latency.

5.1.2 Delta Min

Taking all this into consideration, instead of dealing with measurements' raw data, the historical minimum of the semester was calculated for each (scenario, PoP, measure, destination)-tuple in the dataset. Then a new column was created called delta_min by making the difference between the present measurement and the historical minimum

of its context.

In every scenario and destination pair, the minimum of packet loss was naturally always 0. This happens because it is extremely improbable that in six months every single ping between two fixed destinations will always result in at least one packet being lost. Thus, the delta_min column for rows associated with Packet Loss measurements remained equal to the original measured value.

For availability Data, the minimum strategy was not taken into consideration due to the low granularity of availability data, and also because, analogously to the Packet Loss dimension, every single PoP had at least one day with 100% of availability.

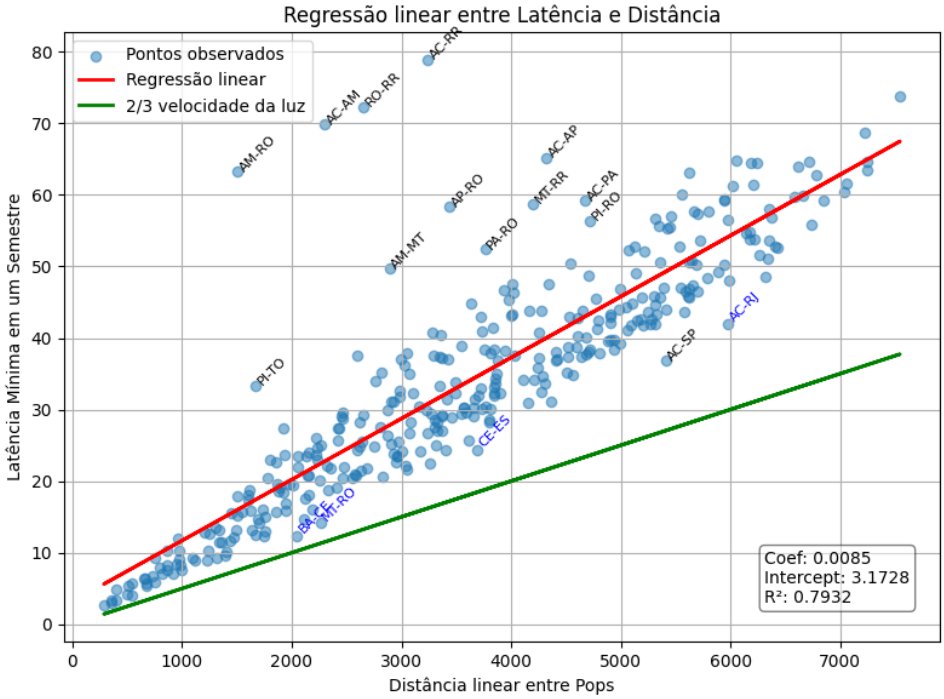

Figure 3: Geographical distance (km) versus RTA historical minimum (ms)

5.1.3 ECDF and The Uni-dimensional Performance

Since network performance involves different axes and dimensions with their respective measures (Latency, Packet Loss, etc.), and due to the diversity of possible source and destinations (both inside and outside of the network itself), it is valuable to first define the performance evaluation considering one single context. As previously defined the word context here has a very precise meaning: a tuple (scenario, measure, PoP, destination). One example would be (P2P, Packet Loss, RJ, SP).

The core of this uni-dimensional case of assessment is the estimation of the ECDF

2.4 of measurements' raw or delta_min values. After the CDF is derived and plotted, a threshold can be determined where values below it are considered ideal, and those above it, undesired.

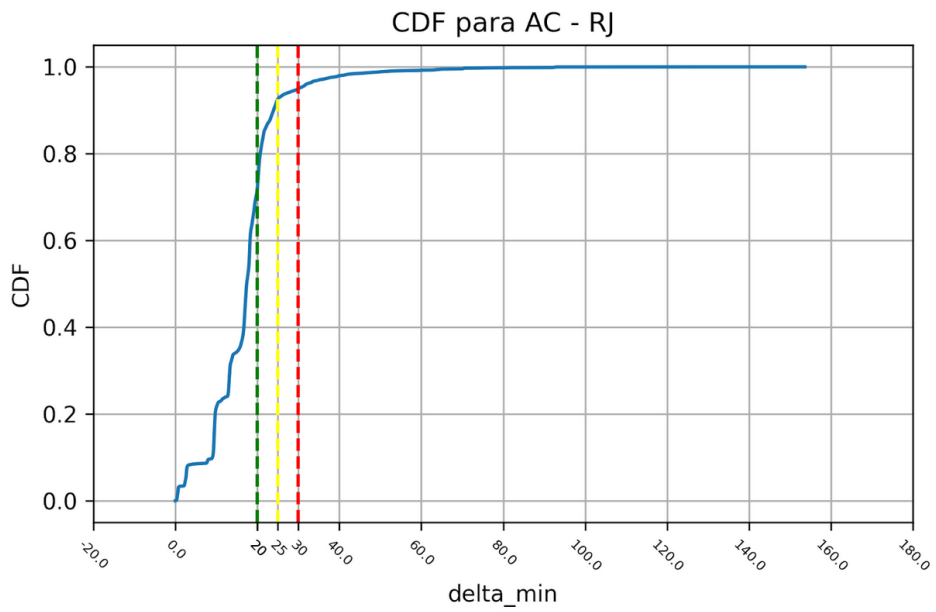

Figure 4: ECDF of delta_min rta values from PoP AC to PoP RJ. Three different thresholds are presented: 20ms (green), 25ms (yellow) and 30ms (red)

Figure 4, for example, shows the derived ECDF from delta_min values from ping_rta measurements between AC (Acre) and RJ (Rio de Janeiro). If we say that for delta_min 20ms the ECDF is 0.75, what should be inferred is that 75% of the measurements in this context are below or equal to 20ms.

In the figure, three possible thresholds are shown as dashed lines, 20, 25 and 30 ms. We can see that 20 is the most exigent of the possibilities since approximately 25% of the data points would not comply to the threshold. On the other hand, the 30 ms threshold is more lenient, since approximately only 5% of the measurements would be deemed unsatisfactory.

Depending on the constraints the network manager may want to impose, different thresholds would better satisfy their needs. Section 6.2 presents an experimental comparison between different threshold choices, either fixed or context-adaptative.

5.1.4 Aggregation Methods

Given the multiplicity of scenarios, flow-level metrics and destinations, one of the challenges the ANPI aims to tackle is that of aggregation. There are several ways of aggregating different individual values, and so in choosing one it is important to think of the desirable properties of the final score.

As defined in the introduction, one of the primary goals of this work is to easily identify problems in the network, so that remediation can be applied. Taking this into consideration, it is important that even if everything else is perfect, bad results in one scenario, destination or dimension of measurement are reflected in the final value. It is paramount that good values in other contexts do not mask local errors in the network, in order not to give the false impression of everything going fine.

Motivated by the outlier-sensitive property of the Harmonic Mean discussed in 2.4, the HM was chosen as a suitable aggregator. In order to validate this sensitiveness within our framework, three possibilities of aggregation were defined, depending on how HM is applied:

- Aggregation level 0: Creating ECDFs for each PoP-Measurement pair, and then performing two layers of harmonic means applications, the first inside each PoP with regard to destinations (yielding one subscore per PoP) and the second between the resulting PoP-specific subscores.
- Aggregation level 1: Creating single ECDFs for each tuple (Scenario, Measure, Pop), regardless of destinations, and then taking the Harmonic Mean between the |Pops| resulting sub-scores.
- Aggregation level 2: a single ECDF is created for each scenario–measure pair, regardless of PoP and destination. In this strategy all data from a single scenario and measure is bundled in one single ECDF. For example all RTA measurements in P2I are aggregated into a single CDF, regardless of their specific source PoPs and destinations.

The official choice for the ANPI was utilizing 2 layers, which is further explained in 6.1 which presents a comparison between all 3 forms of aggregation.

5.2 Formulas

Now that the fundamental design choices have been presented and motivated, we define the ANPI mathematically.

Let a context be defined as a tuple

$$c = (p,d,m,s),$$

where p denotes a PoP, d a destination, m a measure, and s a scenario. For each context c, let X_c be the random variable representing the observed *delta-min* values under that context.

5.2.1 Unidimensional Subscore

Given a threshold T and a context c, the unidimensional subscore is defined as the probability that the metric value satisfies the threshold:

$$score_{\text{uni}}(T,c) = \mathbb{P}(X_c \leq T).$$

Equivalently, for a finite set of observations $\mathcal{X}_c$, this can be estimated empirically through the ECDF

$$score_{\text{uni}}(T,c) = ECDF_{\mathcal{X}_c}(T)$$

When the context c has a well-defined historical minimum, denoted by $\min_c$, we define a multiplicative version of the threshold. In this case, the acceptable threshold is $T \cdot \min_c$, and the score becomes

$$score^{\text{mult}}_{\text{uni}}(T,c) = \mathbb{P}(X_c \leq T \cdot min_c).$$

or,

$$score^{\text{mult}}_{\text{uni}}(T,c) = ECDF_{\mathcal{X}_c}(T \cdot min_c)$$

This gives more robustness to the evaluation, since it dynamically adapts the threshold

according to the particular context. In this sense values in a context are compared having its historical minimum as baseline, through which rich contextual information is added to the evaluation, reducing distortions caused by structural differences between contexts.

5.2.2 Aggregated Subscore

Let $\mathcal{C}$ be a set of contexts associate with a single scenario and measure, whose scores we want to aggregate.

Given a set of scores $score_{\text{uni}}(T,c)$ for all $c \in \mathcal{C}$, the aggregation is performed through the harmonic mean in 2 steps.

- 1) Firstly scores who share the same PoP are aggregated:

$$score_{\text{pop}}(T,p,s) = \frac{|\mathcal{C}_{p,s}|}{\sum_{c \in \mathcal{C}_{p,s}} \frac{1}{score_{\text{uni}}(T,c)}}.$$

  where $\mathcal{C}_{p,s}$ is the set of contexts associated to PoP p in scenario s

- 2) Then all subscores produced in the previous step are aggregated,

$$score(T,s,m) = \frac{|\mathcal{P}|}{\sum_{p \in \mathcal{P}} \frac{1}{score_{\text{pop}}(T,p,s)}}.$$

  where $\mathcal{P}$ is the set of PoPs.

The multiplicative-threshold version is defined analogously.

This process of first calculating a unidimensional-score by ECDFs and subsequently aggregating the subscores with two application of harmonic means will be refered to as 'aggregation level 0'.

5.2.3 PL and AD

Now that we have defined the means through which a set of contexts is reduced to a single score, we will show how that process is applied to produce the PL, AD components of ANPI.

The AD component of the new score is very similar to the analogous component of the previous one. The main difference is that R_{p2p}^{ping} R_{p2i}^{ping}, R_{p2i}^{http}, R_{p2i}^{dns}, R_{p2e}^{ping}, R_{p2e}^{http} and R_{p2e}^{dns}

are now determined not by simple averages but by the correspondent CDFs and harmonic mean aggregation strategies. The weights selected are the merge of utilizing the same weights as the previous score and applying the normalization that keeps RM between 0 and 10/3.

The new PL component also follows the previous one closely at high level, with the main differenc being the way P_{p2p}, P_{p2e}, and P_{p2i} are derived, now from ECDFs and Harmonic means.

For PL, first we calculate the packet loss scores of every scenario:

$$PL_{P2P} = score(T,P2P,PL)$$

,

$$PL_{P2I} = score(T,P2I,PL)$$

and

$$PL_{P2E} = score(T,P2E,PL)$$

Then, these scores are aggregated through a weighted average inherited from the baseline index and rescaled:

$$PL = \frac{10}{3}(0.44 \cdot PL_{P2P} + 0.28 \cdot PL_{P2I} + 0.28 \cdot PL_{P2E})$$

Regarding AD, RTA scores are calculated for all three scenarios, whereas HTTP and DNS scores are calculated for the P2I and P2E scenarios.

$$AD_{P2P}^{RTA} = score(T,P2P,RTA)$$
$$AD_{P2I}^{RTA} = score(T,P2I,RTA)$$
$$AD_{P2I}^{DNS} = score(T,P2I,DNS)$$
$$AD_{P2I}^{HTTP} = score(T,P2I,HTTP)$$
$$AD_{P2E}^{RTA} = score(T,P2E,RTA)$$
$$AD_{P2E}^{DNS} = score(T,P2E,DNS)$$
$$AD_{P2E}^{HTTP} = score(T,P2E,HTTP)$$

Then these scores are scaled and added within each scenario:

$$AD_{P2P} = 1.467 \cdot AD_{P2P}^{RTA}$$

$$AD_{P2I} = 0.311 \cdot (AD_{P2I}^{RTA} + AD_{P2I}^{DNS} + AD_{P2I}^{HTTP})$$

$$AD_{P2E} = 0.311 \cdot (AD_{P2E}^{RTA} + AD_{P2E}^{DNS} + AD_{P2E}^{HTTP})$$

And the Average Delay component of ANPI can be defined as:

$$AD = AD_{P2P} + AD_{P2E} + AD_{P2I}$$

The relative weights assigned to the P2P, P2I, and P2E scenarios in the PL and AD components were inherited from the baseline index. In both components, these scenarios receive approximately 44%, 28%, and 28% of the corresponding score budget, respectively. This weighting assigns greater importance to P2P measurements, which represent connectivity within the RNP backbone and are therefore directly related to its primary role of interconnecting the institutions served by the network.

5.2.4 AA

The new AA component differs from the previous score by avoiding a weighted average over each PoP's mean availability. Instead, it evaluates the probability that a given availability measurement meets or exceeds a threshold T. Equivalently, using the unavailability fraction $X_U = 1 - X_A$, the same score can be computed as the probability that $X_U \leq 1 - T$. This equivalent formulation is used in the experiments so that lower raw values consistently represent better performance. Finally, the score is multiplied by a scaling factor of 10/3.

$$AA = \frac{10}{3} \cdot \mathbb{P}(X_A \geq T)$$

It should also be mentioned that while the Threshold can be altered, the baseline in this work is 99.91%, as defined in the previous score.

5.2.5 Final Score

Finally, the ANPI can be defined as follows:

$$ANPI = AD + PL + AA$$

5.3 Considerations

The goal of this section is to present some details of the different steps of the ANPI calculation giving further detail on how it is performed.

It is worth noting also, that measurements that do not happen on Brazilian business days are excluded from the score calculation. This happens since the RNP network is aimed at catering to Universities and Research, which are way more active during business days than during the rest of the period. In other words, in holidays and weekends the network is rarely stressed, and thus can bring the score up for lack of traffic demand.

As in the baseline index, we exclude from the Packet Loss dataset rows where the loss is too high. This high loss of packets will already be accounted in the Availability related section of the score. Thus, a single failure will be prevented from receiving double penalty.

Finaly, the output of ANPI must represent the quality of performance of the network in a number between 0 and 10. For this goal each of the three dimensions of assessment (Packet Loss, Latency and Availability) must have equal weight and thus are assigned a budget of 10/3 regarding the score. This is ensured by the scaling factors introduced in each of the AD, PL and AA sections of ANPI.

5.4 ANPI Strengths

By building upon the structure of the baseline index and incorporating historical minima, ECDFs, harmonic-mean aggregation, and dynamic thresholds, ANPI preserves the main strengths of the original methodology while introducing additional mechanisms to improve sensitivity to local performance degradation and context-specific irregularities.

As already briefly touched upon, the processing of the data into the delta_min column before the score calculation, together with the dynamic multiplicative thresholds in

consideration the particularities of each context manage to circumvent the rigidness of the previous score calculation, providing each context with its own tailored scoring process.

Furthermore, the new score keeps the previous one's flexibility to add new axis, measurements, and use cases quite naturally. Its flexibility also shows itself in the possibility of evaluating the network not only on a global level, but also locally. We can, if desired, ignore sections of the dataset and focus only in a particular slice concerning a subset of the network. As an example, it is possible to analyze the link between 2 particular PoPs, by considering only data pertaining to their relationship.

ANPI as a whole, but also its core mechanisms (ECDF, HM) are also general in use, being fit for employment in many different networks, easily adding new measures, sources, destinations, scenarios and other data stratification layers.

It is also important to highlight one of the major contributions of the ANPI, which is the more realistic assessment of the network performance by aggregation methods that do not mask local irregularities. The employment of the harmonic mean instead of the arithmetic mean or the median, gives the power of better avoiding false negatives when searching for problems in the network, and thus better aids network managers to intervene when necessary.

The next chapter will further explore ANPI characteristics by means of experimentation.

6 ANPI: Experimental Results

After introducing and describing ANPI, our next goal is to explore its capabilities and possibilities.

This chapter, then, presents some experiments regarding parameter variation. Furthermore, the flexibility of the proposed score is explored, providing different set-ups that may appeal to varied management constraints and objectives.

