UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE ENGENHARIA DE COMPUTAÇÃO

Relatório: Aplicação de Bayesian Changepoint Detection em Sistemas Auto Adaptativos

Bernardo Pandolfi Costa

Disciplina: Tópicos Especiais I
Professor: Roberto Rodrigues Filho

Araranguá
22 de outubro de 2024

Conteúdo

1 Introdução

Neste trabalho, foi feito um estudo sobre o método de Detecção de Pontos de Mucança Bayesiano. A área de estudo de pontos de mudança possui diversas aplicações, como em finanças, predição de clima, biometrias e robótica. No contexto deste trabalho, apliquei a técnica de detecção de pontos de mudança Bayesiana para encontrar os momentos em que o sistema altera seu comportamento bruscamente, possivelmente necessitando uma reconfiguração da composição atual.

2 Background Teórico

2.1 Teoria de Bayes

A Teoria de Bayes é um princípio matemático que nos ajuda a atualizar nossas crenças sobre algo com base em novas informações. A ideia principal da Teoria de Bayes é simples: você tem uma crença inicial sobre algo (chamada de probabilidade a priori), e quando você observa novos dados ou informações, você ajusta essa crença (chamada de probabilidade posterior). O objetivo é melhorar a precisão da sua crença à medida que você recebe mais dados.

2.1.1 A Fórmula de Bayes

A Regra de Bayes pode ser escrita assim:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Onde:

- $P(H|D)$ é a probabilidade posterior: sua crença ajustada, dado os novos dados.
- $P(D|H)$ é a verossimilhança: a probabilidade de observar os dados, dado que sua hipótese H é verdadeira.
- $P(H)$ é a probabilidade a priori: sua crença inicial.
- $P(D)$ é a evidência: a probabilidade de observar os dados, independentemente da hipótese.

2.1.2 Exemplo

Imagine que você está tentando adivinhar se vai chover hoje. Inicialmente, com base na previsão do tempo, você acha que a chance de chover é 30%. Ou seja, sua probabilidade a priori ($P(H)$) é 0,3.

Agora, você sai de casa e vê que as nuvens estão escuras (esse é o dado novo, ou seja, a evidência $P(D)$). Sabendo que quando as nuvens estão escuras, há uma grande chance de chover (isso é a verossimilhança $P(D|H)$, ou seja, a chance de ver nuvens escuras se for chover), você atualiza sua crença.

Depois de aplicar a Regra de Bayes, você conclui que a chance de chover agora é maior, talvez 70% – essa é sua probabilidade posterior ($P(H|D)$), ou seja, sua nova crença ajustada.

2.1.3 Contexto de Detecção de Pontos

No contexto da detecção de pontos de mudança bayesiana, a lógica é semelhante:

- Probabilidade a Priori ($P(H)$): Você começa com uma suposição inicial sobre a frequência com que os pontos de mudança ocorrem.

- Verossimilhança ($P(D|H)$): Conforme você observa novos dados, você calcula quão provável seria ver essas mudanças, dado que realmente houve um ponto de mudança.
- Probabilidade Posterior ($P(H|D)$): A cada novo dado, você ajusta sua crença sobre onde os pontos de mudança estão, refinando sua detecção conforme mais dados chegam.

2.2 *Changepoints*

Primeiramente, precisamos entender o que é um ponto de mudança (changepoint). Pontos de mudança são alterações abrúptas em dados de uma série temporal, em que as características de uma sequência de dados se altera para as próximas iterações. Comumente, isso é facilmente visualizável através de gráficos, onde um conjunto de dados parece ser homogêneo e de uma hora para outra passa a apresentar valores diferentes.

2.3 Detecção *Offline* e *Online* de Pontos de Mudança

Existem diferentes métodos em algoritmos de detecção de pontos de mudança. A metodologia *Offline* faz a análise de todos os dados de um *dataset* completo, analisando retroativamente cada ponto para detectar onde ocorrem as mudanças de estado.

