TÉCNICO LISBOA

Deteção de Ataques de Segurança utilizando Análise de Séries Temporais

Inês Sofia Moreira Alves

Dissertação para obtenção do Grau de Mestre em

Engenharia de Telecomunicações e Informática

Orientadores: Prof. Rui Jorge Morais Tomaz Valadas
Prof. Maria do Rosário De Oliveira Silva

Júri

Presidente: Prof. Ricardo Jorge Fernandes Chaves
Orientador: Prof. Rui Jorge Morais Tomaz Valadas
Vogal: Prof. António Manuel Duarte Nogueira

Dezembro 2021

Abstract

There are an increasing number of connected devices due to the evolution of the Internet of Things. With this evolution, the Internet is now more exposed to security attacks. One of the ways to detect an attack is by analyzing the traffic, trying to distinguish regular traffic from the outliers caused by the attacks. This Msc Dissertation studies methods for detecting security attacks based on time series. This Dissertation surveys the state of the art of anomaly detection methods, presents a dataset to which the methods will be applied and analyse some algoritms to understand which is the best one for the dataset under study. After the dataset is introduced, an analysis of the performance of the algorithms is carried out by varying their parameters. Among the studied methods are an heuristic, Tukey's method, Distance Based-Outlier, Symbolic Aggregate aproXimation, and Tukey's method combined with a Piecewise Aggregate Approximation. Our results indicate that the latter method outperforms the remaining one for the detection of redirection attacks caused by BGP prefix hijacking.

This MSc dissertation was supported by Instituto de Telecomunicações.

Keywords

Outliers, Anomaly Detection, Time Series, Internet Traffic, Distance Based-Outlier, Heuristic, Tukey Method, SAX, PAA.

Resumo

Existe um número cada vez maior de dispositivos ligados em rede devido à evolução da Internet das Coisas. Com esta evolução, a Internet está agora mais exposta a ataques de segurança. Uma das formas de detetar um ataque é através da análise do tráfego, procurando-se distinguir o tráfego regular do tráfego anómalo provocado pelos ataques. Esta Dissertação de Mestrado propõe estudar métodos de deteção de ataques de segurança baseados em séries temporais. Esta Dissertação, faz um levantamento do estado da arte dos métodos de deteção de anomalias, apresenta um primeiro conjunto de dados aos quais os métodos serão aplicados e analisa alguns algoritmos para perceber qual é o mais promissor para o conjunto de dados em estudo. Posteriormente faz-se uma análise ao desempenho dos algoritmos variando os parâmetros dos mesmo. Entre os métodos estudados estão uma heurística, método de *Tukey*, *Distance Based-Outlier*, *Symbolic Aggregate aproXimation* e o método de *Tukey* combinado com o *Piecewise Aggregate Approximation*. Os resultados indicam que o último método supera os restantes para a detecção de ataques de redirecionamento de dados causados por *BGP hijacking*.

Esta Dissertação teve a ajuda do Instituto de Telecomunicações.

Palavras Chave

Anomalias, Deteção de Anomalias, Séries temporais, Tráfego, Distance Based-Outlier, Heurística, Método de Tukey, SAX, PAA.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Acrónimos

IoT Internet das Coisas

LOCI Integral de Correlação Local

MDEF Fator de Desvio de Multigranularidade

ARMA Autorregressivo de Médias Móveis

AR Autorregressivo

MA Média Móvel

IQR Amplitude Interquartil

SAX Aproximação Agregada Simbólica

PAA Aproximação Agregada por Partes

BGP Protocolo para Routers de Fronteira

MITM Man-In-The-Middle

1

Introdução

1.1 Motivação

Os ataques de segurança são cada vez mais comuns e as redes de computadores acabam por ser bastante vulneráveis à medida que a população vai tendo acesso às mesmas. A preocupação pela segurança da Internet é cada vez maior, devido ao crescente número de dispositivos ligados à mesma por consequência da evolução da *IoT* (*Internet of Things*).

Na área das telecomunicações, os operadores de rede cada vez mais sentem a necessidade de utilizar métodos estatísticos capazes de detetar anomalias, como forma de detetar ataques de segurança. No entanto, esta tarefa é bastante árdua, visto que os meios são limitados.

Um tipo de ataque que afeta a Internet é o redireccionamento de tráfego, explorando vulnerabilidades do protocolo *BGP* (*Border Gateway Protocol*). O *BGP* permite que um atacante injete na rede um prefixo de rede legítimo como se fosse seu. Estas informações, uma vez aceites por outras redes, são inseridas nas tabelas de encaminhamento *BGP* e fluem pela Internet, podendo ter efeitos catastróficos.

Uma forma de perceber se existe um ataque de redireccionamento é medir o tempo de ida-e-volta

entre a origem e o destino do tráfego. Quando existe um ataque este tempo será superior ao que se verifica numa situação normal. No entanto, esta diferença pode não ser facilmente detetável, sobretudo quando o atacante está próximo do emissor ou do recetor.

Tendo em consideração o que foi mencionado anteriormente, ao se analisarem as séries temporais do tráfego e usando um método de *Machine Learning* apropriado, consegue-se perceber o que é uma anomalia. Uma anomalia é um valor que se desvia significativamente das outras observações e pode ser designado como um *outlier*. Estas anomalias muitas vezes não são facilmente detetáveis. Por esse motivo, existe a necessidade de combinar várias séries temporais para se conseguir alcançar um bom desempenho relativamente ao método de *Machine Learning* utilizado.

1.2 Objetivos

Esta dissertação tem como objetivo estudar métodos de deteção de ataques de segurança utilizando técnicas de deteção de anomalias em séries temporais. Pretende-se encontrar um método que consiga alcançar um bom desempenho e que consiga detetar ataques, nomeadamente ao *BGP*, em séries temporais. Para além disso, tenciona-se perceber se os métodos aplicáveis a séries temporais são os mais favoráveis ou se outros métodos também se podem aplicar.

1.3 Estrutura

Na organização deste relatório, inicialmente, encontra-se a parte teórica em que se explicam os algoritmos analisados neste trabalho (Capítulo 2). Depois segue-se um breve capítulo sobre ataques de segurança em que se introduz um pouco o *BGP*, protocolo cujo ataque é estudado neste relatório (Capítulo 3). O capítulo seguinte faz uma análise inicial do conjunto de dados que é utilizado nesta Dissertação (Capítulo 4). Nos capítulos seguintes estudam-se em detalhe todos os parâmetros dos algoritmos, entre eles a heurística proposta por Salvador e Nogueira (Capítulo 5), método de *Tukey* (Capítulo 6), *Distance Based-Outlier* (Capítulo 7), método de *SAX* (Capítulo 8) e por fim uma adaptação que agrega o método de *Tukey* e o *PAA* (Capítulo 9). Por último apresenta-se uma breve conclusão (Capítulo 10).

2

Estado da Arte

Com a deteção de anomalias é possível detetar padrões que variam do seu comportamento regular. Uma empresa usualmente gere grandes quantidade de dados, como por exemplo no comércio. Por vezes é relevante conseguir analisar os dados de vendas de determinados produtos de forma simples, barata e rápida. Através dos métodos de deteção de anomalias os negócios podem-se tornar mais lucrativos, pela análise dos dados atuais comparativamente com os dados regulares passados.

Na área da segurança da Internet a deteção de anomalias ajuda a detetar possíveis ataques. Por vezes os atacantes pretendem desviar o tráfego e enganar o utilizador final para aceder a dados pessoais do mesmo. Através da deteção de anomalias é possível detetar estes ataques.

Neste capítulo, inicialmente existe uma definição do que é uma anomalia, bem como os tipos de aprendizagem que um algoritmo pode seguir, como a aprendizagem supervisionada, a aprendizagem semi-supervisionada e a aprendizagem não-supervisionada. Define-se também como se categoriza um algoritmo, isto é, que métricas são utilizadas para se decidir o quão eficaz é o algoritmo a classificar os dados.

Existem diversos métodos para deteção de anomalias que são mencionados neste relatório. Na secção 2.5, apresentam-se os algoritmos baseados na distância como o *Distance Based-Outlier*, o

LOCI e a abordagem do vizinho mais próximo. A secção 2.6 apresenta uma explicação sobre séries temporais. Por fim, na secção 2.7 mencionam-se algoritmos para a deteção de anomalias em séries temporais, como a heurística proposta por Salvador e Nogueira, o método de *Tukey* e o método de *SAX*.

2.1 Anomalias

As anomalias acontecem quando um fenómeno aleatório se afasta do seu comportamento regular. Quando se analisa um problema real um dos objetivos principais centra-se em descobrir estas discrepâncias, ou seja, descobrir observações que se possam desviar do padrão regular da generalidade das observações [2]. Analisando eventos ao longo de um ano, por exemplo, o número de festivais, percebe-se que o número de festivais musicais em Portugal em cada mês é quase sempre baixo, havendo um número elevado de eventos nos meses de Verão, julho e agosto. Se num ano (como por exemplo 2020) isso não acontecer provavelmente é uma anomalia, pois afasta-se do seu comportamento regular dos dados. Em particular, em 2020 em que a maior parte dos festivais de Verão foi cancelada devido à pandemia Covid-19.

Na análise dos dados, inicialmente, considera-se que terá de existir um padrão dentro da aleatoriedade dos dados. Este padrão será considerado o padrão regular dos dados.

2.2 Tipos de Algoritmos de Aprendizagem

Para se definir os tipos de algoritmos existentes, é necessário compreender primeiro o conceito de dados rotulados e dados não rotulados. Quando os dados são rotulados, há uma indicação das classes existentes, e é possível confiar nas mesmas pois provêm de observações verdadeiras, e para cada observação conhece-se a classe a que pertence.

Existem três tipos de algoritmos de aprendizagem possíveis quando se analisam dados: supervisionados, não-supervisionados e semi-supervisionados.

Nos algoritmos de aprendizagem supervisionada, o processo de aprendizagem é feito a partir de um conjunto de dados de treino rotulados. Neste tipo de aprendizagem conhecem-se os valores que caracterizam cada observação (variáveis explicativas) e ainda a verdadeira classe da observação. Ou seja, todos os dados são rotulados, e a partir do mesmo tenciona-se encontrar uma solução que consiga prever corretamente a verdadeira classe de uma nova observação [2].

Quando se trata de dados semi-supervisionados, alguns dados estão rotulados enquanto outros não. Como tal, procura-se agrupar os dados por categoria ou por semelhança entre os mesmos. Usualmente, entre os dados rotulados são consideradas apenas as observações regulares para estimar o

classificador e identificar o mais corretamente possível os dados não rotulados [2].

Nos algoritmos não-supervisionados não se tem conhecimento sobre a verdadeira classe de cada observação, ou seja, os dados não estão rotulados. Neste caso, uma das abordagens possíveis centra-se em agrupar os dados por similaridades entre observações para detetar as possíveis anomalias [2]. Para esta aprendizagem, o mais comum é utilizar métodos de *clustering*. O *clustering* consiste na criação de pequenos subconjuntos de dados consoante a proximidade entre as observações. Dependendo da medida de proximidade escolhida, esta abordagem pode permitir a identificação de áreas com uma grande e baixa densidade de observações. Novas observações que pertencem a áreas de baixa densidade são consideradas candidatas a anómalas.

2.3 Métricas Utilizadas para Avaliação do Desempenho de um Algoritmo

A avaliação do desempenho de um método de deteção de anomalias é de extrema importância para garantir a sua utilização prática. Quando se utilizam metodologias não supervisionadas existe uma maior dificuldade de encontrar uma amostra para a qual se conhece a classe verdadeira (regular ou anómala) a que cada observação pertence. No entanto, para classificar o desempenho do algoritmo é necessário ter dados rotulados.

O procedimento usual é dividir a amostra em dois subconjuntos (escolhidos aleatoriamente). Um destes subconjuntos é usado para treinar o classificador (amostra de treino) e o outro para avaliar o seu desempenho (amostra de teste). A amostra de teste é usada para classificar cada observação que a constitui. Uma vez que a verdadeira classe é conhecida, pode-se cruzar a classe estimada com a classe verdadeira e construir uma tabela chamada de matriz de confusão, onde se consta os quatro padrões possíveis de observações:

- A classe verdadeira e a classe estimada indicam que a observação é anómala. Esta observação é denominada de Verdadeiro Positivo (VP);
- A classe verdadeira e a classe estimada indicam que a observação é regular. Esta observação é denominada de Verdadeiro Negativo (VN);
- A classe verdadeira indica que a observação é anómala, mas o classificador indica erradamente que a observação é regular. Esta observação é denominada de Falso Negativo (FN);
- A classe verdadeira indica que a observação é regular, mas o classificador indica erradamente que a observação é anómala. Esta observação é denominada de Falso Positivo (FP).

Por simplicidade de notação, o número de observações, de cada tipo, na amostra de treino representa-se pela sigla associada a cada padrão. Assim sendo $VP + VN + FN + FP = n_T$, onde n_T representa a dimensão da amostra de treino. A matriz de confusão encontra-se sumariada na Tabela 2.1.

Tabela 2.1: Matriz de confusão.

		Classe Verdadeira	
		Anómalo	Não Anómalo
Classe Estimada	Anómalo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Não Anómalo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Os Falsos Positivos e Falsos Negativos correspondem a erros do classificador. O primeiro termo diz respeito a observações em que a classe verdadeira é a regular. O método previu a observação como anómala. O segundo termo corresponde a observações que na realidade são anómalas, mas que o método as classificou como regulares.

Se o problema com que se está a lidar é um problema de classificação binária de duas classes desequilibradas, é espectável que as observações mais predominantes sejam as observações não anómalas, isto é, o tráfego regular, e que as observações menos predominantes sejam as anómalas.

Este desequilíbrio indicia que além de medidas globais de desempenho do classificador é igualmente importante considerar medidas do desempenho por classe. Por exemplo, se a verdadeira percentagem de anomalias for baixa e se um classificador (absurdo) indica que todas as observações são regulares, a percentagem global de observações mal classificadas coincide com a verdadeira percentagem de anomalias, que se sabe ser baixa. No entanto, a repercussão de ignorar a existência de anomalias pode se catastrófica. Por exemplo, se existir um ataque de segurança e o classificador não o detetar corretamente, o utilizador continua a fornecer os seus dados pessoais ao atacante sem perceber. Nesta situação, o atacante pode adquirir dados pessoais do utilizador como acessos a contas bancárias e isto pode ter graves consequências.

Após se conhecer estas definições e estes valores, calculam-se as métricas para que seja possível avaliar os classificadores em estudo.

A medida global de desempenho de um classificador é a percentagem de observações da amostra de treino bem classificadas, denominada na literatura inglesa por *overall Accurancy* ou simplesmente *Accuracy*. Considerando a Tabela 2.1 pode estimar-se a *Accuracy* por:

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.1}$$

Existem três métricas muito utilizadas para avaliar o quão bem as observações anómalas são detetadas pelo classificador: *Precision*, *Recall* e *F1-score* da classe das anomalias.

A *Precision* contabiliza de entre as anomalias detetadas, qual a percentagem que corresponde a

anomalias verdadeiras [3],

$$Precision = \frac{VP}{VP + FP}. \tag{2.2}$$

Adaptando ao caso das telecomunicações por exemplo, se um classificador indicar que há uma falha de comunicação num certo troço de uma rede, a operadora deve estar segura de que de facto a falha ocorre. No entanto, se não existir falha e o classificador detetar erradamente uma falha (Falso Positivo) este erro pode já não ser tão desagradável para o consumidor final. Neste exemplo, quando um Falso Negativo é muito mais desfavorável para uma operadora, é mais útil usar o *Recall*. O *Recall* de entre as anomalias verdadeiras existentes, identifica a percentagem de anomalias que foram corretamente detetadas [3],

$$Recall = \frac{VP}{VP + FN}. \tag{2.3}$$

Por último, a *F1-score* é a média harmónica das duas métricas acima referidas. Um *F1-score* baixo significa que pelo menos uma das duas métricas, *Precision* ou *Recall*, tem também um valor baixo e, portanto, o método de classificação utilizado não é o mais adequado [3]:

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall}. \tag{2.4}$$

2.4 Abordagens para Deteção de Anomalias

Existem várias abordagens possíveis para deteção de anomalias. A família de abordagens que utilizam distâncias chama-se a abordagem *distance-based*. As outras abordagens possíveis para a deteção de anomalias como a *density-based* e *rank-based* não usam distâncias entre pontos para encontrar observações anómalas e são pouco estudadas neste trabalho [2].

A abordagem *distance-based* considera que as observações mais próximas são mais similares entre si e as observações mais distantes de uma medida de centralidade são possíveis anomalias.

Na abordagem *density-based* considera-se que existe um *cluster* quando existe uma região densa de observações. Neste caso, as observações localizadas em regiões com baixas densidades de observações são consideradas anómalas. Esta abordagem utiliza a densidade local, que pode ser definida como o número de observações numa certa área, ou seja, a densidade de observações existente numa área especifica. Já a abordagem *rank-based* atesta que as observações são anómalas baseando-se em *ranks* das observações pertencentes à vizinhança da observação a ser classificada como anómala ou regular [2].

2.4.1 Abordagens Baseadas na Distância

Quando se estão a definir observações como anómalas ou regulares, usualmente espera-se poder identificar o quão dissimilares ou similares são duas observações. Considera-se que uma observação é bastante dissimilar de outra se a distância ou medida de dissemelhança entre as duas for elevada, ou seja, se existir uma fraca semelhança entre as observações em causa do conjunto de dados. A dissemelhança quantifica o quão diferentes dois objetos são. A dissemelhança entre dois objetos A e B é uma função d_{AB} que verifica as seguintes propriedades: $d_{AB} \geq 0$, $d_{AA} = 0$ e $d_{AB} = d_{BA}$, onde d é a dissemelhança. Uma distância é uma dissemelhança se, para além dos critérios mencionados, também verificar as seguintes condições: $d_{AB} = 0$ se e somente se $A = B$ e se $d_{AB} \leq d_{AC} + d_{CB}$ (desigualdade triângular).

Algumas das medidas de dissemelhanças mais populares entre observações que podem ser modeladas como realizações de vetores aleatoriamente contínuos, isto é, $\mathbf{x}, \mathbf{y} \in R^p$, são a distância de Mahalanobis, a distância Euclidiana, a distância de Minkowski e a distância de Manhattan.

A distância de Mahalanobis entre duas observações representadas por $x = (x_1, x_2, \ldots, x_p)^T$, $y = (y_1, y_2, \ldots, y_p)^T \in R^p$ é definida por:

$$d_{Mahalanobis}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \sum^{-1} (\mathbf{x} - \mathbf{y})}, \tag{2.5}$$

onde $\sum$ é a matriz de covariância amostral estimada a partir dos dados de dimensão $(p \times p)$ e $\sum^{-1}$ é a sua inversa.

Se a matriz de covariância amostral, $\sum$, for igual à matriz de identidade então está-se perante a distância Euclidiana [2]:

$$d_{Euclidiana}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}. \tag{2.6}$$

A distância de Minkowski é uma generalização da distância Euclidiana e é definida por [2]:

$$d_{Minkowski}(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{n} |x_i - y_i|^l)^{\frac{1}{l}}, \tag{2.7}$$

onde $l \in N$ é a ordem da distância. Se $l = 2$, então está-se a calcular a distância Euclidiana, se $l = 1$, então a distância é a distância de Manhattan,

$$d_{Manhattan}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|. \tag{2.8}$$

Tal como o nome indica, a distância da observação **x** a todos os pontos existentes no subconjunto

de dados D define-se por [2],

$$d_{todos}(\mathbf{x}) = \sum_{\mathbf{y} \in D} d(\mathbf{x}, \mathbf{y}), \tag{2.9}$$

onde d é uma qualquer distância escolhida pelo investigador. Esta é conhecida como a distância a todos os pontos.

Os métodos baseados na distância ao vizinho mais próximo baseiam-se na determinação da distância entre si e o seu vizinho mais próximo e definem-se como [2]:

$$d_{proximo}(\mathbf{x}) = min_{\mathbf{y} \in D, \mathbf{x} \neq \mathbf{y}} d(\mathbf{x}, \mathbf{y}), \tag{2.10}$$

sendo D o conjunto de dados.

O cálculo da distância ao k-ésimo vizinho mais próximo é semelhante ao método anterior. No entanto, faz-se uma média das distâncias entre $\mathbf{x} \in R^p$ e os seus k-ésimos vizinhos mais próximos, onde k é menor ou igual ao número total de observações em estudo. Seja $D_k(\mathbf{x})$ o conjunto das k observações pertencentes aos dados iniciais, D, mais próximas de $\mathbf{x}$ então [2],

$$d_{k-proximo}(\mathbf{x}) = \sum_{\mathbf{y} \in D_k(\mathbf{x})} \frac{d(\mathbf{x}, \mathbf{y})}{k}. \tag{2.11}$$

Se $k = 1$, então o resultado será o mesmo que o obtido em (2.10).

2.5 Algoritmos Usando Abordagens Baseadas na Distância

Nesta secção mencionam-se diferentes abordagens de deteção de anomalias baseadas na distância. Inicialmente explica-se a abordagem *Distance Based-Outlier* e fazem-se alguns testes para verificar o desempenho da mesma. Depois menciona-se a abordagem *LOCI*, onde também se explica como funciona a mesma com pseudocódigo. Por último faz-se uma breve referência à abordagem do vizinho mais próximo.

2.5.1 Abordagem Distance Based-Outlier

A abordagem *Distance Based-Outlier*, também é conhecida como $DB(\pi, r) - outlier$. Esta abordagem considera um ponto $\mathbf{x} \in R^p$ com uma vizinhança centrada em $\mathbf{x}$ e de raio r. Sendo $N_{\mathbf{x}}(r)$ a vizinhança, se a esta vizinhança pertencer um número muito baixo de observações então este ponto encontra-se isolado da maioria dos pontos e é considerado uma anomalia. O método tem dois parâmetros: r, o raio da vizinhança e $(1 - \pi)$, a percentagem mínima de pontos fora da vizinhança de $\mathbf{x}$ que leva $\mathbf{x}$ a ser classificado como uma anomalia, dito $DB(\pi, r) - outlier$ [2] [4]. Sendo $D = \{x_1, x_2, \ldots, x_n\}$ o conjunto

das observações recolhidas e n o número de observações pertencentes a $N_{\mathbf{x}}(r)$, se $\frac{N_{\mathbf{x}}}{n} \leq (1-\pi)$ então **x** é classificado como anómalo.

Considera-se que a população em estudo tem uma distribuição Normal univariada de valor esperado μ e variância σ^2, e que uma anomalia é uma observação x que satisfaz a seguinte condição: $|x-\mu| > 3\sigma$. Neste caso, o $DB(\pi, r) - outlier$ correspondente a este critério tem os parâmetros $\pi = 0.9988$ e $r = 0.13\sigma$, e obtem-se $DB(0.9988, 0.13\sigma) - outlier$ [2].

Existem diversas observações anómalas que não se conseguem detetar facilmente, portanto a existência de um método que as detete de forma eficaz e eficiente é essencial. Se por um lado existem observações que se confundem com o conjunto de dados regular, por outro, também existem algumas observações que se distanciam de tal maneira das regulares que facilmente, mas erradamente, são consideradas anómalas sem ser necessária nenhuma abordagem demasiado exaustiva.

Algoritmo 2.1: Algoritmo *DB-outlier*.

Entrada: D, r, π
Saída: Lista de Anomalias
início
  $Outlier\_list = NULL$;
  repita
    $N_{\mathbf{x}}(r) = NULL$;
    repita
      if $dist(\mathbf{x}, \mathbf{y}) \leq r$ **then**
        Inserir **y** em $N_{\mathbf{x}}(r)$;
    até **y** $\in D$;
    if $\#N_{\mathbf{x}}(e) \leq (1-\pi)\#D$ **then**
      Inserir **x** em $Outlier\_list$;
  até **x** $\in D$;

No Algoritmo 2.1 encontra-se um pseudocódigo do algoritmo $DB - outlier$, em que $Outlier\_list$ representa o conjunto de dados que são considerados anómalos pelo algoritmo, $D = \{x_1, x_2, \ldots, x_n\}$ é o conjunto de dados inicial, $(1-\pi)$ é a fração de observações pode estar a uma distância superior a r e $N_{\mathbf{x}}(r)$ é o conjunto de observações **y** que está a uma distância de **x** inferior a r. Os valores de π, r e D são inputs necessários para a utilização deste algoritmo [2]. Para representar o número de elementos de um conjunto utiliza-se o símbolo $\#$.

Considerando um conjunto de dados de uma distribuição Normal padronizada e gerando-se uma amostra de 1000 observações obtém-se as estatísticas sumárias expressas na Tabela 2.2 que descreve a amostra simulada. Como se espera o valor da média e da mediana amostral são semelhantes a zero e o desvio padrão amostral semelhante a um.

Após a utilização do algoritmo com os parâmetros $\pi = 0.9988$ e $r = 0.13$ nos dados gerados, ou seja, em dados que à partida não contêm anomalias obtém-se 972 observações classificadas como regulares e 28 classificadas como anómalas. As estatísticas sumárias dos dados classificados como regulares encontram-se na Tabela 2.3. Nesta tabela são visíveis algumas alterações nas estatísticas

Tabela 2.2: Estatísticas sumárias dos dados sem anomalias.

Mínimo	-3.18
1º Quartil	-0.63
Mediana	0.03
Média	0.01
3º Quartil	0.69
Máximo	3.20
Desvio Padrão	0.95

sumárias, mas a média e mediana amostral continuam semelhantes a zero e o desvio padrão bastante próximo de 1, como era de esperar visto que $r = 0.13\sigma$ e neste caso utilizou-se $\sigma = 1$

Tabela 2.3: Estatísticas sumárias dos dados classificadas como regulares pelo Algoritmo $DB(0.9988, 0.13) - outlier$.

Mínimo	-2.00
1º Quartil	-0.60
Mediana	0.03
Média	0.01
3º Quartil	0.67
Máximo	2.08
Desvio Padrão	0.87

Analisa-se um segundo cenário onde se opta por considerar que 20% de observações da amostra são anómalas e 80% de observações são regulares, tal como se mostra na Figura 2.1. Os dados anómalos seguem uma distribuição Normal de desvio padrão 1 e valor médio 4, portanto a sobreposição dos dados regulares e anómalos é escassa.

Na Tabela 2.4 observam-se os dados correspondentes à Figura 2.1, ou seja, os dados com presença de 20% de anomalias. Através desta tabela consegue-se perceber que o valor da média é superior ao que deveria ser, por interferência do conjunto de dados irregulares. O máximo tem o valor de 7.05, que corresponde ao máximo das observações anómalas e é esperado que este máximo diminua entre as observações classificadas como regulares pelo algoritmo.

Tabela 2.4: Estatísticas sumárias dos dados gerados com 20% de anomalias.

Mínimo	-3.06
1º Quartil	-0.44
Mediana	0.29
Média	0.79
3º Quartil	1.43
Máximo	7.05
Desvio Padrão	1.87

Posteriormente, aplicou-se o algoritmo, com os parâmetros da distribuição Normal mencionados acima, ou seja, $DB(0.9988, 0.13\sigma)$. Definiu-se o desvio padrão, σ, com o valor de 1.

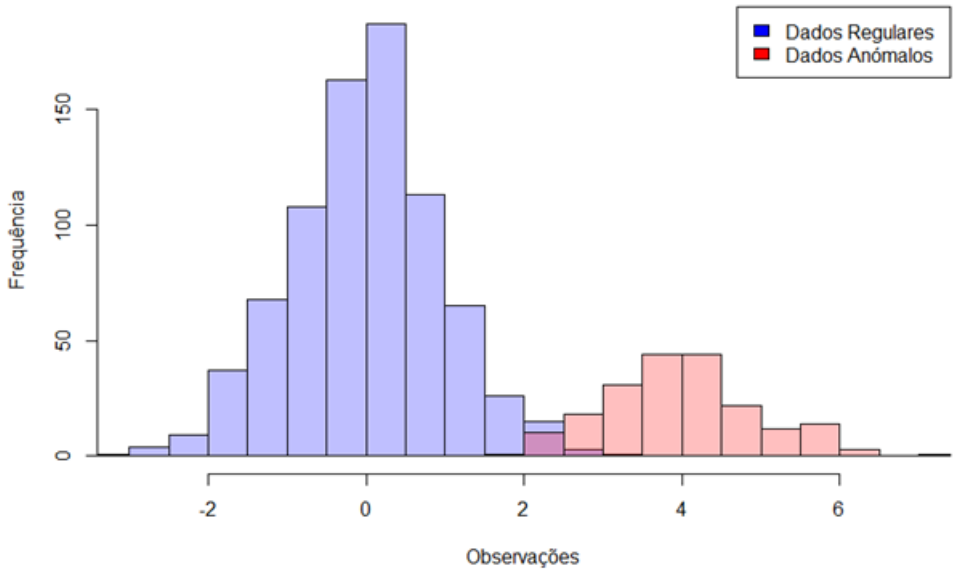

Figura 2.1: Histograma dos dados com presença de 20% de anomalias, geradas por uma Distribuição Normal de valor esperado 4 e variância unitária.

Na Tabela 2.5 consta o resumo das estatísticas sumárias dos dados classificados como regulares pelo algoritmo. Pode-se concluir que os valores ficaram mais próximos dos valores observados da Tabela 2.3, como era de esperar, visto que a Distribuição Normal dos dados regulares tem média zero e desvio padrão unitário, apenas foram introduzidas anomalias para testar o comportamento deste algoritmo perante conjuntos de dados com irregularidades.

Tabela 2.5: Estatísticas sumárias dos dados, com 20% de contaminação, classificados como regulares da amostragem com anomalias após o Algoritmo $DB(0.9988, 0.13) - outlier$.

Mínimo	-1.87
1º Quartil	-0.48
Mediana	0.13
Média	0.34
3º Quartil	0.80
Máximo	4.22
Desvio Padrão	1.34

No entanto, após a observação do histograma da Figura 2.2 com os dados classificados como regulares (839) pelo algoritmo $DB(\pi, r) - outlier$, percebe-se que ainda existiram observações irregulares que não foram detetadas como anomalias.

Tendo em consideração a Tabela 2.6 calcularam-se os valores para a *Accuracy*, *Precision* e *Recall*. Neste caso, grande parte das observações foram corretamente identificadas e como tal obteve-se uma *Accuracy* de 0.885. Adicionalmente, a *Precision* toma o valor 0.764, visto que ainda existiram algumas observações que foram erradamente designadas como anómalas. O *Recall* toma o valor mais baixo

Tabela 2.6: Matriz de confusão com os dados obtidos após a utilização do Algoritmo $DB(0.9988, 0.13) - outlier$.

		Classe Estimada	
		Anómalo	Não Anómalo
Classe Verdadeira	Anómalo	123	77
	Não Anómalo	38	762

de todos, em torno de 0.615, visto que o número de Falsos Negativos foi superior ao número de Falsos Positivos. Perante estes valores, concluiu-se que o desempenho do algoritmo não foi o melhor, pois o *Recall* e a *Precision* tomaram valores relativamente baixos.