6.1 Aggregation Levels

As previously stated, we are interested in comparing 3 forms of aggregation levels, depending on how the CDF and Harmonic Mean are combined. Their definitions are presented in 5.1.4.

Since we are comparing aggregation levels, and dynamic thresholds are only applicable when the aggregation level is 0, every threshold considered in this section was fixed.

Table 8 shows the impact of the aggregation level on the PL component. The score increases from 3.3031 at aggregation level 0 to 3.3033 at level 1 and 3.3036 at level 2. This limited variation is consistent with the high number of zero-valued observations in the packet-loss dataset. Regardless of the aggregation method, the PL component remains close to its upper limit, indicating that the aggregation level has only a negligible influence on this component for the analyzed dataset.

Tables 9 and 10 present, respectively, the P2I subscores for DNS resolution, HTTP, and RTA at a fixed threshold of 30, and the resulting AD component for multiple thresholds. Both tables show that configurations with a greater number of harmonic-mean aggregation steps produce stricter scores.

For instance, the DNS-resolution subscore increases from 0.0657 at aggregation level 0 to 0.3769 at level 2, representing an increase by a factor of approximately 5.7. Similarly,

for a threshold of 30, the AD component increases from 1.6846 at level 0 to 2.1715 at level 2, a difference of 0.4869. The same ordering is observed for every threshold reported in Table 10.

The influence of the aggregation strategy is greater for stricter thresholds. For example, the difference between levels 0 and 2 is 0.8004 at a threshold of 10, but decreases to 0.2390 at a threshold of 100.

These results suggest that aggregation level 0 is more appropriate when greater sensitivity to performance degradation is desired. Aggregation levels 1 and 2 provide more lenient evaluations and may be preferred when a smaller penalty for context-specific degradation is intended.

Table 8: PL by aggregation level

Aggregation level	Threshold	PL
0	0.5	3.3031
1	0.5	3.3033
2	0.5	3.3036

Table 9: Scores by measure and aggregation level for threshold = 30

Scenario	Measure	Threshold	Aggregation level	Score
p2i	`Dns_Resolve`	30	0	0.0657
p2i	`Dns_Resolve`	30	1	0.0706
p2i	`Dns_Resolve`	30	2	0.3769
p2i	HTTP	30	0	0.0003
p2i	HTTP	30	1	0.0104
p2i	HTTP	30	2	0.0666
p2i	RTA	30	0	0.5543
p2i	RTA	30	1	0.7015
p2i	RTA	30	2	0.8713

Table 10: AD results by aggregation level and threshold

Aggregation level	Threshold	AD
0	10	0.7228
1	10	1.1134
2	10	1.5232
0	20	1.3653
1	20	1.5810
2	20	1.9237
0	25	1.5272
1	25	1.7191
2	25	2.0641
0	30	1.6846
1	30	1.9097
2	30	2.1715
0	50	2.1800
1	50	2.2916
2	50	2.4557
0	100	2.5303
1	100	2.6307
2	100	2.7693

6.2 Threshold

Concerning thresholds, there are two types of variation to explore: regarding the strategy, which determines whether the threshold is fixed or dynamic, and the threshold value itself. While it makes sense to vary both when studying the AD component, thresholds for the PL component are always fixed, since its historical minimum is always zero. The AA component does not use historical minima due to the nature and size of its dataset; thus, only the threshold parameter is varied for this component as well.

For the AD component, both thresholding strategies produce the same monotonic effect as the thresholds increase, as shown in Table 11. For fixed thresholds, for example, AD increases from 0.7228 to 1.6846, and then to 2.5303, when the threshold changes from 10 ms to 30 ms and then to 100 ms. Likewise, for dynamic thresholds, AD increases from 2.1028 to 3.0779, and then to 3.2619, when multiplicative factors of 2, 5, and 10 are considered. Therefore, under both strategies, larger thresholds result in more lenient evaluations and higher AD scores.

Table 12 shows that PL does not change given different thresholds in this particular dataset, remaining at 3.3031, which is a consequence of the value 0 being extremely prevalent in the data. Similarly, Table 14 displays how slowly AA changes when we vary thresholds. In fact every single value between 99.76 and 99.99 chosen to be the threshold

Table 11: Effect of fixed and multiplicative thresholds on the AD component at aggregation level 0

Mult	Threshold	AD
False	10	0.7228
False	20	1.3653
False	25	1.5272
False	30	1.6846
False	50	2.1800
False	100	2.5303
True	2	2.1028
True	3	2.4898
True	5	3.0779
True	7	3.1935
True	10	3.2619

Table 12: PL by threshold with `Aggregation level` $= 0$

Aggregation level	Threshold	PL
0	0.5	3.3031
0	1.5	3.3031
0	2.5	3.3031

will yield the same result. Furthermore, comparing 15.00 to 99.99 shows a difference of less than 0.02. Again this stems from the low quantity of samples with availability less than 100.00.

6.3 Weights

In the baseline index, a weight was assigned to each PoP for the calculation of the availability component. These weights represent the relative infrastructural and connectivity relevance of each PoP. The weight of each PoP is obtained through criterion-specific scores derived from the following characteristics:

- Number of links $\geq$ 100 Gb/s
- PoPs $\geq$ 100 Gb/s
- International physical connections
- Customers with IP transit via RNP
- IXPs and SFPs (Gb/s)

- PNI and Cache (public cloud)
- RNP services (CND, IDC, etc.)
- IP transit (providers)

The final PoP weight corresponds to the sum of these criterion-specific scores. Complete scoring ranges and rules are described in (Rede Nacional de Ensino e Pesquisa, 2025) and the resulting weights presented in Table 13.

Table 13: PoP weights by Brazilian region

Region	States and weights
North	AC (8), AP (11), AM (13), PA (13), RO (8), RR (8), TO (8)
Northeast	AL (10), BA (25), CE (33), MA (10), PB (10), PE (21), PI (13), RN (12), SE (8)
Central-West	DF (25), GO (10), MS (8), MT (8)
Southeast	ES (10), MG (20), RJ (26), SP (39)
South	PR (19), RS (25), SC (18)

Source: (Rede Nacional de Ensino e Pesquisa, 2025)

The following experiments assess the influence of these weights on AA and evaluate whether their use should be extended to the AD and PL components of ANPI.

6.3.1 Weighted Availability

Since the new score leverages a CDF instead of an average to aggregate availability measurements, weights cannot be applied as directly as in a weighted mean. Thus, to circumvent this issue, measurements for each PoP were repeated according to the corresponding weight. Therefore, the empirical availability distribution changes, and consequently so does the CDF, giving PoPs with higher weights greater influence on its shape.

Table 14 shows that weighting has a minimal effect on the AA component for this particular dataset. At thresholds of 15, 20, and 50, the absolute differences between the weighted and unweighted scores remain below 0.001. The largest difference is observed at thresholds of 99.76, 99.91, and 99.99, for which AA increases from 3.3136 to 3.3189, corresponding to a difference of only 0.0053.

Table 14: Weighting impact on AA

Threshold	10.00	15.00	20.00	50.00	99.76	99.91	99.99
AA (unweighted)	3.3300	3.3293	3.3286	3.3286	3.3136	3.3136	3.3136
AA (weighted)	3.3300	3.3296	3.3292	3.3292	3.3189	3.3189	3.3189

6.3.2 Weighted Latency

For aggregation levels 0 and 1 of Latency, the addition of weights is straightforward, since they can be directly incorporated into the harmonic means used to aggregate the PoP scores. Aggregation level 2 presents an additional challenge, since it does not use a mean at the PoP aggregation stage. The solution was therefore to adopt the same strategy used for the Availability component.

Here, the analysis is restricted to aggregation level 0. For fixed thresholds, Table 16 shows that the largest difference between the weighted and unweighted versions is 0.0535, with AD increasing from 0.7228 to 0.7763 at a threshold of 10 ms. Likewise, for dynamic thresholds, Table 15 shows that the largest difference is 0.0979, with AD decreasing from 2.1028 to 2.0049 at a multiplicative threshold of 2. Therefore, although the differences are larger than those observed for the Availability component, weighting produces only limited changes in the aggregated AD component for the analyzed dataset.

Table 15: Weighting impact on AD with `Aggregation level` = 0 and dynamic thresholds

Aggregation level	Threshold	Mult	Weighted	AD
0	2	True	False	2.1028
0	2	True	True	2.0049
0	3	True	False	2.4898
0	3	True	True	2.4067
0	5	True	False	3.0779
0	5	True	True	3.0678
0	7	True	False	3.1935
0	7	True	True	3.1869
0	10	True	False	3.2619
0	10	True	True	3.2608

Table 16: Weighting impact on AD with `Aggregation level` = 0 and fixed thresholds

Aggregation level	Threshold	Mult	Weighted	AD
0	10	False	False	0.7228
0	10	False	True	0.7763
0	20	False	False	1.3653
0	20	False	True	1.3970
0	25	False	False	1.5272
0	25	False	True	1.5504
0	30	False	False	1.6846
0	30	False	True	1.7037
0	50	False	False	2.1800
0	50	False	True	2.2331
0	100	False	False	2.5303
0	100	False	True	2.5644

6.3.3 Weighted Packet Loss

Because of its similar data structure, the application of weights to the PL component is analogous to their application to the AD component. Table 17 shows that, across all three aggregation levels, weighting increases PL only slightly, from approximately 3.3031 to 3.3100. This result follows the previously observed pattern of negligible variation in the PL component for the analyzed dataset.

Table 17: Weighting impact on PL

Aggregation level	Threshold	Weighted	PL
0	0.5	False	3.3031
0	0.5	True	3.3096
1	0.5	False	3.3033
1	0.5	True	3.3098
2	0.5	False	3.3036
2	0.5	True	3.3100

6.3.4 Weighted Score

From the definition of the weighted version of each component, it is now possible to explore the effect of weighting on the final score.

Table 18 shows that, for the stricter dynamic thresholds of 2 and 3, weighting has a slightly negative impact, decreasing ANPI from 8.7196 to 8.6335 and from 9.1066 to 9.0352, respectively. These changes correspond to reductions of 0.0861 and 0.0714. For the remaining dynamic thresholds, the combined influence of weighting on PL, AD, and

AA produces small increases ranging from 0.0017 to 0.0106. A similarly limited effect can be observed for fixed thresholds, with the largest difference being 0.0653 at a threshold of 10 ms. Therefore, the effect of weighting is not consistently positive or negative and remains limited across all evaluated configurations.

Table 18: Weighting impact on ANPI with fixed PL and AA thresholds

Aggr. AD	Aggr. PL	PL thr	AA thr	AD thr	AD mult	Weighted	ANPI
0	0	0.5	99.91	2	True	False	8.7196
0	0	0.5	99.91	2	True	True	8.6335
0	0	0.5	99.91	3	True	False	9.1066
0	0	0.5	99.91	3	True	True	9.0352
0	0	0.5	99.91	5	True	False	9.6946
0	0	0.5	99.91	5	True	True	9.6963
0	0	0.5	99.91	10	True	False	9.8787
0	0	0.5	99.91	10	True	True	9.8893
0	0	0.5	99.91	10	False	False	7.3395
0	0	0.5	99.91	10	False	True	7.4048
0	0	0.5	99.91	20	False	False	7.9820
0	0	0.5	99.91	20	False	True	8.0255
0	0	0.5	99.91	30	False	False	8.3013
0	0	0.5	99.91	30	False	True	8.3322
0	0	0.5	99.91	50	False	False	8.7968
0	0	0.5	99.91	50	False	True	8.8616

6.4 Baseline Comparison

When computing the baseline index with the same dataset we get the following results:

$$Baseline\ Index = AD + PL + AA = 33.2344 + 33.3000 + 32.6299 = 99.1643$$

which scaled to yield a score from 0 to 10 is 9.9164.

This is higher than any of the derived ANPI values in Table 18, which corroborates a higher leniency for performance faults.

This discrepancy, however, does not mean one of the two approaches is objectively wrong, since there is no ground truth. Furthermore, the aim of the present work should not be viewed through the lenses of accuracy, since the evaluation itself is being discussed.

Our goal is not one of numerical error minimization, but of devising an evaluation method which is more sensitive to local abnormality, enlarges the window for improvement, and ensures a solid methodological basis to performance assessment.

6.5 Takeaways and Recommendations

The experiments suggest that methodological soundness, as theoretically defended in 5, is also empirically supported. The evidence points to the desired properties of local sensitivity and amplification of enhancement possibility. Nonetheless, specific parameters still need to be determined.

Considering that aggregation level 0 showed itself to give the highest penalties for failures to comply, it seems to be the most suitable for the identification of irregularities. This, combined with the possibility of dynamic thresholding, motivates its suitability.

Regarding thresholds, the dynamic approach provides a highly valuable degree of context customization. Furthermore, it provides a new layer of historical minimum information which better captures local particularities and reduces context-related bias.

In terms of magnitude, threshold selection depends on operational objectives. Stricter requirements should lead to lower thresholds, while more lenient settings should favor higher ones.

Therefore, in the remaining chapters, ANPI will be applied with the following parameters:

- AD and PL Aggregation Level: 0
- AD Threshold: 2, multiplicative
- PL Threshold: 0.5
- AA Threshold: 99.91
- Weighted: False

Regarding AD, the multiplicative threshold of 2 was chosen due to its strictness and contextual flexibility, both desirable qualities for this work's goals. PL and AA thresholds, on the other hand, were inherited from the baseline index, while weighting was abandoned since it presented almost no impact on the present dataset.

7 Outlier Detection: The CCBOS

After addressing general performance assessment through the ANPI, we now focus on identifying irregular monitoring data and the spatiotemporal contexts in which they occur, specifically through the introduction of CCBOS.

The main goal of outlier detection is to distinguish common events from those that deviate from normality and therefore require closer inspection. In this classification paradigm, two main types of errors may occur:

1. **False positives:** normal behavior is labeled as outlier.
2. **False negatives:** outlier behavior is labeled as normal.

False negatives are generally more serious, since actual problems may go undetected. However, false positives are also an issue, as they may significantly increase the number of data points requiring analysis. In extreme cases, this number may exceed the available operational capacity, increasing the risk that correctly identified outliers are ignored when selecting which positive cases, whether true or false, should be investigated. Since many outlier detection systems must operate in real time, deal with extensive quantities of data and limited resources, complexity is also an important factor to consider when selecting suitable outlier detection methods.

In this chapter, we first describe the necessary data alignment process for the application of such detection techniques on the dataset of interest, motivating data processing choices and preferences.

We then present the CCBOS (Complementary Cumulative-Based Outlier Score), a novel outlier detection technique which combines HBOS's interval and density-based calculation with the cumulative characteristics of ECOD.

7.1 Problem Definition

Given the measurement datasets used to derive ANPI, we aim to identify tuples of the form

(source PoP,destination,timestamp)

associated with measurements that contribute most strongly to performance degradation. The objective is to identify the spatiotemporal measurement contexts most strongly associated with performance degradation, thereby supporting network managers in making informed decisions about network improvement.

To evaluate each spatiotemporal context, the measurements from the different performance dimensions must first be aligned according to source PoP, destination, and timestamp. Each aligned row then represents a specific spatiotemporal context and contains the measurements associated with that context. These measurements jointly contribute to its outlier score.

An outlier-scoring method is subsequently applied to each row of the aligned dataset. The k rows with the highest scores are then classified as outliers. This subset represents the spatiotemporal contexts that should receive the highest priority for further inspection, helping network managers determine which locations, time intervals, and measurements should be investigated to improve network performance.

7.2 Data Alignment

7.2.1 P2P

Our initial strategy consisted of associating, for each scenario, each pair of PoPs with a time series composed of a set of simultaneous measurements.

For P2P, alignment was done in the following way: inside the same context of source and destination, RTA and packet loss measurements were paired through temporal closeness. Since each ping measurement happened approximately in intervals of 1 minute, the maximum window considered for pairing was 29 seconds.

Spatiotemporal contexts for which some matching metric was not found were paired with the value zero. This choice was made because we aim to identify the contexts associated with the highest levels of performance degradation. By assuming that missing measurements correspond to perfect performance, represented by zero, we avoid artificially

increasing the outlier score of the corresponding context. If, however, a context is still classified as an outlier under this assumption, its available measurements already provide sufficient evidence of degradation.

Table 19 presents examples of rows of the aligned P2P data.

Table 19: Example of the P2P Aligned Dataframe

Time	Pop	Dest.	Month	PL	Norm_delta_min	Unavailability
2025-01-01 00:00:22.679355189+00:00	RJ	SP	1	0.0	0.142216	0.0
2025-01-01 00:00:22.688940409+00:00	RJ	DF	1	0.0	0.245079	0.0
2025-01-01 00:00:22.699233175+00:00	RJ	SC	1	0.0	0.094241	0.0
2025-01-01 00:00:22.709253171+00:00	RJ	PE	1	0.0	0.028581	0.0
2025-01-01 00:00:22.759197487+00:00	RJ	AC	1	0.0	0.319027	0.0

7.2.2 P2I

Regarding P2I, RTA, packet loss and dns resolve measurements were aligned in the same fashion as described in 7.2.1. HTTP data, however, does not possess the same destinations as the RTA, packet loss and dns resolve (IPs). Because of that, each different URL in P2I HTTP measurements was assigned its own column, and these columns were merged into the aligned dataset by source PoP and timestamp. Again, missing data was imputed by 0.