A detecção *Online*, por outro lado, tenta detectar os pontos de mudança em tempo real, então à medida que novos dados chegam, ele calcula a probabilidade daquele ponto, baseado em seu valor, ser um ponto de mudança de estado.

2.4 Detecção de Pontos de Mudança Bayesiana *Online*

Pense em uma sequência de dados, como o preço de uma ação ao longo do tempo. Se o preço de repente mudar de um comportamento de estabilidade para uma queda, esse momento em que o comportamento mudou é o que chamamos de ponto de mudança. A detecção bayesiana de pontos de mudança é uma técnica que usa a probabilidade para tentar detectar esses momentos à medida que os dados chegam.

A detecção "bayesiana" usa a Teoria de Bayes, que nos permite calcular uma nova "crença" sobre onde os pontos de mudança estão conforme novos dados são observados. Ou seja, começamos com uma ideia inicial de onde podem estar os pontos de mudança e vamos atualizando essa ideia à medida que observamos mais dados.

2.4.1 Como Funciona a Detecção Bayesiana *Online*

O conceito central dessa técnica é o *run length* r_t (comprimento da corrida), que é o número de passos de tempo desde o último ponto de mudança. Por exemplo, se em um determinado momento os dados seguem um padrão estável por 5 pontos consecutivos, o run length seria 5. Quando um ponto de mudança ocorre, o run length volta a zero, e o algoritmo começa a contar novamente desde o último ponto de mudança.

De forma simplificada, o algoritmo roda os seguintes passos:

1. Começa supondo que o run length é zero ($r_t = 0$), ou seja, acredita que o primeiro ponto de dados é o início de uma nova sequência sem pontos de mudança.

2. Quando um novo ponto de dado chega, faz uma previsão de como ele deveria se comportar com base no comportamento observado até agora.

3. Com base no padrão de comportamento do conjunto de dados de r_t, calcula a probabilidade de que o novo dado siga o mesmo padrão (ou seja, não houve ponto de mudança).

4. Calcula a probabilidade de que o novo dado represente o início de um novo padrão (ou seja, um ponto de mudança acabou de ocorrer).

5. Com base nessas probabilidades, atualiza nossa crença sobre qual é o run length atual.

6. Repete a partir do segundo passo para os próximos dados.

2.5 Detecção de Pontos de Mudança Bayesiana *Offline*

Na abordagem *offline*, o objetivo ainda é detectar pontos em uma sequência de dados onde há uma mudança no comportamento (como na média ou na variância). A diferença é que agora você tem todos os dados disponíveis desde o início, então a técnica pode ser mais sofisticada e usar a informação completa para identificar os pontos de mudança de forma retrospectiva.

Pense em uma sequência de dados históricos, como a temperatura registrada ao longo dos anos. Se houver um período onde a média da temperatura subitamente muda, o objetivo da detecção de pontos de mudança offline é encontrar esses momentos de mudança com o benefício de poder "olhar para trás" e analisar o comportamento global dos dados.

2.5.1 Como Funciona a Detecção Bayesiana *Offline*

1. Começa assumindo que não sabe onde os pontos de mudança estão. Sua a priori pode ser que mudanças sejam raras ou que ocorram em intervalos regulares.

2. O algoritmo explora todas as maneiras possíveis de dividir os dados em diferentes segmentos. Para cada divisão possível, ele calcula a verossimilhança de que aquela divisão explica os dados corretamente.

3. Para cada divisão, calcula a verossimilhança. Por exemplo, se supõe que há um ponto de mudança no meio da sequência, calcula o quão bem os dados antes e depois desse ponto se ajustam às distribuições modeladas.

4. Usando o Teorema de Bayes, ele combina a verossimilhança observada com sua distribuição a priori e calcula a distribuição posterior, que é a probabilidade de que um ponto de mudança tenha ocorrido em cada ponto da sequência.