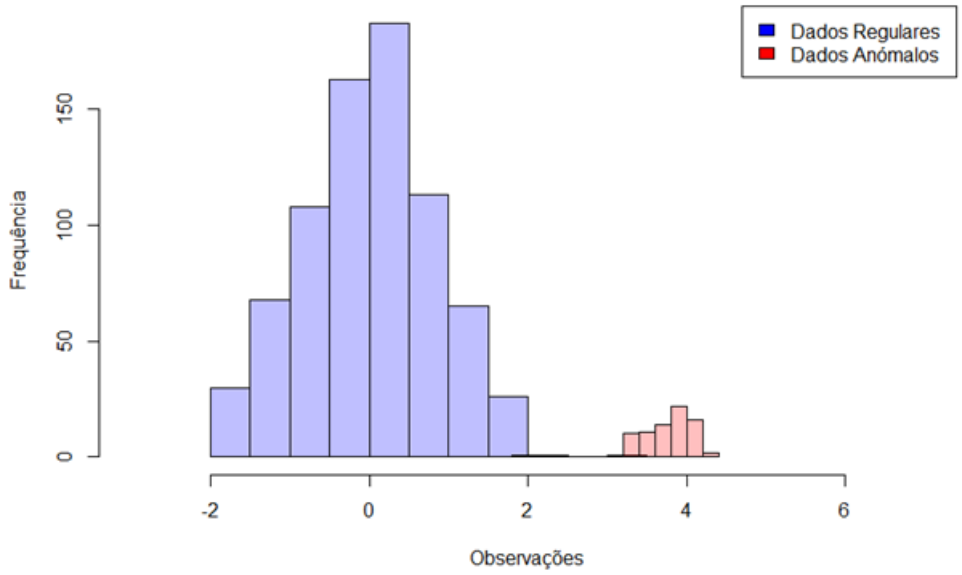

Figura 2.2: Histograma dos dados classificados como regulares pelo algoritmo $DB(0.9988, 0.13) - outlier$, quando aplicado à amostra gerada com contaminação.

Posteriormente, realizaram-se algumas comparações após 500 utilizações do algoritmo para cada valor de percentagem de contaminação $\epsilon = 0, 0.05, 0.1, 0.15, 0.2$ onde os dados regulares seguem uma Distribuição Normal com valores de 0 para a média e de 1 para o desvio padrão. No que diz respeito às observações anómalas utilizou-se, tal como referido anteriormente, uma Distribuição Normal com desvio padrão 1 e média 4. Calculou-se a *Accuracy*, *Precision* e *Recall* para cada amostra gerada e posteriormente utilizou-se a média das mesmas, para cada valor de ϵ, para perceber o quão fidedigno é este algoritmo variando o nível da contaminação dos dados.

Na Figura 2.3 observa-se que a *Accuracy* vai diminuindo ligeiramente à medida que ϵ aumenta. Verifica-se também que a *Accuracy* nunca chega ao valor 1, nem para um valor de $\epsilon = 0$. Este algoritmo é propenso a uma redução da cauda, isto é, os valores extremos usualmente são eliminados.

No que diz respeito à Figura 2.3 observa-se que a *Precision*, da classe das anomalias, para valores muito baixos de ϵ tende a tomar valores perto de 0. A *Precision* calcula a percentagem de anomalias

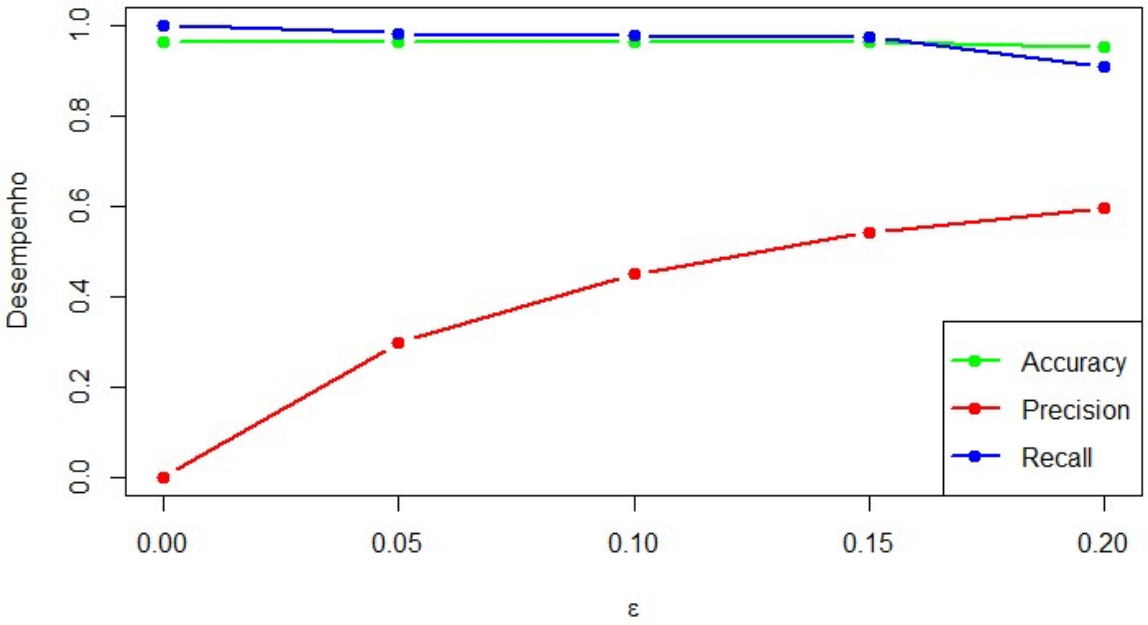

Figura 2.3: Métricas em função de ϵ, percentagem de contaminação nos dados.

reais de entre as observações que foram consideradas anómalas. À medida a percentagem de anomalias aumentam no conjunto de dados a *Precision* tende a aumentar. No entanto, o valor ideal para a *Precision* deveria situar-se bastante perto do valor 1, excetuando o caso em que não existem anomalias. Uma *Precision* tão baixa como a da figura abaixo revela um algoritmo que classifica bastantes observações como anómalas quando na realidade são regulares.

Observando a Figura 2.3, percebe-se que o *Recall* da classe das anomalias para $\epsilon = 0$ tem o valor 1. No entanto, este valor deve-se à inexistência de anomalias. Os restantes valores, exceto para $\epsilon = 0.2$, são bastante elevados o que revela que a maior parte das anomalias são corretamente detetadas. Tal como observado na Tabela 2.5, espera-se que a média e mediana das observações classificadas como regulares se aproximem do valor da média da Distribuição Normal regular, retirando assim as observações anómalas. Portanto seria de esperar que fossem corretamente detetadas uma grande quantidade de observações irregulares.

Numa outra experiência fixou-se o ϵ para se considerar um conjunto de dados em que 20% das observações são anómalas. Seguidamente, variou-se a média da Distribuição Normal da anomalias entre 1 e 9 para perceber o quão eficaz este algoritmo é, tanto em casos em que as médias são muito próximas, como em casos que as médias das duas Distribuições Normais, regular e anómala, são bastante diferentes. Por exemplo, quando a Distribuição Normal anómala tem média 9 as "caudas" das Distribuições já nem se sobrepõem, visto que a média da Distribuição Normal regular é 0. Pela análise da *Accuracy*, presente na Figura 2.4, percebe-se que quanto mais elevado for o valor da média, maior

é *Accuracy*, ou seja, a classificação da classe estimada vai-se aproximando da classe verdadeira e os Falsos Negativos e Falsos Positivos vão decrescendo. Pode observar-se que por exemplo com uma média de 4 na Distribuição Normal regular obtém-se uma *Accuracy* de 0.84.

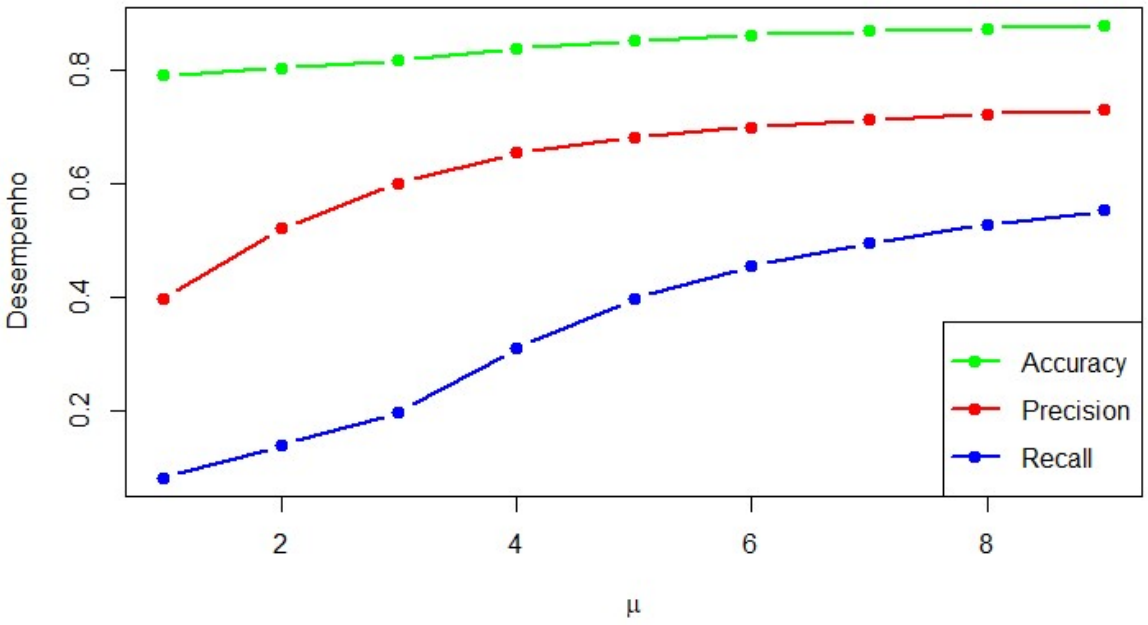

Figura 2.4: Métricas em função da média da Distribuição Anómala, μ, dados gerados com 20% de anomalias.

Adicionalmente, na Figura 2.4 a *Precision* também aumenta com o aumento da média, ou seja, aumenta à medida que as médias vão sendo mais distintas. Como as observações irregulares se distanciam significativamente das restantes observações, torna-se mais simples a identificação das mesmas.

Na Figura 2.4 percebe-se que apesar de o *Recall* toma valores cada vez mais elevados, o seu valor nunca é superior a 0.6. Ou seja, aproximadamente 60% das observações que foram detetadas como anómalas, correspondem a observações regulares na classe verdadeira.

2.5.2 LOCI

O LOCI (*Local Correlation Integral*) é uma abordagem para detetar anomalias que é facilmente adaptável e eficaz, pois faz os cálculos necessários em apenas um passo [2] [5].

O método LOCI utiliza valores calculados através do MDEF (*Multi-granularity Deviation Factor*) para escolher as observações que são anómalas. Se **x** for uma observação do conjunto de dados pode-se definir o MDEF como:

$$MDEF(\mathbf{x}, r, \alpha) = 1 - \frac{n(\mathbf{x}, \alpha r)}{\hat{n}(\mathbf{x}, r, \alpha)}, \tag{2.12}$$

onde $0 < \alpha < 1$ é um parâmetro predeterminado, $n(\mathbf{x}, \alpha r)$ é o número de observações cuja distância a **x** é menor ou igual a αr. Considera-se $\hat{n}(\mathbf{x}, r, \alpha)$ como a média das observações de **y** : $\mathbf{y} \in N_{\mathbf{x}}(r)$ e $N_{\mathbf{x}}(r)$ é o conjunto de todas as observações pertencentes à vizinhança centrada em **x** de raio r.

O MDEF pode ser positivo ou negativo e consoante o resultado podem-se tirar conclusões relativamente à irregularidade das observações. Se o coeficiente for positivo, então a observação será uma candidata a ser classificada como anómala. Se por outro lado o MDEF tiver um valor negativo, então **x** é classificado como regular.

O valor r pertence ao intervalo $[r_{min}, r_{max}]$, sendo que $r_{max} \approx \alpha^{-1} max_{\mathbf{x},\mathbf{y} \in D} \delta(\mathbf{x}, \mathbf{y})$, em que D é o conjunto de dados e $\delta(\mathbf{x}, \mathbf{y})$ representa a distância entre **x** e **y**. O valor r_{min} é escolhido de forma a que os vizinhos mais próximos contenham cerca de 20 observações.

Se o desvio padrão de $n(\mathbf{x}, \alpha r)$ se definir por $\sigma_{\hat{n}}(\mathbf{x}, r, \alpha)$ então

$$\sigma_{MDEF}(\mathbf{x}, r, \alpha) = \frac{\sigma_{\hat{n}}(\mathbf{x}, r, \alpha)}{\hat{n}(\mathbf{x}, r, \alpha)}. \tag{2.13}$$

De acordo com este método, considera-se que **x** é anómalo se

$$MDEF(\mathbf{x}, r, \alpha) > k_\sigma \times \sigma_{MDEF}(\mathbf{x}, r, \alpha). \tag{2.14}$$

Autores em [6] sugerem $\alpha = \frac{1}{2}$ e $k_\sigma = 3$, apesar de k_σ poder tomar outros valores não negativos e $0 < \alpha < 1$.

Na Figura 2.5 observa-se um exemplo da utilização do algoritmo para descobrir as observações a uma distância de **x** menor ou igual a αr, ou seja, uma visualização gráfica de como utilizar a fórmula $n(\mathbf{x}, \alpha r)$. Neste caso $\hat{n}(x_0, r, \alpha) = \frac{1+4+4+1+3}{5} = \frac{13}{5} = 2.6$ e $MDEF(x_0, r, \alpha) = 1 - \frac{1}{2.6} = 0.61538$. Note-se que a vizinhança r de x_0 contém outras 4 observações, x_1, x_2, x_3 e x_4. A vizinhança αr de x_0 contém apenas 1 observação, que é o próprio x_0. No que diz respeito a x_1, x_2, x_3 e x_4 contêm 4, 4, 1 e 3 observações, respetivamente.

Na Algoritmo 2.2 encontra-se o pseudocódigo do algoritmo LOCI onde $N_{\mathbf{x}}(r_{max})$ é o conjunto de observações com uma distância do **x** inferior a r_{max}, $\delta(\mathbf{x}, x_{m-NN})$ é a distância entre **x** e o m-ésimo vizinho mais próximo de **x**, $n(\mathbf{x}, \alpha r)$ é o número de observações no conjunto de dados e $n(\mathbf{x}, \alpha r)$ é a média dos vizinhos mais próximos de **x** [2]. MDEF é o fator de desvio de multigranularidade e $\sigma_{MDEF}(\mathbf{x}, r, \alpha)$ é o

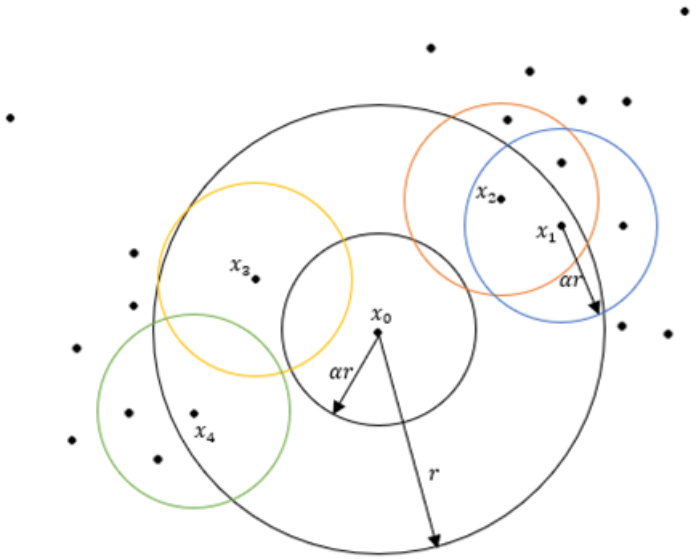

Figura 2.5: Representação Gráfica do algoritmo LOCI para um pequeno conjunto de dados.

desvio padrão do mesmo.

Algoritmo 2.2: Algoritmo *Local Correlation Integral*.

início
  repita
    Encontrar $N_{\mathbf{x}}(r_{max})$;
    Calcular $\delta(\mathbf{x}, x_{m-NN})$ e $\delta(\mathbf{x}, \alpha x_{m-NN})$ para cada $1 \leq m \leq N$;
    Ordenar a lista das distâncias por ordem crescente;
    Por cada r na lista ordenada, calcular $n(\mathbf{x}, \alpha r)$ e $\hat{n}(\mathbf{x}, r, \alpha)$;
    Calcular $MDEF(\mathbf{x}, r, \alpha)$ e $\sigma_{MDEF}(\mathbf{x}, r, \alpha)$;
    if $MDEF(\mathbf{x}, r, \alpha) > k_\sigma \sigma_{MDEF}(\mathbf{x}, r, \alpha)$ **then**
      Sinalizar **x** como uma potencial anomalia;
  até each $\mathbf{x} \in D$;

2.5.3 Abordagem do Vizinho Mais Próximo

Pode-se considerar que uma observação anómala é uma observação cujo vizinho mais próximo se encontra a uma distância consideravelmente grande [2]. Contrariamente às abordagens mencionadas em 2.5.1 e 2.5.2, em que as observações eram classificadas como anómalas segundo a quantidade de observações numa vizinhança do ponto com um determinado raio, esta abordagem baseia-se diretamente na distância entre observações para decidir quais as observações anómalas.

Mais especificamente, é calculada a distância entre **x** e cada uma das restantes observações do conjunto de dados. O mínimo destas distâncias representa a distância entre **x** e o seu vizinho mais próximo. Esta abordagem foi desenvolvida para uma maior eficiência a calcular $D^k(\mathbf{x})$, que é a distância do k vizinho mais próximo de **x**. Se o valor $D^k(\mathbf{x})$ for elevado, então a observação será classificada como anómala [7].

Seja

$$\alpha(\mathbf{x}) = min_{\mathbf{y} \in D \setminus \{\mathbf{x}\}} d(\mathbf{x}, \mathbf{y}) \tag{2.15}$$

a distância entre **x** e o seu vizinho mais próximo. **x** diz-se uma anomalia se $\alpha(\mathbf{x})$ é muito elevado quando comparado com $\alpha(\mathbf{y}), \mathbf{y} \in D \setminus \{\mathbf{x}\}$. Em [2] é dado um exemplo onde um objeto anómalo ou um ponto anómalo é considerado o centro de um cluster bem definido. Nesse sentido, os autores sugeriram considerar a distância ao k vizinho mais próximo.

Seja $\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \ldots, \alpha_k(\mathbf{x})$ as k menores distâncias entre **x** e **y** $\in D \setminus \{\mathbf{x}\}$. A ideia é usar uma medida de localização de $\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \ldots, \alpha_k(\mathbf{x})$ que nos permite decidir se uma observação é ou não anómala. O mais usual é usar a média da $\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \ldots, \alpha_k(\mathbf{x})$ ou a mediana destas distâncias, que se sabe ser mais robusta.

No entanto, este critério não funciona corretamente se existir uma grande quantidade de *clusters* no conjunto de dados, especialmente se a densidade dos mesmos for significativamente diferente. Para tal foi proposta uma medida baseada no peso de **x**:

$$\sum_{i=1}^{k} d_i(\mathbf{x}), \tag{2.16}$$

onde **x** é a observação, tal como mencionado anteriormente.

2.6 Noções Gerais de Séries Temporais

Assume-se que um conjunto de dados é uma série temporal quando se tem uma sequência de observações ordenadas cronologicamente, em que cada observação está associada a um instante de tempo [8]. No caso de uma série temporal univariada, a cada instante corresponde uma observação univariada, ou seja, um único valor. No caso multivariado, a cada instante corresponde uma observação multivariada, ou seja, um vetor de valores. Note-se que observações consecutivas não têm necessariamente de ocorrer em instantes igualmente espaçados entre si, embora tal aconteça em muitos casos.

Um objetivo importante da análise de séries temporais é encontrar métodos capazes de descrever os dados. Na literatura sobre o tema, surgem diversos métodos estatísticos visando a caracterização de uma série temporal em termos de aspetos como dependência em relação às observações passadas, tendência, comportamento sazonal ou cíclico, e visando também remover ruído aleatório presente nos dados [2] [9].

Um dos modelos bastante usados em séries temporais é o ARMA (*Autoregressive Moving Average*) cuja designação identifica as características do modelo. “AR” indica que é autorregressivo, “MA” refere-se a médias móveis (*Moving Arevage*). Admita-se que $x_1, x_2, \ldots, x_n$ são as n observações de um processo estocástico, em que x_t é o valor observado no instante $t (t = 1, 2, \ldots, n)$.

Um modelo autorregressivo (AR) de ordem p, AR(p), significa que valor atual da série, x_t, pode ser explicado como uma função de $p(p > 0)$ valores passados [10],

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t. \tag{2.17}$$

Sendo que ϵ_t é o erro, um ruído branco que segue uma Distribuição Normal com média 0 e variância $\sigma^2(\epsilon_t \ N(0, \sigma^2))$ e $\phi_i(i = 1, 2, \ldots, p)$ são constantes reais, $\phi_p \neq 0$. A expressão (2.17) corresponde a uma série com média zero.

Por exemplo, um modelo autorregressivo AR(1) escreve-se como

$$x_t = \phi_1 x_{t-1} + \epsilon_t. \tag{2.18}$$

Num modelo de médias móveis (MA) assume-se que o valor atual da série, x_t, pode ser expresso como uma combinação linear de ruídos passados. Assim, um modelo de médias móveis de ordem $q(q > 0)$, MA(q), pode ser escrito como

$$x_t = \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}, \tag{2.19}$$

onde $\epsilon_w \ N(0, \sigma_w^2)$, com $w = t - q, \ldots, t, \theta_i(i = 1, 2, \ldots, q), \theta_q \neq 0$ são constantes reais. Esta série é estacionária em covariância, isto é, a média e a autocovariância são constantes, não variando no tempo. Assim, para $q = 1$ o modelo MA(1) fica:

$$x_t = \epsilon_t + \theta \epsilon_{(t-1)}. \tag{2.20}$$

Quando se combina um modelo AR(p) com um modelo MA(q) tem-se o modelo ARMA(p,q) em que o parâmetro p é o número de termos autorregressivos, ou seja, o número de atrasos que são necessários como previsores e q é o número de erros de previsão passados.

O modelo ARMA pode ser representado por:

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t, \tag{2.21}$$

com $\theta_q \neq 0$, $\phi_p \neq 0$, $\epsilon_w \ N(0, \sigma_w^2)$. Esta expressão corresponde a uma série com média zero.

2.7 Abordagens de Deteção de Anomalias em Séries Temporais

Na deteção de anomalias em séries temporais, podem-se destacar dois tipos de problemas diferentes. Por vezes, o objetivo é distinguir séries temporais anómalas entre diferentes séries e outras vezes o objetivo é distinguir subsequências anómalas pertencentes a uma única série temporal. Para além disso, pode-se referir também o problema da deteção de anomalias online, em que se pretende detetar anomalias à medida que as mesmas vão aparecendo, admitindo-se que o processo de geração de dados possa variar ao longo do tempo.

Quando o objetivo é distinguir séries anómalas entre várias séries temporais, pode-se considerar todo o período de tempo em que se têm dados das múltiplas séries e aplicar algoritmos que permitam assinalar as séries anómalas. Por exemplo, se as séries puderem ser representadas satisfatoriamente por modelos ARMA(p,q), podem ser calculadas distâncias entre os parâmetros que definem as séries e serem definidos limiares para estas distâncias que permitam identificar séries anómalas.

Para determinar o quão similares são duas séries devem-se ter em conta certos aspetos:

- Duas séries podem estar sobrepostas em intervalos de tempo diferentes, considerando-se para efeitos de cálculo a interseção dos dois intervalos de tempo;
- Se existirem missing values e não existir qualquer tipo de informação a esse respeito, assume-se que os missing values podem ser preenchidos por uma interpolação das observações dos dados existentes.

Uma das abordagens plausíveis para classificar a similaridade entre uma série temporal e um conjunto de outras séries temporais é definir a distância como a média das distâncias ponto a ponto [2].

No entanto, as séries temporais podem ser bastante diferentes umas das outras apenas num determinado intervalo de tempo. Neste caso, pode-se querer identificar quer a série com comportamento anómalo, quer o período de tempo relevante em que o comportamento é diferente das restantes séries [2].

Analisando o comportamento de uma mesma serie temporal, é importante perceber o que varia do seu comportamento regular e se existe algum motivo aparente para esse acontecimento. Nestes casos, uma observação é candidata a anómala quando se desvia significativamente do resto dos dados desse mesmo conjunto de dados. É importante identificar o momento em que os dados começaram a variar relativamente ao comportamento regular e se foi apenas uma anomalia pontual ou uma sequência anómala.

Quando se tentam identificar anomalias numa mesma série temporal, podem surgir dois tipos de anomalias: anomalias taxa e anomalias contextuais [2]. Considera-se que existe uma anomalia taxa quando os valores, se observados individualmente parecem assumir valores regulares, mas o ritmo a que a alteração foi feita parece anómalo. Por outro lado, na anomalia contextual, as observações

não parecem anómalas tendo em conta toda a gama de valores possíveis no passado, mas apenas relativamente às observações imediatamente anteriores.

Quando se está perante um algoritmo de deteção de anomalias online é esperado que o mesmo consiga detetar o mais rápido possível a presença de uma anomalia. Tal como foi referido anteriormente, os algoritmos de deteção de anomalias online têm de consegui detetar as novas anomalias no conjunto de dados, mesmo que o comportamento dos dados varie ao longo do tempo. Caso os dados variem substancialmente e se tenha utilizado um algoritmo de aprendizagem para determinar os parâmetros do modelo, é necessário treinar de novo o algoritmo, adicionando novos dados de treino e descartando alguns dos dados antigos. Desta forma, os parâmetros do modelo são atualizados regularmente de modo a que novas anomalias possam ser detetadas corretamente [2].

2.7.1 Algoritmos Aplicáveis a Séries Temporais

A deteção de anomalias em séries temporais, seja qual for o tipo de problema abordado, requer métodos que tenham em consideração a dependência temporal destes dados, que formam uma sequência de observações determinada pelos instantes de ocorrência. Portanto, são necessários métodos de deteção de anomalias aplicáveis a séries temporais.

O problema da deteção de anomalias em series temporais também pode ser abordado recorrendo a métodos aplicáveis a dados não temporais, tais como os algoritmos descritos nas secções 2.5.1, 2.5.2 e 2.5.3, adaptando estes métodos a séries temporais. Outra forma de abordar este problema baseia-se na construção de um modelo específico para séries temporais como o ARMA, descrito na secção 2.6, com base em valores passados da série. Este modelo permite calcular resíduos correspondentes às diferenças entre os valores observados e os estimados pelo modelo e assim assinalar anomalias com base, por exemplo, num limiar pré-estabelecido [11]. Para além disso, segundo [12], outra abordagem possível visa determinar os instantes em que ocorrem mudanças nas características da série, tais como tendência e sazonalidade, para desta forma detetar anomalias. Adicionalmente, a deteção de subsequências de observações anómalas em séries temporais tem motivado o desenvolvimento de algoritmos que identificam anomalias com base na similaridade entre subsequências como é o caso do método SAX (*Symbolic Aggregate aproXimation*) [13] [14].

De seguida, são descritos três algoritmos particularmente relevantes tendo em conta o conjunto de dados que se irá analisar na Dissertação de Mestrado. Assim, é apresentada a heurística proposta por Salvador e Nogueira em [6], o método de *Tukey* de deteção de anomalias e o método *SAX*.

2.7.1.A Abordagem Heurística

Esta abordagem descrita em [6] propõe o uso de médias móveis das observações passadas e o uso do RTT médio (avgRTT). O RTT é definido como o tempo de ida-e-volta entre um *host* de origem e um

host sob vigilância [1]. Tendo em conta esta heurística, uma observação é declarada como anómala se um determinado número de $k = 10$ observações consecutivas exceder o limite ϵ, considerando $\epsilon = 1.2$ multiplicado pela média das $h = 480$ observações passadas.

2.7.1.B Método de *Tukey*

O método de deteção de anomalias de *Tukey* recorre aos 1º e 3º quartis amostrais e à diferença entre eles para definir um limite inferior e um limite superior, fora dos quais as observações são potenciais anomalias [15] [16]. Sejam Q_1 e Q_3, respetivamente, o 1º e 3º quartis amostrais e seja a amplitude interquartil (IQR) a diferença entre eles:

$$IQR = Q_3 - Q_1. \tag{2.22}$$

Segundo este método, uma observação será potencialmente uma anomalia quando se encontrar na região $x : Q_3 + \delta IQR < x \vee x < Q_1 - \delta IQR$. O valor de k pode variar de acordo com o conjunto de dados. É habitual uma observação ser considerada uma anomalia severa quando $\delta = 3$ e ser considerada de possível anomalia quando $\delta = 1.5$.

Este método, permite também que o limite superior e limite inferior sejam definidos por $Q_3 + \delta IQR$ e $Q_1 - \delta IQR$ respetivamente, sendo que δ é um valor considerado apropriado para o conjunto de dados.

Este método pode ser aplicado no contexto de séries temporais univariadas considerado, por exemplo, uma janela deslizante de *n* observações para as quais são calculados limiares segundo o método de *Tukey* e a observação no instante seguinte ao intervalo é comparada com estes limiares para decidir se é uma anomalia ou não.

Este procedimento é análogo ao adotado na heurística descrita em 2.7.1.A. No entanto, a heurística calcula o limiar para assinalar as anomalias com base na média, um estimador que não é robusto na presença de anomalias. O método de *Tukey* tem vantagem em termos de robustez, por definir os limiares com base no 1º e 3º quartil e amplitude interquartil.

2.7.1.C Método *SAX*

O número de dispositivos com acesso à Internet é cada vez maior e, como tal, a quantidade de tráfego também. Portanto, ao analisar séries temporais no contexto do estudo do tráfego na Internet, são particularmente interessantes os métodos que permitem a redução do tamanho da série.

O *SAX* (*Symbolic Aggregate aproXimation*) é um método de deteção de anomalias que representa uma série temporal através de símbolos, isto é, por um conjunto de letras [12] [14]. A primeira etapa do *SAX* envolve uma transformação feita pelo *PAA* (*Piecewise Aggregate Approximation*), um algoritmo usado para diminuir, em tempo, o tamanho de uma série temporal mediante a respetiva divisão em

partes de tamanho igual. Nesta etapa a quantidade de observações é significativamente reduzida. Na segunda etapa, a cada parte é atribuído um símbolo que é uma aproximação do valor original [17].

Seja $x_1, x_2, \ldots, x_n$ uma série temporal de tamanho n. Tipicamente, antes de aplicar o *PAA*, a série é normalizada para ter média zero e desvio padrão unitário, fazendo a padronização:

$$c_j = \frac{x_j - \mu_x}{\sigma_x}, \tag{2.23}$$

onde c_j são as observações originais da serie temporal $(j = 1, \ldots, n)$, μ_x é a média das observações e σ_x é o desvio padrão das observações. A normalização é essencial nos problemas de comparação de séries, pois só normalizada é que a comparação faz sentido [13].

Posteriormente, aplica-se o *PAA* dividindo o conjunto de dados em segmentos do mesmo tamanho para tornar o conjunto de dados mais pequeno e mais simples, visto que usualmente os conjuntos de dados são bastante complexos pelo elevado número de observações existentes nos mesmos. Assim, a série normalizada de tamanho n vai ser transformada numa série de tamanho $w < n$ (desejavelmente $w \ll n$), pela divisão em segmentos com o mesmo número de observações.

Seja $c_1, c_2, \ldots, c_n$ a série temporal normalizada e w o número de divisões feitas nesta série para obter a representação reduzida da série, ou seja, o número de segmentos da série temporal $\overline{c}_1, \overline{c}_2, \ldots, \overline{c}_w$, onde $\overline{c}_i$ é dado por [12]:

$$\overline{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j. \tag{2.24}$$

Isto significa que cada segmento é representado pela média das observações em cada segmento. Por simplicidade, assume-se que w é um divisor de n. Após determinar os segmentos da representação *PAA* será atribuído um símbolo a cada segmento, de acordo com a média calculada acima. É utilizado um alfabeto $\alpha_1, \alpha_2, \ldots, \alpha_{af}$, de tamanho $af > 2$, onde cada α_j é uma letra do alfabeto $(\alpha_1 = a, \alpha_2 = b, \ldots)$. A série resultante, $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n$ é tal que: $\hat{c}_i = \alpha_j$, se e só se $\beta_{j-1} \leq \overline{c}_i < \beta_j$, onde $i = 1, \ldots, w, j = 1, \ldots, af$ e $\beta_1, \beta_2, \ldots, \beta_{af-1}$ é uma sequência de pontos tais que a área sob a curva da distribuição N(0,1) entre β_i e β_{i+1} é $\frac{1}{af}$ e β_0 e β_{af-1} são definidos com $-\infty$ e $+\infty$. Em [14], estes pontos são definidos com base na distribuição N(0,1).