Table 20 presents examples of rows of the aligned P2I data.

Table 20: Example of the P2I Aligned Dataframe

Time	Pop	Dest.	PL	Norm_delta_min	Amazon DNS	Mconf	RNP.br	ViaIPE	Unavailability
2025-01-01 00:00:12.099487903+00:00	AC	200.19.16.53	0.0	0.022510	0.180723	1.239067	0.934884	0.572770	0.0
2025-01-01 00:00:12.111131089+00:00	AC	200.137.53.53	0.0	0.021678	0.168675	1.239067	0.934884	0.572770	0.0
2025-01-01 00:00:12.650234860+00:00	AL	200.19.16.53	0.0	0.054784	0.348485	0.231034	0.118605	0.251208	0.0
2025-01-01 00:00:12.663217974+00:00	AL	200.137.53.53	0.0	0.325573	2.682540	0.231034	0.118605	0.251208	0.0
2025-01-01 00:00:13.163272788+00:00	AM	200.19.16.53	0.0	0.373800	0.500000	0.216015	0.799692	1.289377	0.0

7.2.3 P2E

P2E alignment was performed similarly to P2I for RTA, packet loss and http measurements. Three different IPs were shared by them as resolvers. Dns resolve measurements, on the other hand, were separated by URL: Amazon, Microsoft, Facebook and Google. Google, Microsoft and Facebook all possessed all three aforementioned destinations in their subsets of measurements, however, the first of them possessed an extra destination, while the other two possessed 2. Each of those URLs was assigned a column for measurements that shared destinations with the already aligned dataset, and thus merged

by PoP, timestamp and destination. For each of the extra URL-destination pairs a new column was created and merged to the aligned dataset by PoP and time only.

In the case of Amazon, only two destinations were represented (both shared by rta, etc.), for each of them a new column was created and merged by PoP, time and destination.

The complete list of columns in the aligned P2E dataset is:

- `_time`
- `_value_pl`
- `pop`
- `dest`
- `norm_delta_min`
- `norm_delta_min_apple`
- `norm_delta_min_aws`
- `norm_delta_min_facebook`
- `norm_delta_min_google`
- `norm_delta_min_gov.br`
- `norm_delta_min_dns_facebook`
- `norm_delta_min_dns_microsoft`
- `norm_delta_min_dns_facebook_200_137_53_53`
- `norm_delta_min_dns_google_200_137_53_53`
- `norm_delta_min_dns_microsoft_200_137_53_53`
- `norm_delta_min_dns_facebook_200_19_16_53`
- `norm_delta_min_dns_microsoft_200_19_16_53`
- `norm_delta_min_dns_amazon_1_1_1_1`
- `norm_delta_min_dns_amazon_8_8_8_8`
- `indisp_frac`

7.2.4 Availability Alignment

Since lower values of RTA and packet loss indicate better performance, whereas higher availability indicates better performance, we decided to use the unavailability fraction instead of the availability fraction. This transformation keeps the interpretation of all metrics consistent: higher values indicate worse network conditions.

The original data provided daily availability fraction per PoP. Therefore, each row in the P2I and P2E aligned datasets was assigned the unavailability fraction corresponding to its date.

This implies the approximation that availability is uniformly distributed throughout the day. Higher data granularity could weaken this hypothesis, assigning different values throughout the same day depending on specific downtime event timestamps.

For the P2P scenario, the same procedure was applied, but the availability values of both the source and destination PoPs were combined, thus forming link availability data.

7.3 CCBOS

CCBOS (Complementary Cumulative-Based Outlier Score) is based on HBOS, with an additional step after the bin counts are computed: each bin receives, in addition to its sample count, the quantities of every bin that comes after it. In this way quantities are monotonically decreasing, regarding bin order. So, after height normalization and inversion, bins that come first will necessarily be associated to lower outlier scores. Conversely latter bins will receive higher scores, by design.

Now, not only density information is considered for scoring, but also the magnitude of the values of every single attribute. This was done for, in our data context, outliers necessarily happen in the right tail of the distributions.

This also prevents lower density regions with small metric values from being assigned high outlier probability, avoiding the undesired effect of the original HBOS.

7.3.1 Mathematical Description

The formula for the CCBOS is as follows:

For each attribute i, a histogram is computed yielding bins with height H_j^i where j is

the number of the bin.

Then an accumulated histogram is created with bins of height C_j^i in the following manner:

$$C_j^i = \sum_{k \geq j} H_k^i$$

In this way (since all heights are non-negative) the new heights of the bins are monotonically non-increasing, and actually monotonically decreasing if every bin has at least one sample.

It is also worth noting that the cumulative process happens backwards, resulting in a discretization of the unnormalized CCDF.

Then (as in HBOS) the heights are normalized so that the highest one is equal to 1.

After the bins have been accumulated and normalized, each sample p (i.e. each aligned row) has its CCBOS score computed as:

$$\text{CCBOS}(p) = -\sum_{i=1}^{d} \log(\text{C}_i^{norm}(p))$$

where d is the number of attributes considered and C_i^{norm} is the normalized height of the accumulated bin where the i-th attribute of p lies.

Even though CCBOS and ECOD both leverage cumulative distributions for considering contexts where outliers happen in the tails, CCBOS keeps HBOS feature of binning, which will prove itself useful in experiments and is motivated in 7.4.2.

7.4 Design Choices

7.4.1 Delta Min Factor and Global Density

Figure 5 highlights an interesting phenomenon observed at the beginning of the analysis: when degraded performance becomes the norm, performance improvements may be incorrectly identified as outliers if the algorithm considers only density information (such as HBOS). In the plot, we can see that most of the data samples are concentrated after ANPI's threshold. If only density was considered for outlier scoring, this abnormal peak would be assigned low attribute scores, while data related to good performance would

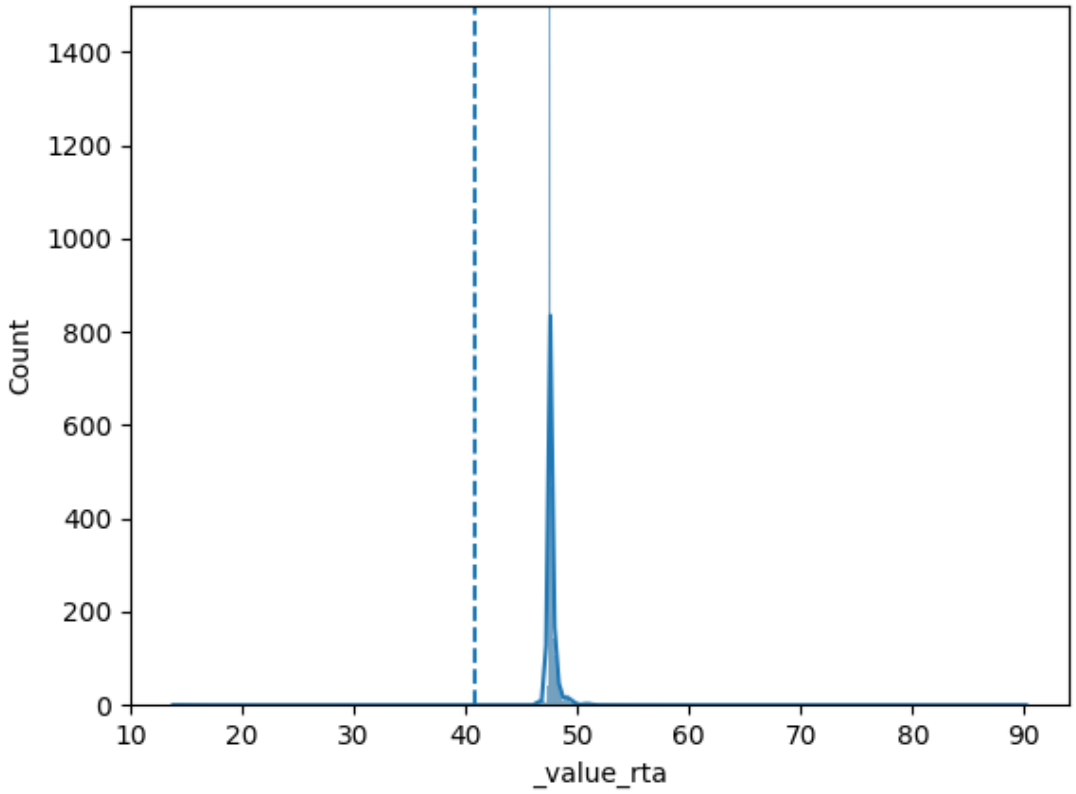

Figure 5: This graphic presents a phenomenon where the values of a measurement possess a peak after the threshold, thus making HBOS assign bigger outlier scores to values under the threshold

strongly contribute to its context to be deemed problematic.

Although this behavior was observed, it occurred only in a limited number of cases and was not representative of the dataset as a whole. Still, detection frameworks must be robust enough to handle such instances, since they may encounter datasets where this kind of undesired behavior is the norm.

Based on this observation, instead of fitting scoring models for each PoP and scenario, a single model was trained for each scenario using data aggregated across all PoPs. The motivation behind this choice was to contextualize these problematic samples at a broader scale. In this aggregated setting, such samples remain in low-density regions, while the large mass of normal data from other PoPs concentrates around the expected behavior. Cases where this unwanted behavior happens are then recontextualized with a large amount of data, which shifts high density regions to the left of the threshold. Of course, this only happens since most of the data do not present this oddity.

This approach requires feature values to be comparable across different PoPs. However, raw latency values are inherently affected by factors such as geographical distance, making direct comparison inappropriate.

To address this issue, the delta-min factor was introduced. Instead of using raw RTA, dns resolve and http values, each observation is subtracted its historical minimum and this difference is expressed as a multiple of the same minimum. This transformation produces a dimensionless quantity that normalizes structural differences between PoPs, enabling meaningful comparisons across them, and achieving its lowest level and best score at 0.

7.4.2 Binning Choices

In the original HBOS paper, histograms could be computed in 2 ways, either by choosing the number of bins and assigning each one the same width (fixed-bin method), or by assigning approximately k values to each bin, thus yielding bins of same area (Dynamic method).

In this work, we adopted a third strategy: bins in the same histogram had a fixed length, but determined by the length itself and not the number of total bins.

Since, in ANPI, every value up to a certain threshold (e.g. 2) is considered as 'perfect', for the outlier detection technique to be well coupled with the index, it was deemed that values that lie before the respective threshold should not contribute to the outlier score, and the contrary for values that surpassed this threshold. Since every bin, but the first, has positive score associated to them (excluding edge cases where no measurements were made lower than the threshold) the bin size was then chosen to be exactly the threshold utilized for the respective metric of the histogram.

The defining features of the chosen (and proposed) outlier detection method are: data alignment, accumulation and threshold-consistent binning. It is worth noting that if the context demands, CCBOS can be utilised with other binning strategies and even consider both tails in the same pipeline as ECOD.

7.4.3 Threshold Selection

After deriving outlier scores for each sample it is necessary to define which of the samples are outliers and which possess normal behavior. In order to reach a good classification, with high precision, but most importantly high recall, a good threshold must be defined

to separate the two groups of samples.

It is important to notice that while false negatives are a bigger problem due to permitting irregular behavior to go unnoticed, a high rate of false positives is also problematic due to the limited resources in analyzing each sample individually in a supervised way.

To balance both requirements many diverse threshold techniques have already been developed. (Komadina *et al.*, 2024) gives a review of many such strategies, providing a categorical structure and performance comparisons.

For the present work, the methodology chosen was simply to get the top k highest scores for different values of k, or use the ECDF of the scores and take the threshold to which a certain percentage of rows are deemed outliers.

8 Outlier Detection and Network Quality Assessment

This chapter investigates the relationship between the two main methodological fronts developed in this work: ANPI and CCBOS.

Although these two components were presented separately in the previous chapters, they are closely related. Together, they address two complementary goals: assessing network performance and identifying the spatiotemporal measurement contexts most strongly associated with performance degradation. While ANPI provides a structured measure of service quality, the outlier detection framework supports the recognition of abnormal behavior and adds a greater degree of explainability to the index.

Therefore, the results presented in this chapter aim to validate the dual objective of this work: to support network managers not only in evaluating quality of service, but also in identifying where problems may occur, thereby guiding future remediation efforts.

8.1 Impact of Outlier Removal on ANPI Components

The first experiment consisted of calculating the ANPI while outliers were gradually removed from the dataset.

The removal process was carried out as follows. For each row of the aligned datasets for P2P, P2E, and P2I, an outlier score was determined. The three sets of outlier scores were then considered jointly as a single set, and thresholds were produced from it by selecting the top-k highest scores, for different values of k. In the following, these top-k values are presented as percentages of the total number of scores.

Using these thresholds, we classified spatiotemporal tuples as outliers. Measurements from the original datasets, prior to alignment, that were associated with these spatiotemporal contexts were subsequently removed before the ANPI was calculated. It is important to note that all measurements associated with a selected context were removed, including

those that did not individually contribute to the score penalization.

The main objective of this experiment was to observe how ANPI changes as more data related to irregular contexts (as determined by CCBOS) is removed. The desired behavior is that this removal would result in better ANPI scores, which would indicate that CCBOS is correctly identifying contexts that degrade performance.

Such contexts, then, would merit particular analysis for an effective understanding of pathologies in the network. In turn, network managers would be rewarded with valuable information when deciding how to optimize quality of service.

In order to validate both CCBOS and its binning choices in this task, five methods were considered for comparison:

- CCBOS
- HBOS
- ECOD
- CCBOS with double-sized bins
- CCBOS with half-sized bins

8.1.1 CCBOS

Figure 6 shows the evolution of the ANPI score as a higher percentage of outliers is removed. Not only does the score increase monotonically but it also goes from approximately 8.7 to over 9.6 with less than 20% of samples being classified as outliers and removed. This validates the usage of CCBOS as an outlier detection method and its consistency with ANPI.

When ANPI is calculated separately for each PoP, Figure 7 shows that the range of scores goes approximately from 9.1–9.5 to 9.8–10.0. This indicates improvement for the respective worst- and best-performing PoPs. Furthermore, the narrowing of ANPI's interval indicates lower performance variability across the evaluated PoPs after the removal procedure.