5. Após calcular a distribuição posterior para todos os pontos da sequência, o algoritmo identifica os pontos onde a probabilidade de mudança é mais alta. Esses são os pontos onde é mais provável que uma mudança nos dados tenha ocorrido.

6. Se houver incerteza sobre onde exatamente a mudança ocorreu (por exemplo, se a mudança foi gradual), o algoritmo pode refinar sua estimativa iterativamente.

3 Como Executar o Projeto

Com o repositório clonado, basta instalar as dependências do Python com o comando "pip install -r requirements.txt" e, em seguida, no arquivo Trabalho.ipynb, você pode executar as células do notebook individualmente.

4 Apresentação dos Casos

Neste trabalho, foram aplicados os algoritmos de detecção de pontos de mudança Bayesiano *Online* e *Offline* em seis casos diferentes. Começaremos com dois exemplos mais claros, cada um com um total de 400 requisições, 100 para cada configuração do sistema. O objetivo da aplicação do algoritmo Bayesiano é detectar os pontos em que houve a alteração na composição do Distributor, como por exemplo mudando do Local para o Sharding.

4.1 Caso 1

O primeiro Dataset grava o comportamento do sistema com a lista de tamanho 2. O sistema inicia na composição *propagate*, e alterna entre as outras composições no decorrer do experimento.

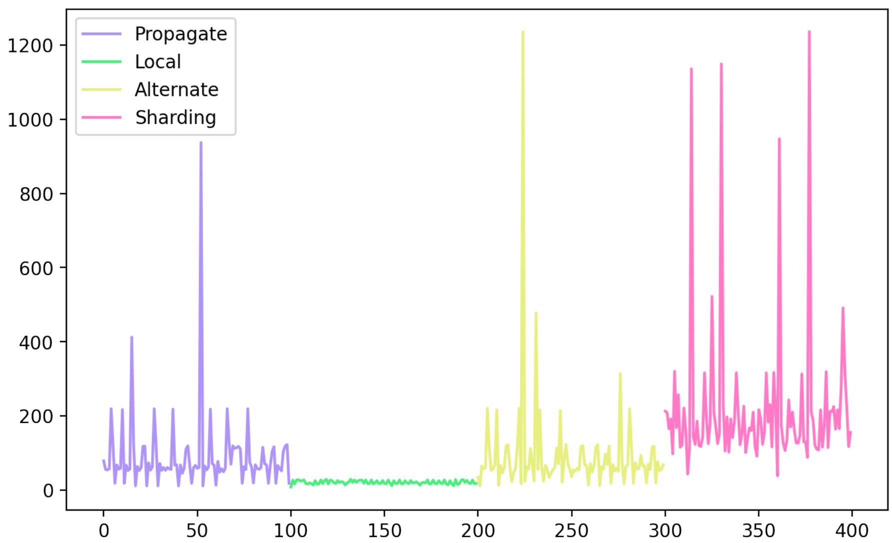

Figura 1: Comportamento das configurações com tamanho 2.

4.2 Caso 2

O segundo dataset considera uma lista de tamanho 36. O processo aqui é o mesmo, o sistema inicia em uma configuração e a cada 100 iterações ele altera o estado.