A grande diferença entre o *SAX* e o *PAA* é que o *SAX* atribui um símbolo a cada segmento, portanto torna-se mais fácil quantificar em termos de semelhança e dissemelhança, pela facilidade de pesquisa de padrões em sequências de caracteres. Para além disso, com um *SAX* reduz-se a dimensão do conjunto de dados, em amplitude, para apenas af possibilidades. Por exemplo, um troço em que a primeira letra é a letra "a", poderá ser mais similar a outro troço que também comece pela letra "a" [18]. Além disso, quando se está a lidar com grandes quantidades de dados, o tempo de procura de padrões diminui e torna-se mais simples.

A Figura 2.6 é uma figura ilustrativa da aplicação do método de *SAX*. Verifica-se que, para cada

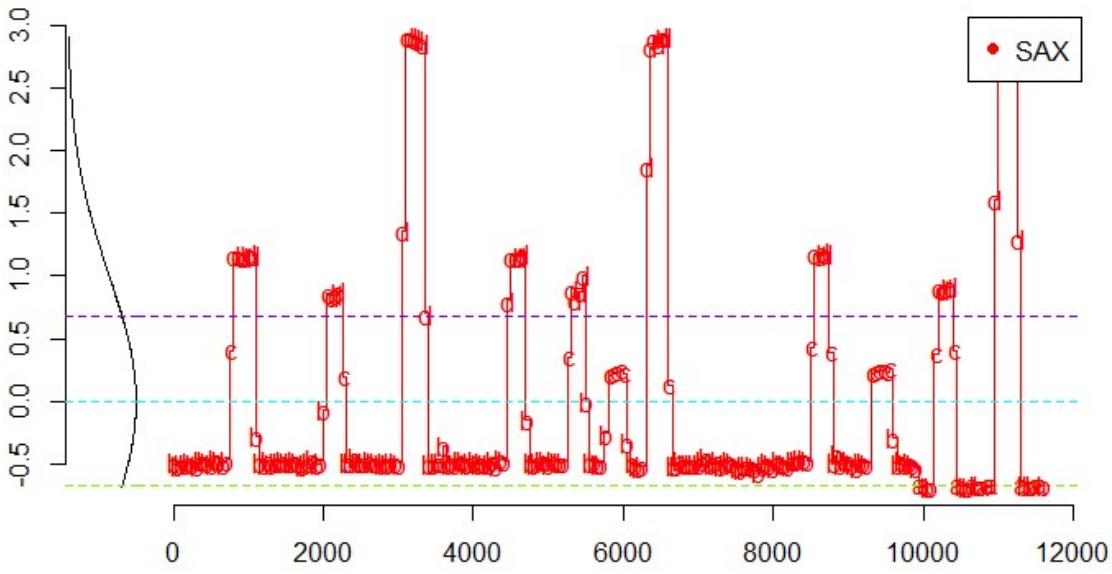

Figura 2.6: Figura ilustrativa relativa ao método de *SAX*.

símbolo existe alguma variedade de valores possíveis, algo que o diferencia do *PAA* visto que com a utilização do *SAX* apenas existem 4 resultados possíveis, a, b, c e d. Note-se que a Figura 2.6 utiliza dados normalizados.

3

Ataques de Redirecionamento BGP

Os ataques que, neste relatório, se tentam detetar são de redirecionamento *BGP*. Este ataque é semelhante a um *man-in-the-middle*. No ataque *MITM* a comunicação entre servidor-cliente é alterada, começando a ser desviada por uma atacante. Desta forma o atacante pode transmitir os dados sem fazer nenhuma alteração, ou pode alterar informações. Neste ataque normalmente o objetivo é interceptar informações confidenciais e utilizá-las para fins pessoais.

BGP significa *Border Gateway Protocol* e é um protocolo que permite que o tráfego circule através da Internet a partir de um IP de origem e até a um IP de destino. Cada router *BGP* armazena uma tabela de reencaminhamento de dados com as melhores rotas entre Sistemas Autónomos, *AS*. Eles são constantemente atualizados, permitindo assim que o tráfego circule sempre pelo percurso mais curto e direto. O *BGP* também possibilita o crescimento em larga escala da Internet, daí a importância em detetar e prevenir ataques às tabelas de reencaminhamento *BGP*.

Um ataque *BGP* geralmente acontece quando um sistema autónomo espalha prefixos IP que não lhe pertencem. Esta informação posteriormente espalha-se pela Internet e são adicionadas novas rotas às tabelas de encaminhamento *BGP*. Muitas vezes, estes ataques são de difícil deteção porque os *AS* podem estar "camuflados".

O *BGP* pode ser *eBGP*, quando acontece entre dois *routers BGP* de diferentes Sistemas Autónomos (*AS*), e *iBGP*, quando a comunicação *BGP* ocorre entre o mesmo *AS*. O *eBGP* é implementado em *routers* de fronteira e é responsável pela conexão entre diferentes organizações, [19].

As informações de *routing* são mantidas num tabela de encaminhamento por cada um dos *routers*. Esta tabela contém todas as rotas incluindo rotas estáticas ou rotas pelo próprio protocolo *BGP* (*iBGP* e/ou *eBGP*), [20]. A tabela de encaminhamento *BGP* também é útil para resolução de endereços, sendo assim deve estar sempre atualizada com as novas rotas. Se uma entrada desta tabela de encaminhamento estiver errada, isto é, se outro router anunciar uma rota erradamente pode ter resultados catastróficos. Sendo assim, é importante detetar falhas na tabela de encaminhamento o mais rápido possível.

Os *BGP hijacks*, ou seja, as alterações nas tabelas *BGP* por parte de um atacante, são um grande problema na Internet. As consequências de ataques *BGP hijacking* são variadas e não existe um sistema de confiança que descarte automaticamente falsos anúncios que afetam negativamente a tabela de encaminhamento de dados. Muitas vezes o *BGP hijacking* é uma ação recorrente e algumas redes utilizam diferentes prefixos ao longo de vários anos [21].

4

Análise inicial do conjunto de dados

Nesta Dissertação de Mestrado é analisado um conjunto de dados relativo a ataques de redireccionamento de tráfego na Internet [6]. Estes ataques são provocados pelo envenenamento do protocolo *BGP* (*Border Gateway Protocol*).

O *BGP* é o protocolo utilizado na Internet para encaminhamento entre Sistemas Autónomos. Este protocolo não possuí mecanismos de segurança. Utilizando o protocolo *BGP*, os *routers* situados na fronteira entre Sistemas Autónomos anunciam prefixos de rede e rotas para esses prefixos, sendo uma rota uma sequência de Sistemas Autónomos.

Um *router* malicioso pode anunciar um prefixo de rede que não pertence ao seu Sistema Autónomo, como por exemplo o prefixo do YouTube, provocando o redireccionamento do tráfego destinado a esse prefixo para si próprio.

Em [6] foi proposta uma metodologia para deteção de ataques de redireccionamento baseando-se num conjunto de sondas (*probes*) espalhadas pelo globo. Com base nesta infraestrutura, foram efetuadas medições e foi obtido o conjunto de dados que é apresentado neste capítulo.

Foram colocados 4 *targets* em locais diferentes, 12 *probes* e 4 *relays*. Um *target* é um recetor de dados que posteriormente os envia à respetiva *probe*, uma *probe* é um local a partir do qual se enviam

pacotes de dados e se recebem, para fazer os cálculos e se decidir se uma observação será anómala ou não, e por fim um *relay* é um atacante que faz o desvio dos dados.

Na Tabela 4.1 encontram-se todos os *targets*, *probes* e *relays* existentes no conjunto de dados e na Figura 4.1 encontra-se a distribuição das *probes*, *targets* e *relays* pelo mapa do mundo, para uma visualização mais clara do que se está a retratar.

Tabela 4.1: Targets, probes e relays do conjunto de dados.

Targets	**Probes**			**Relays**
Chicago1	Amesterdam	Iceland	SaoPaulo2	LA1
Frankfurt1	Chicago2	Israel	Johannesburg1	Madrid
HongKong	VdM	LA2	Johannesburg2	Moscow
London	Frankfurt2	Milan	Sweden	SaoPaulo1

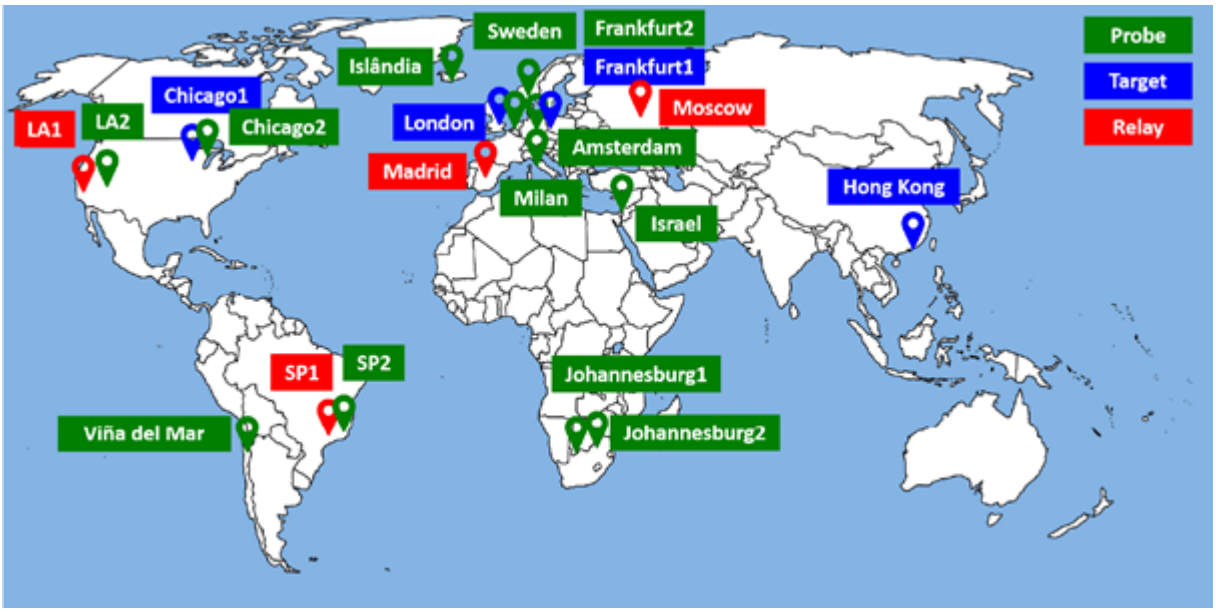

Figura 4.1: Localização das probes, targets e relays. Figura retirada de [1], com permissão.

A deteção dos ataques de redireccionamento é efetuada com base em medições de *RTT* (*Round-Trip-Time*). Neste caso, o *RTT* é o intervalo de tempo que os pacotes demoraram a fazer o caminho *probe-target-probe*. São enviados 10 pacotes de 2 min em 2 min e pretende-se, a partir dos dados, distinguir o que é anómalo do que é regular. O tráfego regular é o tráfego que faz o percurso *probe-target-probe* e o tráfego anómalo é o tráfego que percorre o percurso *probe-relay-target-probe*. De entre estes 10 pacotes selecionou-se o mínimo, máximo, média, mediana e desvio padrão para considerar como observações para o conjunto de dados. Na Figura 4.2 encontra-se um exemplo de trajetória de *probe-target-probe* e de uma trajetória *probe-relay-target-probe*.

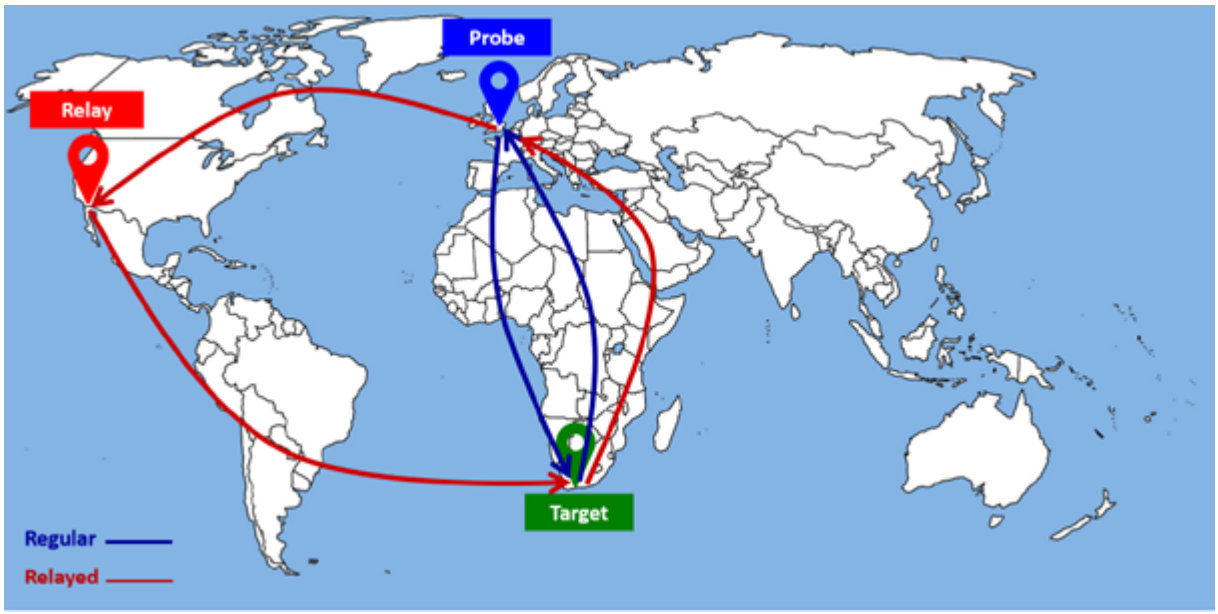

Figura 4.2: Trajeto do tráfego regular e do tráfego anómalo com *probe* em Londres, *target* em Johannesburg e *relay* em Los Angeles. Figura retirada de [1], com permissão.

5

Heurística de Salvador e Nogueira

Este capítulo é referente à heurística proposta por Salvador e Nogueira [6].

A heurística de Salvador e Nogueira utiliza médias móveis das observações passadas e o *RTT* médio. O *RTT* médio é definido como a média de 10 medições de *RTT* realizadas num dado instante. Cada *RTT* é o tempo de ida e volta desde uma origem, *probe*, até um local sob observação, *target*. Usualmente considera-se que para uma observação ser considerada anómala, tem de estar no seguimento de 10 observações anómalas, ou seja, $k = 10$. O k ajuda a prevenir que uma observação regular isolada e com um *RTT* médio bastante elevado seja classificada como anómala. Neste capítulo também são feitos vários testes ao conjunto de dados para identificar qual o melhor valor de k, tendo em conta as diferentes métricas abordadas no Capítulo 2, ou seja, *Accuracy*, *Precision*, *Recall* e *F1-Score*.

A heurística de Salvador e Nogueira descrita em [6] também utiliza outros valores de entrada, tal como mencionado na secção 2.7.1.A. Estes parâmetros serão avaliados neste capítulo para encontrar os valores que conduzem a melhores níveis de desempenho. O parâmetro h define o comprimento da janela deslizante. Isto é, designa o número de *RTT* médios passados necessários para descrever se a observação excede o limite ou não, sendo que o valor considerado adequado pela heurística de Salvador e Nogueira é de $h = 480$. O limiar usado para decidir se uma observação é anómala é definido

utilizando um valor ϵ. Em [6] sugere-se a utilização de um parâmetro $\epsilon = 1.2$. Sendo assim obtém-se a seguinte expressão para o limiar:

$$limiar = \epsilon \times \overline{x}_h, \tag{5.1}$$

sendo que $\overline{x}_h$ corresponde à média das h observações passadas não classificadas como anomalias. O valor de h que mais se adapta ao conjunto de dados é estudado na secção 5.3.

O conjunto de dados em análise dispõe de medições realizadas por 12 *probes*. Sendo que segundo a heurística de Salvador e Nogueira apenas é necessário que 50% das *probes* votem a favor de uma anomalia, para se considerar que o respetivo instante está a sofrer um ataque.

Neste capítulo é feito um estudo para perceber se todas as *probes* são necessárias ou se há certas *probes* mais úteis e outra *probes* dispensáveis e irrelevante, dependendo nomeadamente da sua localização relativamente ao *target*.

5.1 Preparação do Conjunto de Dados

Numa observação inicial dos *RTT* médios correspondentes a uma *probe* do conjunto de dados, por exemplo na Figura 5.1, *probe* Amesterdam na monitorização ao *target* Chicago1, percebe-se que existem algumas observações que tomam valores significativamente elevados e que podem causar ruído desnecessário. Sendo assim, as observações com um valor elevado são eliminadas da análise. Após a análise de todas as *probes* percebeu-se que os dados relevantes nunca excediam o valor 600, sendo assim todas as observações acima de 600 foram excluídas. Observa-se que não são informações relevantes em termos de *RTT* médio e perturbam significativamente a média global do conjunto de dados.

Observando um *target* mais problemático e que contém *avgRTT* mais elevados, como é o caso de Hong Kong, representado na Figura 5.2, percebe-se que a informação útil e importante não passa do valor 600 e o mesmo acontece para os restantes pares *probe-target*. Sendo assim, mais uma vez comprova-se que o valor estabelecido anteriormente é adequado.

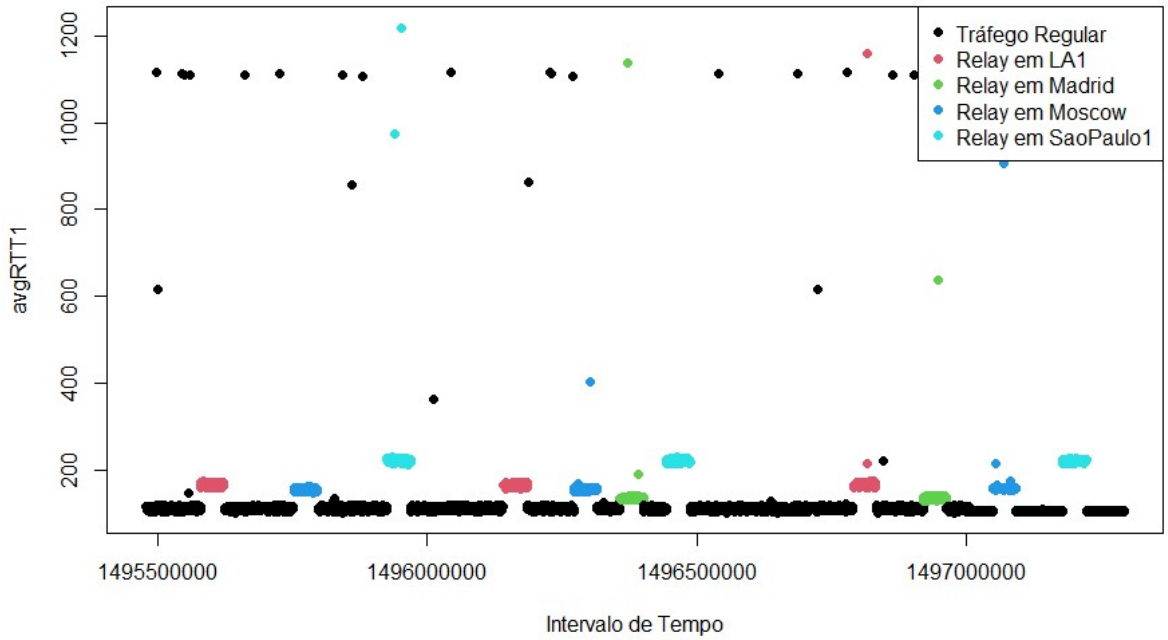

Figura 5.1: Visualização dos *RTT* médios do tráfego entre o *target* 1, Chicago1, e a *probe* 1, Amesterdam.

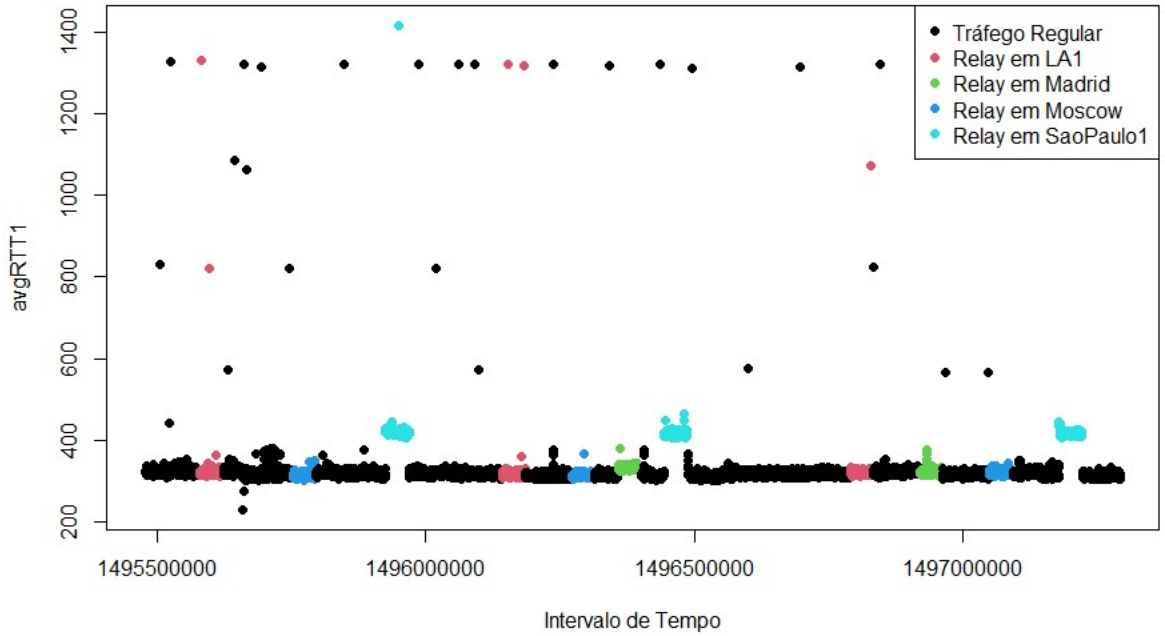

Figura 5.2: Visualização dos *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 1, Amesterdam.

5.2 Estudo da Variação do parâmetro ϵ

A heurística de Salvador e Nogueira definida em [6] pressupõe para ϵ o valor 1.2. Este estudo pretende verificar se o mesmo se adequava ao presente conjunto de dados ou se existia um ϵ cujas métricas

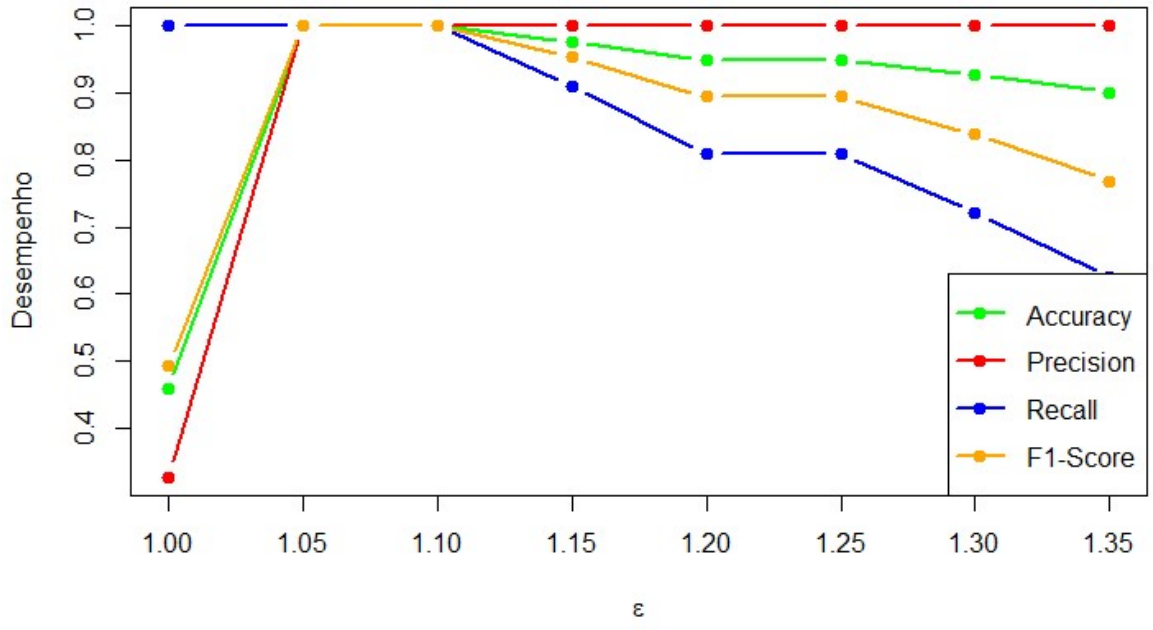

Figura 5.3: Métricas para o *target* 1, Chicago1, considerando $h = 480$, $k = 10$ e variando ϵ.

alcançassem valores mais elevados. Note-se que ϵ é o peso atribuído à média dos *avgRTT*, tal como definido na equação 5.1.

Optou-se por variar o ϵ entre os valores 1 e 1.35, sendo que o valor de 1 significa que o limiar em questão é igual ao valor da média dos *RTTs* passados. O valor de 1.35 significa que a observação só será classificada como anómala se exceder em 35% a média dos *RTTs* passados.

Considerando a Figura 5.3, referente ao *target* Chicago1, consegue-se perceber que os melhores valores para a variável ϵ estão entre 1.05 e 1.10. Para estes valores todas as métricas, *Accuracy*, *Precision*, *Recall* e *F1-Score*, tomam um valor muito próximo de 1. O mesmo sucede com a Figura 5.4, referente ao *target* Frankfurt1. No entanto, a gama de melhores valores já contem o ϵ de 1.20. Recorde-se que este ϵ foi o considerado adequado pela heurística de Salvador e Nogueira.

No *target* 4, Londres, uma alteração no valor de ϵ não altera o desempenho do mesmo. Este *target* evidencia desempenho elevado para qualquer valor de ϵ entre 1.05 e 1.35. Neste caso, também se pode utilizar como base o valor 1.2 para a variável de ϵ.

No que diz respeito ao *target* 3, de Hong Kong, apresentado na Figura 5.5, percebe-se que o único valor de ϵ cujas métricas são simultaneamente elevadas é 1.05. No entanto, para esse valor de ϵ apenas cerca de 80% das observações realmente anómalas são consideradas como tal.

A partir das Figuras 5.6 e 5.7, referentes ao *target* Hong Kong e à *probe* LA2, conclui-se que o *Recall* toma um valor em torno de 0.8 porque existe uma grande dificuldade por parte das *probes* em detetar o *relay* de LA1. Note-se que a Figura 5.7 diz respeito à classificação do tráfego após a votação. No

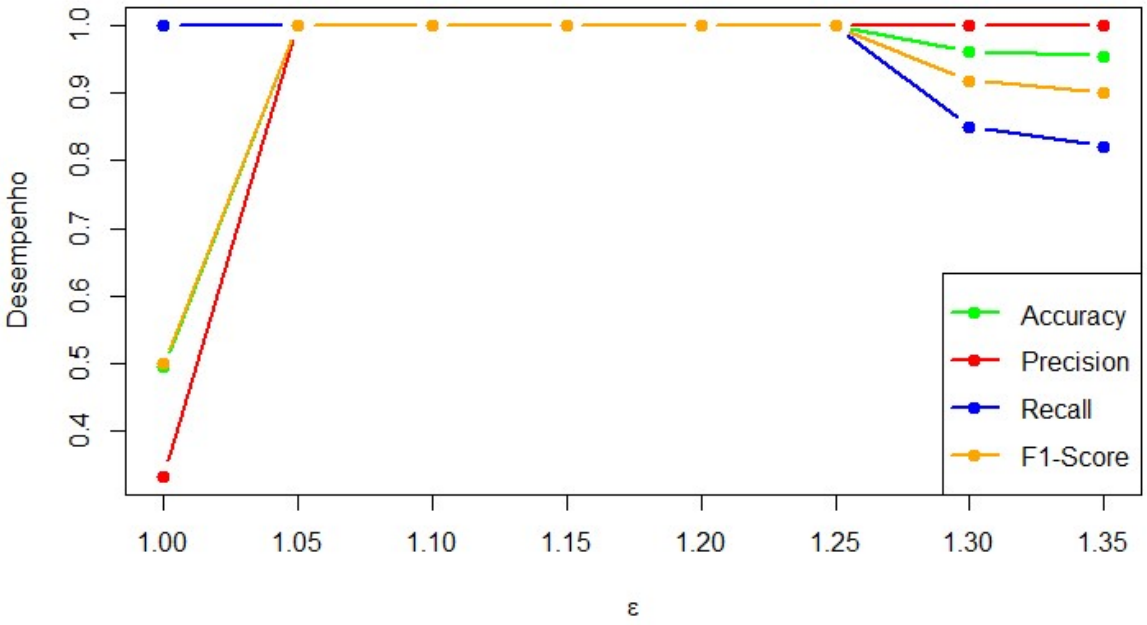

Figura 5.4: Métricas para o *target* 2, Frankfurt1, considerando $h = 480$, $k = 10$ e variando ϵ.

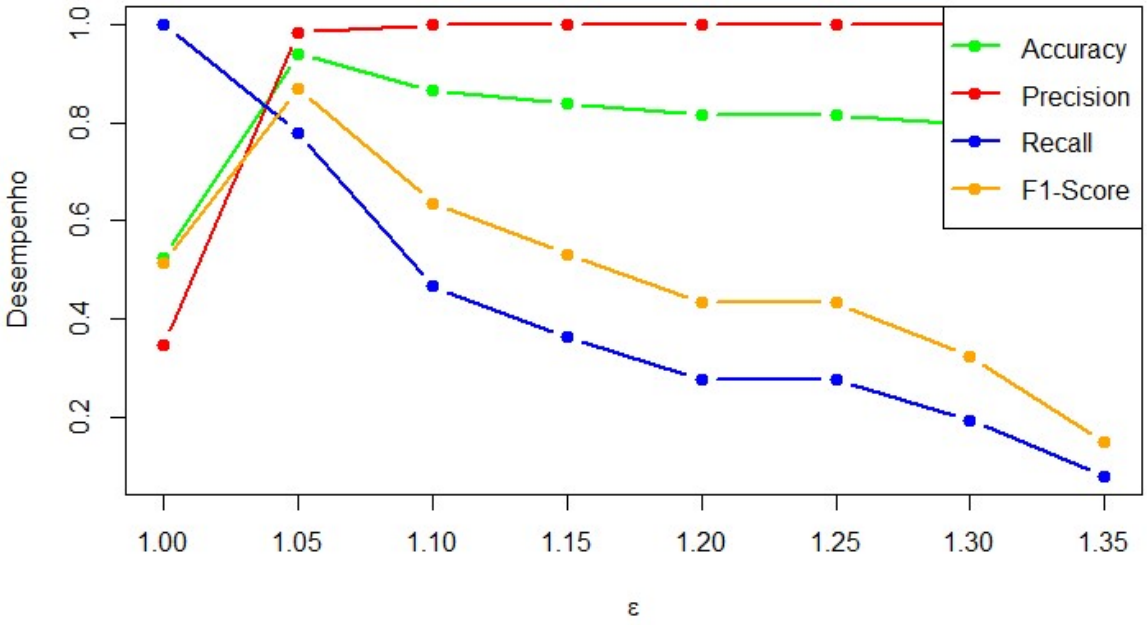

Figura 5.5: Métricas para o *target* 3, Hong Kong, considerando $h = 480$, $k = 10$ e variando ϵ.