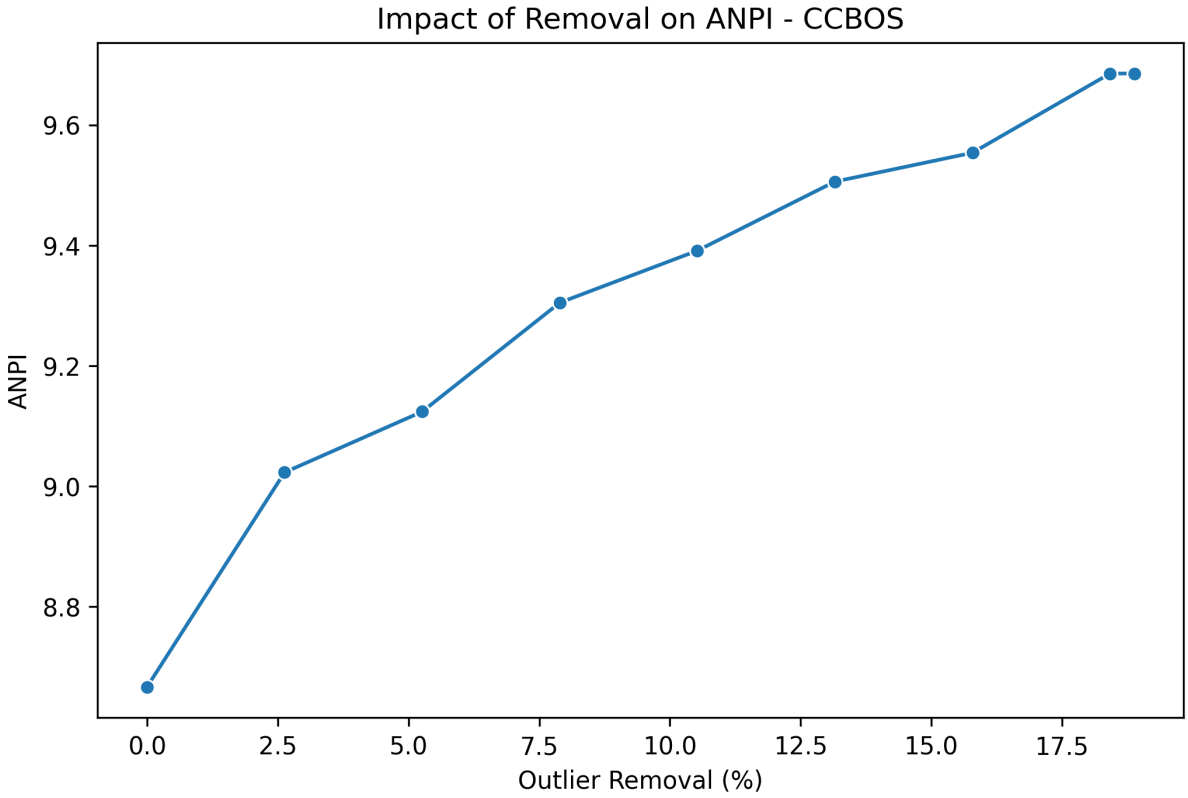

Figure 6: CCBOS outlier removal impact on ANPI

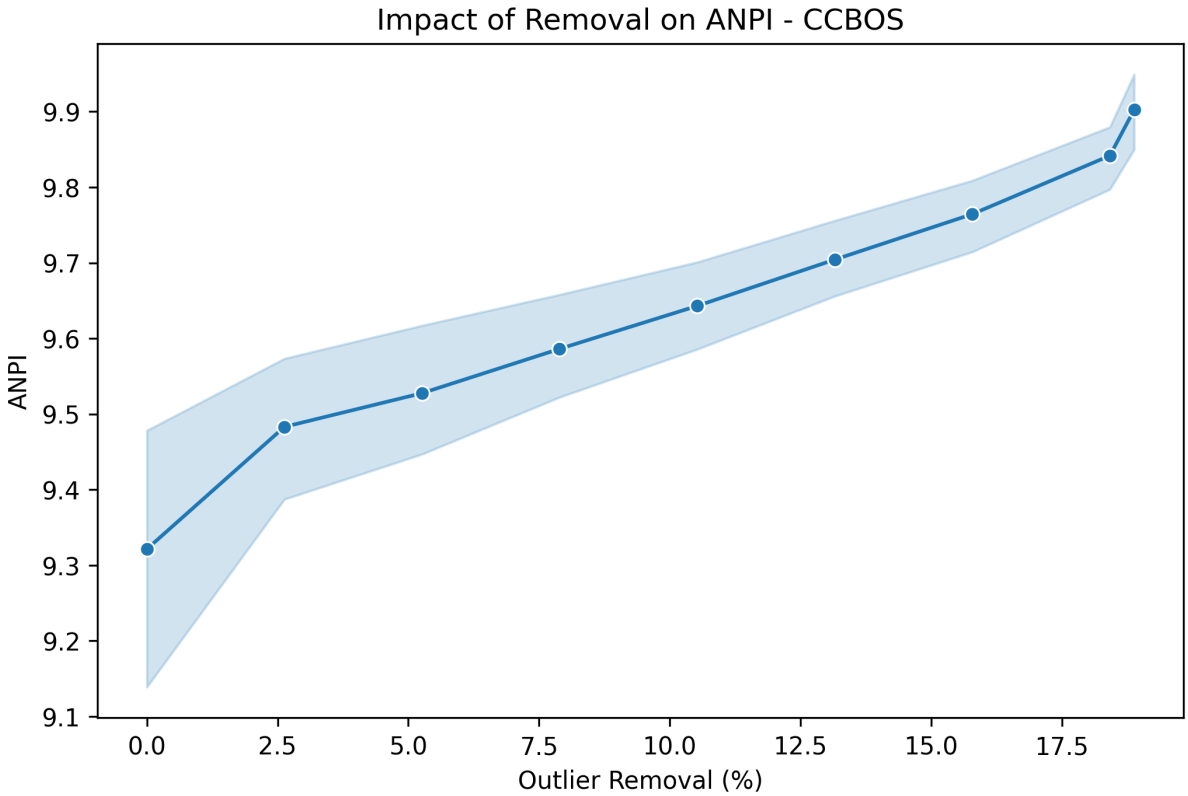

Figure 7: CCBOS outlier removal impact on PoP-specific ANPI. Here ANPI was calculated for each PoP individually, and the interval at each transversal section of this plot represents the interval between the worst and best pop-specific ANPI's. The lineplot in the center represents their arithmetic mean, not the global ANPi

.

8.1.2 HBOS

Figures 8 and 9 show that both in the aggregated and in the PoP segregated cases, HBOS and CCBOS achieve practically the same results.

This can be explained by the correction presented in 7.4.1. The dataset's general healthy profile guarantees the efficiency of this rehabilitation. However, if the oddity to be remediated is largely prominent through different contexts, global re-contextualization of density will not correct local occurrences of the issue.

Here, nonetheless, HBOS histograms became virtually monotonic, even without the accumulation process. In this way, CCBOS's intrinsic effect was also found in HBOS due to the specific and circumstantial properties of the analyzed dataset.

CCBOS, however, possesses the capabilities of maintaining good results even when data does not behave as well, due to the guaranteed monotonicity stemming from its cumulative nature.

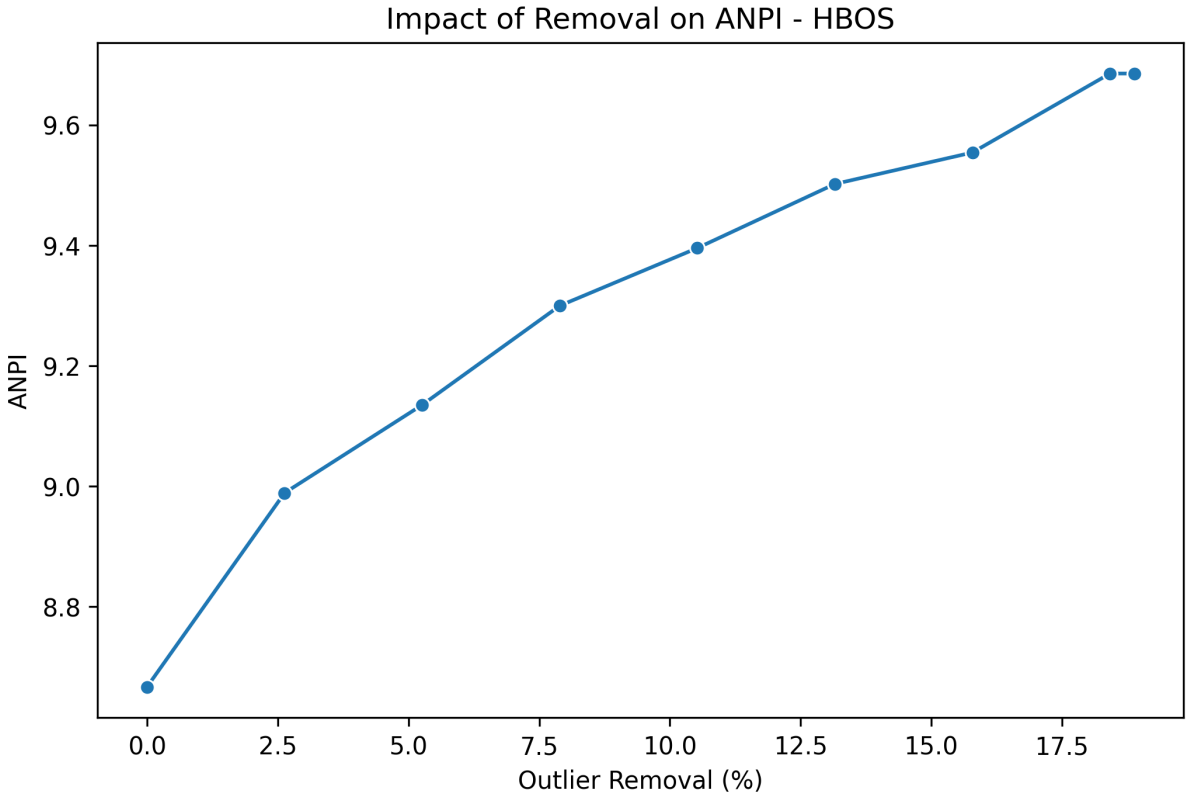

Figure 8: HBOS outlier removal impact on ANPI

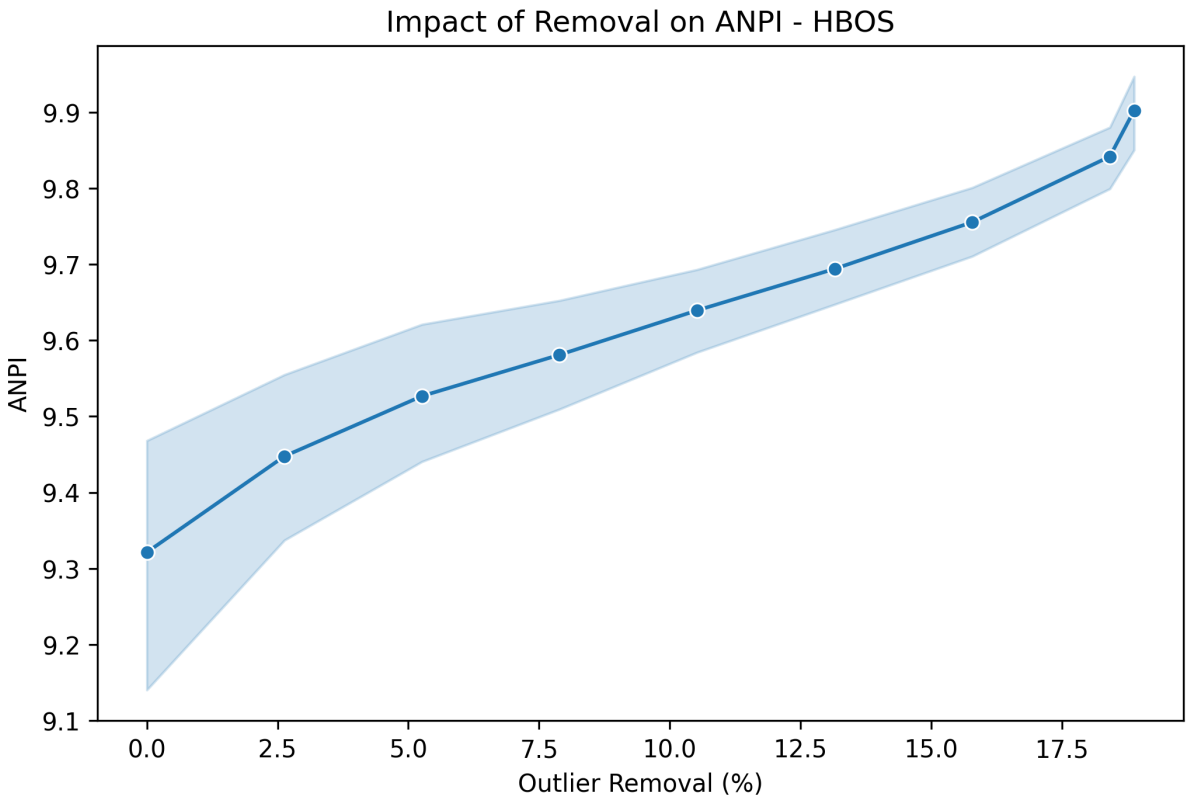

Figure 9: HBOS outlier removal impact on PoP-specific ANPI

8.1.3 ECOD

Figures 10 and 11 present ECOD's removal evolution. Both the drop around 20% in the former and the retention of a large gap between worst- and best-performing PoPs in the latter indicate the lack of stability and uniformity brought about by the absence of discretization by bins, which further motivates HBOS's influence on CCBOS.

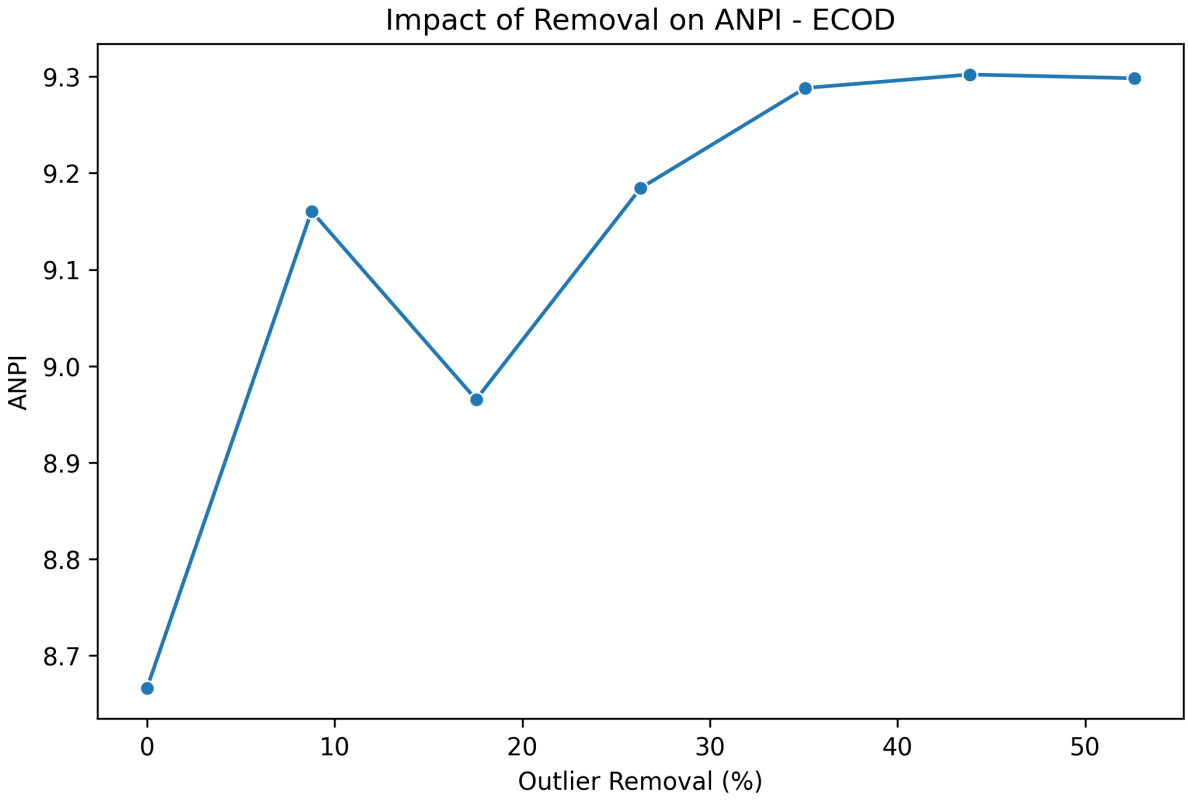

Figure 10: ECOD outlier removal impact on ANPI

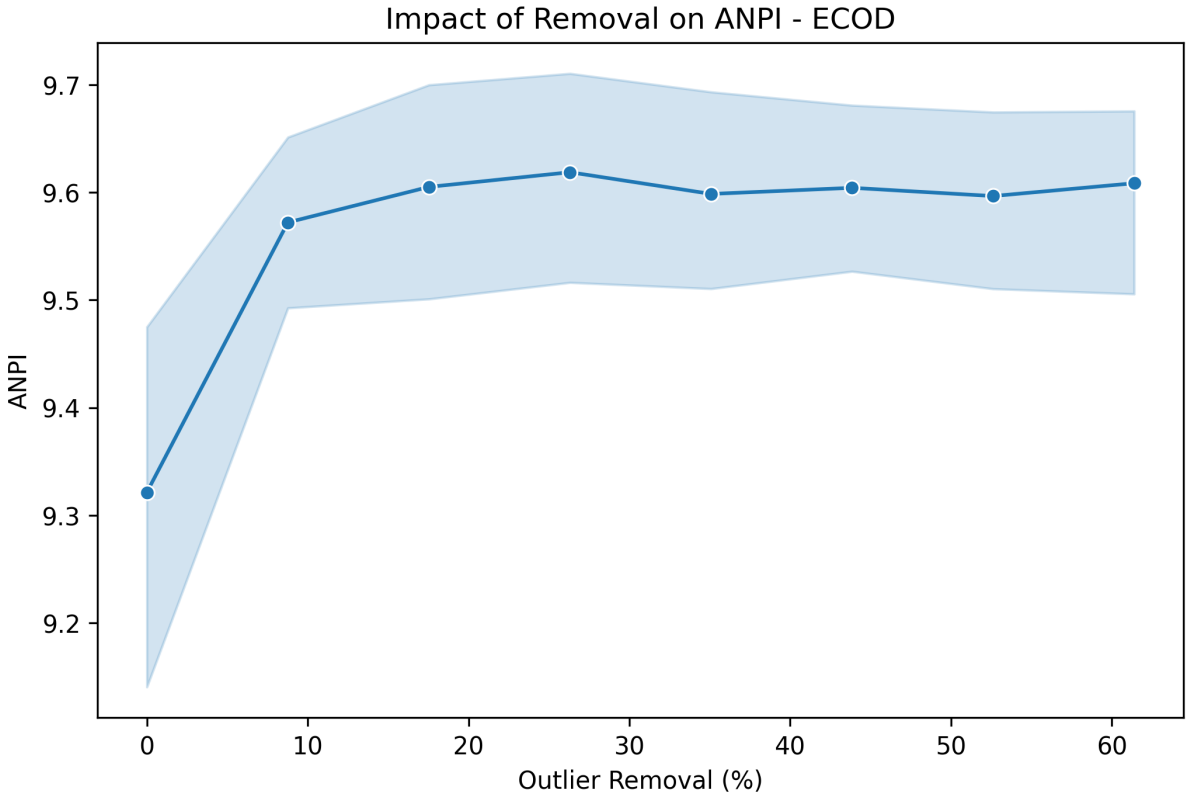

Figure 11: ECOD outlier removal impact on PoP-specific ANPI

8.1.4 CCBOS Double-Sized Bins

When we double the size of the bins, CCBOS performance degrades. This can be explained by samples which fail to comply to ANPI's respective thresholds, but are now contained in the larger first bin.