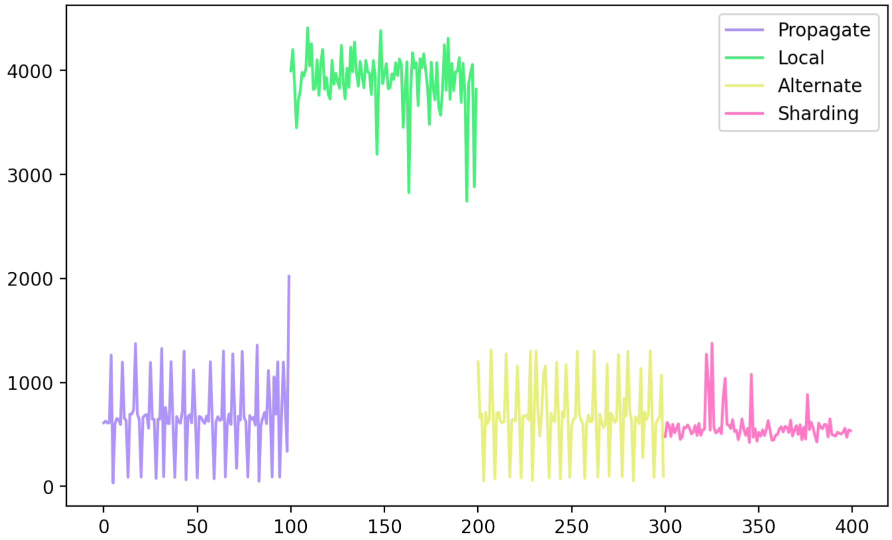

Figura 2: Comportamento das configurações com tamanho 36.

4.3 Caso 3 e 4

Para este e o próximo caso, foram feitas alterações no tamanho da lista para as ações Local e Sharding. Estes datasets simulam o sistema rodando em tempo real em que mudanças nas requisições acontecem de forma não linear.