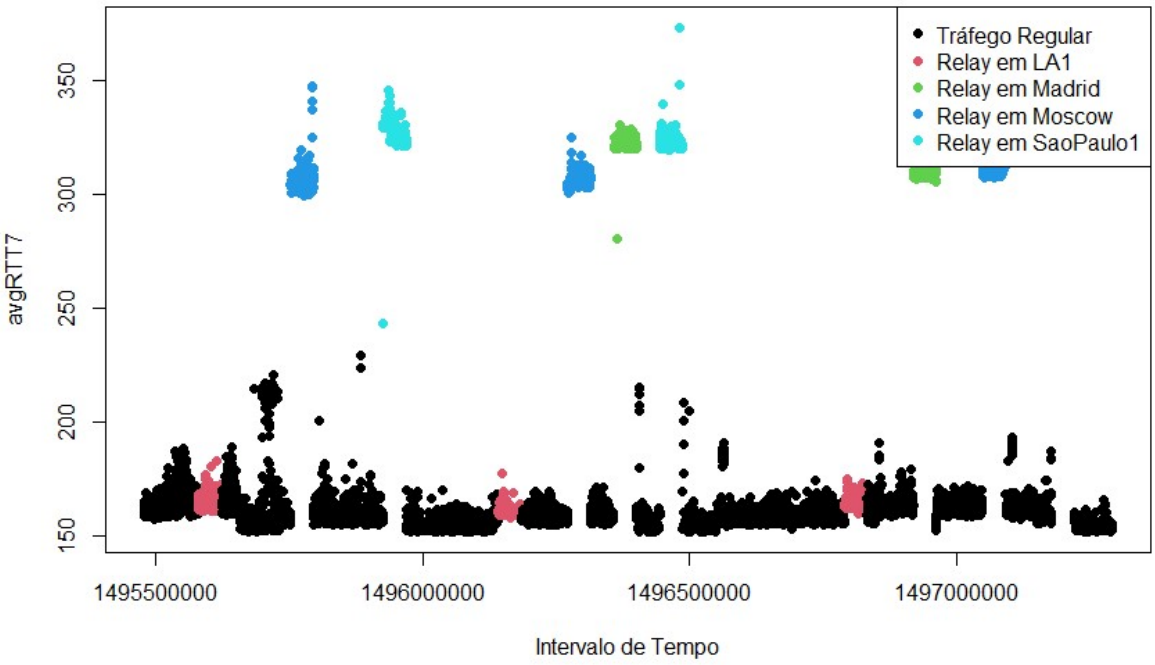

Figura 5.6: Visualização dos *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 7, LA2.

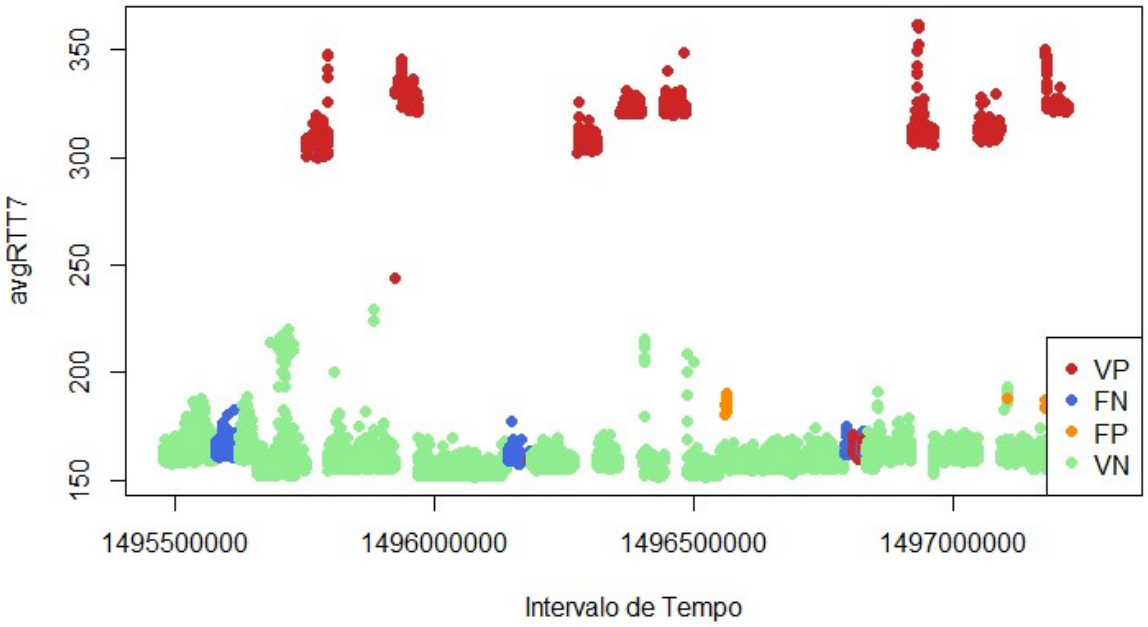

Figura 5.7: Visualização dos *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 7, LA2, após a classificação pela heurística de Salvador e Nogueira com $\epsilon = 1.05$.

entanto, esta *probe* individualmente também não consegue detetar LA1 como *relay* e o mesmo sucede com muitas outras *probes*. Este acontecimento é justificado pelo facto de as observações regulares

terem um *avgRTT* semelhante às observações com ataques em LA1.

De forma a concluir, após analisar o valor de ϵ para todos os *targets*, neste relatório usa-se sempre $\epsilon = 1.05$, independentemente do *target*. Observou-se que é o único valor comum a todos os *targets* que conduz a métricas de desempenho elevadas.

5.3 Estudo da Variação do Comprimento da Janela Deslizante, h

A variável h diz respeito à quantidade de observações passadas que são usadas para calcular a média dos *avgRTT*, que posteriormente é comparada com o *RTT* atual tendo em conta o parâmetro ϵ. O valor estipulado segundo esta heurística é $h = 480$, o que corresponde a uma janela de 480 observações. Pretende-se incluir na janela deslizante apenas observações regulares. À medida que vão surgindo novas observações o algoritmo utiliza sempre as últimas h observações que considerou regulares. Decidiu-se estudar este valor, aumentando e diminuindo a quantidade de observações. Optou-se por escolher os valores 300 e 600, pois são valores que não afetam significativamente o desempenho do algoritmo.

Decidiu-se também testar ainda para um quarto caso em que não existiam janelas deslizantes. Para este último caso, todas as observações passadas que o algoritmo considera regulares são utilizadas para calcular a média do *avgRTT*.

Em relação aos *targets* Chicago1, Frankfurt1 e Londres verificou-se que a escolha do h não interfere em nada com as métricas. Todas as métricas têm o valor 1, ou seja, todas as anomalias são corretamente detetadas. Com uma *Precision* e um *Recall* a tomar o valor 1 não existem nem Falsos Positivos nem Falsos Negativos, o que torna este algoritmo bastante apelativo para o conjunto de dados em estudo.

No que diz respeito ao *target* Hong Kong é melhor quanto menos forem as observações passadas. Quando não se utilizam janelas deslizantes observam-se 9 285 observações regulares. No entanto, existem apenas 8 389 observações regulares na classe verdadeira do conjunto de dados, após a remoção das observações superiores a 600. O *target* de Hong Kong deteta um número considerável de Falsos Negativos e por este motivo o *Recall* toma valores significativamente baixos, aproximadamente 0.67, quando não se utilizam janelas deslizantes. Note-se que utilizando janelas deslizantes o *Recall* é em torno de 0.78, um valor significativamente mais elevado.

Após se analisar cada *target*, decidiu-se optar pelo valor de h considerado por Salvador e Nogueira em [6], ou seja $h = 480$. Não existe necessidade de se optar por outro valor, visto que o mais promissor é efetivamente o valor adoptado por esta heurística de Salvador e Nogueira. Apesar de se considerar 480 como o melhor valor, nos capítulos seguintes faz-se uma análise relativamente aos restantes parâmetros deste algoritmo, para tentar melhorar o desempenho do algoritmo.

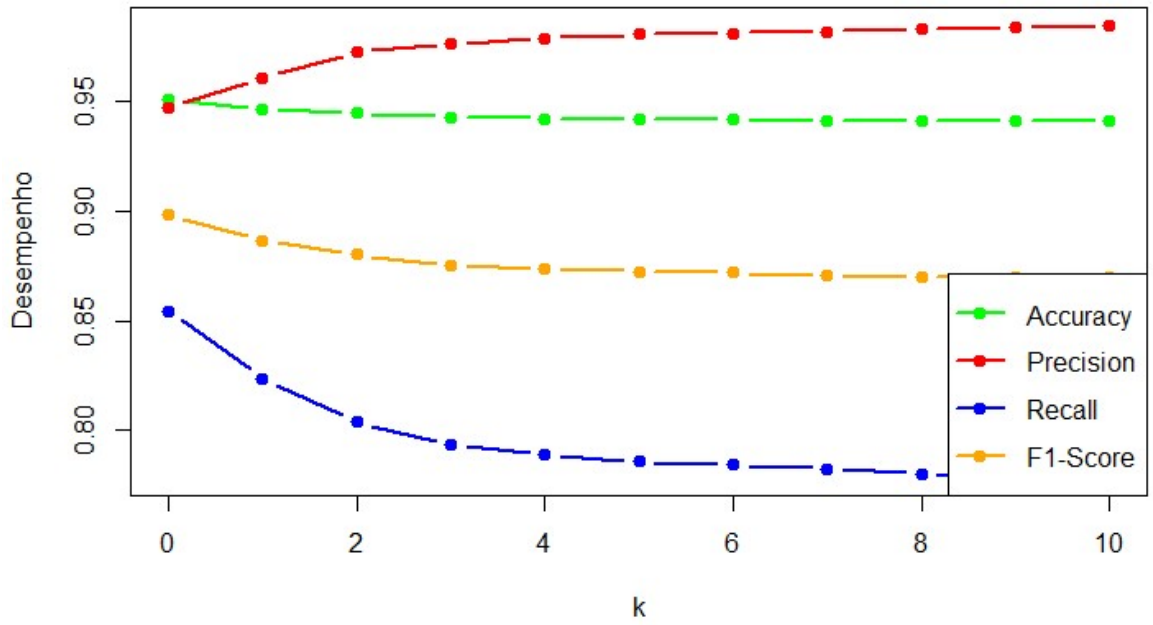

Figura 5.8: Métricas para o *target* 3, Hong Kong, considerando $\epsilon = 1.05$, $h = 480$ e variando k.

5.4 Estudo da Variação do Número de Observações Consecutivas Anómalas, k

O parâmetro k diz respeito à quantidade de observações anómalas consecutivas necessárias para o algoritmo realmente classificar a observação como anómala. Variou-se o k entre 0 e 10, sendo que 0 significa que não existe passado associado. Isto significa que assim que se considera uma observação como anómala, o algoritmo classifica a mesma como anómala, não interessando se as anteriores ou seguintes são regulares ou não. Um $k = 0$ significa que, se por um instante, o tráfego for ligeiramente mais lento, pode ser logo classificado como anomalia.

Quando se observam as *probes* como um todo, não é prudente tirar conclusões quando o $k = 0$.

Com um $k = 0$ tem-se métricas a variar entre 0.9990 e 1, para o *target* de Chicago1, sendo que após $k = 1$ todas as métricas tomam o valor 1. No que diz respeito ao *target* de Frankfurt1, algo semelhante acontece. A partir do valor de $k = 3$ todas as métricas tomam o valor 1. Relativamente ao *target* de Londres, as métricas são sempre 1 independentemente do valor de k.

Quanto ao *target* mais problemático, Hong Kong, verifica-se na Figura 5.8 que o *Recall* vai diminuindo gradualmente à medida que o k aumenta. Como o *avgRTT* regular é bastante semelhante ao *avgRTT* anómalo existe uma grande dificuldade em distinguir o tráfego. Sendo assim, a probabilidade de o algoritmo classificar k observações consecutivas como anómalas vai diminuindo à medida que k aumenta. Anteriormente, verificou-se que a partir de $k = 3$ os restantes *targets* tomavam valores de

1 para todas as métricas. Posto isto, e não havendo uma diferença expressiva entre $k = 3$ e $k = 10$, decidiu-se manter $k = 10$ para todos os *targets*, tal como sugerido em [6].

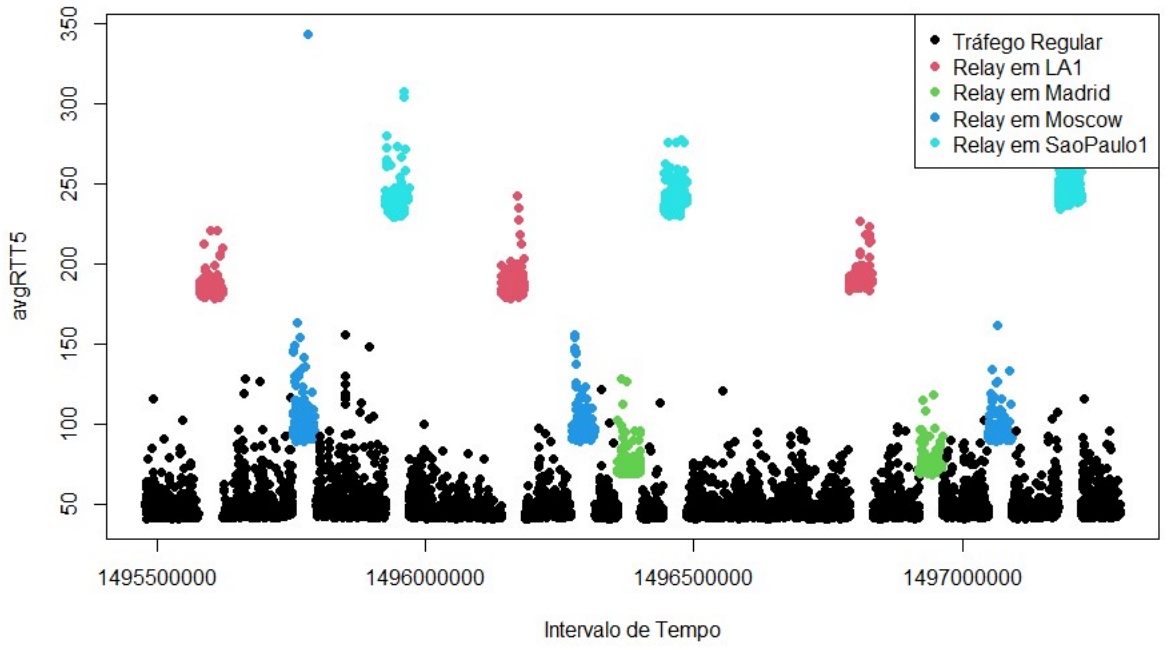

Figura 5.9: Observações para o *target* 4, Londres, e *probe* 5, Islândia.

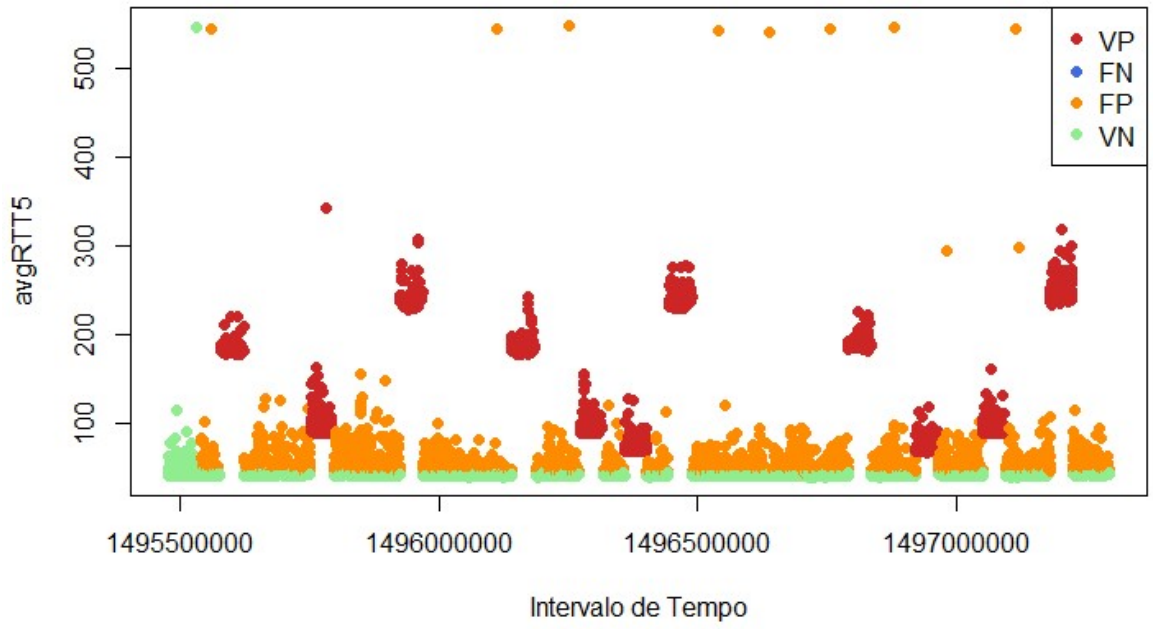

Figura 5.10: Observações para o *target* 4, Londres, e *probe* 5, Islândia, após a utilização da heurística com $k = 0$.

Analisando cada *target* individualmente verifica-se utilizar $k = 0$ não é uma boa opção tendo em

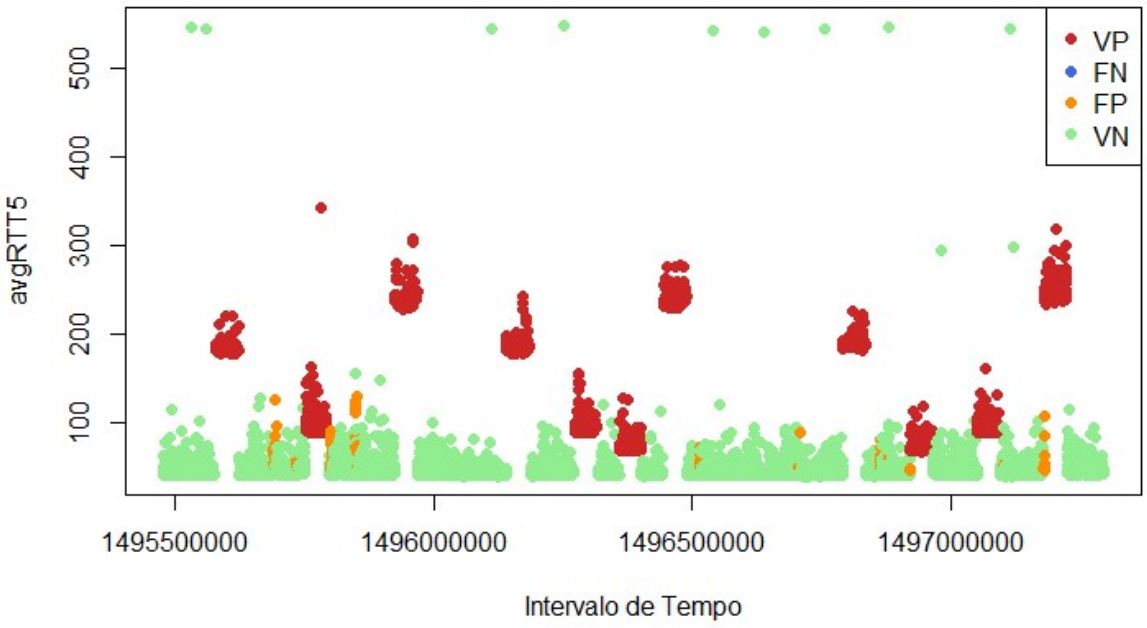

Figura 5.11: Observações para o *target* 4, Londres, e *probe* 5, Islândia, após a utilização da heurística com $k = 10$.

conta os restantes *targets*, como se pode ver nas Figuras 5.9, 5.10 e 5.11. A Figura 5.9 diz respeito à classificação real do tráfego para o *target* de Londres e *probe* da Islândia, a Figura 5.10 tem a classificação do respetivo tráfego com $k = 0$ e a Figura 5.11 tem a mesma classificação mas considerando $k = 10$. Neste exemplo é perceptível a necessidade de utilizar um valor de k superior a 0, pois o número de Falsos Positivos diminuí substancialmente.

5.5 Estudo da Variação da Percentagem de *Probes* Necessárias para uma Observação ser Classificada Anómala, γ

Posteriormente optou-se por variar a percentagem de *probes* que têm de assinalar uma anomalia exatamente no mesmo instante para se decidir que está a ocorrer um ataque. Segundo a heurística de Salvador e Nogueira é necessário que 50% ou mais *probes* detetem observações anómalas naquele instante para decidir se está a ocorrer um ataque, ou seja, é necessário $\gamma = 50\%$. Tendo em conta que existem certas *probes* que são bastante irregulares optou-se por variar entre $\gamma = 10\%$, em que apenas é necessário que 2 *probes* consideram que o *avgRTT* naquele instante é anómalo, e $\gamma = 70\%$, em que são necessárias aproximadamente 9 *probes* para classificar o instante como anómalo. Optou-se por considerar valores de γ entre 10% e 70%, visto que para certos casos se torna bastante complicado que todas as *probes* classifiquem o tráfego exatamente da mesma maneira. Por exemplo, para o *target*

de Hong Kong e *probe* de SaoPaulo2 é quase impossível que não existam Falsos Negativos, tal como é perceptível na Figura 5.12.

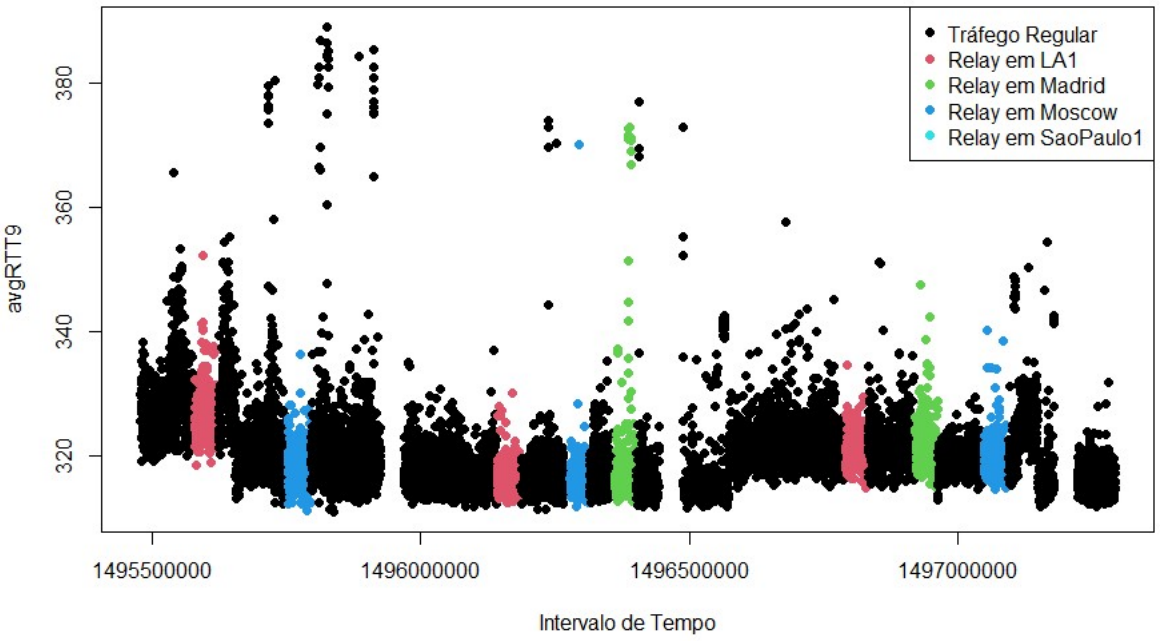

Figura 5.12: Visualização dos *avgRTT* do tráfego entre o *target* 3, Hong Kong, e *probe* 9, SaoPaulo2.

Quanto às Figuras 5.13, 5.14 e 5.16, correspondentes aos *targets* Chicago1, Frankfurt1 e Londres, observa-se que, a partir do momento em que $\gamma = 30\%$ as métricas tomam valores em torno de 1. Note-se que $\gamma = 30\%$ são 4 *probes*.

Através da Figura 5.15, *target* de Hong Kong, percebe-se que os Falsos Positivos têm um decrescimento significativo à medida que há a necessidade de mais *probes* detetarem uma anomalia no mesmo instante. À medida que existe uma quantidade de *probes* a votar bastante superior a $\gamma = 40\%$, ou seja, 5 *probes*, é imediata a percepção que o número de Verdadeiros Positivos diminui e por sua vez essas observações passam a ser detetadas como Falsos Negativos.

No caso de Hong Kong, há muitas *probes* que não conseguem detetar os *avgRTT* anómalos. Com os valores de *avgRTT* medidos pelas *probes* a monitorizar o *target* Hong Kong, coloca-se a hipótese de *probes* correspondentes a *RTT* inferiores serem mais aptos para detetar os desvios de *RTT*. Investiga-se esta hipótese no Capitulo 5.6.

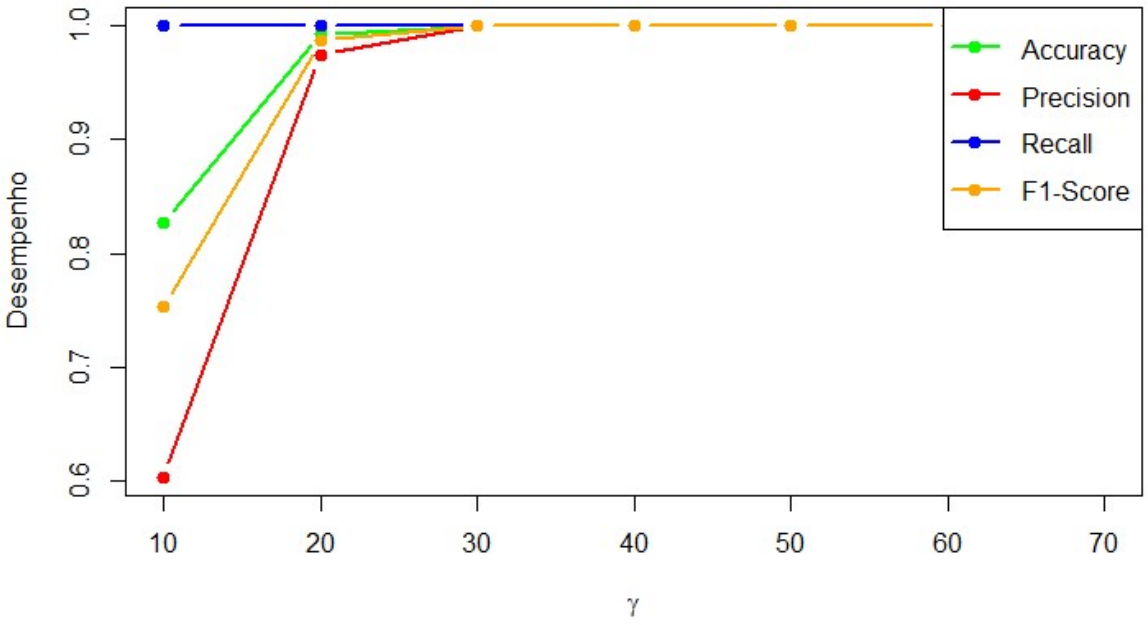

Figura 5.13: Métricas para o *target* 1, Chicago1, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando γ.

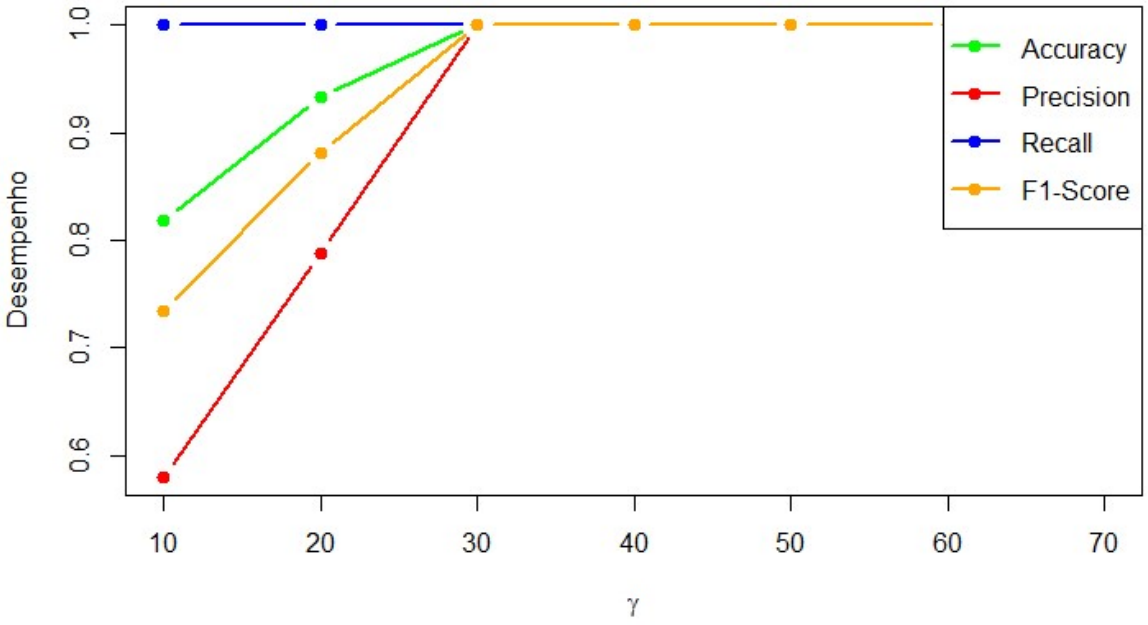

Figura 5.14: Métricas para o *target* 2, Frankfurt1, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando γ.

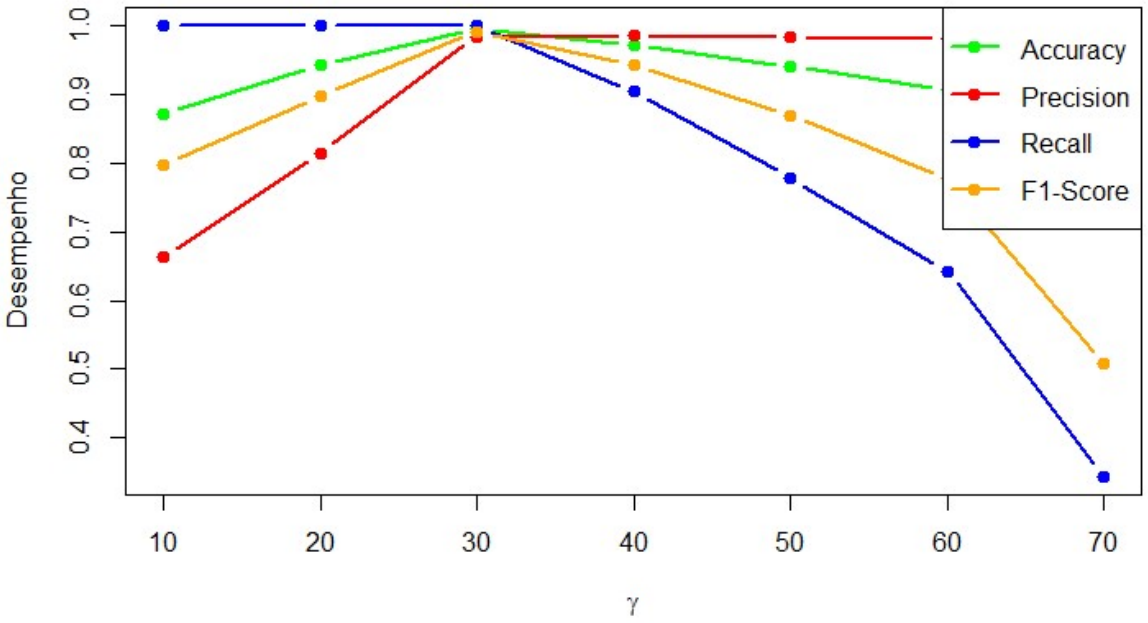

Figura 5.15: Métricas para o *target* 3, Hong Kong, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando γ.