Such points receive an outlier sub-score of 0, and consequently do not contribute to the penalization of its context. Not only does this create bias, but also allows irregular contexts to receive a CCBOS of 0.

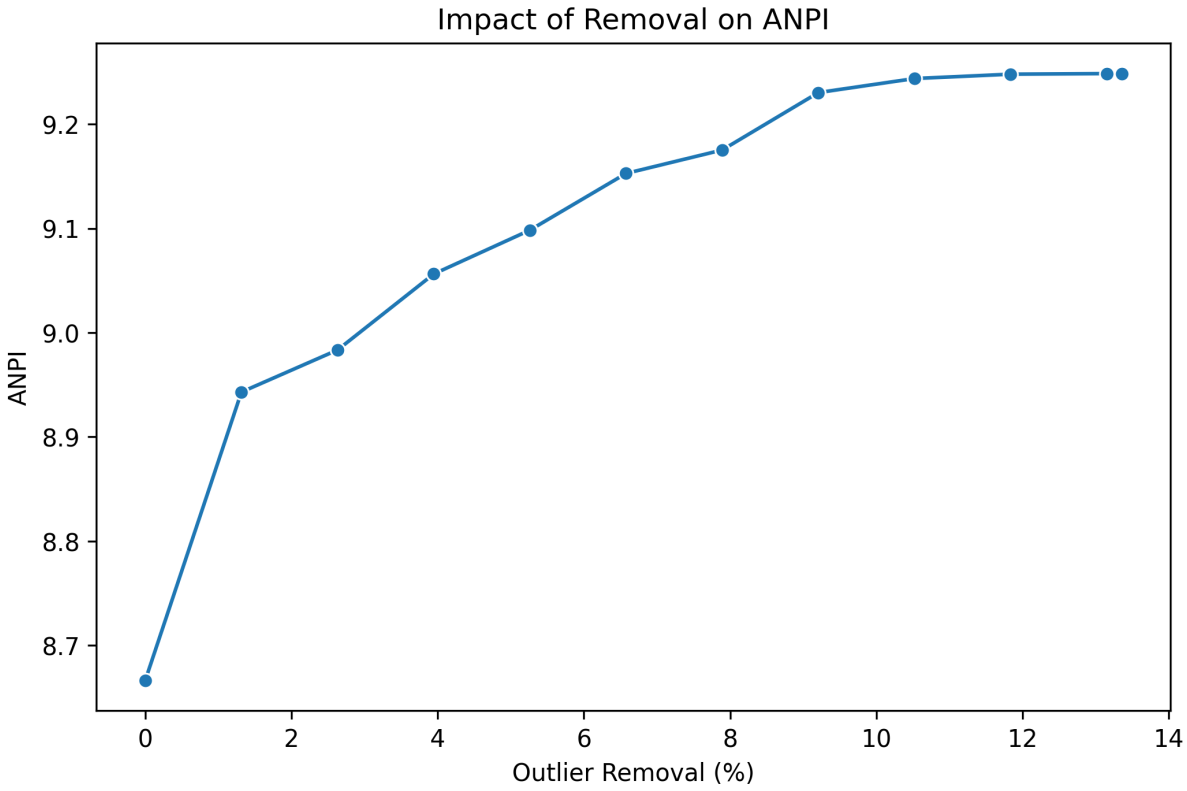

Figure 12: CCBOS with Double-Sized bins outlier removal impact on ANPI

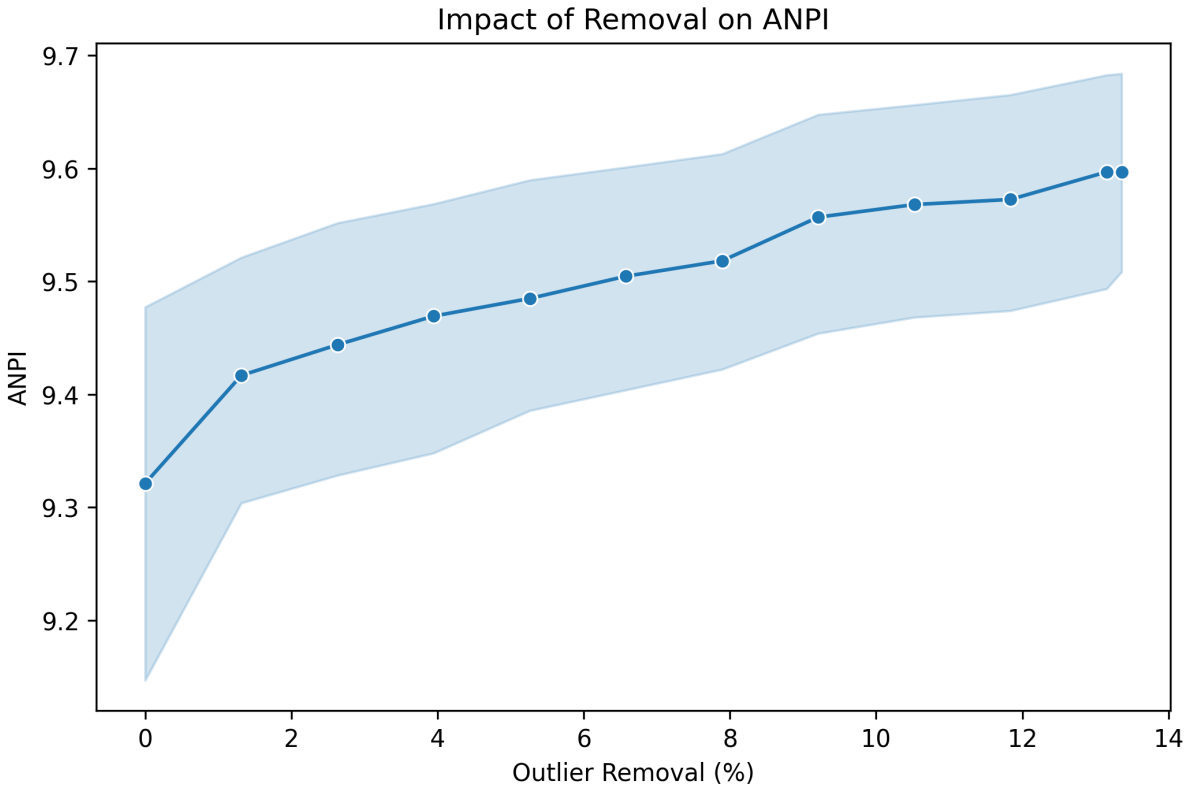

Figure 13: CCBOS with Double-Sized bins outlier removal impact on PoP-specific ANPI

8.1.5 CCBOS Half-Sized Bins

When bins have their size reduced in half, the results are more erratic and take longer and need a higher percentage of outlier removals.

This can be attributed to the significant, albeit small, penalization of samples which comply with the respective threshold, but are not present, however, in the first bin of its histogram.

These removal experiments indicate that indeed the CCBOS is capable of locating network irregularities (temporally and spatially), which can perhaps even aid in data cleansing for machine learning or statistical analysis.

Our goal, however, is to provide network workers with a tool for network management. After assessing the highest-scoring spatiotemporal contexts, the administrator can then see the specific measurements of the chosen contexts and better understand what may be happening. With the new acquired information, they can then perform more guided interventions in the network, either through hardware or software.

In this sense, the joint utilization of ANPI and CCBOS not only facilitates the work

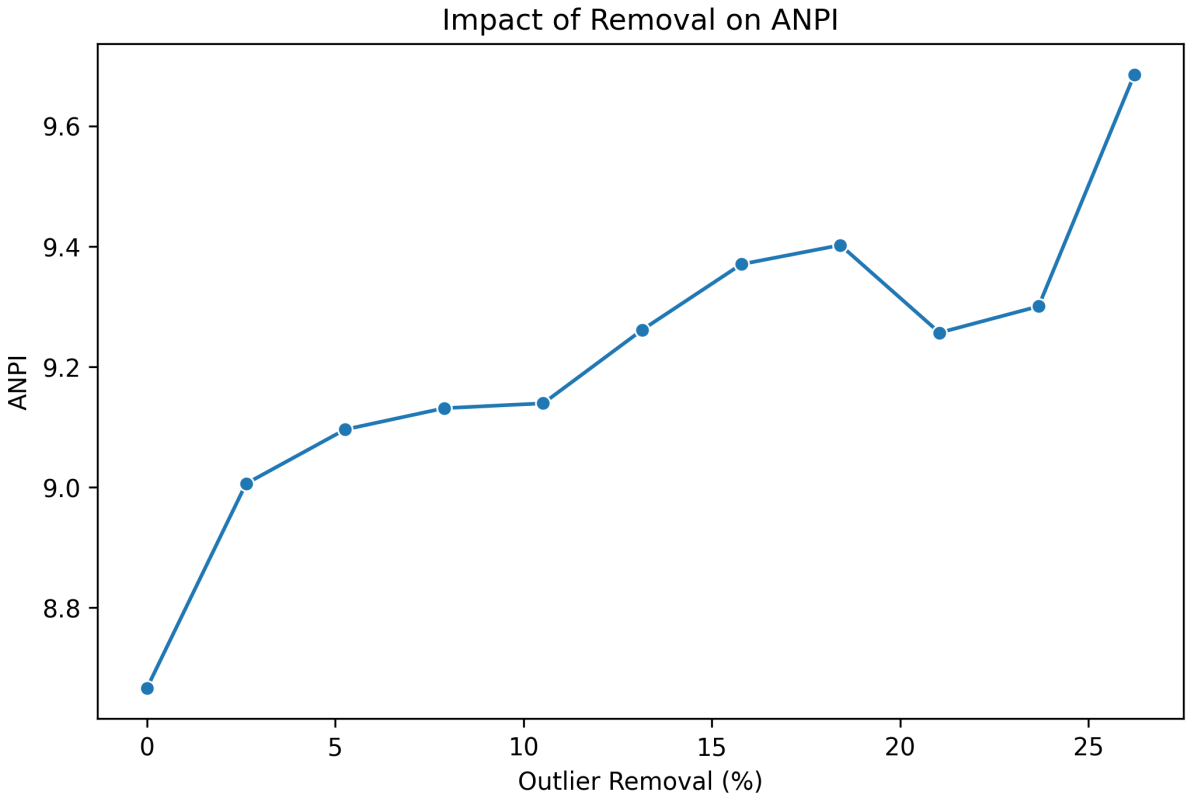

Figure 14: CCBOS with Half-Sized bins outlier removal impact on ANPI

of going through enormous datasets, but also provides a framework for analyzing the set of metrics together as context's state.

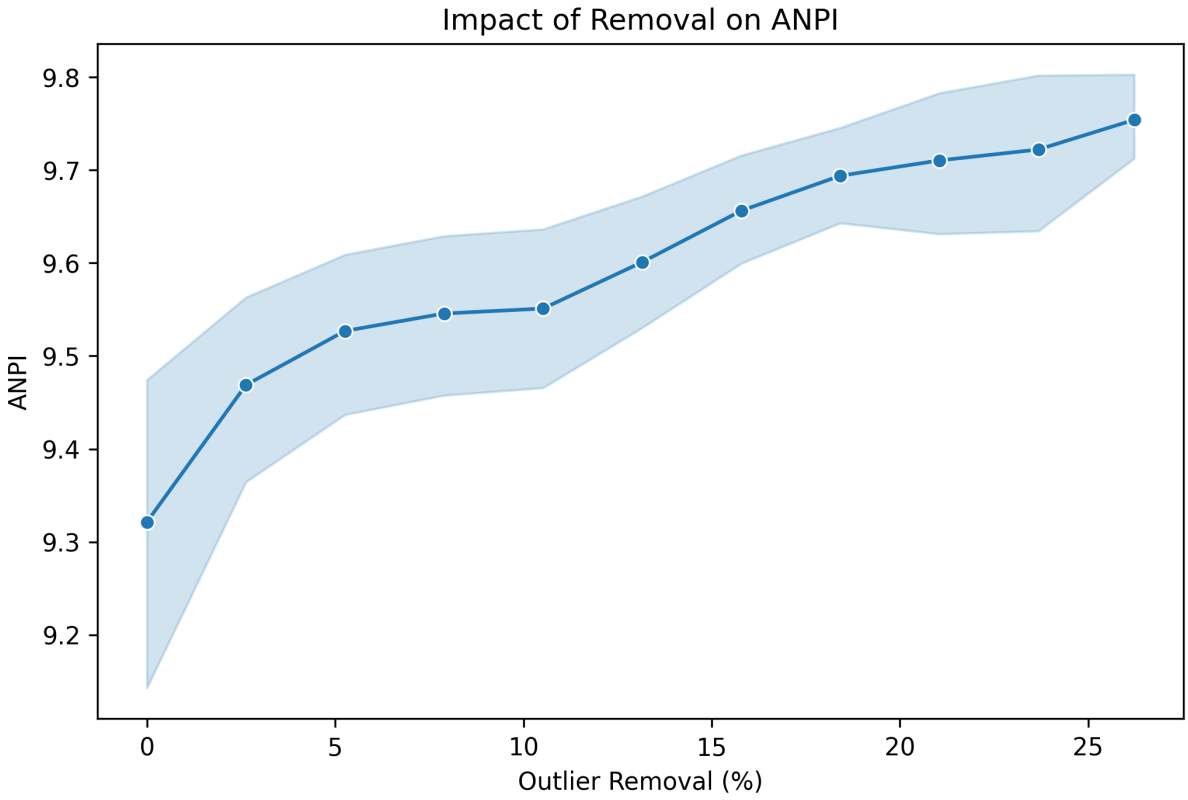

Figure 15: CCBOS with Half-Sized bins outlier removal impact on PoP-specific ANPI

8.2 Outlier Count vs. Network Quality

This section analyzes the relationship between the number of detected outliers and the ANPI. Specifically, we investigate whether a higher number of HBOS outliers necessarily translates into poorer network quality scores.

For this purpose, outlying samples were defined considering the entire semester. The results were then separated by PoP, allowing the number of outliers detected in each group to be compared with the corresponding ANPI. Figure 16 shows the relationship between the outlier count and the respective single-PoP ANPI.

It can be easily seen from the figure that there is a negative linear correlation between ANPI and outlier count. The more outliers there are, the lower the ANPI is. This further validates the relationship between ANPI and CCBOS as a valuable mechanism for information mining through data segregation.

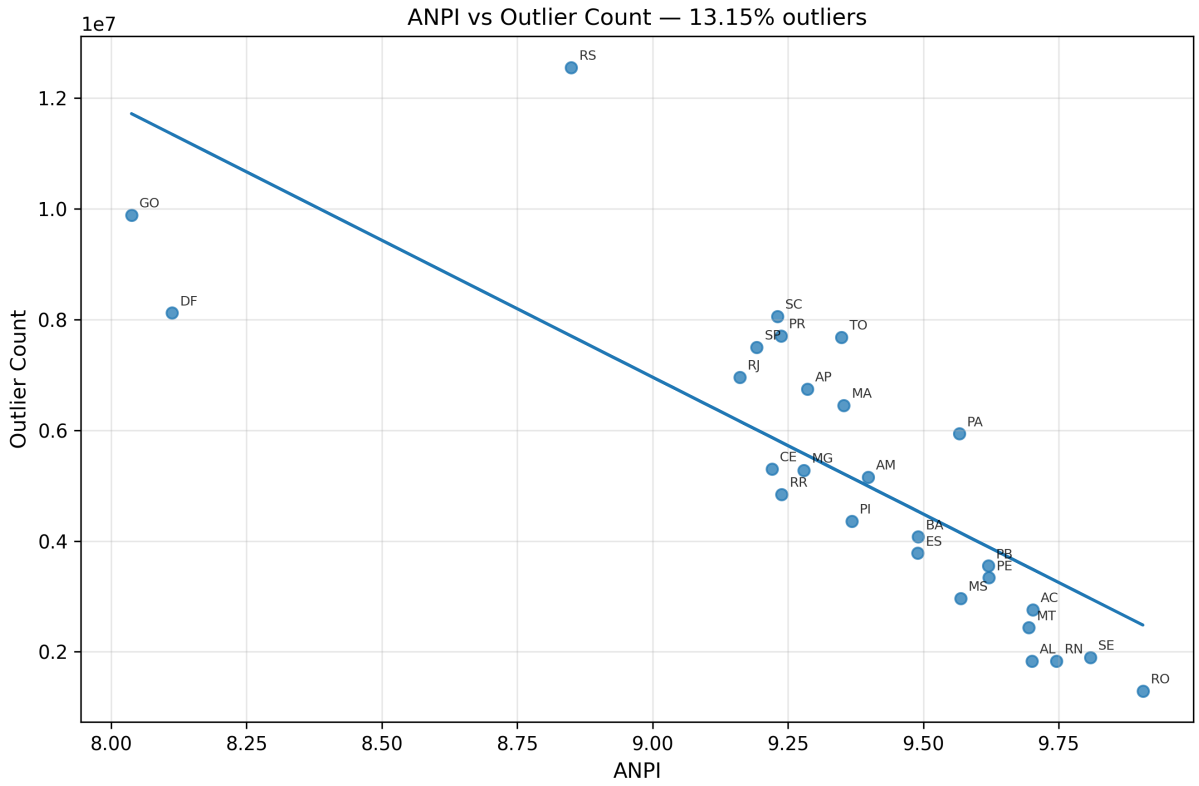

Figure 16: Outlier Count Vs. ANPI

8.3 PoP-Level Impact of Outlier Removal

Finally, we present a further detailed PoP-segregated analysis of the effects of outlier removal on the ANPI.