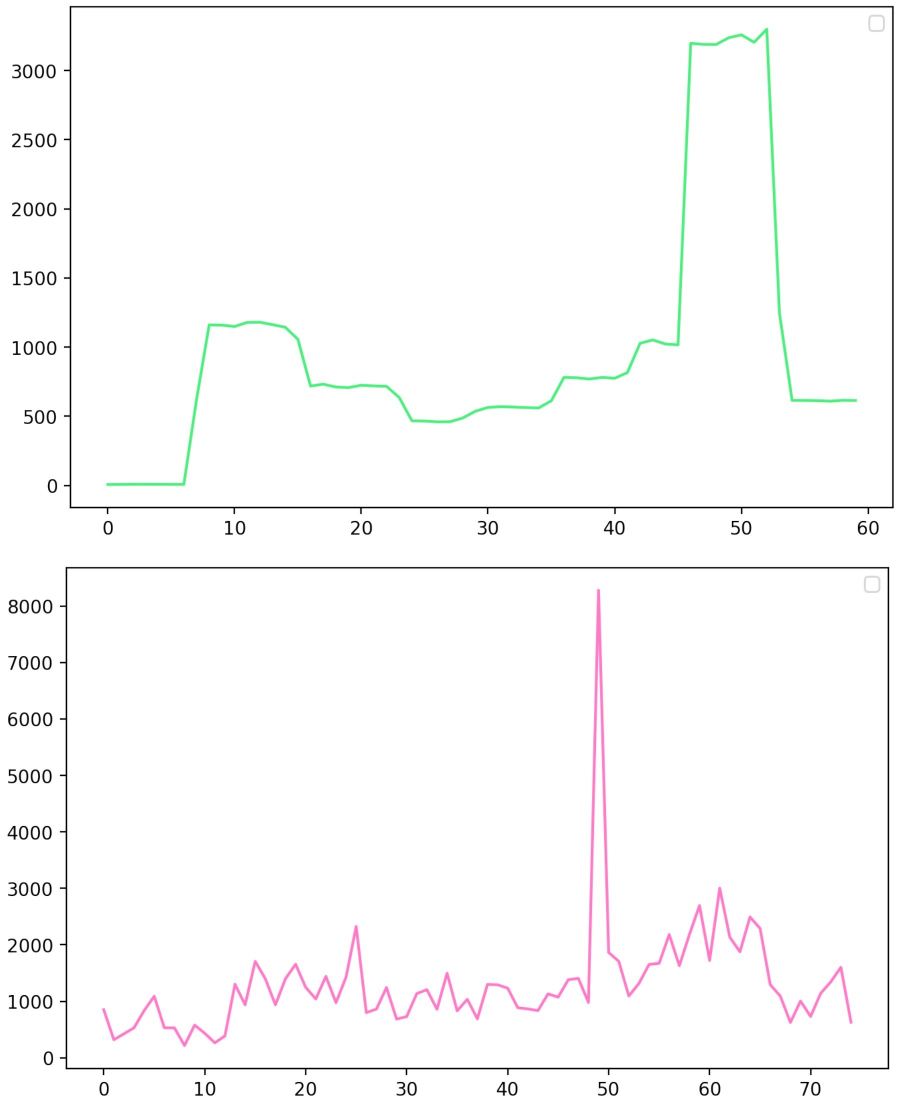

Figura 3: Composições Local e Sharding com alterações não lineares.

4.4 Caso 5

O quinto caso analisa um sistema de dois agentes que tentam configurar o sistema distribuído automaticamente, alternando entre as configurações Local e Sharding em um ambiente de crescimento linear do tamanho da lista.

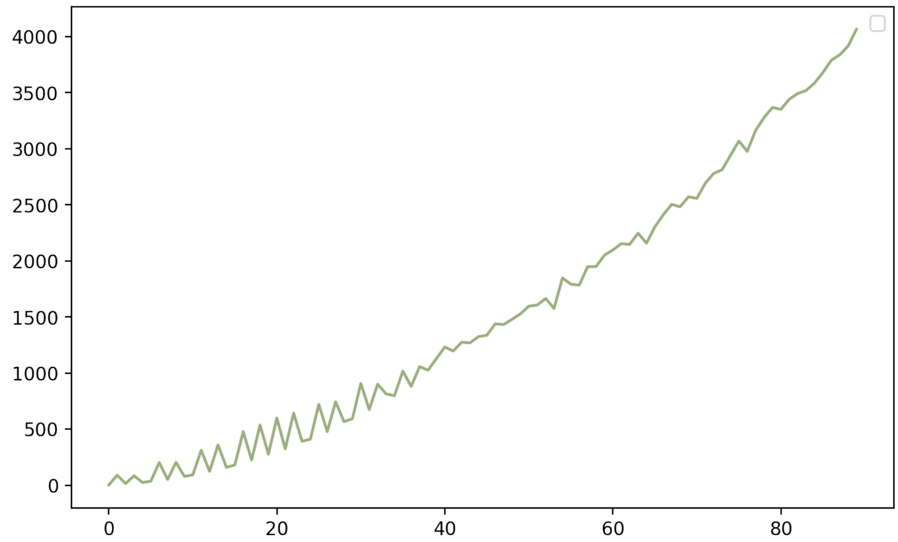

Figura 4: Composições Local e Sharding com crescimento linear.

4.5 Caso 6

O sexto caso analisa o mesmo sistema de agentes do caso 5, mas em um ambiente de decréscimo de tamanho de lista.