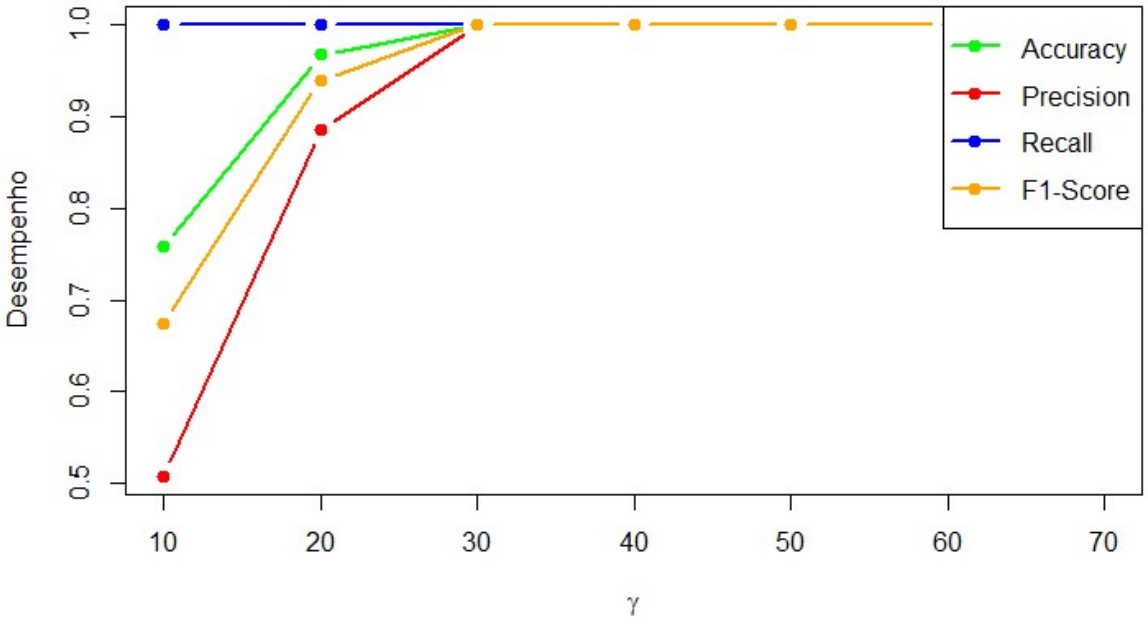

Figura 5.16: Métricas para o *target* 4, Londres, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando γ.

5.6 Estudo da Variação do Número de Melhores *Probes*, τ

Nesta secção verifica-se se é viável utilizarem-se apenas certas *probes*, tendo em conta um *avgRTT* médio mais baixo. Como os resultados obtidos no *target* de Hong Kong não foram os ideais, apurou-se a possibilidade de utilizar um número total de *probes* menor do que as 12 disponíveis, selecionando as *probes* a considerar com base na ordenação dos valores de *avgRTT*. Este estudo procura explorar a hipótese de que *probes* que apresentam valores de *avgRTT* mais baixos na monitorização de um certo *target* possam ter maior capacidade de detetar ataques.

Para ordenar as *probes* que monitorizando um dado *target*, calculou-se a média do *avgRTT* para as primeiras 480 observações. Sendo que na secção 5.3 se considerou que o h usado nas seguintes secções seria 480, nesta secção também é esse valor de referência que se utiliza. Note-se que sempre se considerou que as primeiras h observações eram regulares para ter dados iniciais com que trabalhar.

Começando pelo *target* de Chicago1, cujos valores se encontram na Tabela 5.1, facilmente se percebe que a *probe* mais próxima do *target* Chicago1 é a de Chicago2. Tendo em conta a Figura 5.17, referente ao *target* Chicago1, percebe-se que apenas 1 *probe* não é suficiente para classificar corretamente todas as observações. Este fenómeno deve-se ao facto de qualquer observação que se afaste ligeiramente do valor 0.12 é anómala, pois o ϵ é 1.05 portanto valores superiores a 0.126 já são anómalos, se num período igual a 10 observações todas tomarem valores superiores a 0.126. Note-se que na Figura 5.17 um valor no eixo dos x corresponde a só se considerar essa quantidade de *probes*, selecionando as *probes* de acordo com a ordenação apresentada na Tabela 5.1 e descartando as restantes *probes*. Note-se que continua a ser utilizada a regra de apenas ser necessário que 50% dessas *probes* detetem que a observação como anómala para decidir que está a ocorrer um ataque. Por exemplo, se o valor utilizado for 6, quer dizer que se estão a utilizar as *probes* 2, 7, 4, 1, 12 e 8 e que apenas 3 dessas *probes* necessitam de detetar um *timestamp* como anómalo para o mesmo ser classificado como tal. No entanto, quando se está perante apenas 1 *probe* o resultado das métricas é simplesmente o resultados das métricas para essa mesma *probe*. Quando se observam as 2 melhores *probes* para um instante ser considerado anómalo é suficiente que uma das *probes* o classifique como anómalo.

Analisando a Tabela 5.2 do *target* de Frankfurt1 é possível perceber que a *probe* de Frankfurt2 é a mais próxima do *target*. Neste caso, verifica-se que apenas com esta *probe* os resultados são bastante elevados, como se comprova através das métricas apresentadas na Figura 5.18. Concluí-se que esta *probe* é bastante estável e o tráfego regular não altera muito. No entanto, se a opção fosse utilizar apenas este *probe*, qualquer mudança de nível seria logo considerada um ataque. Desta forma e apesar de o tráfego ser sempre classificado corretamente apenas com uma *probe*, não é uma boa prática.

A informação relativa ao *target* de Hong Kong encontra-se na Tabela 5.3 e na Figura 5.19. A tabela já contém valores mais elevados. No entanto, a figura 5.19 não aparenta ter resultados melhores quando

Tabela 5.1: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Chicago1, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
2 - Chicago2	0.12
7 - LA2	66.54
4 - Frankfurt2	108.28
1 - Amesterdam	111.23
12 - Sweden	121.94
8 - Milan	122.04
5 - Iceland	139.83
9 - SaoPaulo2	141.63
3 - VdM	157.65
6 - Israel	166.19
10 - Johannesburg1	254.46
11 - Johannesburg2	263.64

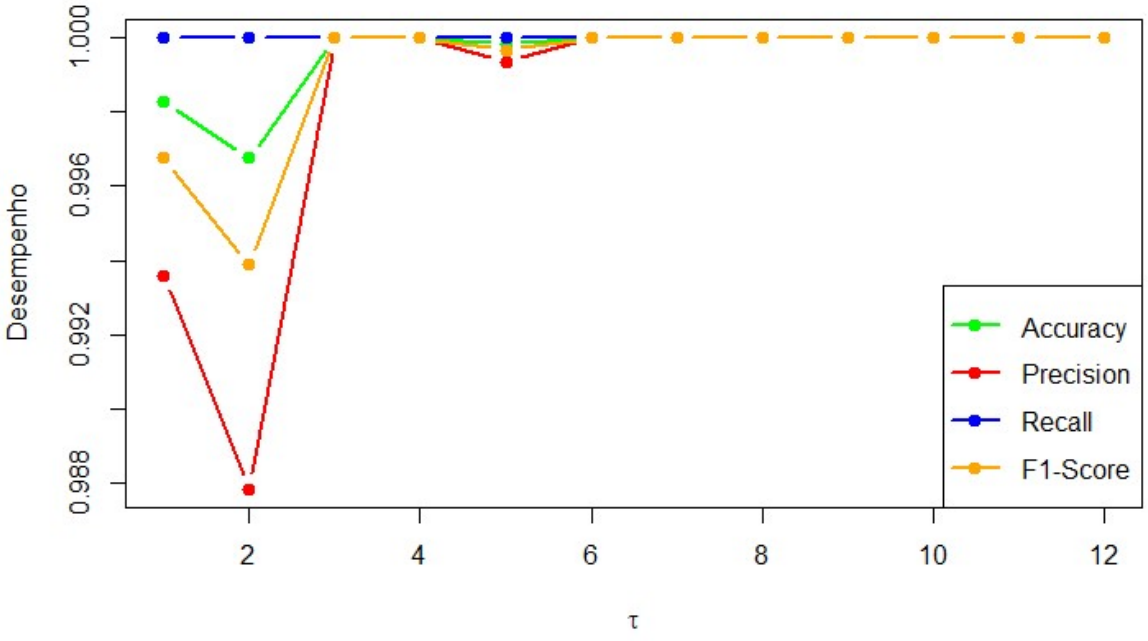

Figura 5.17: Métricas para o *target* 1, Chicago, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando τ.

Tabela 5.2: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Frankfurt1, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
4 - Frankfurt2	0.10
1 - Amesterdam	9.54
8 - Milan	18.25
12 - Sweden	23.20
5 - Iceland	60.27
6 - Israel	63.68
2 - Chicago2	108.28
7 - LA2	154.69
10 - Johannesburg1	186.08
11 - Johannesburg2	186.65
9 - SaoPaulo2	213.37
3 - VdM	225.91

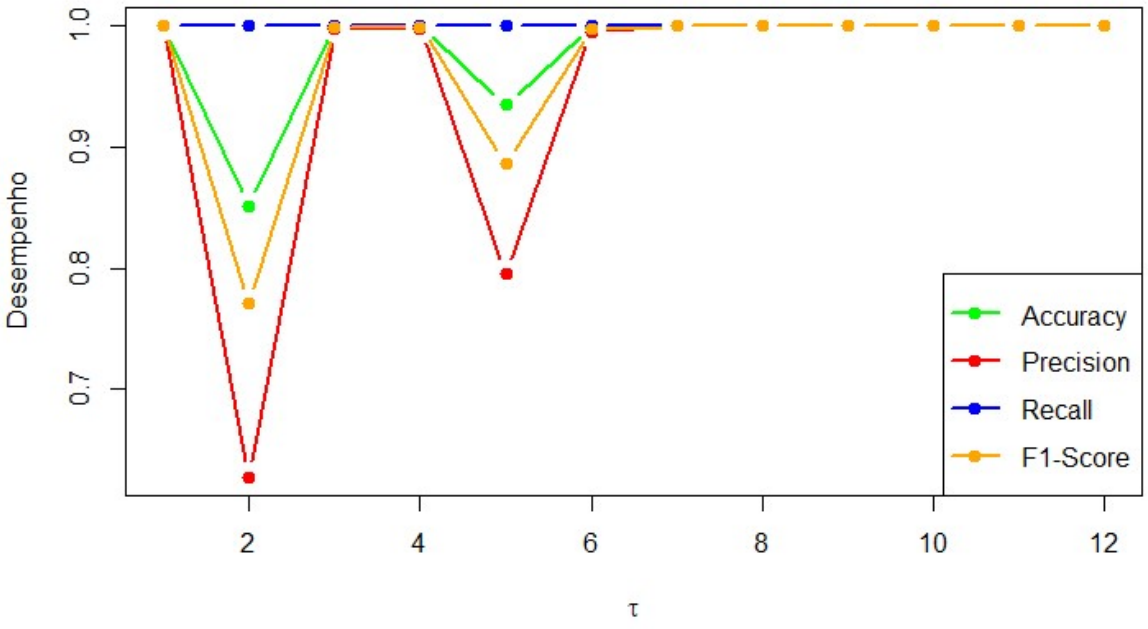

Figura 5.18: Métricas para o *target* 2, Frankfurt1, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando τ.

Tabela 5.3: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Hong Kong, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
7 - LA2	163.17
2 - Chicago2	192.82
8 - Milan	268.44
6 - Israel	277.78
12 - Sweden	291.62
4 - Frankfurt2	318.55
1 - Amesterdam	324.18
5 - Iceland	324.53
9 - SaoPaulo2	326.73
3 - VdM	349.68
10 - Johannesburg1	448.21
11 - Johannesburg2	448.34

Tabela 5.4: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Londres, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
1 - Amesterdam	9.80
4 - Frankfurt2	15.71
8 - Milan	24.96
12 - Sweden	31.26
5 - Iceland	46.84
6 - Israel	77.33
2 - Chicago2	90.59
7 - LA2	142.29
11 - Johannesburg1	162.86
11 - Johannesburg2	190.34
9 - SaoPaulo2	199.63
3 - VdM	220.40

se está perante apenas 1 *probe*. Os valores em que as métricas têm valores mais elevados são em torno de 7 e 9, ou seja quando é necessário que 4 a 5 *probes* detetem os *avgRTTs* anómalos para decidir que há um ataque, o mesmo resultado obtido na secção 5.5. Sendo assim, não se verifica a necessidade de utilizar este método comparativamente ao método referido na secção 5.5.

Quanto ao *target* 4, de Londres, não há muito a acrescentar. Mais uma vez na Figura 5.20 mostra-se que utilizar apenas 1 *probe* com *RTT* muito baixo, como se observa na Tabela 5.4 não é fiável. Com apenas 1 *probe* é bastante complicado separar o tráfego regular do tráfego anómalo, visto que qualquer aumento de *RTT* já é classificado como anomalia. Portanto, concluí-se que o melhor é utilizar várias *probes*. Para os *targets* 1, 2 e 4, respetivamente Chicago1, Frankfurt1 e Londres, bastariam as 3 melhores *probes*. Para o *target* 3, Hong Kong, o ideal é utilizar as 8 melhores *probes*, mantendo o valor de ϵ em 1.05 tal como mencionado na secção 5.2. Esta abordagem não é unânime para todos os *targets* portanto, tal como mencionado anteriormente optou-se pela abordagem referido em 5.5.

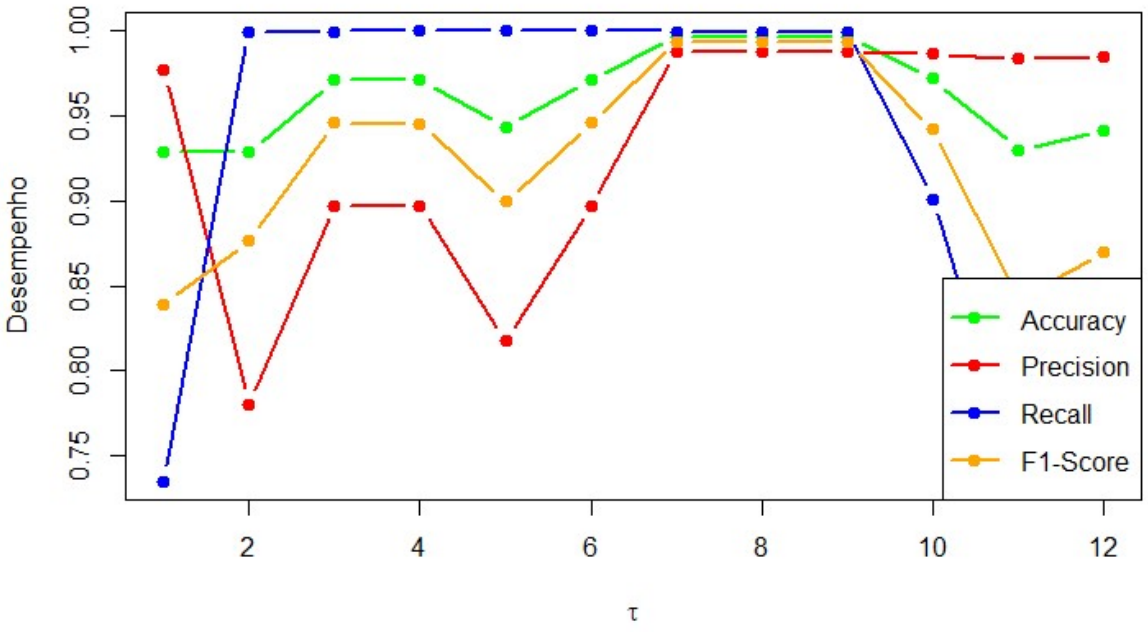

Figura 5.19: Métricas para o *target* 3, Hong Kong, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando τ.

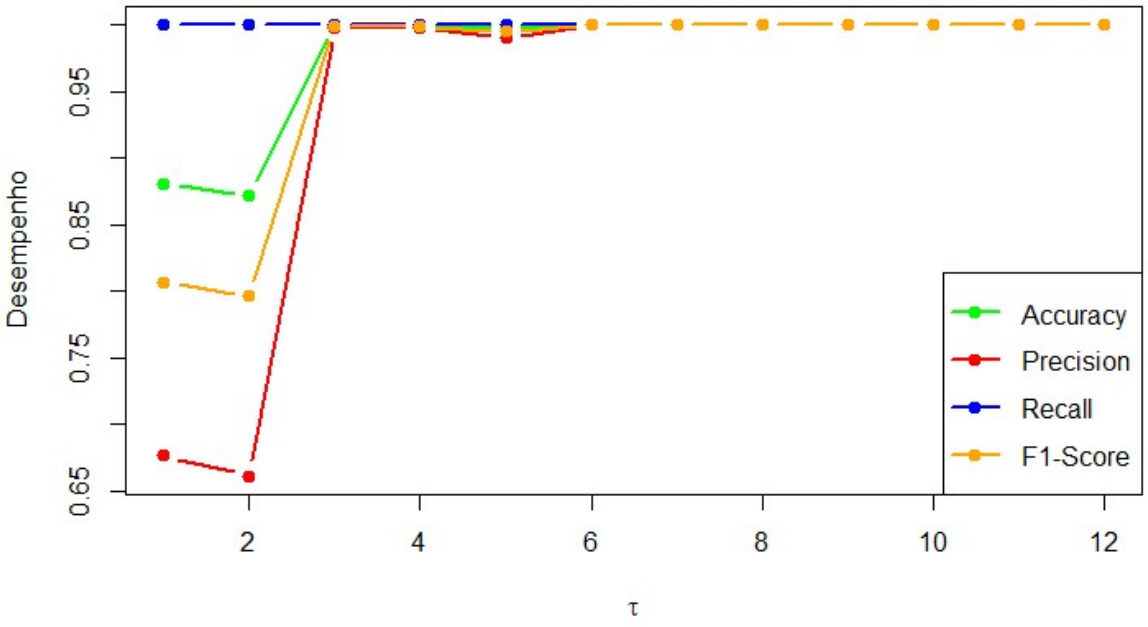

Figura 5.20: Métricas para o *target* 4, Londres, considerando $\epsilon = 1.05$, $h = 480$, $k = 10$ e variando τ.

Concluí-se também que *probes* associadas a *avgRTTs* mais baixos conduzem a melhores resultados quando combinados entre si, pois com $\epsilon = 1.05$ é muito fácil uma observação regular ser erradamente classificada como anómala. No entanto, quando se combinam este efeito é atenuado.

5.7 Conclusões

Após a analise de todas as componentes deste algoritmo, percebe-se que ainda não se encontrou algo que consiga detetar a 100% as *relays* no que diz respeito ao *target* de Hong Kong. No entanto, para os 4 *targets* alcançaram-se as métricas presentes na Figura 5.21, com os valores de $h = 480$, $k = 10$, $\epsilon = 1.05$ e $\gamma = 30\%$, tal como concluído na secção 5.5. Optou-se por usar a metodologia da secção 5.5 em vez da secção 5.6 por ser mais similar à própria heurística e por ser mais unânime entre todos os *targets*.

Verifica-se que este algoritmo deteta corretamente a maioria do tráfego, apesar de não ser tão robusto como o pretendido, pois considera-se $\gamma = 30\%$ e este valor com outro conjunto de dados pode conduzir a resultados muito diferentes, ou seja, é bastante sensível a pequenas alterações no conjunto de dados.

A Figura 5.22 apresenta as métricas sem a remoção de observações superiores a 600ms.

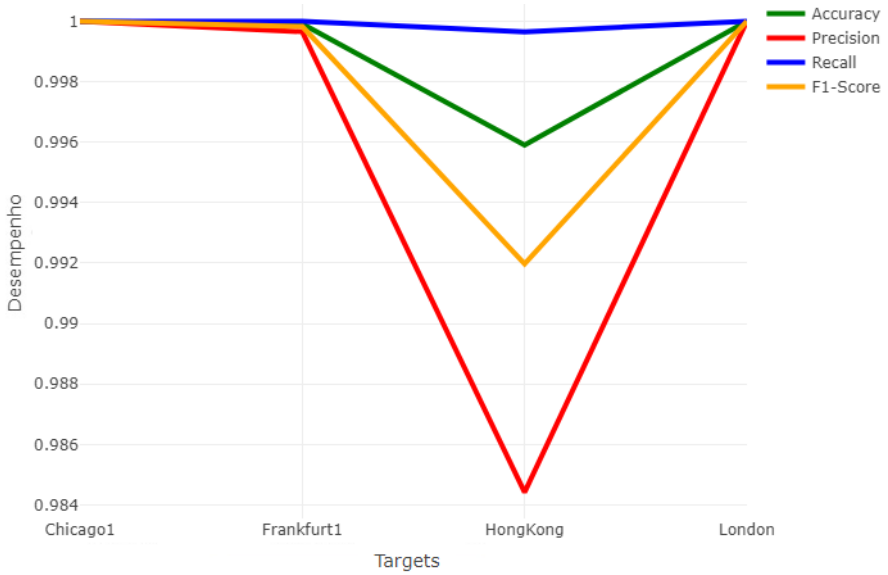

Figura 5.21: Métricas finais para a heurística, excluindo as observações com *avgRTT* superior a 600ms.

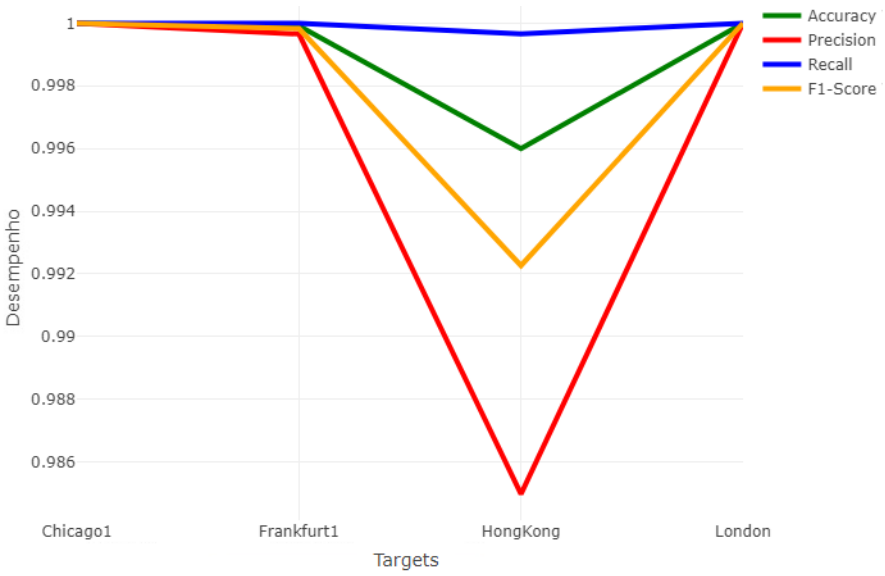

Figura 5.22: Métricas finais para a heurística, incluindo as observações com *avgRTT* superior a 600ms.

6

Método de *Tukey*

O Método de *Tukey* baseia-se na amplitude interquartil (δ) para definir se uma observação é anómala ou não. A amplitude interquartil é definida como a diferença entre o 3º e 1º quartis amostrais:

$$\delta = Q_3 - Q_1. \tag{6.1}$$

Tendo em conta que em geral o tráfego anómalo tem um *avgRTT* superior ao *avgRTT* do tráfego regular, apenas é considerado anómalo o tráfego que exceda $Q_3 + \delta IQR$, sendo que δ é o peso da amplitude interquartil. Sendo assim, o tráfego inferior ao 1º quartil amostral é sempre considerado regular.

Tendo em conta a dependência temporal dos dados, optou-se por considerar uma janela deslizante de h observações passadas classificadas como regulares, à semelhança do que foi proposto pela heurística de Salvador e Nogueira, descrita anteriormente na secção 2.7.1.A e no Capítulo 5. Como a heurística utiliza um valor pré-definido de $h = 480$, também se utiliza esse valor na aplicação do método de *Tukey*, sendo que este valor h corresponde ao número de observações da janela deslizante . Foi avaliado por meio de um método descrito na secção 6.2, para verificar se outro valor podia conduzir a

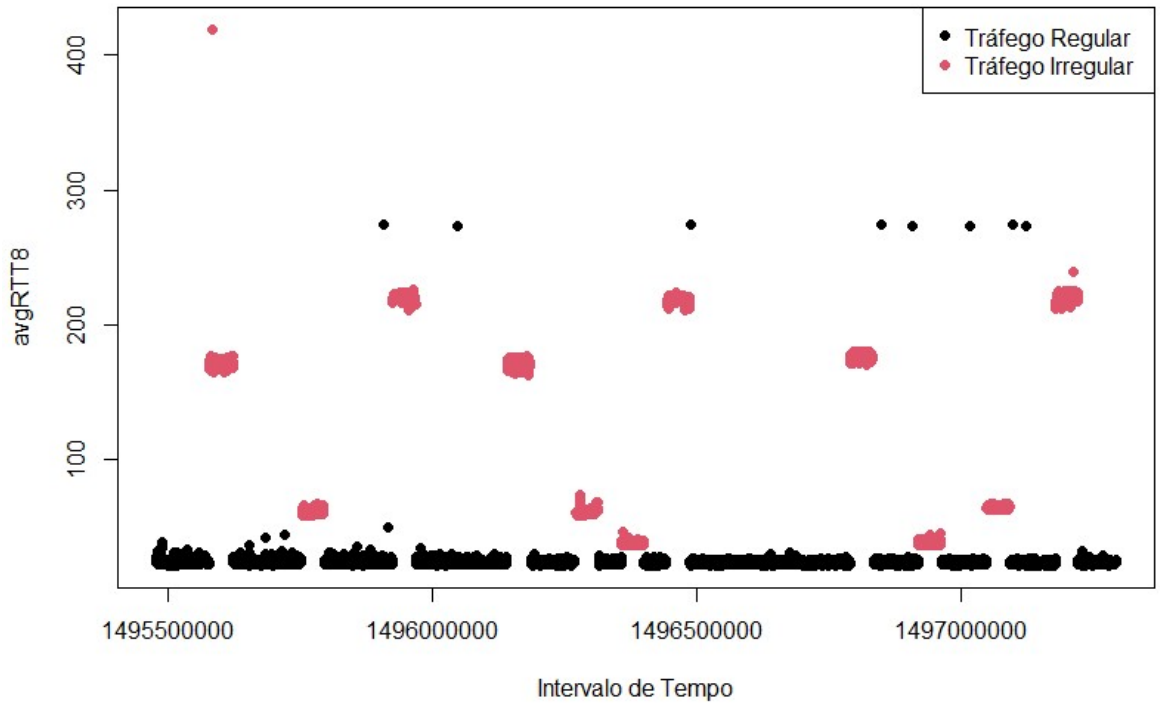

Figura 6.1: Visualização do *avgRTT* do tráfego entre o *target* 4, Londres, e a *probe* 8, Milan, utilizando $k = 10$ e $h = 480$.

melhores resultados.

Tendo em conta que este método considera que todas as observações que excedem o limiar superior estabelecido de acordo com o método de *Tukey* são anómalas, tornou-se necessário e útil considerar que uma observação só é anómala se existir uma sequência de observações anómalas, sendo que a dimensão a considerar para a sequência k é estudada na secção 6.3. No entanto, inicialmente optou-se por fixar o k em 10, visto que, mais uma vez, é o valor sugerido pela heurística proposta por Salvador e Nogueira.

Na Figura 6.1, correspondente ao *target* de Londres e *probe* de Milão, observa-se um exemplo da necessidade da utilização de um $k \neq 0$. Quando o k toma o valor 0, as observações com valores de *avgRTT* em torno de 300ms, que são realmente observações regulares, nunca seriam consideradas regulares. Na figura foi utilizado um $k = 10$ para que estas observações em torno de 300 fossem consideradas regulares e não anómalas.

Apresentam-se em seguida os estudos que avaliam o impacto da escolha de diferentes valores para os parâmetros necessários à aplicação deste método.

6.1 Estudo da Variação do Peso da Amplitude Interquartil, δ

Inicialmente foi feito um estudo no que diz respeito à amplitude interquartil, tal como referido anteriormente. Em 2.7.1.B referiu-se que os pesos, δ, atribuídos à amplitude interquartil na definição dos limiares que são mais utilizados são 1.5, para assinalar possíveis anomalias, e 3, para assinalar anomalias severas. Contudo, são valores gerais e podem não ser apropriados ao conjunto de dados em estudo. Sendo assim, decidiu-se verificar qual seria o valor mais adequado para δ. Optou-se por variar entre 0 e 4, com saltos de 0.5 em 0.5, sendo que 0 significa que todas as observações acima do 3º quartil amostral são consideradas anómalas.

Quanto aos *targets* Chicago1, Frankfurt1 e Londres, qualquer valor cujo peso para amplitude interquartil seja superior a 1 é suficiente para detetar corretamente todas as observações. No entanto, após a análise da Figura 6.2, relativa ao *target* de Hong Kong, percebe-se que os melhores valores para o peso da amplitude interquartil, δ são 1 e 1.5. Sendo que o método de *Tukey*, utilizado em [15] e [16], se baseia no valor 1.5 e que este é efetivamente um dos melhores para todos os *targets*, optou-se por fixar este parâmetro no valor 1.5. Note-se que o *Recall* e a *Precision* são referentes à classe das anomalias.

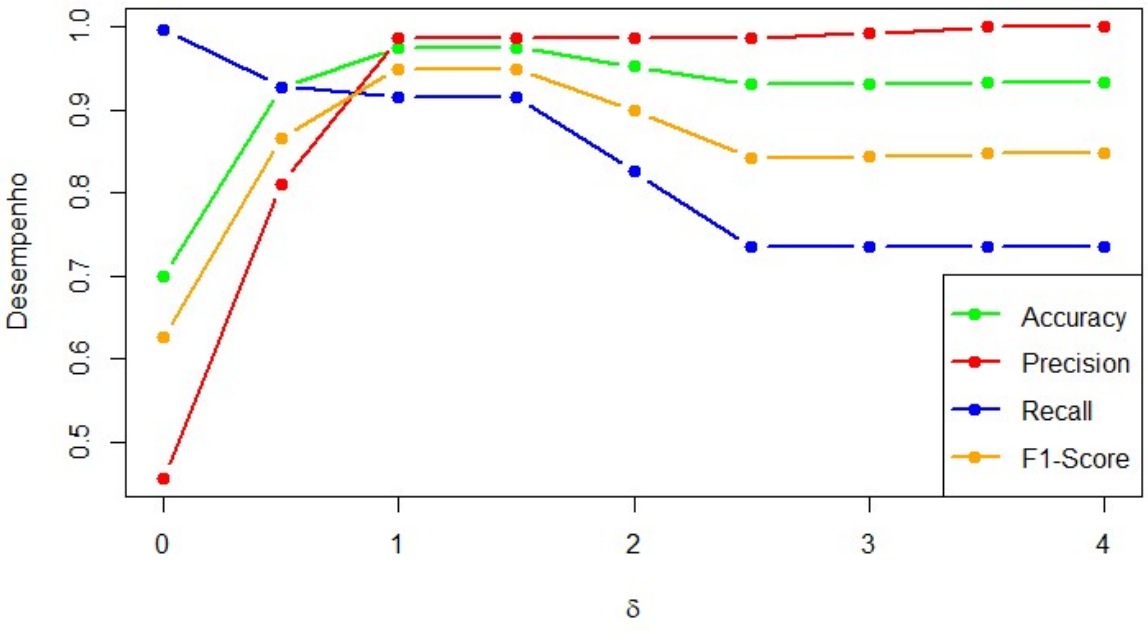

Figura 6.2: Métricas para o *target* 3, Hong Kong, considerando $h = 480$, $k = 10$ e variando δ.