The results are presented using polar graphs, where an increase in area represents higher scores. This visualization makes it possible to compare, for each PoP, the network quality profile before and after outlier removal, as well as to identify which components are most affected by outlying samples.

8.3.1 CCBOS

Figures 17, 18, 19 and 20 show the PoP-specific influence of gradual CCBOS-outlier removal.

Not only do they highlight almost perfect results with 11.88% removal, (with visible gaps restricted to MG, RJ, SP, PR, RS, SC, and DF), but the path to that result begins by first improving PoPs with the worst results (GO, DF, RS and RR), filling in any concavities present in the polar graph. This shows not only gradual global improvement,

but also some sense of fairness in the process.

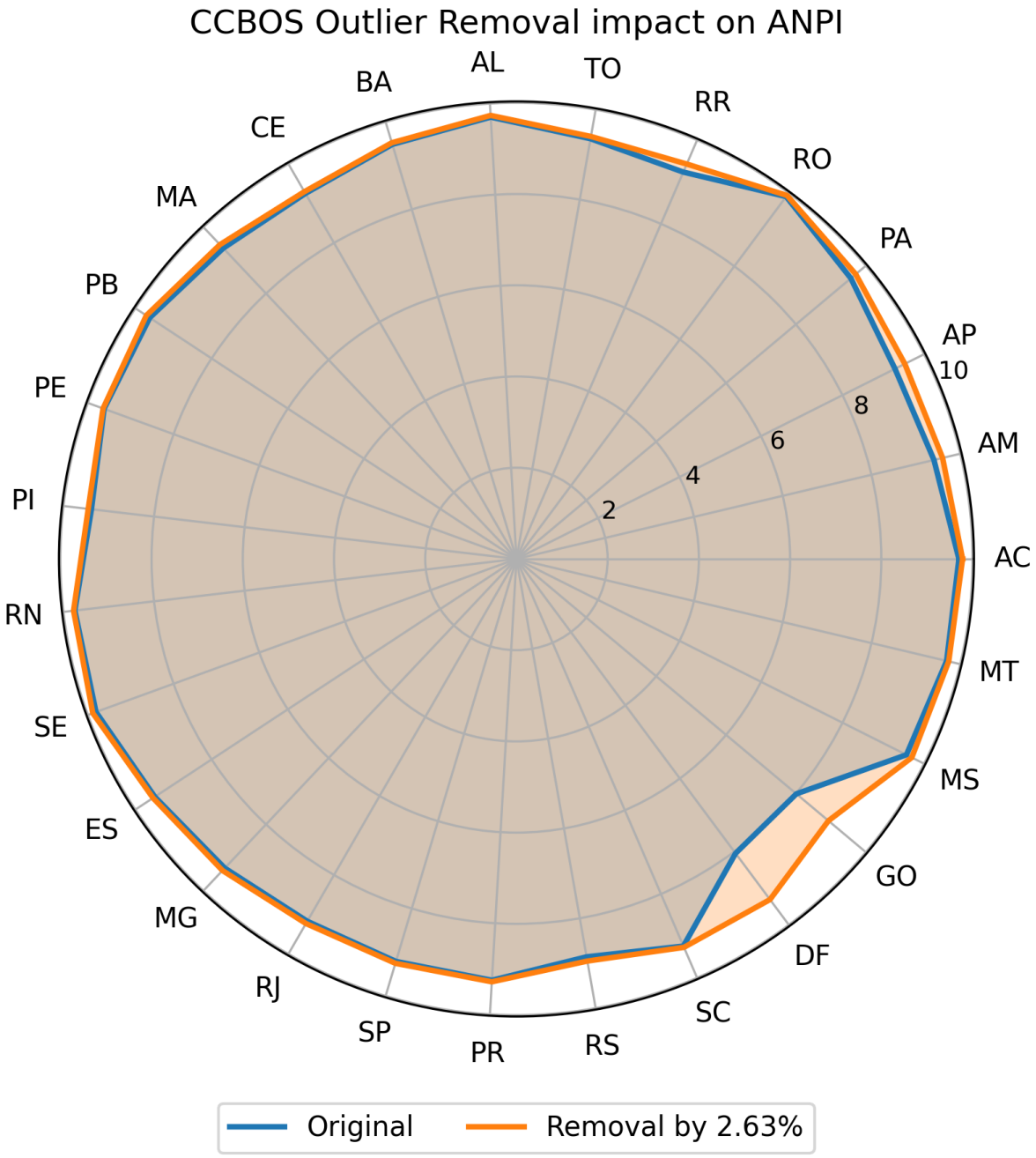

Figure 17: PoP-Level impact of CCBOS Outlier Removal

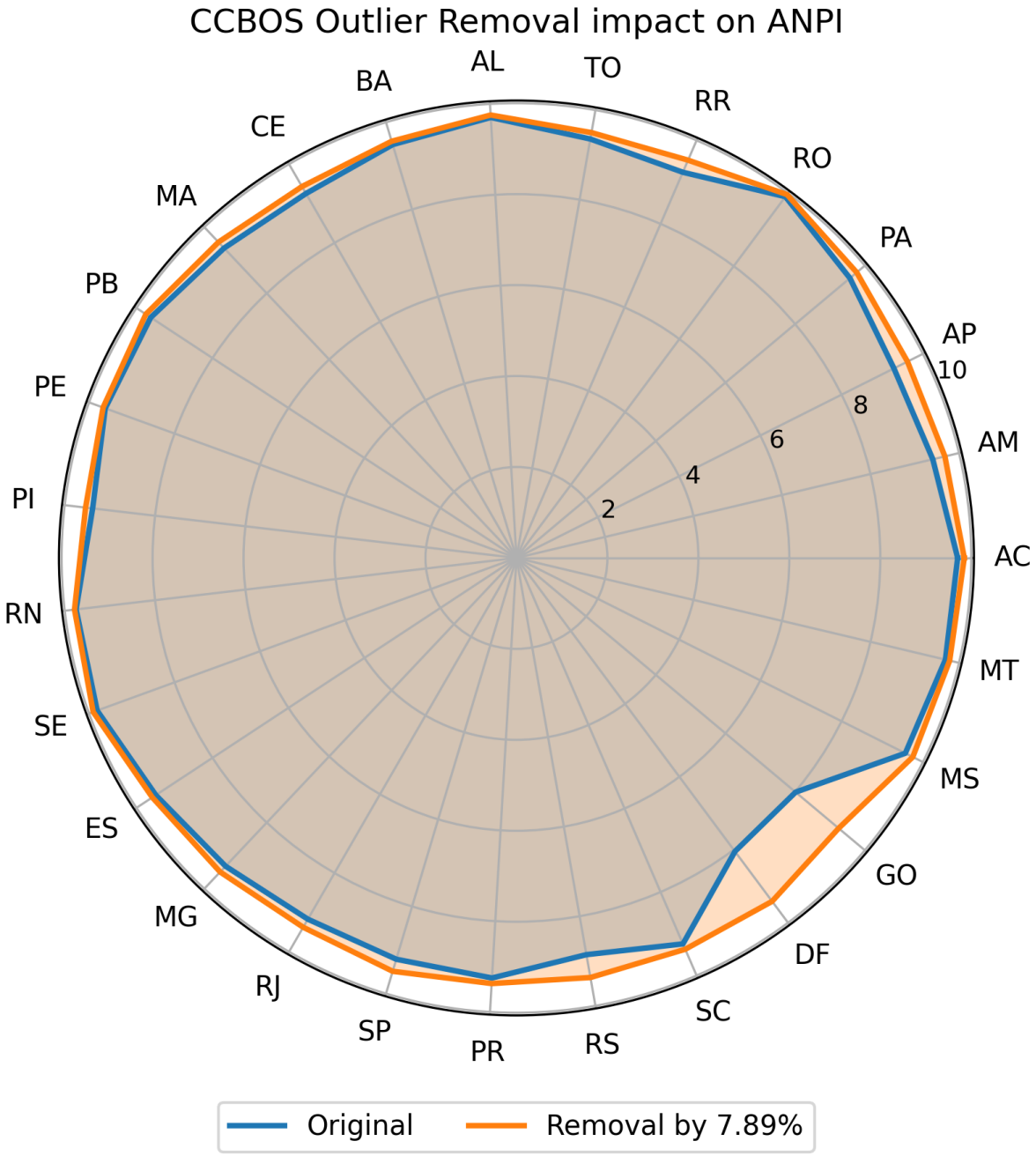

Figure 18: PoP-Level impact of CCBOS Outlier Removal

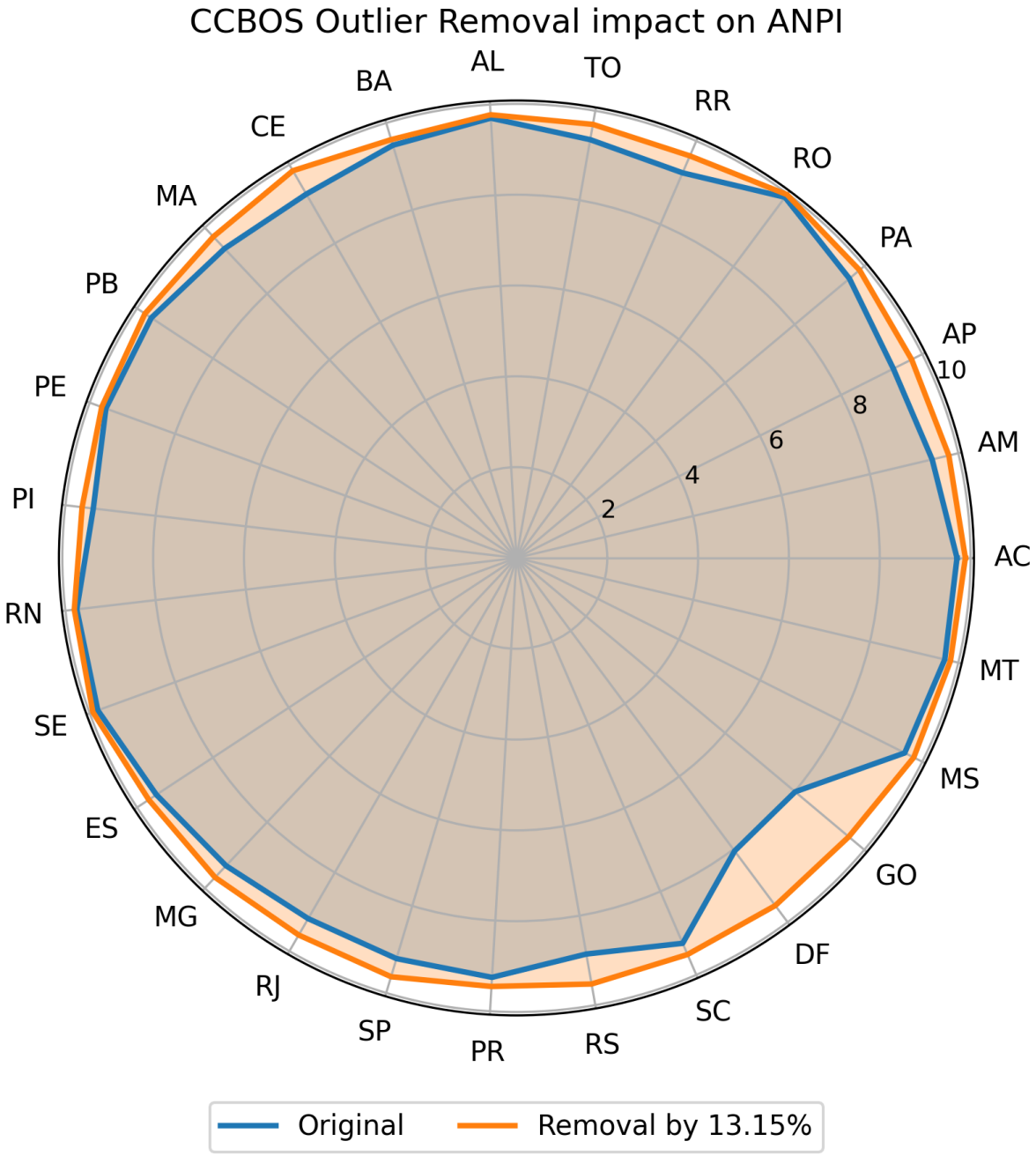

Figure 19: PoP-Level impact of CCBOS Outlier Removal

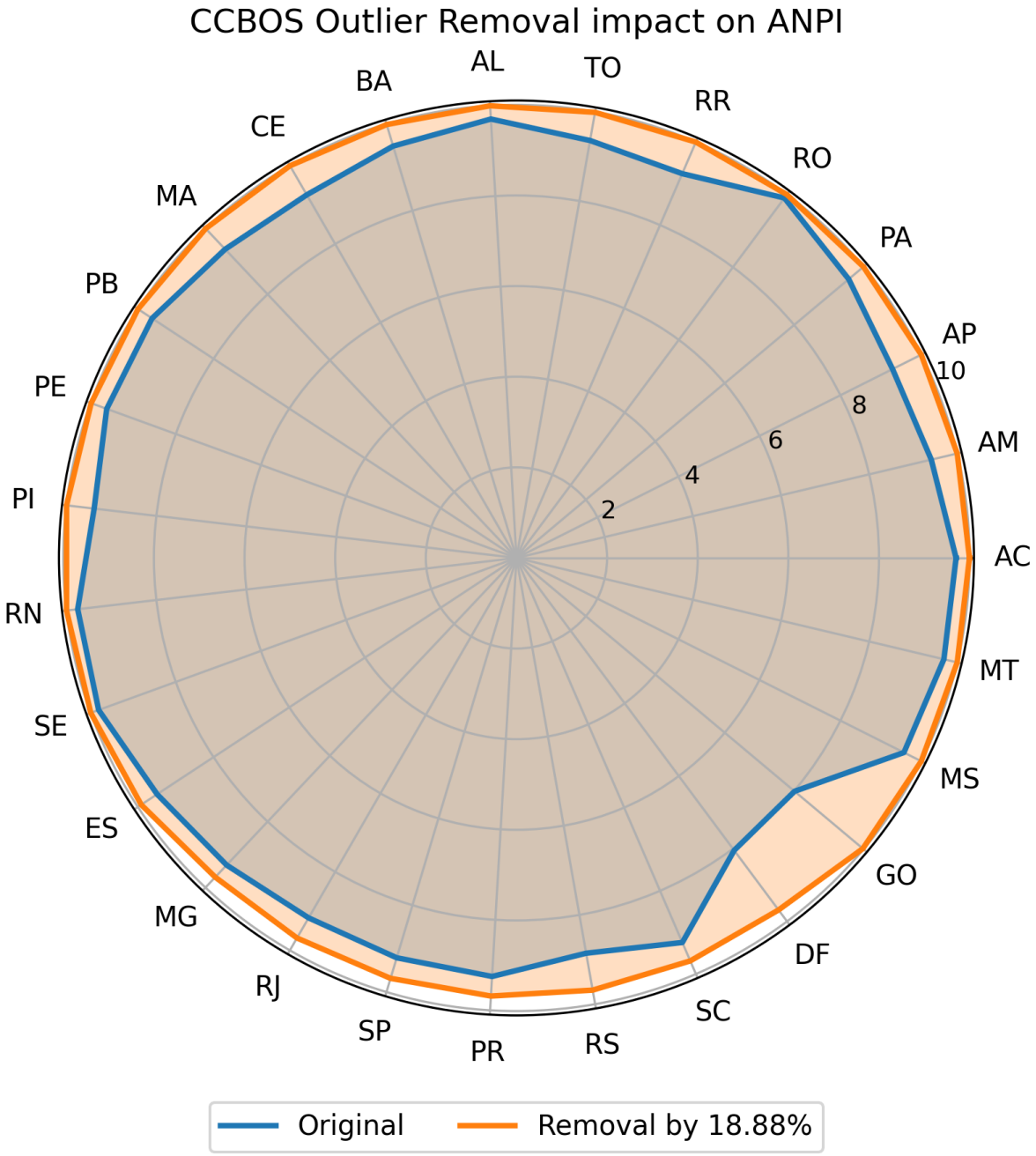

Figure 20: PoP-Level impact of CCBOS Outlier Removal

8.3.2 HBOS

Again, Figure 21 shows that results from HBOS mirror those of CCBOS, for the same reasons presented in subsection 8.1.2.