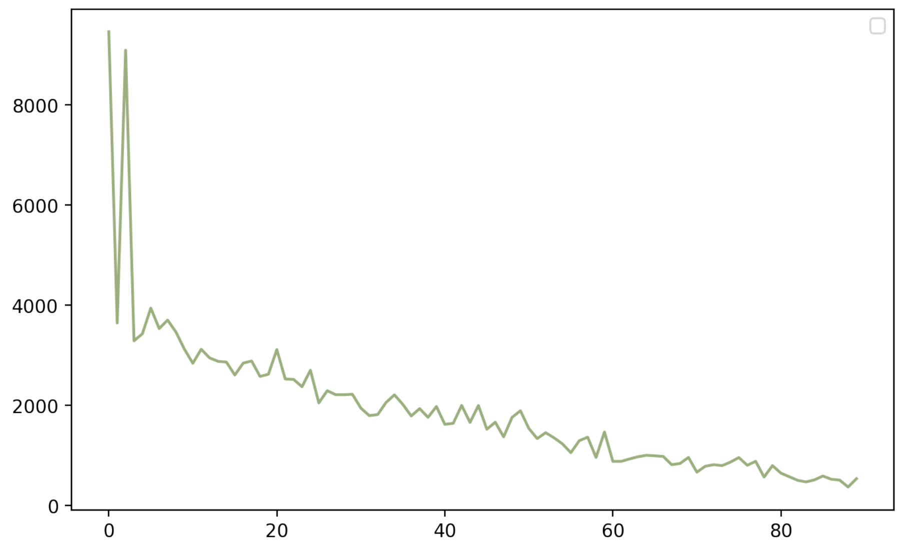

Figura 5: Composições Local e Sharding com descrescimento linear.

5 Resultados

Nesta seção, serão apresentados os gráficos com os resultados da aplicação dos algoritmos de detecção de pontos para os casos 1-4.

5.1 Aplicação do Algoritmo *Offline*

Para o algoritmo *offline*, apenas um gráfico é gerado, o da probabilidade de cada ponto ser um ponto de mudança. O gráfico (em azul) apresenta o resultado do algoritmo, logo abaixo do gráfico do dataset original para comparação.