6.2 Estudo da Variação do Comprimento da Janela Deslizante, h

Decidiu-se adaptar o Método de *Tukey* adicionando um h que tem exatamente o mesmo significado que o descrito em 5.3, para a heurística de Salvador e Nogueira, ou seja, descreve o comprimento da janela deslizante. Neste caso, à semelhança do capitulo anterior, consideraram-se 3 valores para o número de observações h nas janelas deslizantes, 300, 480 e 600. Isto significa que, para classificar um instante, é necessário ter como base as últimas h observações consideradas regulares. Tal como aconteceu para a heurística de Salvador e Nogueira, também se considerou um outro valor, sem janelas deslizantes, que tem como base todas as observações anteriores que foram consideradas regulares.

A análise relativamente ao h é bastante simples. Considerando os *targets* de Chicago1 e Londres, não há qualquer variação nas métricas à medida que se varia o comprimento da janela deslizante.

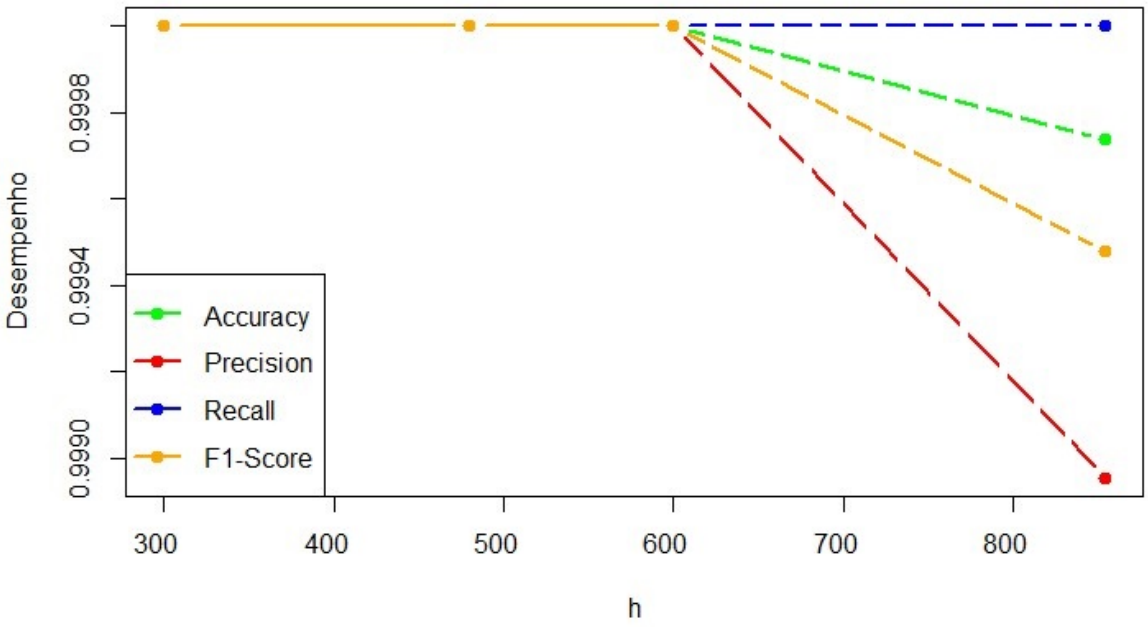

Figura 6.3: Métricas para o *target* 2, Frankfurt1, considerando $k = 10$, $\delta = 1.5$ e variando h.

A Figura 6.3 diz respeito ao tráfego do *target* de Frankfurt1. Através desta figura percebe-se que o algoritmo tem melhor desempenho quando não se utiliza todo o passado, além de que em termos computacionais também se torna mais eficiente quanto menor o comprimento da janela deslizante.

Relativamente ao *target* de Hong Kong, representado na Figura 6.4 verifica-se que os melhores valores de h estão associados a janelas deslizantes com o passado mais recente, não se verificando qualquer alteração entre utilizar 300, 480 ou 600 observações passadas.

Posto isto, decidiu-se manter o valor 480 para uma maior similaridade entre os valores escolhidos para a heurística de Salvador e Nogueira e para o Método de *Tukey*.

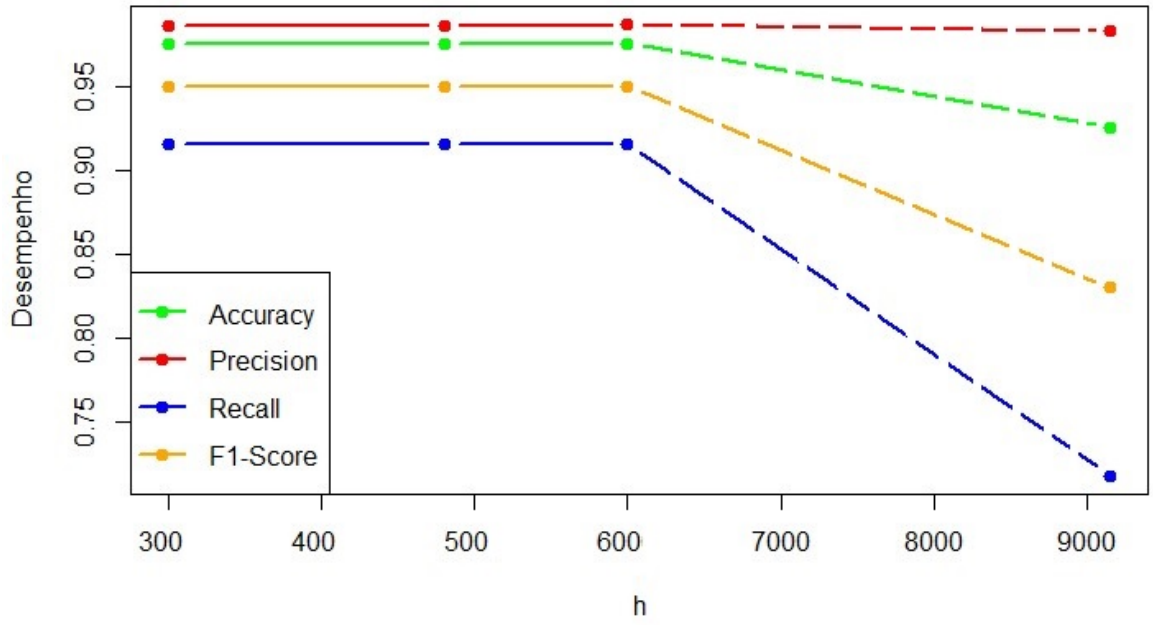

Figura 6.4: Métricas para o *target* 3, Hong Kong, considerando $k = 10$, $\delta = 1.5$ e variando h.

6.3 Estudo da Variação do Número de Observações Consecutivas Anómalas, k

Após a observação da evolução das métricas à medida que o k vai aumentando, percebeu-se que o número de Falsos Positivos diminuía. Ou seja, com $k = 0$, detetaram-se várias observações regulares que foram classificadas como observações anómalas. Note-se que um $k = 0$ significa que uma observação com um *avgRTT* muito elevado já é classificada como anómala, independentemente de as observações anteriores ou posteriores serem regulares. Sabe-se que um ataque para este conjunto de dados é constituído por diversas observações anómalas em instantes consecutivos. Sendo assim, tirou-se vantagem desta informação ser previamente conhecida para melhorar o método e o tornar o melhor possível.

Relativamente aos diferentes *targets* percebe-se que para o *target* de Chicago1 e para o *target* de Londres não há nenhum valor que seja prejudicial para o conjunto de dados. Para o *target* de Frankfurt1, ou seja, para a Figura 6.5 qualquer valor cujo k seja superior a 5 tem todas as métricas a 1, ou seja, deteta corretamente todas as observações do conjunto de dados. Quanto ao *target* de Hong Kong, representado na Figura 6.6, mais uma vez, torna-se difícil que o algoritmo classifique corretamente todas as observações. No entanto quanto maior o k, maior é a *Precision*, ou seja, menor é a quantidade de Falsos Positivos. Sendo assim o k que utilizado em experiências futuras é $k = 10$. É de relembrar que foi exatamente o mesmo que foi selecionado em para a heurística na secção 5.4 e o

mesmo sugerido em [6].

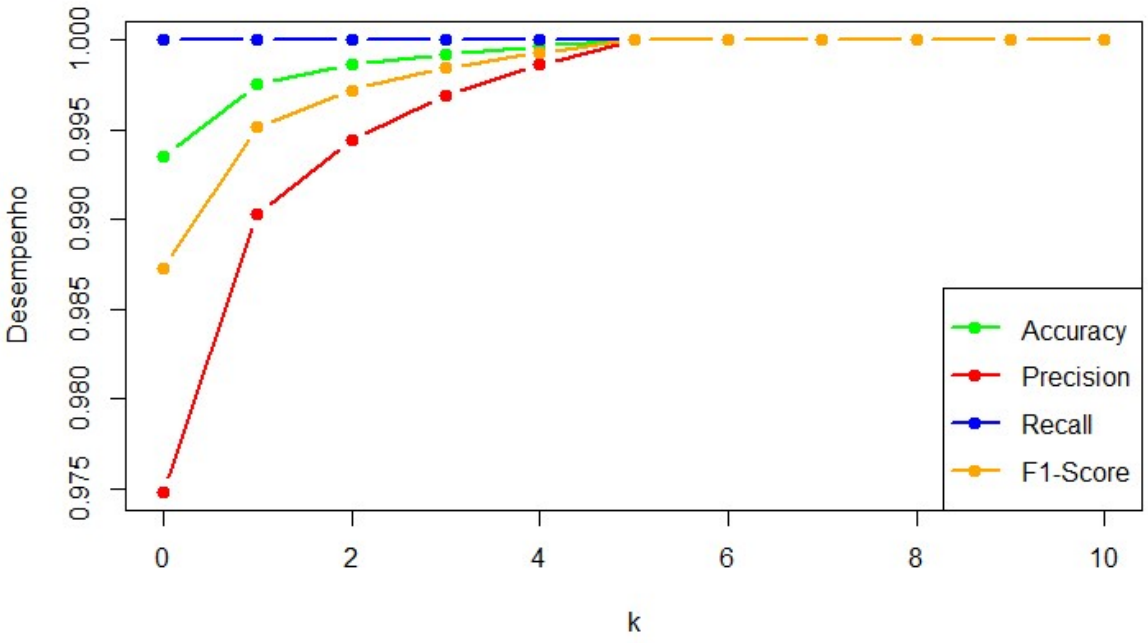

Figura 6.5: Métricas para o *target* 2, Frankfurt1, considerando $h = 480$, $\delta = 1.5$ e variando k.

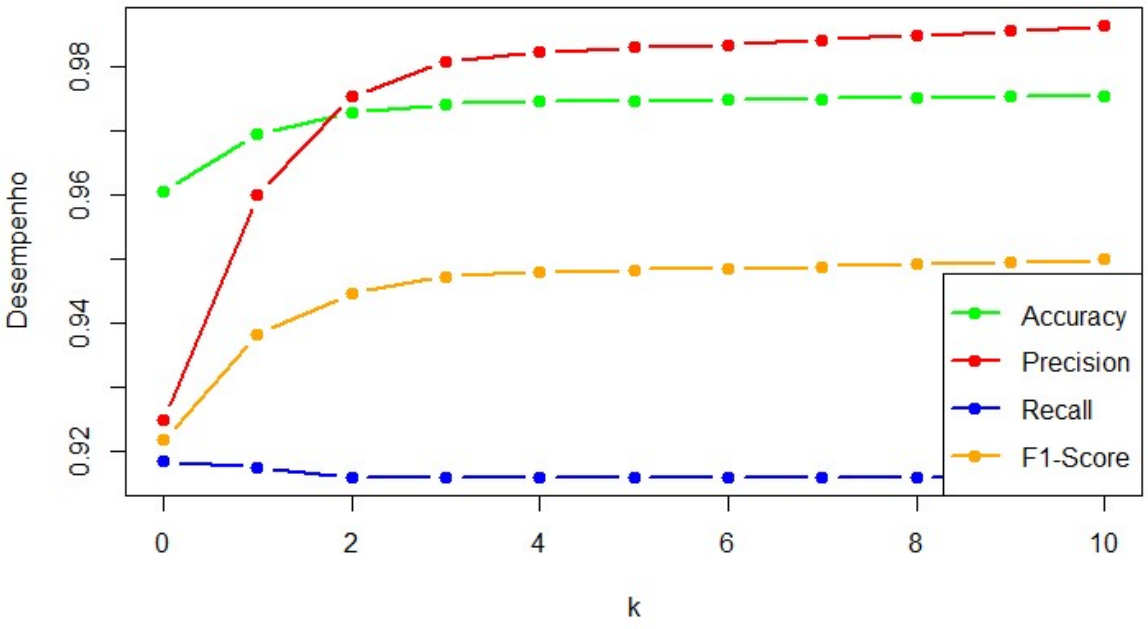

Figura 6.6: Métricas para o *target* 3, Hong Kong, considerando $h = 480$, $\delta = 1.5$ e variando k.

6.4 Estudo da Variação da Percentagem de *Probes* Necessárias para uma Observação ser Classificada Anómala, γ

Em seguida tenta-se verificar se uma alteração na quantidade de *probes* é benéfica ou prejudicial. A mesma análise já foi feita para a heurística proposta por Salvador e Nogueira na secção 5.5. É de recordar que quando se considera apenas 1 *probe* considera-se que apenas é necessário que 1 *probe* detete o *timestamp* como anómalo para o mesmo ser classificado como anómalo.

Relativamente à variação do número de *probes*, verificou-se que com o Método de *Tukey*, os melhores resultados, em termos gerais, situam-se entre $\gamma = 40\%$ e $\gamma = 60\%$. Ou seja, sensivelmente metade das *probes* precisam de detetar uma anomalia no mesmo instante para o mesmo ser classificado como anómalo. Sendo assim, manteve-se o que foi definido inicialmente. Ou seja, manteve-se γ em 50%. Note-se que neste caso se utilizam todas as *probes*, num caso podem ser as *probes* 1,2,3,4,5,6 a classificar um instante como anómalo e noutros casos podem ser as *probes* 2,4,6,8,10,12. Todas as probes estão a ser consideradas.

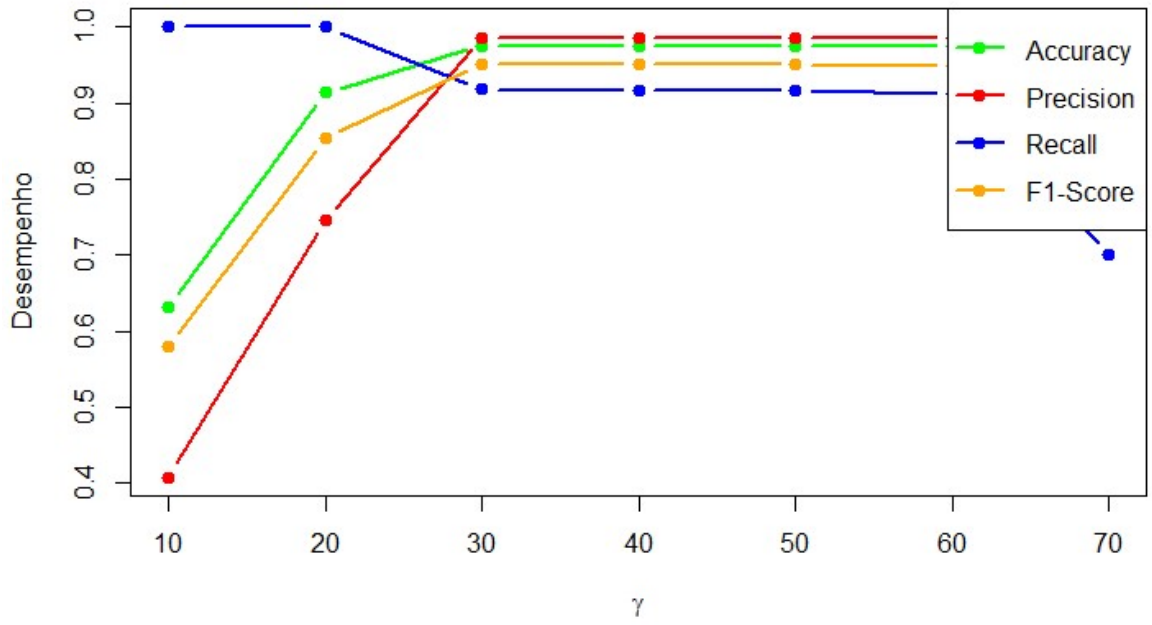

Figura 6.7: Métricas para o *target* 3, Hong Kong, considerando $h = 480$, $k = 10$, $\delta = 1.5$ e variando γ.

Se for feita uma análise individual ao *target* mais problemático, Figura 6.7, relativa a Hong Kong, os melhores valores situam-se em entre $\gamma = 30\%$ e $\gamma = 60\%$. Note-se que recuando à secção 5.5, relativa à heurística, verifica-se que o melhor valor se situava em torno de $\gamma = 30\%$, ou seja, 4 *probes*. Neste caso, já existe uma ampla gama de valores cujas métricas tomam valores bastante semelhantes.

Para concluir, neste trabalho é utilizado $\gamma = 50\%$. Ou seja, são utilizadas 6 *probes*. Comparativa-

mente com a heurística, o Método de *Tukey*, para o *target* de Hong Kong, apresenta um *Recall* mais baixo, mesmo selecionado a quantidade de *probes* que suscita as melhores métricas. No entanto, não é tão sensível a pequenas alterações relativamente ao valor de γ.

6.5 Estudo da Variação do Número de Melhores *Probes*, τ

Contrariamente à secção anterior, em que se utilizavam todas as *probes*, nesta secção são descartadas as *probes* com um *avgRTT* mais elevado. Mais uma vez pretende-se testar se o *avgRTT* tem influência ou não na fiabilidade da solução.

Tal como na secção 5.6, calcula-se o *avgRTT* médio para as primeiras 480 observações.

Começando pela Figura 6.8, que diz respeito ao *target* de Chicago1, é visível que com apenas 1 *probe* os resultados são bastante eficazes. Um *avgRTT* bastante baixo, neste caso, tornou os resultados bastante fiáveis. Note-se que o *avgRTT* da melhor *probe* é 0.12 e está descrito na Tabela 6.1. Pela análise do gráfico percebe-se que independentemente do valor de *probes* as métricas são sempre bastante elevadas. É importante frisar que quando se considera apenas a melhor *probe*, as métricas refletem as métricas da própria *probe*. Quando se escolhem as 2 melhores *probes*, uma observação é classificada como anómala quando pelo menos 1 das *probes* a considera anómala. Note-se que, para este mesmo caso, com a heurística, só 1 *probe* não era suficiente devido ao valor de ϵ ser bastante baixo.

Relativamente à Figura 6.9, de Frankfurt1 qualquer um dos valores é bastante satisfatório. Consegue-se ver o *avgRTT* médio na Tabela 6.2 e facilmente se percebe que o Método de *Tukey* funciona bastante bem quando o *avgRTT* médio é muito baixo, contrariamente à heurística de Salvador e Nogueira.

No que diz respeito à *probe* 3, de Hong Kong, já não é algo tão simples. O *Recall* da Figura 6.10 a partir de 6 *probes* é sempre em torno de 0.9 e as restantes métricas melhoram significativamente. No entanto, a Tabela 6.3 refere que a melhor *probe* é LA2 com um *avgRTT* médio bastante elevado, em torno de 163.17. Como este *avgRTT* médio já é bastante elevado apenas com 1 *probe* torna-se impossível classificar corretamente as anomalias, existindo portanto uma grande quantidade de Falsos Positivos como se pode provar pelas Figuras 6.11 e 6.12. A Figura 6.11 diz respeito à classificação real do tráfego entre o *target* de Hong Kong e a *probe* de LA2. A Figura 6.12 diz respeito à classificação do mesmo tráfego utilizando apenas a *probe* de LA2.

Após a análise da Figura 6.13, referente ao *target* de Londres, percebe-se que qualquer número de melhores *probes* funciona, pois o *avgRTT* médio das melhores *probes* é bastante baixo. Portanto qualquer desvio do tráfego regular é bastante perceptível. Note-se que o *avgRTT* médio para as 12 *probes* se encontra ordenado na Tabela 6.4.

Tabela 6.1: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Chicago1, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
2 - Chicago2	0.12
7 - LA2	66.54
4 - Frankfurt2	108.28
1 - Amesterdam	111.23
12 - Sweden	121.94
8 - Milan	122.04
5 - Iceland	139.83
9 - SaoPaulo2	141.63
3 - VdM	157.65
6 - Israel	166.19
10 - Johannesburg1	254.46
11 - Johannesburg2	263.64

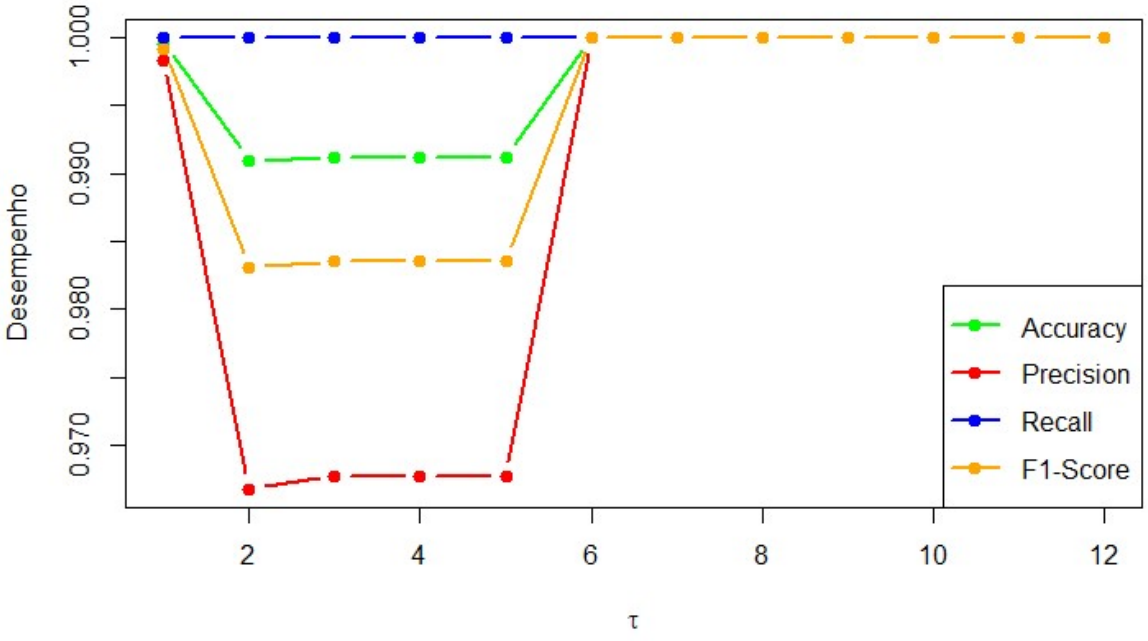

Figura 6.8: Métricas para o *target* 1, Chicago1, considerando $h = 480$, $k = 10$, $\delta = 1.5$ e variando τ.

Tabela 6.2: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Frankfurt1, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
4 - Frankfurt2	0.10
1 - Amesterdam	9.54
8 - Milan	18.25
12 - Sweden	23.20
5 - Iceland	60.27
6 - Israel	63.68
2 - Chicago2	108.28
7 - LA2	154.69
10 - Johannesburg1	186.08
11 - Johannesburg2	186.65
9 - SaoPaulo2	213.37
3 - VdM	225.91

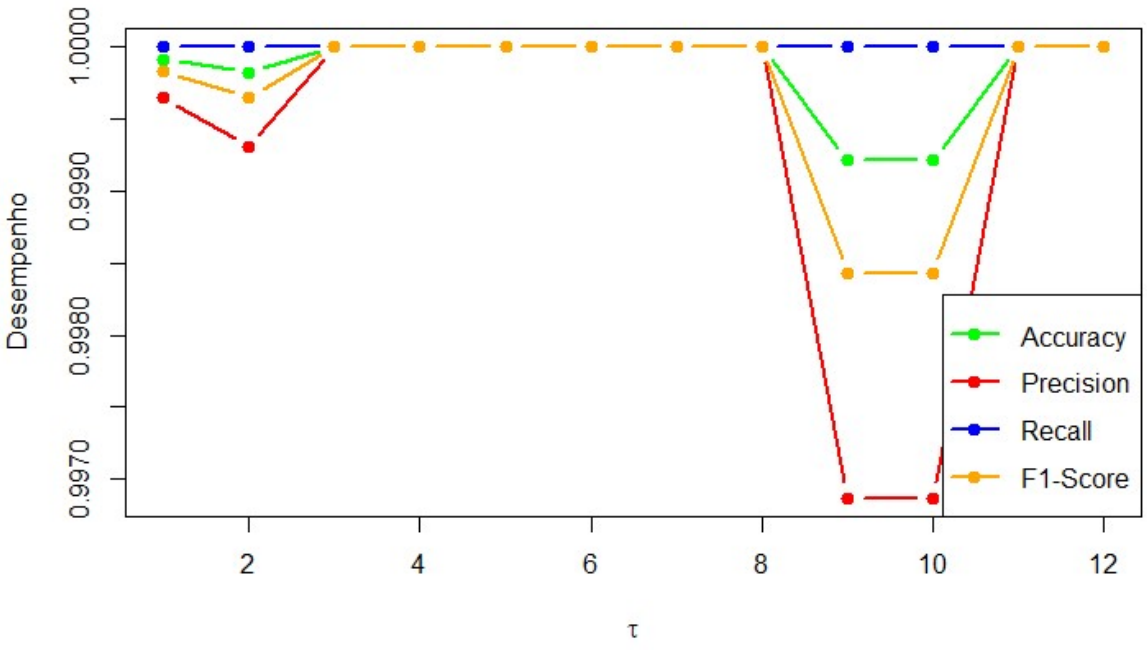

Figura 6.9: Métricas para o *target* 2, Frankfurt1, considerando $h = 480$, $k = 10$, $\delta = 1.5$ e variando τ.

Tabela 6.3: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Hong Kong, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
7 - LA2	163.17
2 - Chicago2	192.82
8 - Milan	268.44
6 - Israel	277.78
12 - Sweden	291.62
4 - Frankfurt2	318.55
1 - Amesterdam	324.18
5 - Iceland	324.53
9 - SaoPaulo2	326.73
3 - VdM	349.68
10 - Johannesburg1	448.21
11 - Johannesburg2	448.34

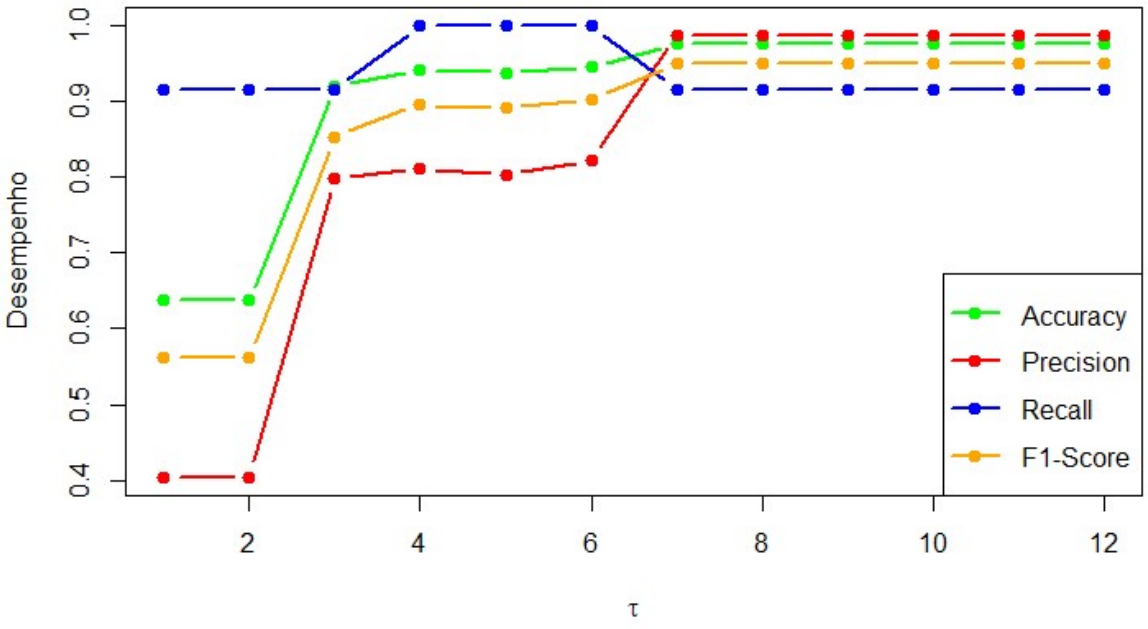

Figura 6.10: Métricas para o *target* 3, Hong Kong, considerando $h = 480$, $k = 10$, $\delta = 1.5$ e variando τ.

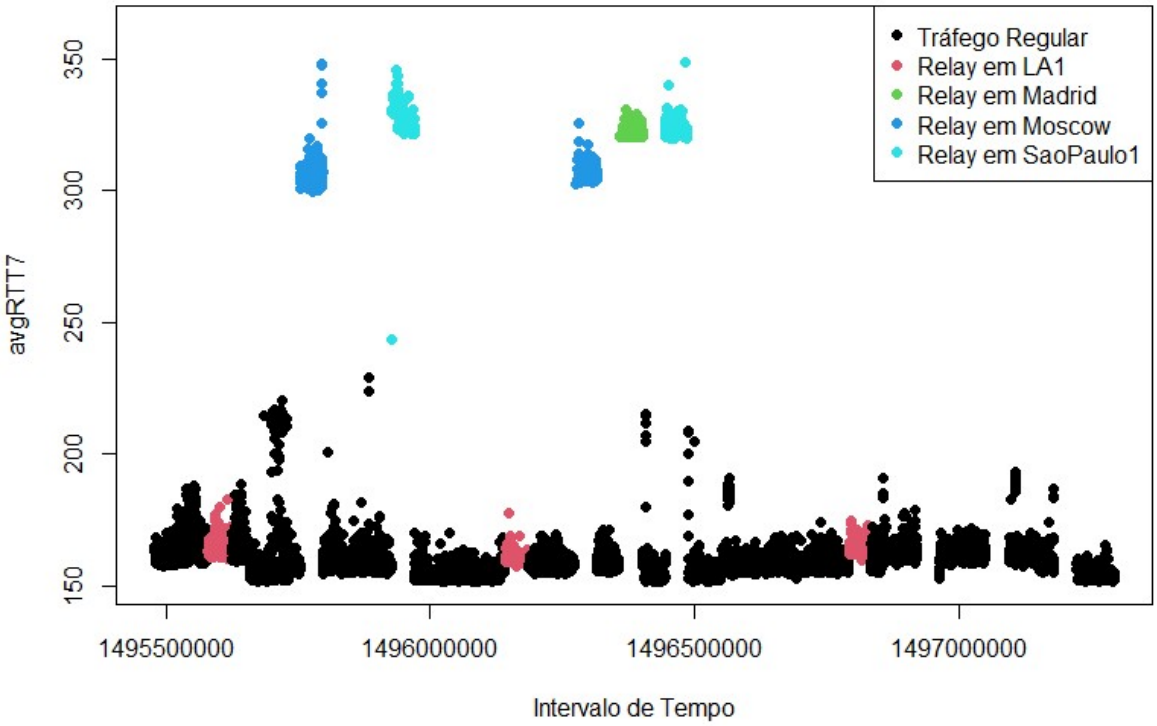

Figura 6.11: Visualização do avgRTT do tráfego entre o *target* 3, Hong Kong, e LA2.