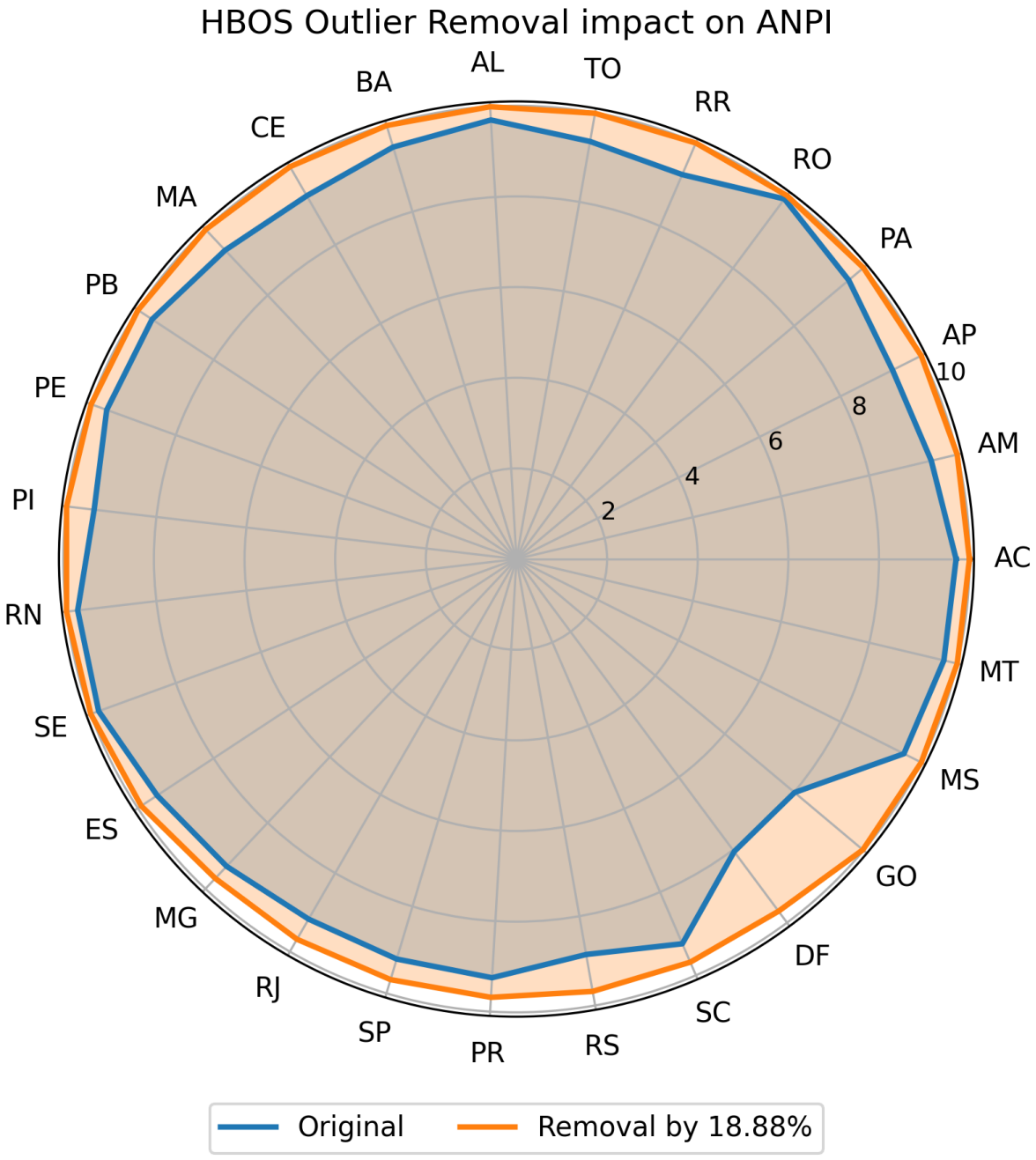

Figure 21: PoP-Level impact of HBOS Outlier Removal

8.3.3 ECOD

For ECOD, considering a comparable percentage, results were not as good, as shown in Figure 22. Even when higher outlier ratios were considered and removed, results did not seem to get better. ANPI's behavior actually became more erratic, as evidenced by Figure 23 with more than 50% removal.

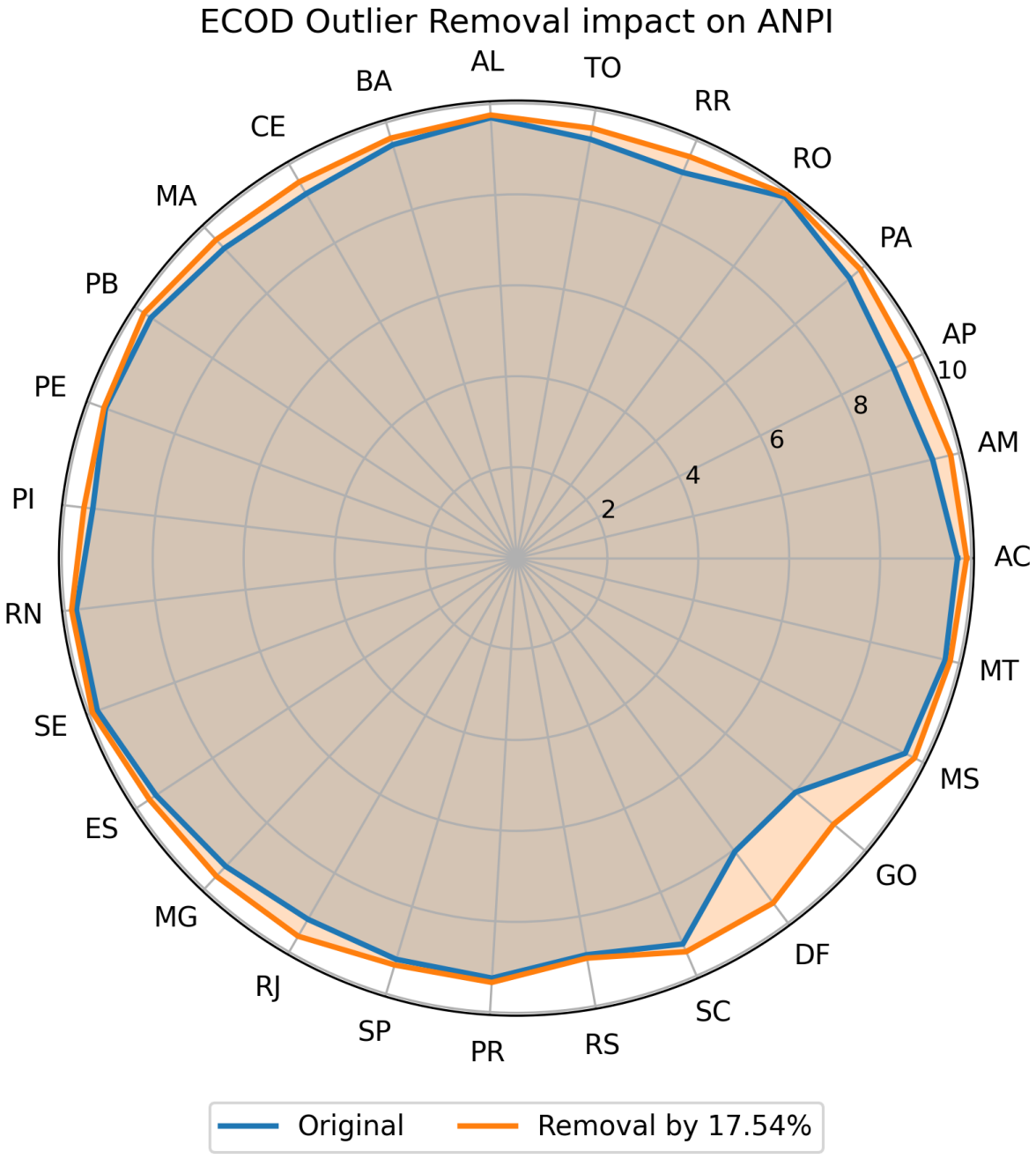

Figure 22: PoP-Level impact of ECOD Outlier Removal

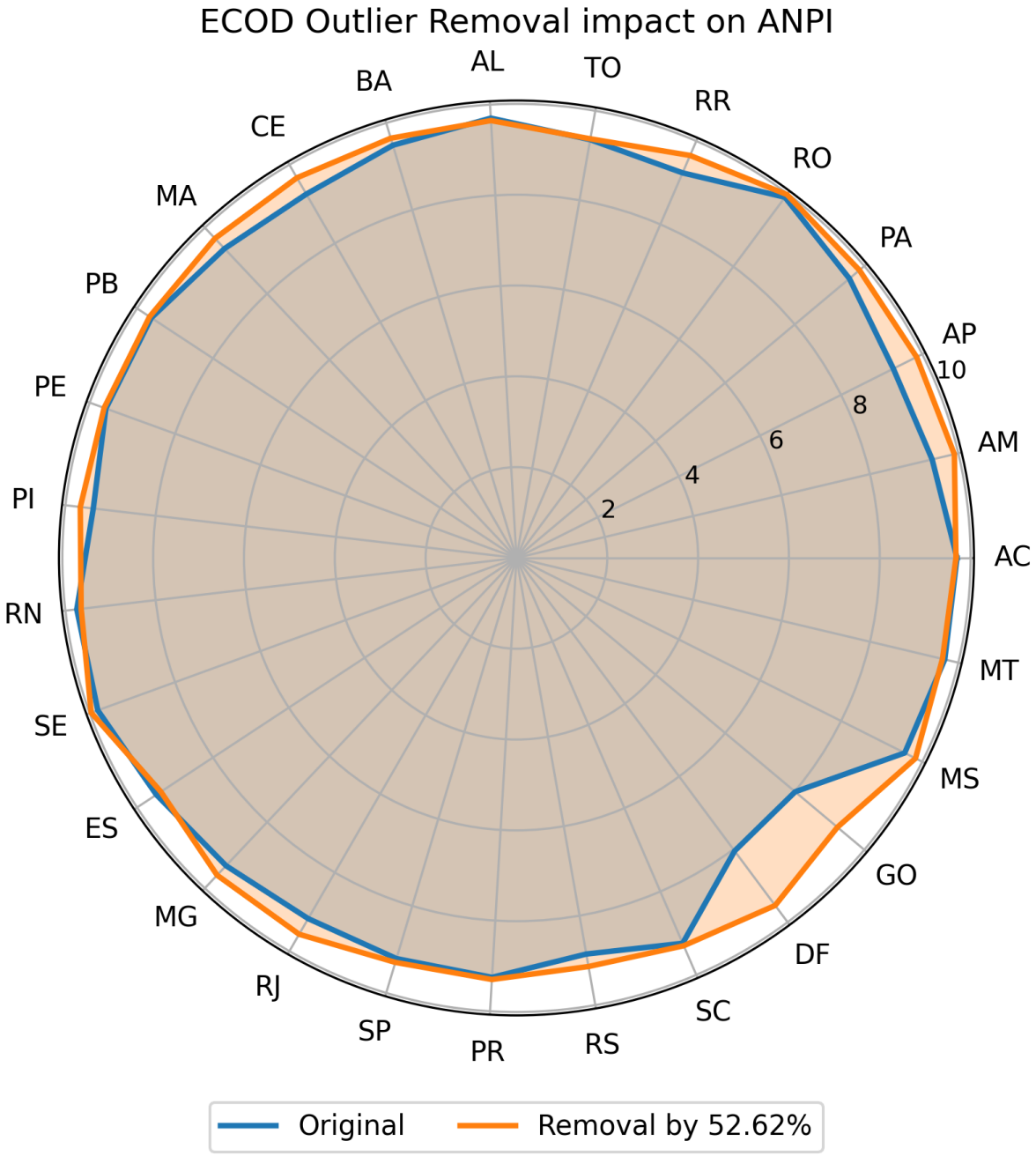

Figure 23: PoP-Level impact of ECOD Outlier Removal

8.3.4 CCBOS with Double-Sized Bins

With Double-Sized Bins, the highest possible percentage of removal was 13.36%, since the rest of the data possessed outlier scores of 0.

Even with this ratio, Figure 24 presents underwhelming results when compared with the original CCBOS, for the reasons discussed in Subsection 8.1.4.

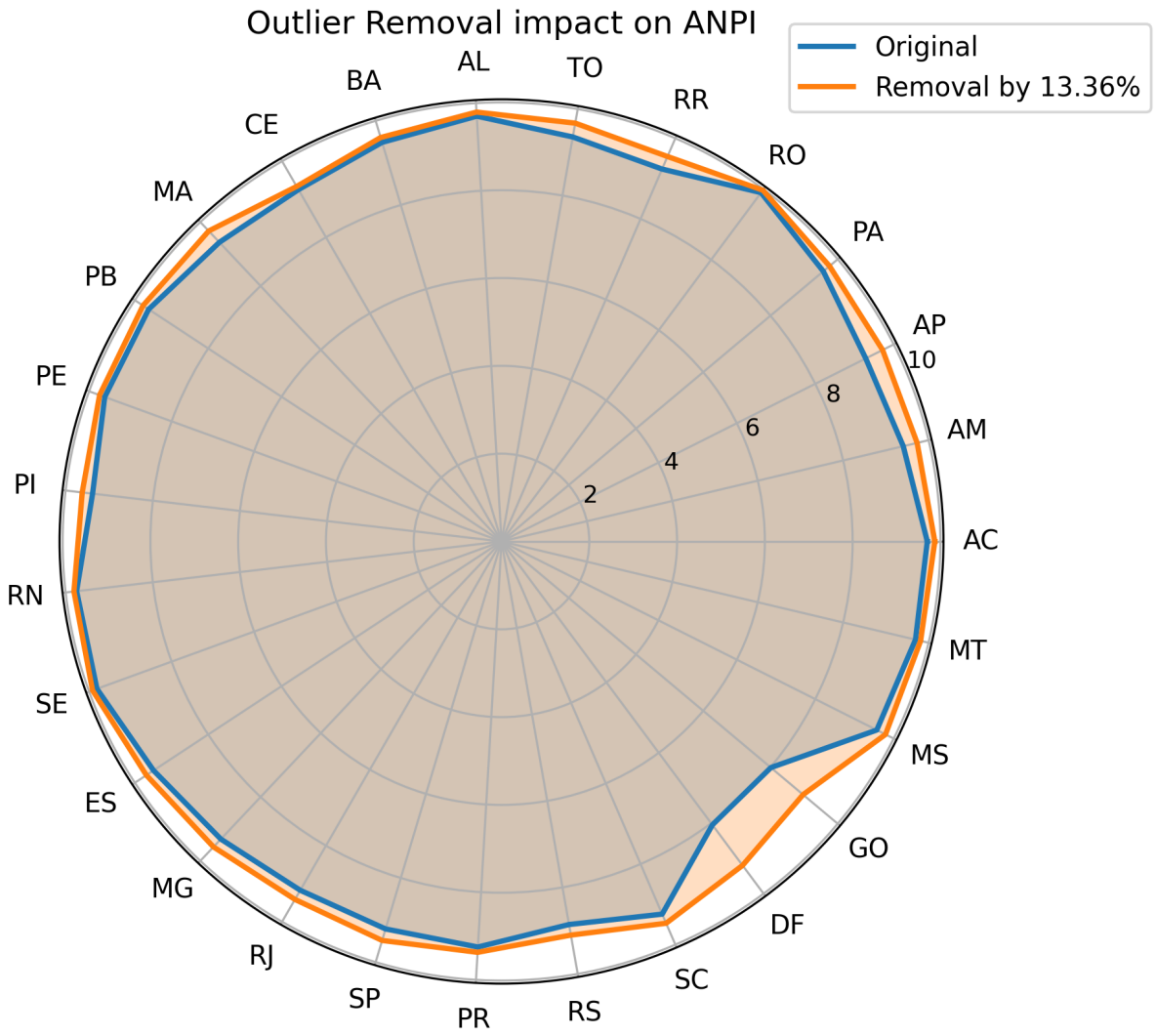

Figure 24: PoP-Level impact of Double-Sized Bins CCBOS Outlier Removal

8.3.5 CCBOS with Half-Sized Bins

While Half-Sized Bins produced more satisfactory results than Double-Sized, as shown in Figure 25, it still did not achieve the performance of the original CCBOS, even with a higher removal rate, 26.22% compared with the original's 18.88%. The reasoning behind this behavior is the same as that detailed in Section 8.1.5.

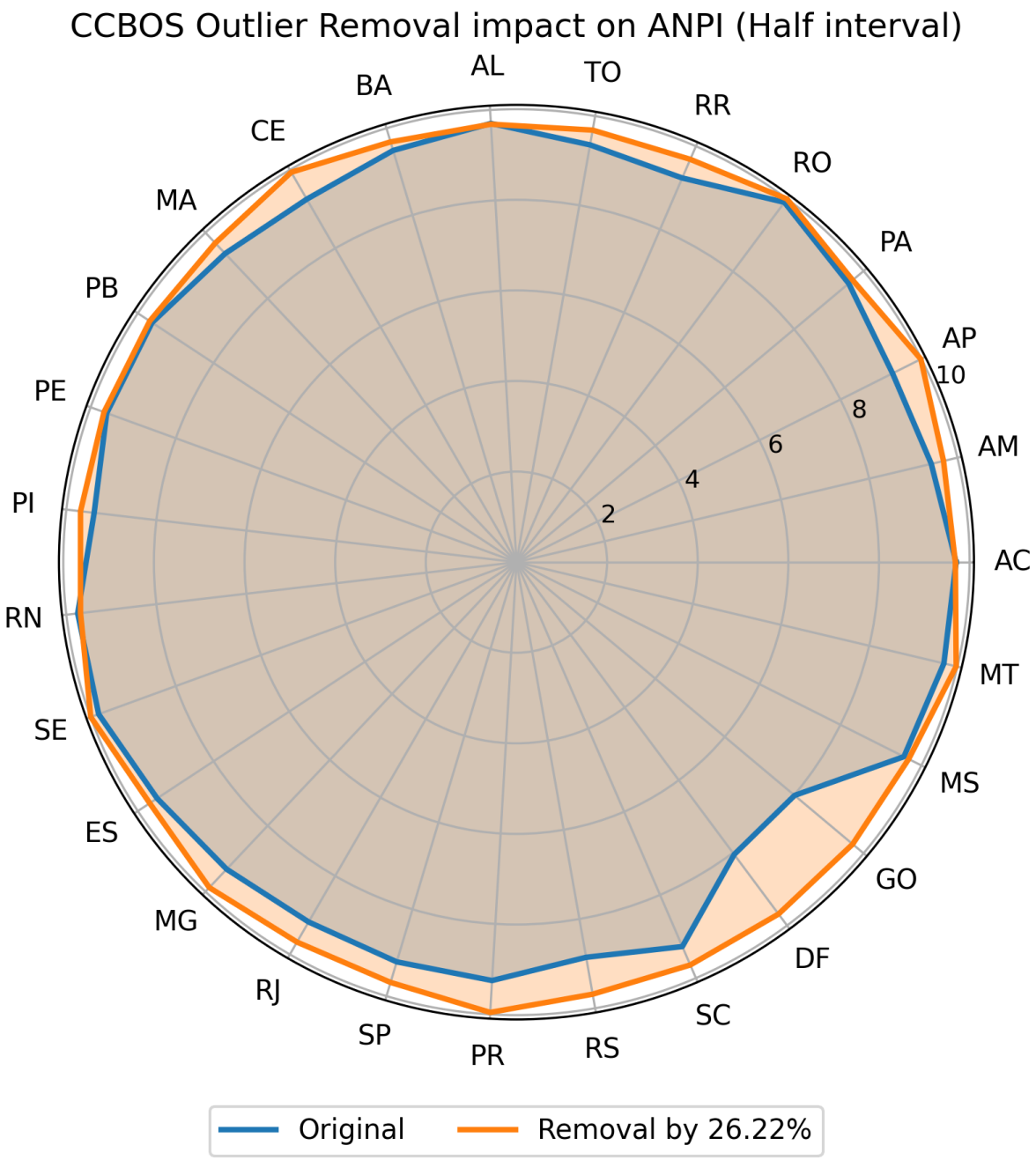

Figure 25: PoP-Level impact of Half-Sized Bins CCBOS Outlier Removal

8.4 Takeaways

The empirical outcomes presented throughout this chapter provide relevant insights both into the proposed methods and into the RNP monitoring dataset. From a general perspective, four main insights can be synthesized:

- CCBOS's use of bin-based discretization improves stabilization and PoP homogeneity in the context ordering generated by the scores. This is manifested in the comparisons between ECOD and CCBOS's results.
- The double and half bin-size variations of CCBOS indicate the importance and efficacy of deriving bin sizes from ANPI's thresholds. Both larger and smaller bins were shown to produce undesirable effects.
- When data is well behaved, the coupling of global re-contextualization of density with the "delta_min factor" pre-processing technique is sufficient for CCBOS's monotonic property to emerge in HBOS. CCBOS's efficacy, however does not rely on strong dataset assumptions, since it derives monotonicity by design.
- Through cumulative-based data aggregation and segregation, respectively, ANPI and CCBOS can be integrated to address the complementary tasks of network performance evaluation and the identification of spatiotemporal measurement contexts associated with performance degradation. The results suggest that this integration may support network managers in both performance analysis and the prioritization of contexts for further investigation, which may, in turn, inform targeted intervention. Because both methods rely on limited assumptions about the underlying data distribution, they may also be applicable to other network environments and, potentially, to domains beyond computer networks.

9 Conclusion

The present work had the dual goal of contributing to the complementary tasks of computer network performance evaluation and the identification of spatiotemporal measurement contexts associated with performance degradation.

The work first surveyed different network evaluation methods, focusing on the diversity of strategies applied to the task. The field of anomaly/outlier detection was also examined, with an initial focus on detection in computer networks, differentiating outlier and anomaly detection by malicious intent and general applied techniques. In addition, outlier detection was explored in relation to the use of cumulative techniques, mostly based on CDFs, with the contextual assumption that outliers lie in the tails of distributions.

The dataset employed in this study was then described and detailed. This description was followed by the introduction of background concepts from the realm of statistics and network monitoring, further providing the backdrop for the proposed artifacts.

This was followed by the introduction of a baseline performance index, both to motivate the proposal introduced in this work, and to provide a basis for comparison with the proposed index. ANPI (Aggregated Network Performance Index) was then proposed and detailed as a novel performance index based on the aggregation of metrics across different contexts and monitoring dimensions. The index aggregation process was based on both CDFs and the Harmonic Mean, taking advantage of data preprocessing steps for the removal of contextual bias.

An exploration of different ANPI parameter configurations was performed, in order to better understand its behavior and motivate certain design and parameter choices.

After motivating, introducing, and experimenting with ANPI, we turned our focus to the question of outlier detection. After evaluating the network performance, how can we aid network managers in improving network performance? To this end, outlier scoring methods were used as inspiration for the definition of irregular spatial and time frames, which among the extensive monitoring dataset, required closer analysis. Based

on methods such as HBOS and ECOD, the CCBOS was derived, which considered both density and magnitude information for deriving outlier scores in a context where outliers are known to be in the right tails of distributions. This method also leveraged HBOS discretization through bins, motivated by ANPI's thresholding nature.

In this sense, CCBOS and ANPI exhibited complementary behavior in the empirical evaluation. Experiments based on global and PoP-specific outlier removal, together with the analysis of outlier counts, provided evidence that the spatiotemporal contexts assigned higher CCBOS scores were associated with performance degradation as measured by ANPI. This investigation not only evaluated each proposed artifact, but also their joint work in helping the assessment and maintenance of network performance.

Although the results were positive, further investigation is needed to optimize the importance ordering provided by CCBOS. Additional datasets should also be investigated for assessing the expected generality of both the CCBOS and ANPI. Finally, future work should be performed with labeled datasets, to investigate CCBOS's possible capability for anomaly detection, that is, of deliberate attacks to the network.

References

ALI, Saqib *et al.* Detecting anomalies from end-to-end internet performance measurements (PingER) using cluster based local outlier factor. *In:* IEEE. 2017 IEEE international symposium on parallel and distributed processing with applications and 2017 IEEE international conference on ubiquitous computing and communications (ISPA/IUCC). [*S. l.: s. n.*], 2017. p. 982–989.

ANGIULLI, Fabrizio. CFOF: a concentration free measure for anomaly detection. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM New York, NY, USA, v. 14, n. 1, p. 1–53, 2020.

ARIFUZZAMAN, Md; ISLAM, Shafkat; ARSLAN, Engin. Towards generalizable network anomaly detection models. *In:* IEEE. 2021 IEEE 46th conference on local computer networks (LCN). [*S. l.: s. n.*], 2021. p. 375–378.

BALDONI, Sara; BATTISTI, Federica. Histogram-based network traffic representation for anomaly detection through PCA. **Computer Networks**, Elsevier, v. 265, p. 111276, 2025.

BICSKI, Balint; PEKAR, Adrian. Early detection of network service degradation: An intra-flow approach. *In:* IEEE. 2024 20th International Conference on Network and Service Management (CNSM). [*S. l.: s. n.*], 2024. p. 1–5.

BOUMAN, Roel; BUKHSH, Zaharah; HESKES, Tom. Unsupervised anomaly detection algorithms on real-world data: how many do we need? **Journal of Machine Learning Research**, v. 25, n. 105, p. 1–34, 2024.

DURAJ, Agnieszka; SZCZEPANIAK, Piotr S. Outlier detection in data streams—A comparative study of selected methods. **Procedia Computer Science**, Elsevier, v. 192, p. 2769–2778, 2021.

GOLDSTEIN, Markus; DENGEL, Andreas. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. **KI-2012: poster and demo track**, Saarbruecken, Germany, v. 1, p. 59–63, 2012.

GOMES, Ian José Agra *et al.* Diagnóstico de QoS em Redes a partir de Análise conjunta de Sobrevivência e Mudanças Estatísticas em Séries Temporais de

Desempenho. *In:* SBC. WORKSHOP em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance). [*S. l.: s. n.*], 2025. p. 49–60.

HAN, Songqiao *et al.* Adbench: Anomaly detection benchmark. **Advances in neural information processing systems**, v. 35, p. 32142–32159, 2022.

JÄNTSCHI, Lorentz. A test detecting the outliers for continuous distributions based on the cumulative distribution function of the data being tested. **Symmetry**, MDPI, v. 11, n. 6, p. 835, 2019.

KAYA, Muhammed Onur; OZDEM, Mehmet; DAS, Resul. A new hybrid approach combining GCN and LSTM for real-time anomaly detection from dynamic computer network data. **Computer Networks**, Elsevier, v. 268, p. 111372, 2025.

KHANAL, Bipal; KUMAR, Chandan; ANSARI, Md Sarfaraj Alam. Real-time anomaly detection framework to mitigate emerging threats in software defined networks. **Journal of Network and Systems Management**, Springer, v. 33, n. 2, p. 26, 2025.

KIRAN, Mariam *et al.* Detecting anomalous packets in network transfers: investigations using PCA, autoencoder and isolation forest in TCP. **Machine Learning**, Springer, v. 109, n. 5, p. 1127–1143, 2020.

KOMADINA, Adrian *et al.* Comparing threshold selection methods for network anomaly detection. **IEEE access**, IEEE, 2024.

KOUGIOUMTZIDIS, Georgios *et al.* A survey on multimedia services QoE assessment and machine learning-based prediction. **Ieee Access**, IEEE, v. 10, p. 19507–19538, 2022.

LATIF-MARTÍNEZ, Hamid *et al.* GAT-AD: Graph Attention Networks for contextual anomaly detection in network monitoring. **Computers & Industrial Engineering**, Elsevier, v. 200, p. 110830, 2025.

LI, Angela; LI, David. Spline-Interpolated Density and Dependency-Aware Outlier Scoring for Lightweight Deployment. *In:* IEEE. 2025 IEEE International Conference on Data Mining (ICDM). [*S. l.: s. n.*], 2025. p. 427–436.

LI, Angela; WANG, Yu; LI, David. IDOS: Lightweight Nonparametric Outlier Detection Algorithm for Resource-Constrained Applications. *In:* IEEE. 2025 59th Annual Conference on Information Sciences and Systems (CISS). [*S. l.: s. n.*], 2025. p. 1–6.

LI, Zheng; ZHAO, Yue, *et al.* Copod: copula-based outlier detection. *In:* IEEE. 2020 IEEE international conference on data mining (ICDM). [*S. l.: s. n.*], 2020. p. 1118–1123.

LI, Zhong; ZHU, Yuxuan; VAN LEEUWEN, Matthijs. A survey on explainable anomaly detection. **ACM Transactions on Knowledge Discovery from Data**, ACM New York, NY, v. 18, n. 1, p. 1–54, 2023.

LIU, Yating *et al.* MSCA: An unsupervised anomaly detection system for network security in backbone network. **IEEE Transactions on Network Science and Engineering**, IEEE, v. 10, n. 1, p. 223–238, 2022.

MARNERIDES, Angelos K; SCHAEFFER-FILHO, Alberto; MAUTHE, Andreas. Traffic anomaly diagnosis in internet backbone networks: A survey. **Computer Networks**, Elsevier, v. 73, p. 224–243, 2014.

MICHALAK, Marcin *et al.* Outlier Detection in Network Traffic Monitoring. *In:* ICPRAM. [*S. l.: s. n.*], 2021. p. 523–530.

MING, Tong; MAN, Zhu A. Research on Network Quality Evaluation System Based on the Entropy Weight Method. *In:* PROCEEDINGS of the 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology. [*S. l.: s. n.*], 2024. p. 51–59.

OHLSEN, Lai Yi; SERMPEZIS, Pavlos; NEWCOMB, Melissa. Poster: The Internet Quality Barometer Framework. *In:* PROCEEDINGS of the 2025 ACM Internet Measurement Conference. [*S. l.: s. n.*], 2025. p. 1064–1065.

PANAHI, Parsa Hassani Shariat; JALILVAND, Amir Hossein; NAJAFI, M Hassan. Bridging Subjective and Objective QoE: Operator-Level Aggregation Using LLM-Based Comment Analysis and Network MOS Comparison. **arXiv preprint arXiv:2506.00924**, 2025.

PASCO, Carlos David R; BERNARDINI, Flavia; ANTÔNIO, A De A. In-network Latency Nowcast Using Data Stream Learning Models. *In:* IEEE. 2023 International Wireless Communications and Mobile Computing (IWCMC). [*S. l.: s. n.*], 2023. p. 1214–1219.

RAO, Nageswara S *et al.* **Detecting outliers in network transfers with feature extraction**. [*S. l.*], 2018.

REDE NACIONAL DE ENSINO E PESQUISA. **QIM – Indicadores 3 e 4: Qualidade do Serviço de Conectividade na Rede Ipê**. [*S. l.*], July 2025.

REZAEI, Hadiseh; TAHERI, Rahim; SHOJAFAR, Mohammad. FedLLMGuard: A federated large language model for anomaly detection in 5G networks. **Computer Networks**, Elsevier, v. 269, p. 111473, 2025.

SAMARIYA, Durgesh; THAKKAR, Amit. A comprehensive survey of anomaly detection algorithms. **Annals of Data Science**, Springer, v. 10, n. 3, p. 829–850, 2023.

SILVEIRA, Fernando *et al.* ASTUTE: Detecting a different class of traffic anomalies. **ACM SIGCOMM Computer Communication Review**, ACM New York, NY, USA, v. 40, n. 4, p. 267–278, 2010.

SUN, Ming wei; ZHANG, Qing wei; ZHAO, Hai yuan. Evaluation and Prediction of Network QoS Based on Multidimensional Data. *In:* PROCEEDINGS of the 2022 International Conference on Cyber Security. [*S. l.: s. n.*], 2022. p. 46–51.

THOTTAN, Marina; JI, Chuanyi. Anomaly detection in IP networks. **IEEE Transactions on signal processing**, IEEE, v. 51, n. 8, p. 2191–2204, 2003.

WAKUI, Taku; KONDO, Takao; TERAOKA, Fumio. GAMPAL: an anomaly detection mechanism for Internet backbone traffic by flow size prediction with LSTM-RNN. **Annals of telecommunications**, Springer, v. 77, n. 5, p. 437–454, 2022.

YANG, Mingzhe *et al.* Survey on QoE assessment approach for network service. **Ieee Access**, IEEE, v. 6, p. 48374–48390, 2018.

YILMAZ, Ali; DAS, Resul. A novel hybrid approach combining GCN and GAT for effective anomaly detection from firewall logs in campus networks. **Computer Networks**, Elsevier, v. 259, p. 111082, 2025.

ZHANG, Kaisa *et al.* Cellular QoE prediction for video service based on causal structure learning. **IEEE Transactions on Intelligent Transportation Systems**, IEEE, v. 24, n. 7, p. 7541–7551, 2022.

ZHU, Ye *et al.* CDF Transform-and-Shift: An effective way to deal with datasets of inhomogeneous cluster densities. **Pattern Recognition**, Elsevier, v. 117, p. 107977, 2021.