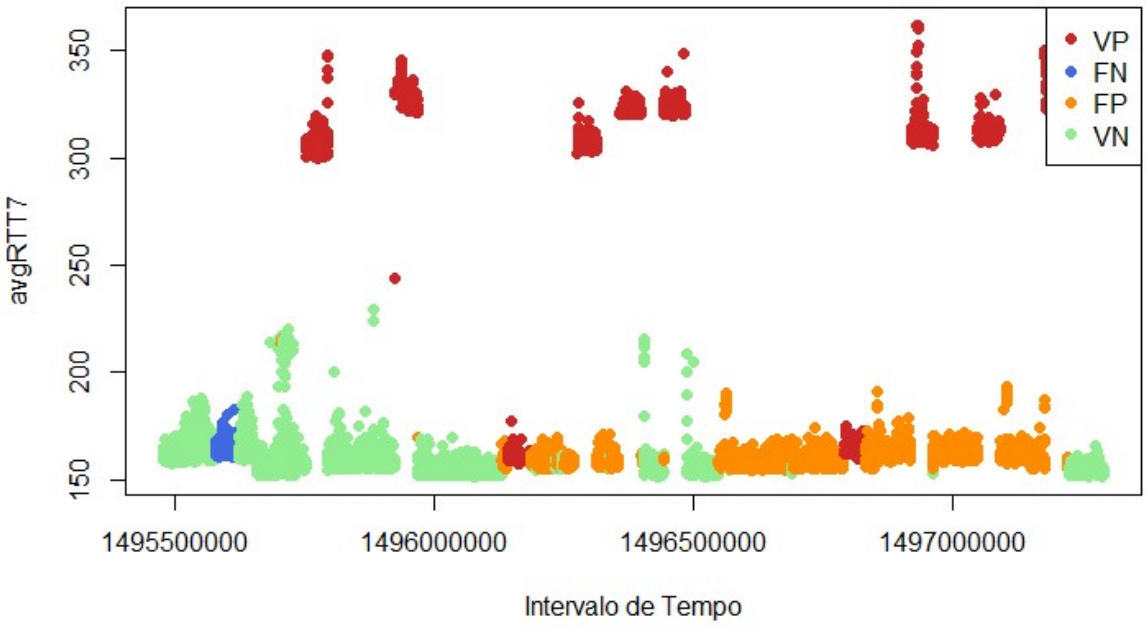

Figura 6.12: Visualização do avgRTT do tráfego entre o *target* 3, Hong Kong, e a *probe* 7, LA2, utilizando o Método de *Tukey* apenas com 1 *probe* .

Tabela 6.4: *Probes* ordenadas pela média do *avgRTT*, para o *target* de Londres, considerando as primeiras 480 observações.

Probe	Média do *avgRTT*
1 - Amesterdam	9.80
4 - Frankfurt2	15.71
8 - Milan	24.96
12 - Sweden	31.26
5 - Iceland	46.84
6 - Israel	77.33
2 - Chicago2	90.59
7 - LA2	142.29
11 - Johannesburg1	162.86
11 - Johannesburg2	190.34
9 - SaoPaulo2	199.63
3 - VdM	220.40

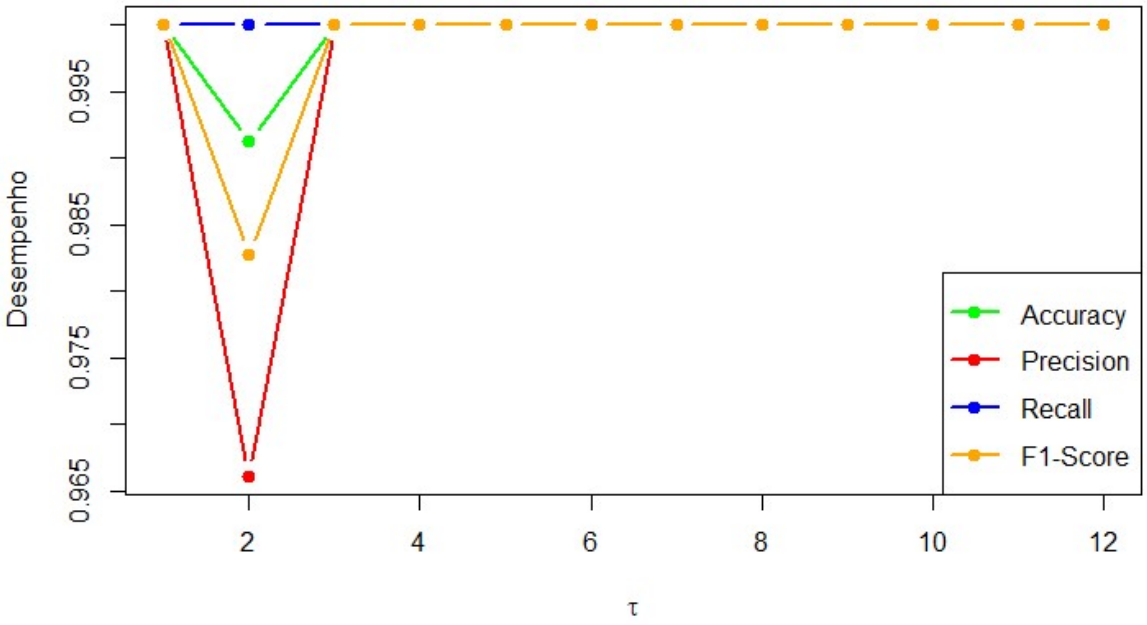

Figura 6.13: Métricas para o *target* 4, Londres, considerando $h = 480$, $k = 10$, $\delta = 1.5$ e variando τ.

6.6 Conclusões

Após a análise de todos os parâmetros pré-definidos necessários à aplicação do método de *Tukey*, chegou-se à conclusão que os valores que suscitam melhores resultados são $k = 10$, $h = 480$ e $\delta = 1.5$. Quando ao número de *probes*, decidiram-se utilizar as *probes* todas e considerar $\gamma = 50\%$. Contrariamente ao verificado na heurística, neste caso com $\gamma = 50\%$ consegue-se obter o melhor resultado possível para todos os *targets*. Note-se que para a heurística se utilizou $\gamma = 30\%$, ou seja, 4 *probes*.

Um ponto negativo é que as métricas não conseguem superar as alcançadas pela heurística proposta por Salvador e Nogueira, tal como se pode observar na Figura 6.14. A Figura 6.15 apresenta as métricas incluindo as observações com um *avgRTT* superior a 600.

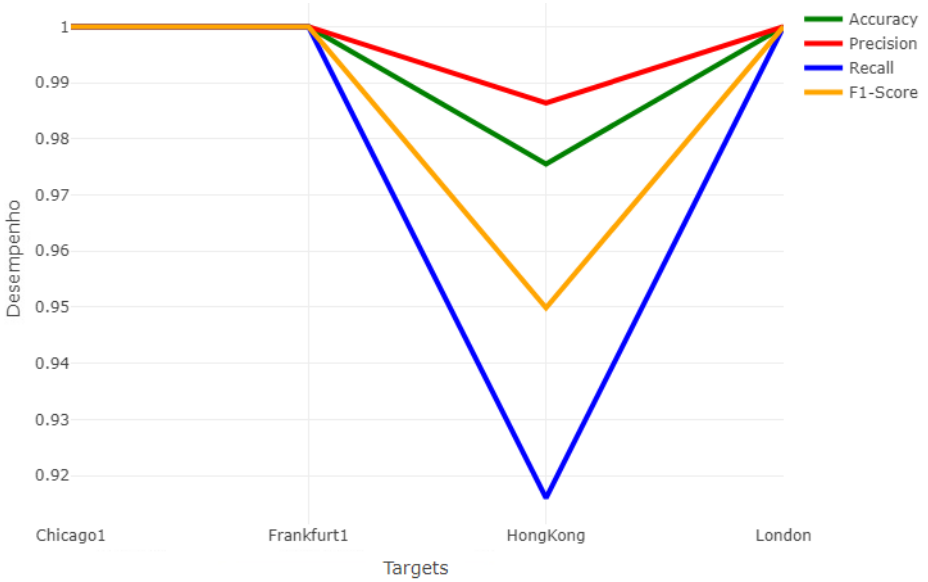

Figura 6.14: Métricas finais para o Método de *Tukey*, excluindo as observações com *avgRTT* superior a 600ms.

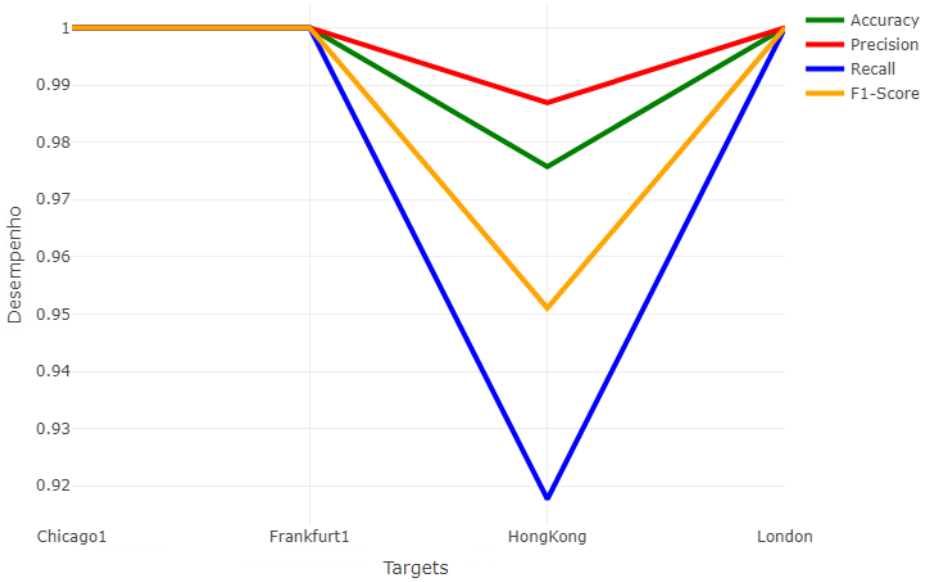

Figura 6.15: Métricas finais para o Método de *Tukey*, incluindo as observações com *avgRTT* superior a 600ms.

7

Distance Based-Outlier

Em problemas reais, espera-se que as anomalias tenham um padrão muito diferente dos dados regulares e sejam relativamente raras. Assim, se pensarmos que cada observação é representadas por um ponto no $\Re^p$, espera-se que em geral a concentração de observações em torno de uma observação regular seja bem maior do que a concentração de observações em torno de uma observação anómala. Neste caso, em que se tem uma quantidade de dados regulares muito superior à dos dados anómalos, as observações podem ser classificadas como regulares se em torno de uma mesma observação existir uma grande quantidade de observações. O *Distance Based-Outlier* considera que os dados regulares têm à sua volta uma grande concentração de observações. Isto é, se a percentagem de vizinhos nas proximidades for superior a um valor previamente estipulado, considera-se que a observação é regular. Neste capítulo, é feita uma análise no que diz respeito aos parâmetros do algoritmo *Distance Based-Outlier*. O objetivo é perceber o desempenho deste algoritmo para o conjunto de dados em análise e comparar os resultados obtidos com o Método de *Tukey* e a heurística de Salvador e Nogueira, previamente estudados.

7.1 Estudo da Variação do Comprimento da Janela Deslizante, h

À semelhança do que foi estudado para a heurística de Salvador e Nogueira e para Método de *Tukey*, para *Distance Based-Outlier* também se estudou a variação do desempenho do método com várias dimensões das janelas deslizantes. Utilizou-se o parâmetro h que representa a quantidade de observações passadas que se utiliza como input do algoritmo. O h tal como mencionado no método de *Tukey*, Capítulo 6, e na heurística de Salvador e Nogueira, Capítulo 5, representa o comprimento da janela deslizante, ou seja, o número de observações passadas com que se vai comparar a observação atual. Neste caso, para os parâmetros considerou-se que a observação atual tem pelo menos $\pi = 70\%$ observações numa vizinhança, centrada na observação atual e de raio r, de entre as h observações passadas. Note-se que π corresponde à percentagem de observações vizinhas que tem de estar dentro do raio r e que o raio r corresponde a $2 \times média$. Visto que o *Distance Based-Outlier* é um algoritmo menos eficaz e mais lento perante grandes quantidades de dados optou-se por $h \in \{180, 280, 380, 480\}$.

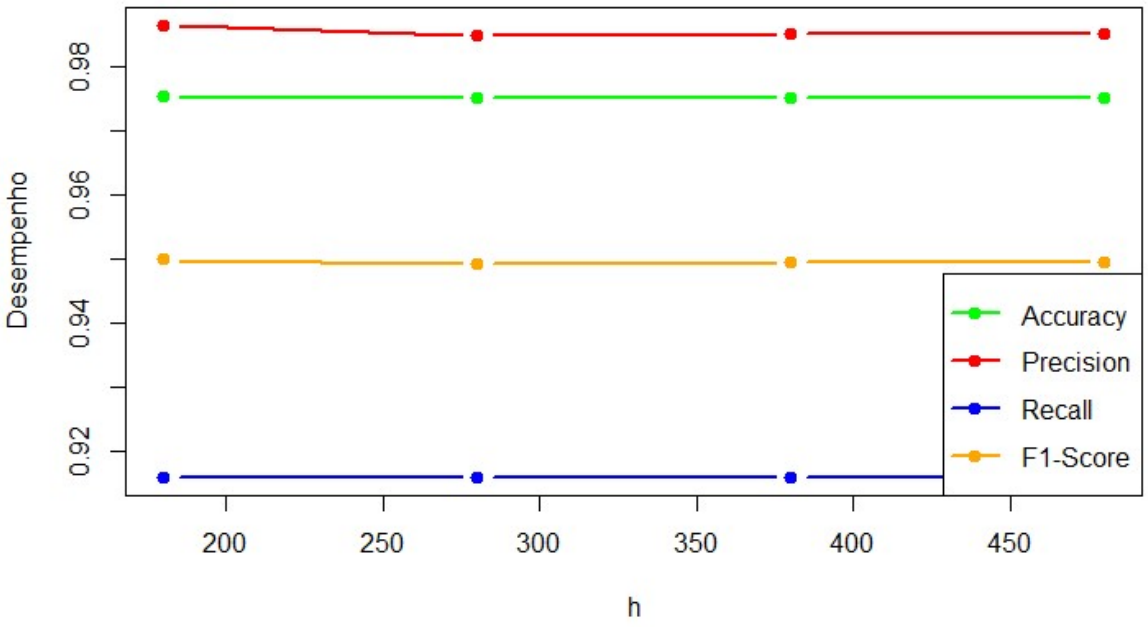

Figura 7.1: Métricas para o *target* 3, Hong Kong, considerando que $\pi = 70\%$, $k = 10$ e variando h.

No que diz respeito aos *targets* menos problemáticos, ou seja, Chicago1, Frankfurt1 e Londres, todas as métricas obtiveram valores perfeitos, isto é, todas as observações regulares e anómalas foram bem identificadas. Relativamente ao *target* mais desafiante, Hong Kong, representado na Figura 7.1, observou-se um *Recall* mais baixo. Este acontecimento deve-se ao facto de as observações se encontrarem todas muito próximo umas das outras e as *probes* muito longe do *target*. Assim sendo, o tráfego anómalo encontra-se muito perto do tráfego regular e o algoritmo classifica-o erradamente como regu-

Tabela 7.1: Métricas obtidas para o *target* 3, Hong Kong por *probe*.

Probe	Accuracy	Precision	Recall	F1-Score	Verdadeiros Negativos	Falsos Negativos	Falsos Positivos	Verdadeiros Positivos
1 - Amesterdam	0.8305	0.9258	0.3596	0.5180	8307	1822	82	1023
2 - Chicago2	0.9562	0.8666	0.9775	0.9187	7961	64	428	2781
3 - Viña del Mar	0.9251	0.9579	0.7367	0.8329	8297	749	92	2096
4 - Frankfurt2	0.8700	0.9028	0.5455	0.6801	8222	1293	167	1552
5 - Iceland	0.9477	0.9494	0.8383	0.8904	8262	460	127	2385
6 - Israel	0.9838	0.9399	1.0000	0.9690	8207	0	182	2845
7 - LA2	0.9140	0.8405	0.8151	0.8276	7949	526	440	2319
8 - Milan	0.9367	0.8000	1.0000	0.8889	7678	0	771	2845
9 - SaoPaulo2	0.7396	0.2802	0.0179	0.0337	8258	2794	131	51
10 - Johannesburg1	0.9306	0.9769	0.7434	0.8443	8339	730	50	2115
11 - Johannesburg2	0.9389	0.9603	0.7916	0.8678	8296	593	93	2252
12 - Sweden	0.9412	0.8115	1.0000	0.8959	7728	0	661	2845

lar. Note-se que k diz respeito à quantidade de observações anómalas consecutivas necessárias para o algoritmo classificar que um determinada *timestamp* é anómalo e o raio r corresponde a $2 \times média$.

Para esclarecer estes resultados que são ligeiramente diferentes dos resultados da heurística de Salvador e Nogueira e do Método de *Tukey*, decidiu-se estudar *probe* a *probe* os dados e perceber quantos Verdadeiros Positivos e Verdadeiros Negativos existiam na classificação. Optou-se por verificar com um h=180, visto que é o que utiliza uma menor quantidade de observações passadas diminuindo assim o tempo computacional. Posto isto, obteve-se a Tabela 7.1.

Verifica-se que a *probe* 8, Milan, por exemplo, tem muitos Falsos Positivos, isso aconteceu sobretudo porque este algoritmo não detetou pequenas mudanças a nível do *avgRTT*, como se observa na Figura 7.2. Este aumento no *avgRTT* do tráfego regular foi automaticamente detetado como tráfego anómalo. Concluí-se então que este algoritmo acaba por ser pouco susceptível a mudanças de nível. Na Figura 7.3, relativa à *probe* 12, Sweden, acontece algo semelhante nos mesmos *timestamps*. A partir destes 2 casos percebe-se a necessidade de conjugar *probes* em diferentes locais, conseguindo assim compensar estes Falsos Positivos.

Relativamente aos Falsos Negativos, uma das *probes* com uma maior quantidade de Falsos Negativos é a *probe* 1, de Amesterdam. Observando a Figura 7.4, que contém a classificação após a aplicação do algoritmo, antes da votação de todas as *probes*, percebe-se que existiu alguma dificuldade em classificar o tráfego anómalo porque o *avgRTT* é muito semelhante. O mesmo acontece em muitas outras *probes*, nem a junção de todas as *probes* consegue excluir este tráfego. Na Figura 7.4 a azul encontram-se os Falsos Positivos. No entanto, sem esta diferença de cor não se distingue o tráfego anómalo do tráfego regular. Com esta *probe* o tráfego é erradamente classificado, relembrando assim a necessidade de conjugar *probes*, pois se por um lado contém muitos Falsos Negativos, por outro são raros os Falsos Positivos existentes.

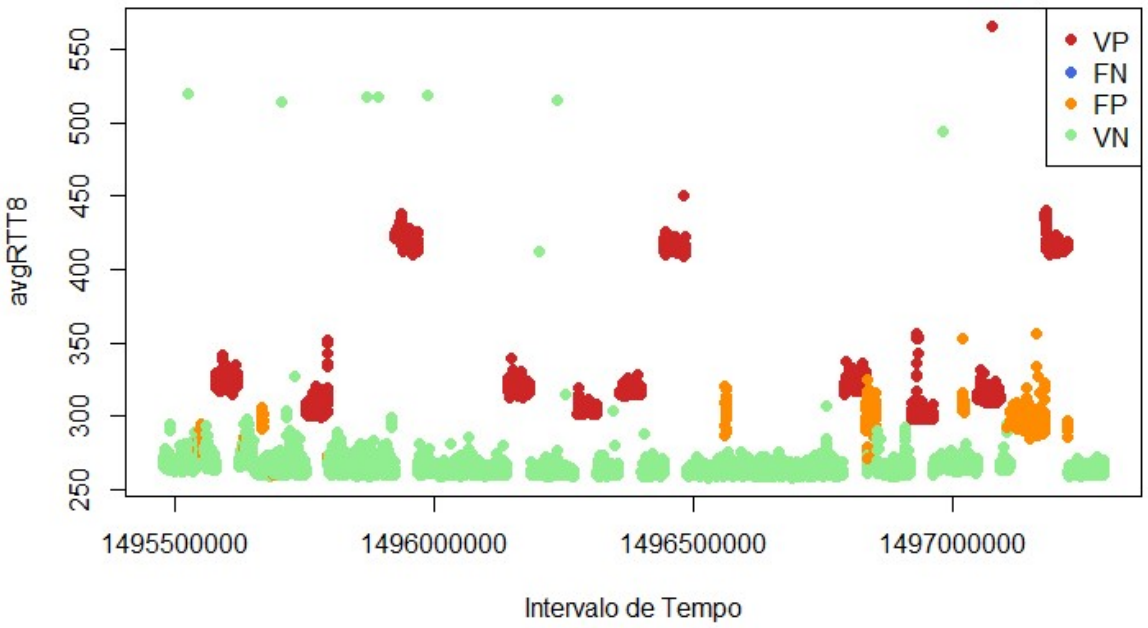

Figura 7.2: Visualização do *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 8, Milan, após a aplicação do *Distance Based-Outlier* sem votações entre *probes*.

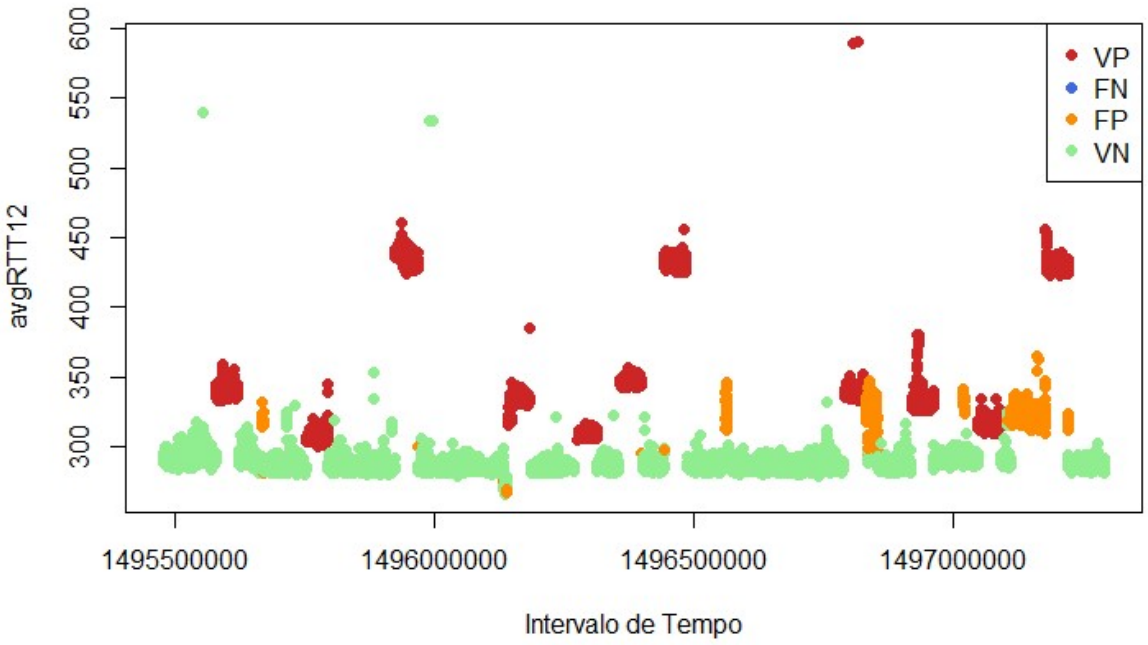

Figura 7.3: Visualização dos *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 12, Sweden, após a aplicação do *Distance Based-Outlier* sem votações entre *probes*.

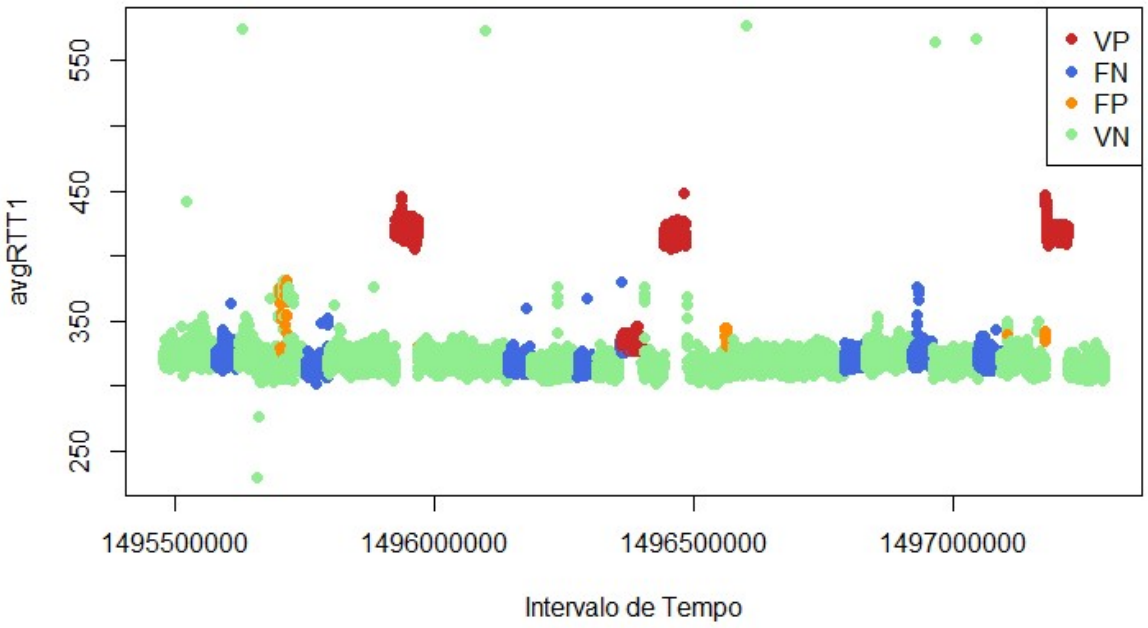

Figura 7.4: Visualização do *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 1, Amesterdam, após a aplicação do *Distance Based-Outlier* sem votações entre *probes*.

7.2 Estudo da Variação da Percentagem de Vizinhos Próximos, π

Nesta secção decidiu-se variar o valor de π. Tal como mencionado em 2.5.1, π é a fração de observações vizinhas que tem de estar a uma distância inferior a r para uma observação ser considerada regular e r corresponde ao dobro da média da observações anteriores consideradas regulares. Considerou-se que as janelas deslizantes contém $h = 180$ observações, pois em termos de eficiência operacional é a mais rápida e em termos de métricas não se verificava qualquer diferença significativa relativamente aos restantes valores de h.

Tendo em conta que a percentagem de observações regulares no conjunto de dados é cerca de 75%, decidiu-se variar o valor de π entre 70% e 90%. Note-se que a janela deslizante apenas tem observações consideradas regulares pelo algoritmo. Portanto, a observação em causa apenas tem de estar perto de π das restantes 180 observações da janela deslizante para ser classificada com regular.

Relativamente ao *targets* de Chicago1, Frankfurt1 e Londres o *Distance Based-Outlier* mostrou-se realmente eficaz, com resultados perfeitos para todas as métricas. No que diz respeito ao *target* de Hong Kong, representado na Figura 7.5, percebe-se que para valores de π superiores a 80% a *Precision* tem um decréscimo acentuado, verificando-se um aumento de Falsos Positivos, acompanhando um ligeiro aumento do *Recall*, significando que mais observações são classificadas como anómalas e por isso o método acerta mais na classificação das observações anómalas.

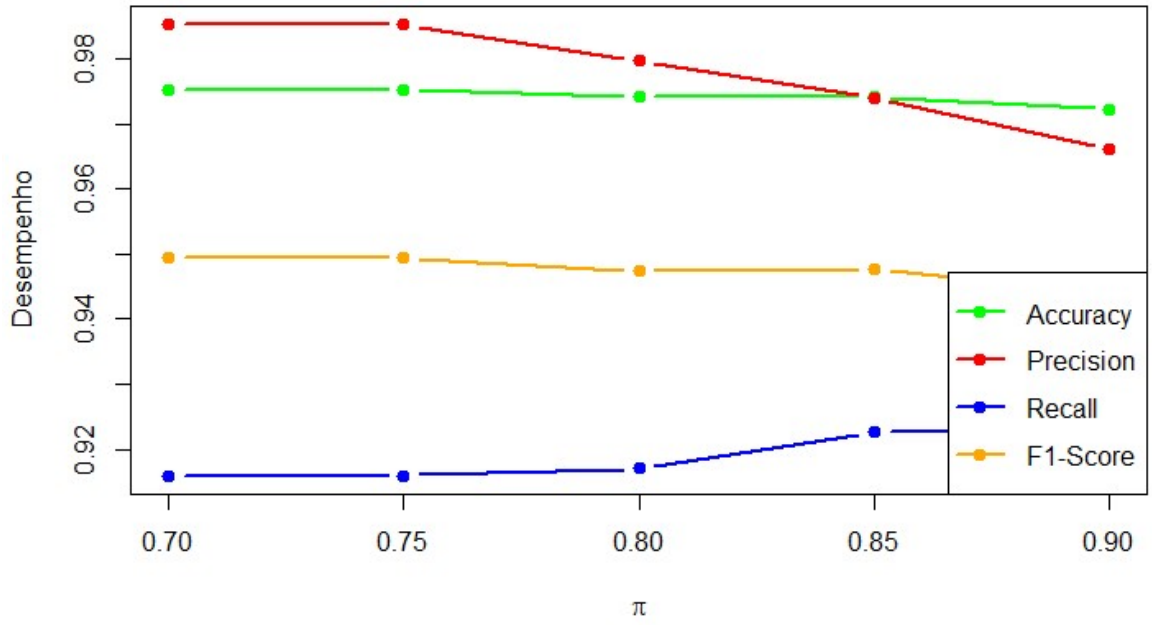

Figura 7.5: Métricas para o *target* 3, Hong Kong, variando π e com $h = 180$.

7.3 Conclusões

Tal como referido anteriormente optou-se por fixar o h em 180, por ser menos exigente em termos de tempo operacional e devido ao facto de não se verificar qualquer tipo de beneficio face aos restantes valores de h. Quanto ao valor de π considerou-se 75%, pois é o valor mais elevado que consegue manter a *Precision* perto 1. Após isto, as métricas finais obtidas sem observações com um *avgRTT* superior a 600ms encontram-se na Figura 7.6. A Figura 7.7 apresenta as métricas com todas as observações.

Comparativamente aos outros métodos mencionados, o resultado é muito semelhante ao método de *Tukey*. No entanto, é relativamente mais lento, tal como mencionado anteriormente utilizou-se uma janela deslizante de apenas 180 observações, devido ao elevado tempo computacional necessário para executar o algoritmo. Tendo em conta que o pretendido é classificar tráfego em tempo real quanto mais rápido melhor.

Comparativamente à heurística de Salvador e Nogueira, Capítulo 5, as métricas são mais baixas, excepto a *Precision*. No entanto, é preciso relembrar que na heurística o algoritmo classifica uma observação como anómala quando apenas 4 a detetam como tal. Portanto, possivelmente com outro conjunto de dados pode não funcionar tão bem sendo que foram necessárias mais modificações.

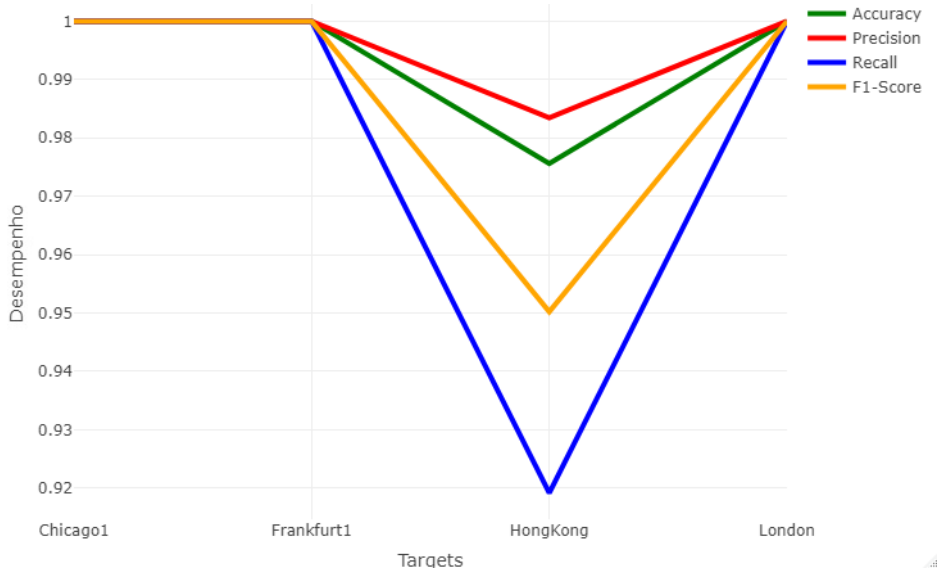

Figura 7.6: Métricas finais para o *Distance Based-Outlier* excluindo observações com um *avgRTT* superior a 600ms.

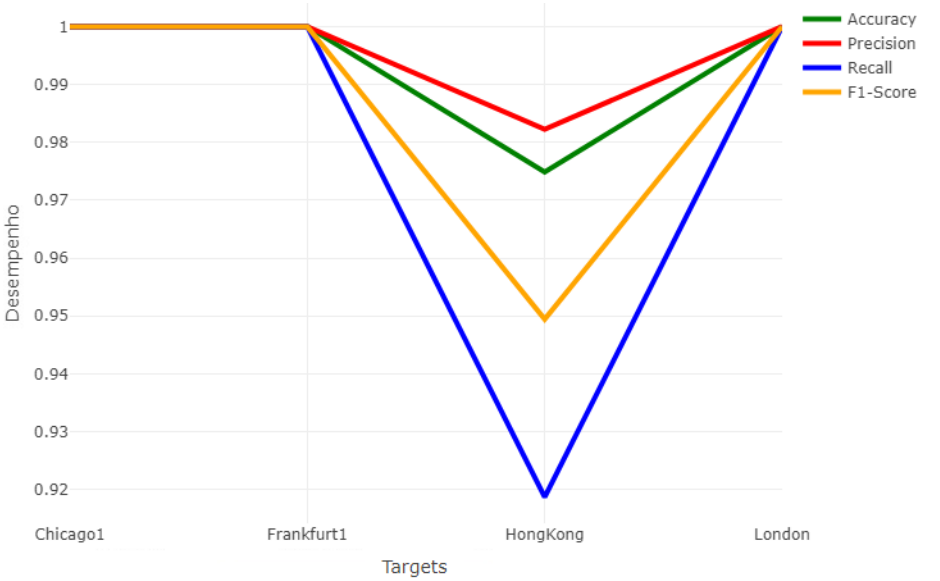

Figura 7.7: Métricas finais para o *Distance Based-Outlier* incluindo observações com um *avgRTT* superior a 600ms.

8

Método SAX

O método de *SAX* (*Symbolic Aggregate aproXimation*), tal como foi mencionado na secção 2.7.1.C, é um método que se baseia em símbolos para fazer a classificação de uma série temporal. Primeiramente, aplica-se o *PAA* (*Piecewise Aggregate Approximation*) para reduzir consideravelmente o tamanho da série temporal, depois a cada valor obtido através do *PAA* é atribuída uma letra. Note-se que na redução da dimensão feita através do *PAA*, a observação obtida é a média de um determinado número de observações, número esse que é estudado neste Capítulo.

Contrariamente aos restantes algoritmos, em que se utilizam janelas deslizantes, neste algoritmo é necessário um conjunto de dados fixo, isto é, não é possível adicionar novas observações à medida que o tempo evoluí. O conjunto de dados é dividido equitativamente em w subconjuntos, sendo possível que após a aplicação do *PAA*, um determinado valor w_i contenha simultaneamente observações regulares e observações anómalas. Isto é, supondo que o conjunto de dados tem 1000 observações e se quer reduzir para 100 observações, agrupam-se as observações 10 a 10 e calcula-se a média. Ou seja, considerada-se $\varphi = 10$, sendo φ o número de observações a agrupar. Este valor obtido pode conter 6 observações regulares e 4 observações anómalas. Sendo assim não é um algoritmo em que se consigam alcançar métricas competitivas, sem qualquer modificação ao mesmo.

Em seguida é feito um estudo que visa a encontrar o melhor desempenho possível para este método. É estudada a dimensão do conjunto de dados, a percentagem de *probes* necessárias para classificar um *timestamp* como anómalo, bem como a redução em amplitude, tendo em conta o valor de α. O α traduz o número de símbolos a utilizar, ou seja, a quantidade de valores que uma observação pode ter em amplitude.

8.1 Estudo das Variáveis e Melhorias Efetuadas

Inicialmente testou-se a nível de redução da dimensão do conjunto de dados. No entanto, verificou-se que quanto menos observações se reduzissem, melhor era o resultado em termos de métricas. Sendo que o objetivo era reduzir consideravelmente o conjunto de dados optou-se por considerar $\varphi = 50$, transformando 50 observações apenas numa observação. Sendo a dimensão do novo conjunto de dados é w, sendo que $w = \frac{n}{50}$ e n é a dimensão do conjunto de dados inicial.

Relativamente à percentagem de *probes* que detetam um *timestamp* como anómalo, testes já efetuados na heurística de Salvador e Nogueira, no Método de *Tukey* e no *Distance Based-Outlier*, decidiu-se manter em 50%. Esta percentagem não influenciava de modo algum os valores das métricas. Não existiu nenhum valor que se mostrou benéfico relativamente aos restantes, neste caso consideraram-se 6 *probes* para se classificar o *timestamp* como anómalo.

O *SAX* utiliza uma outra variável, α, que designa a quantidade de símbolos que são necessários para classificar o conjunto de dados da melhor forma possível. Neste caso, considerou-se que as primeiras 50 observações determinam qual seria a letra do alfabeto a partir do qual um instante é classificado anómalo. Ou seja, se a média das primeiras observações classificam o *timestamp* como a letra b, qualquer letra superior a b é anómala. Variou-se o α entre 3 e 9 e optou-se por escolher o valor 4, que pareceu o mais promissor, ou seja, utilizam-se as 4 primeiras letras do alfabeto. Note-se que este método usualmente é utilizado para comparar dias diferentes em que é suposto o conjunto de dados ter o mesmo comportamento ao longo do dia. Neste trabalho, o conjunto de dados tem de ter o comportamento das primeiras observações, pois são as observações que se consideraram regulares. Sendo assim, fixou-se a primeira letra obtida e declarou-se que todas as letras superiores representam irregularidades no conjunto de dados.

No entanto, como no *SAX* as observações regulares com valores demasiados altos já se tornam problemáticas, é muito difícil que o algoritmo perceba que é algo momentâneo. Decidiu-se também fazer a mesma experiência com a mediana. Os resultados obtidos para o *target* de Chicago1 encontram-se na Tabela 8.1, enquanto que os do *target* de Hong Kong se encontram na Tabela 8.2. Inicialmente esperava-se que a mediana melhorasse consideravelmente as métricas apresentadas. No entanto, as tabelas já apresentam a combinação de todas as *probes*. Sendo assim estes valores acima de 600 já

Tabela 8.1: Métricas finais utilizando a média em comparação com a mediana, para o *target* de Chicago1.

		Accuracy	Precision	Recall	F1-Score
Sem observações superiores a 600ms	Média	0.985	0.950	0.998	0.973
	Mediana	0.987	0.956	0.998	0.976
Com observações superiores a 600ms	Média	0.955	0.867	0.981	0.920
	Mediana	0.957	0.872	0.981	0.923

Tabela 8.2: Métricas finais utilizando a média em comparação com a mediana, para o *target* de Hong Kong.

		Accuracy	Precision	Recall	F1-Score
Sem observações superiores a 600ms	Média	0.917	0.794	0.910	0.848
	Mediana	0.920	0.800	0.917	0.858
Com observações superiores a 600ms	Média	0.878	0.722	0.856	0.784
	Mediana	0.877	0.725	0.840	0.778

são suavizados quando se faz a classificação após a votação de todas as *probes*.

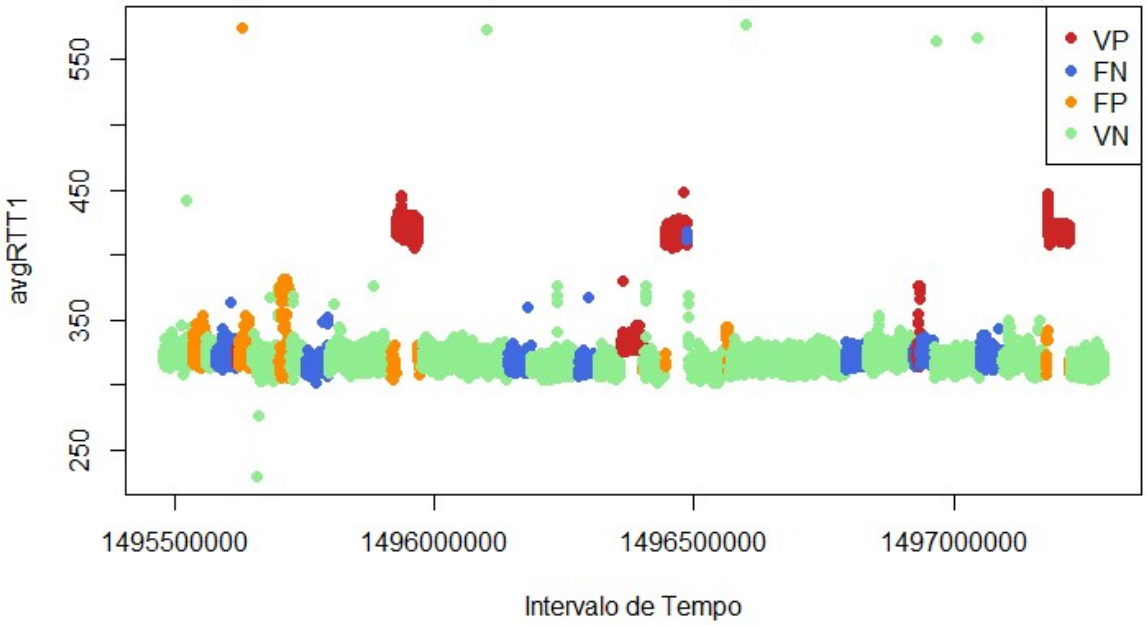

Figura 8.1: Visualização do *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 1, Amesterdam, após a aplicação do *SAX* sem votações entre *probes*.

Posteriormente decidiram-se ver algumas *probes* mais detalhadamente, relativas ao *target* de Hong Kong, o mais problemático. Uma das *probes* com mais Falsos Positivos no *Distance Based-Outlier* foi a probe 1, Amesterdam. Sendo assim, tornou-se relevante perceber se com o método de *SAX* o número de Falsos Negativos nesta *probe* seria menor. Analisando a Figura 8.1 é perceptível que a quantidade de Falsos Negativos é elevada tal como na Figura 7.4 relativa ao *Distance Based-Outlier*. Para além disso, neste algoritmo observa-se uma elevada quantidade de Falsos Positivos que nem com

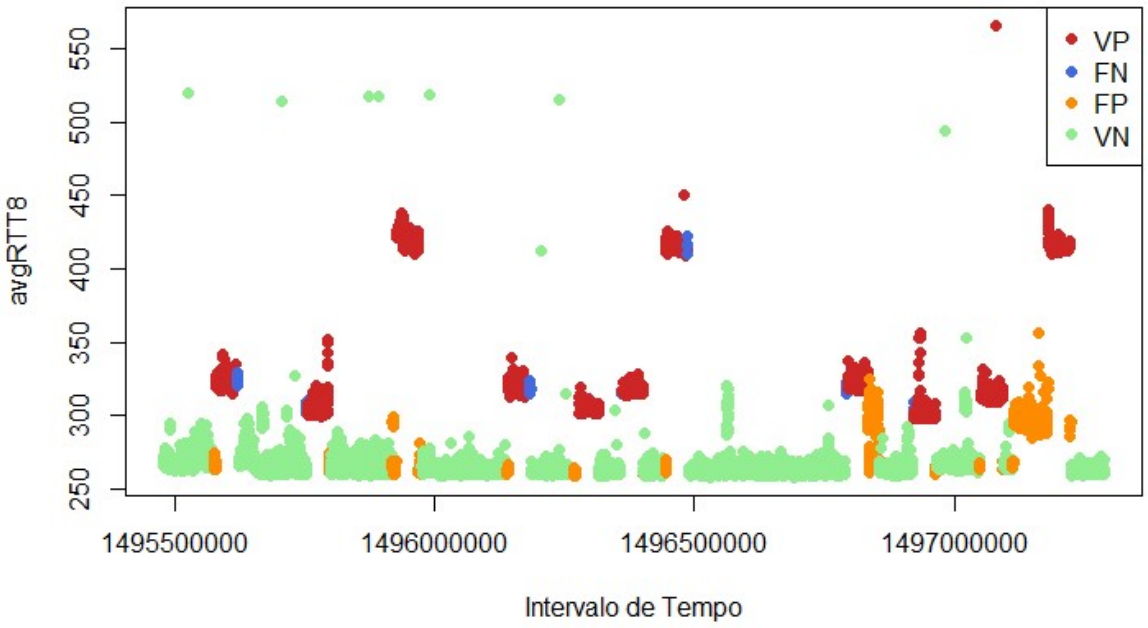

Figura 8.2: Visualização do *avgRTT* do tráfego entre o *target* 3, Hong Kong, e a *probe* 8, Milan, após a aplicação do *SAX* sem votações entre *probes*.

a combinação de todas as *probes* conseguem ser atenuados como se comprou através da Tabela 8.2. Como se pode ver na Figura 8.2 o número de Falsos Positivos, relativamente à mesma *probe*, Milan, no *Distance Based-Outlier* também aumentou, justificando mais uma vez o facto de a *Precision* tomar valores, no melhor dos casos de 80%.

8.2 Conclusões

Apesar de as métricas serem consideravelmente elevadas, verifica-se um grande decréscimo quando se acrescentam observações muito superiores a 600ms. Portanto verifica-se que o método não é muito robusto comparativamente ao método de *Tukey*, pois pequenas alterações traduzem um decréscimo na *Precision* a rondar os 10% como se pode comprovar pela Tabela 8.1.

No entanto, verificou-se muito promissor utilizar o *PAA* e conjugá-lo com outros métodos, nomeadamente o método de *Tukey* que se verificou ser o método mais robusto e com um tempo computacional relativamente mais baixo do que o *Distance Based-Outlier.* A redução da dimensão contribui para uma maior eficiência do algoritmo visto que a quantidade de observações é bastante inferior.

9

Aplicação do *PAA* com o Método de *Tukey*

Posteriormente decidiu-se criar outro algoritmo, que agregava o melhor do *SAX* e o melhor do método de *Tukey*. Para este algoritmo utiliza-se o *PAA* para reduzir a dimensão dos dados e desta forma tornar-se mais eficiente. Incorporou-se também o método de *Tukey*, que se mostrou bastante eficaz, com resultados bastante competitivos, e que em termos computacionais se revelou particularmente promissor.

9.1 Estudo da Redução do Conjunto de Dados, φ

Relativamente ao *PAA* decidiu-se variar φ, isto é, a quantidade de observações a reduzir. Para a realização da média das observações é importante utilizar uma quantidade de observações w muito inferiores ao número total de observações iniciais, h, considerando que não se perdem informações relevantes. Para este caso, também se faz uma comparação entre a média e a mediana, para todos os

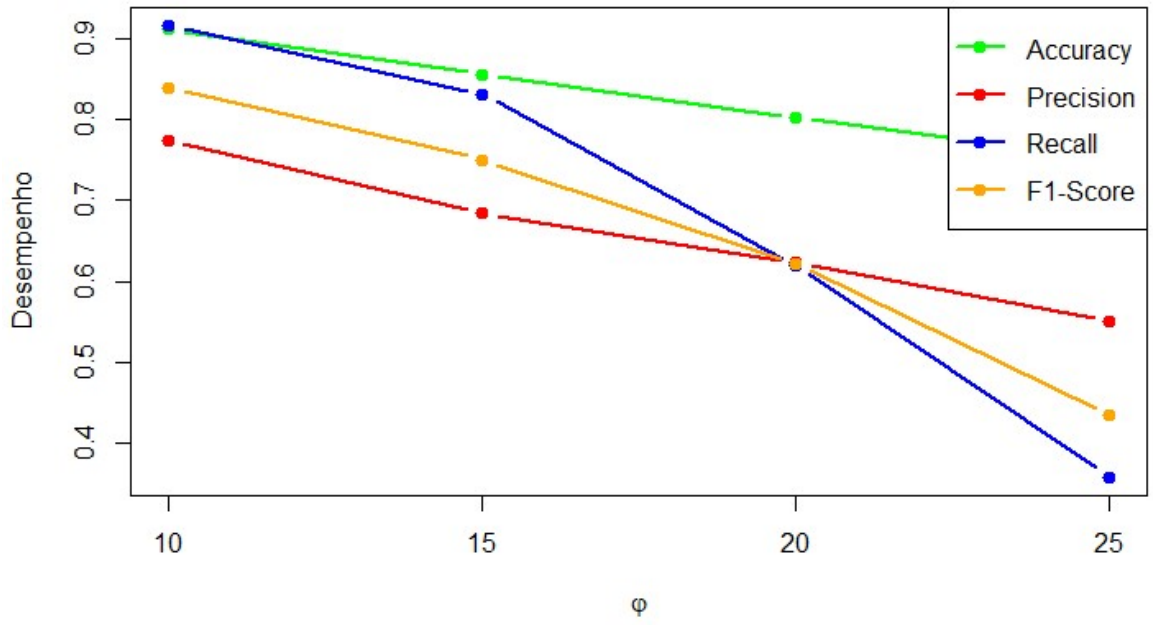

Figura 9.1: Variação de φ para o *target* Hong Kong, com $k = 10$, $h = 48$ e $\gamma = 50\%$.

targets, pois as observações regulares muito elevadas podem suscitar dúvidas quando se utilizam os valores da média. Sendo assim, neste Capítulo investiga-se a possibilidade de a mediana desencadear métricas mais elevadas.

Relativamente ao valor da redução do *PAA*, decidiu-se variar a quantidade de observações que são reduzidas à média das mesmas. Para tal, consideraram-se valores de φ entre 10 e 25 com um intervalo de 5 observações entre cada valor. Após análise de todos os *targets* percebeu-se que os resultados foram melhores para um valor de φ mais baixo, devido ao facto de se reduzir a probabilidade de existirem várias observações anómalas e não anómalas numa mesma observação w.

Na Figura 9.1, referente ao *target* de Hong Kong, é visível que os melhores valores são para $\varphi = 10$. Sendo assim, fixou-se $\varphi = 10$ para as seguintes experiências.

9.2 Estudo da Variação da Percentagem de *Probes* Necessárias para uma Observação ser Classificada Anómala, γ

Para este algoritmo utilizaram-se quase todos os valores de método de *Tukey* que se consideraram os melhores no Capítulo 6. O valor de k manteve-se em 10, ou seja, o número de observações consecutivas que teriam de ser anómalas para o algoritmo as classificar como tal. Relativamente ao h, número de observações anteriores com que compara o *avgRTT*, optou-se por reduzir o mesmo, visto

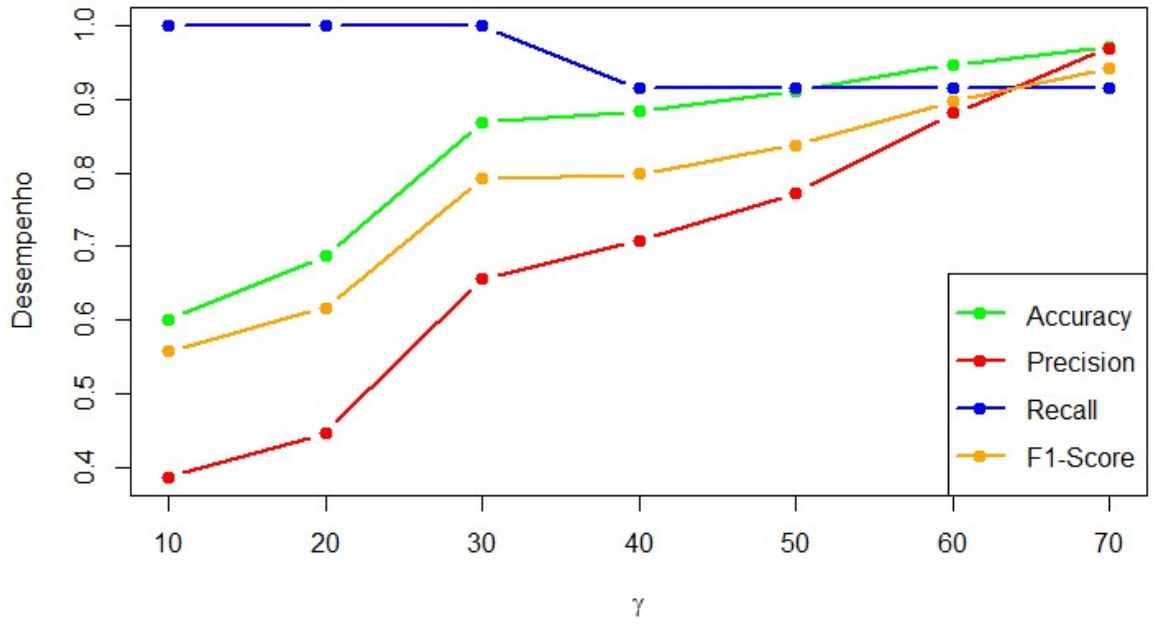

Figura 9.2: Variação de γ com $k = 10$, $h = 48$ e $\varphi = 10$, referente ao *target* Hong Kong.

que também se reduziram as observações do conjunto de dados. Note-se que o tamanho do conjunto de dados passou a ser $w = \frac{n}{\varphi}$, sendo $\varphi = 10$, como descrito na secção 9.1. Sendo assim, em vez de $h = 480$ utilizou-se $h = 48$. O peso da amplitude interquartil, δ, também se tornou irrelevante alterar. No entanto, tornou-se bastante promissor alterar a percentagem de *probes* que votam que uma observação é anómala para a mesma ser considerada anómala, γ. Na Figura 9.2 percebe-se a necessidade de alterar este valor e fixar $\gamma = 70\%$, pois os resultados são visivelmente melhores para o *target* de Hong Kong. Neste método, agrupam-se φ observações consecutivas, sendo φ a quantidade de observações que são agrupadas apenas numa. Sendo assim, nestas observações podem estar simultaneamente valores regulares e anómalos portanto é de prever que sejam necessárias mais *probes* para a correta deteção de uma anomalia.

9.3 Conclusões

Tal como mencionado anteriormente fez-se uma comparação entre a média e mediana. Na Figura 9.3 encontra-se esta comparação após a eliminação das observações superiores a 600. Inicialmente removeram-se estas observações por não contribuírem para a análise dos dados e não conterem informação relevante. Como se pode observar, não se verifica nenhuma diferença entre as métricas utilizando a média ou a mediana. A Figura 9.4 já contém todos os dados do conjunto de dados, incluindo observações superiores a 600. Neste caso, já se verificam algumas alterações nos valores das

Figura 9.3: Comparação entre a média e a mediana excluindo observações superiores a 600.

métricas. Conclui-se que com a média se conseguem alcançar os melhores resultados possíveis. Mais uma vez, apura-se que o valor médio suscita melhor métricas que o valor mediano.

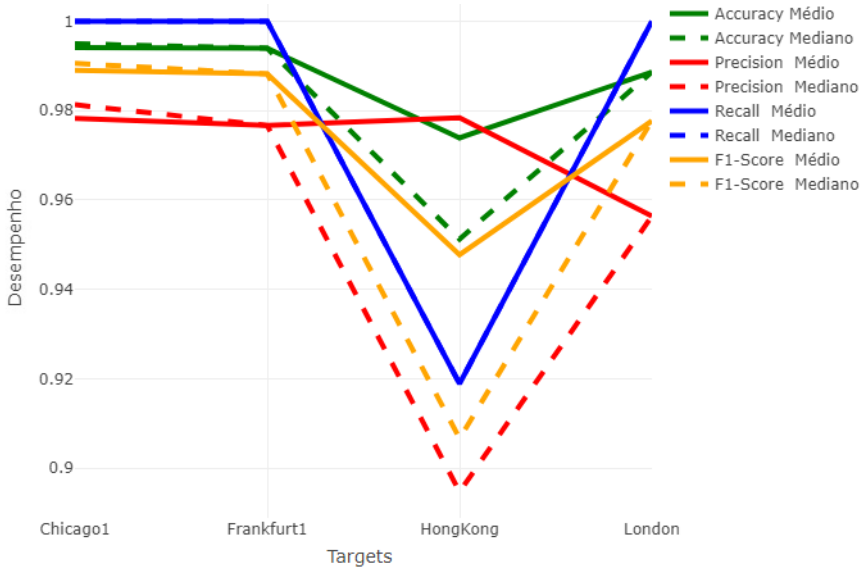

Figura 9.4: Comparação entre a média e a mediana incluindo observações superiores a 600.

10

Conclusão

Neste relatório foram analisados e testados vários algoritmos para deteção de anomalias em séries temporais. Neste caso, pretendeu-se detetar ataques às tabelas de encaminhamento *BGP* (*Border Gateway Protocol*). Os algoritmos estudados foram a heurística de Salvador e Nogueira, Capítulo 5, o método de *Tukey*, Capítulo 6, o *Distance Based-Outlier*, Capítulo 7, o *SAX* (*Symbolic Aggregate aproXimation*), Capítulo 8 e uma variação agregando o *PAA* (*Piecewise Aggregate Approximation*) e o método de *Tukey*, Capítulo 9. As métricas estudas para avaliar o desempenho dos algoritmos foram a *Accuracy*, *Precision*, *Recall* e *F1-Score*.

O primeiro algoritmo analisado neste trabalho foi a heurística proposta por Salvador e Nogueira. Inicialmente mostrou-se bastante promissora devido ao facto de utilizar janelas deslizantes. No entanto, foi necessário ajustar vários parâmetros para alcançar um resultado favorável. Nesse sentido, tornou-se um algoritmo muito sensível a mudanças nos diferentes parâmetros estudados.

Em seguida estudou-se o método de *Tukey*, que também utiliza janelas deslizantes. Verificou-se que este método é mais robusto e não é tão susceptível a mudanças nos valores dos parâmetros em estudo. O método de *Tukey* alcança valores elevados para todas as métricas de desempenho, sendo um dos métodos a considerar para experiências futuras.

Tabela 10.1: Comparação entre os algoritmos estudados neste trabalho.

Parâmetro	Heurística	Tukey	DB-outlier	SAX	PAA + Tukey
Accuracy	0,996	0,98	0,98	0,88	0,97
F1-Score	0,99	0,95	0,95	0,78	0,95
h	480	480	180	————	48

O *Distance Based-Outlier* revelou-se melhor do que o esperado, visto que não é um método que tenha em conta dependências temporais. Este método não utiliza janelas deslizantes. Sendo assim, existiu uma adaptação e uma criação de janelas deslizantes que se adaptassem ao mesmo. No entanto, as métricas de desempenho obtidas foram muito semelhantes às métricas do método de *Tukey* e o tempo computacional do *Distance Based-Outlier* é bastante superior ao método de *Tukey*. Sendo assim, o melhor até este ponto continua a ser o método de *Tukey*.

O algoritmo que que se analisou em seguida foi o *SAX*. O *SAX* é mais apropriado para uma comparação diária ou até mesmo semanal e não utiliza janelas deslizantes, algo que se revelou ser bastante importante para o conjunto de dados. Na Tabela 10.1 o parâmetro h corresponde ao comprimento da janela deslizante e o *SAX* é o único método que não utiliza janelas deslizantes. As janelas deslizantes atenuam pequenas mudanças de nível, por outro lado tornam-se úteis quando o objetivo é detetar ataques em tempo real, porque é possível comparar a observação atual com um pequeno conjunto de observações passadas. Posto isto, não é o método mais adequado. No entanto, o *SAX* incluí uma transformação feita pelo *PAA*, que ajuda na dimensão do conjunto de dados. Sendo assim, posteriormente adaptou-se o *PAA* ao método de *Tukey*, que foi considerado o melhor até agora.

A adaptação do *PAA* com o método de *Tukey* não demonstrou diferenças significativas em termos de desempenho, comparativamente com a utilização do método de *Tukey* sem o *PAA*, como se comprova a partir dos dados relativos à *Accuracy* e ao *F1-Score* apresentados na Tabela 10.1. No entanto, o *PAA* torna o algoritmo mais rápido, visto que se reduz significativamente o conjunto de dados. Sendo assim, concluiu-se que de todos os métodos analisados neste relatório, o melhor para o conjunto de dados em estudo é método de *Tukey* com a redução feita pelo *PAA*.

Bibliografia

[1] A. Subtil, “Latent class models in the evaluation of biomedical diagnostic tests and internet traffic anomaly detection,” 2020.

[2] H. H. Kishan G. Mehrotra, Chilukuri K. Mohan, *Anomaly Detection Principles and Algorithms*, ser. Series Editor. Springer, 2017.

[3] V. Rodrigues. (2019, Abril) Métricas de avaliação: acurácia, precisão, recall... quais as diferenças? [Online]. Available: https://medium.com/@vitorborbarodrigues/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c

[4] R. T. N. Edwin M. Knox, “Algorithms for mining distance-based outliers in large datasets,” pp. 392–403, 1998.

[5] P. B. G. C. F. Spiros Papadimitriou, Hiroyuki Kitagawa, “Loci: Fast outlier detection using the local correlation integral,” pp. 315–326, 2003.

[6] A. N. Paulo Salvador, “Customer-side detection of internet-scale traffic redirection,” 2014.

[7] K. S. Sridhar Ramaswamy, Rajeev Rastogi, “Efficient algorithms for mining outliers from large data sets,” pp. 427–438, 2000.

[8] D. S. M. David Ruppert, *Statistics and Data Analysis for Financial Engineering with R examples*, ser. Series Editor. Springer, 2015.

[9] D. S. S. Robert H. Shumway, *Time Series Analysis and Its Applications with R Examples*, ser. Series Editor. Springer, 2017.

[10] T. Yiu. (2020, Abril) Understanding arima (time series modeling). [Online]. Available: https://towardsdatascience.com/understanding-arima-time-series-modeling-d99cd11be3f8

[11] S. L. D. Yufeng Yu, Yuelong Zhu, “Time series outlier detection based on sliding window prediction,” 2014.

[12] M. Silva, “Modelação da incerteza e deteção de outliers para melhoria do diagnóstico de perdas em sistemas de abastecimento de Água,” 2016.

[13] A. F. Eamonn Keogh, Jessica Lin, “Hot sax: Finding the most unusual time series subsequence: Algorithms and applications,” 2005.

[14] L. W. S. L. Jessica Lin, Eamonn Keogh, “Experiencing sax: a novel symbolic representation of time series,” pp. 107–144, 2007.

[15] L. C. V. T. W. S. K. S. Chengwei Wang, Krishnamurthy Viswanathan, “Statistical techniques for online anomaly detection in data centers,” 2011.

[16] L. C. P. S. A. U. N. V. Georgy Shevlyakov, Kliton Andrea, “Robust versions of the tukey boxplot with their application to detection of outliers,” pp. 6506–6510, 2013.

[17] V. Krish. (2018, Fevereiro) Piecewise aggregate approximation. [Online]. Available: https://vigne.sh/posts/piecewise-aggregate-approx/

[18] ——. (2018, Junho) Symbolic aggregate approximation. [Online]. Available: https://vigne.sh/posts/symbolic-aggregate-approximation/

[19] (2020, Agosto) Symbolic aggregate approximation. [Online]. Available: https://www.geeksforgeeks.org/difference-between-ebgp-and-ibgp/

[20] T. L. E. S. H. E. Y. Rekhter, Ed., “A border gateway protocol 4 (bgp-4),” Janeiro.

[21] A. K. A. D. D. C. Cecilia Testart, Philipp Richter, “Profiling bgp serial hijackers: Capturing persistent misbehavior in the global routing table,” Outubro.