5.1.1 Caso 1

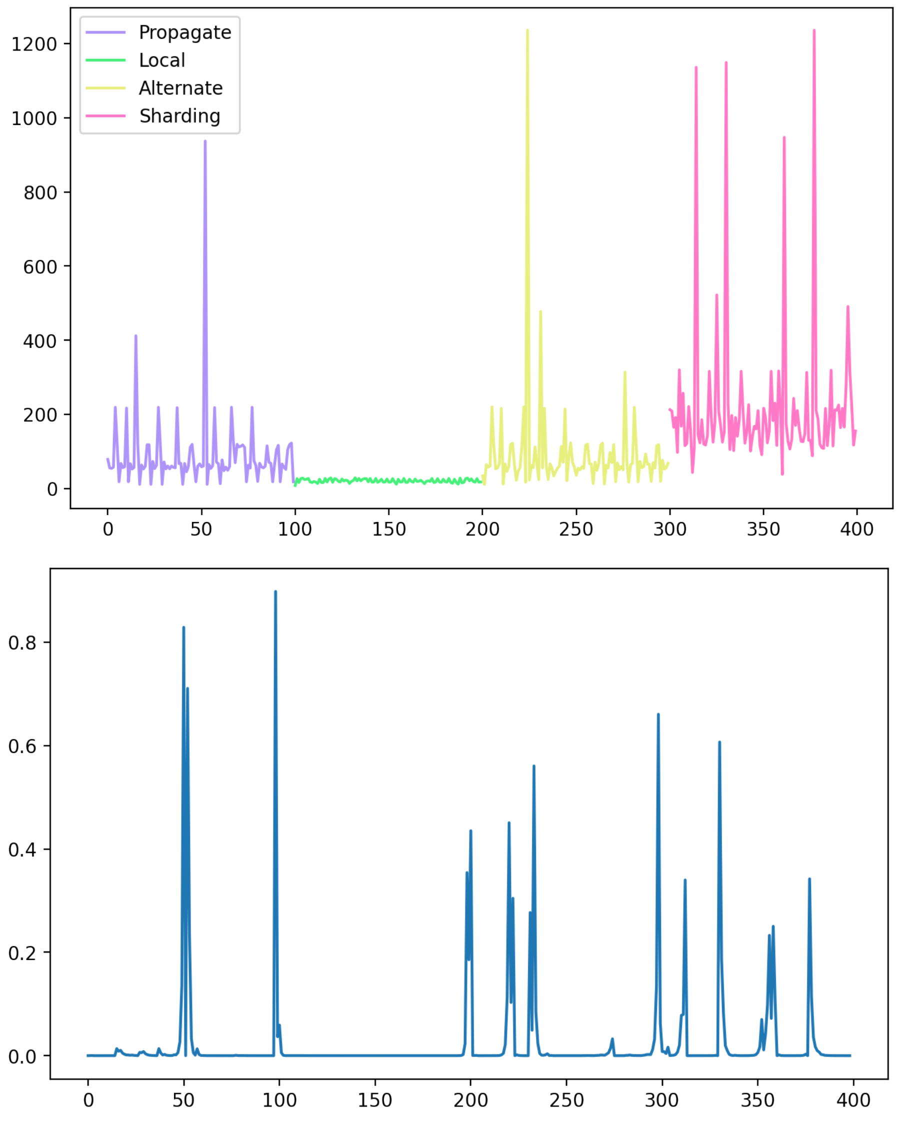

Figura 6: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.1.2 Caso 2

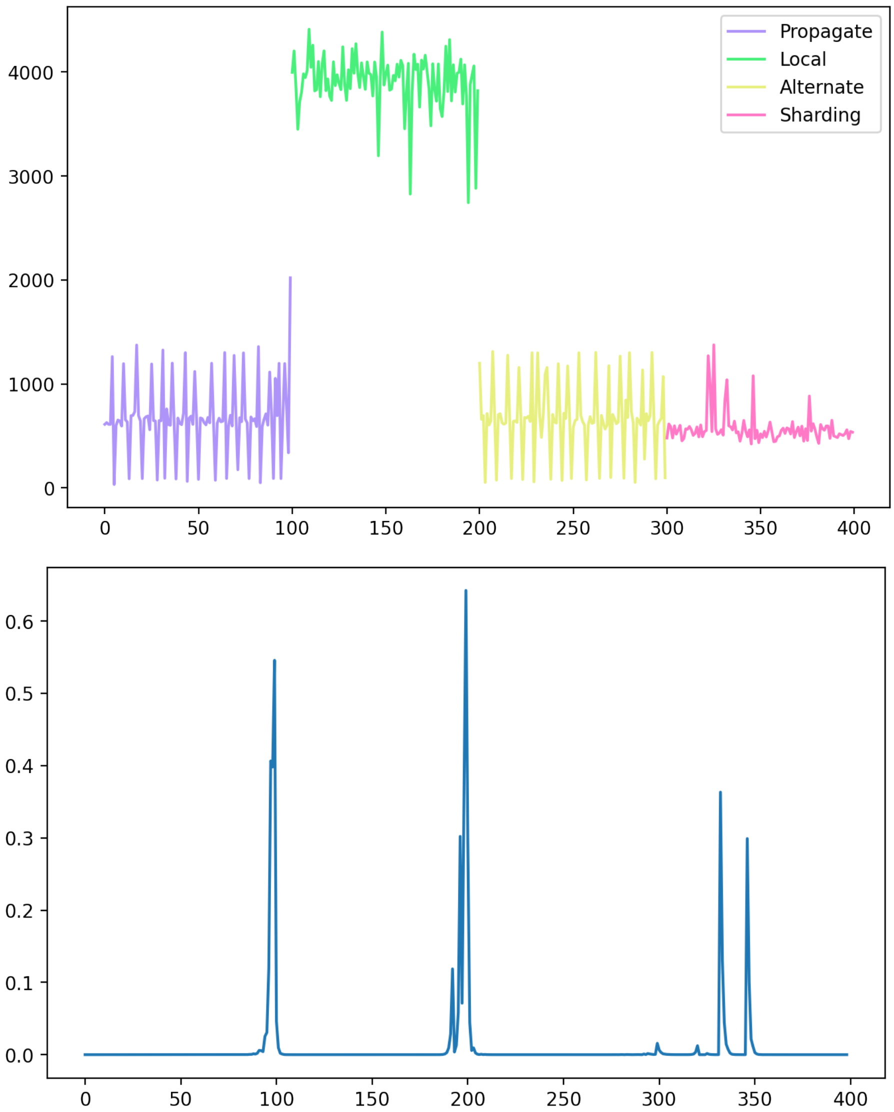

Figura 7: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.1.3 Caso 3

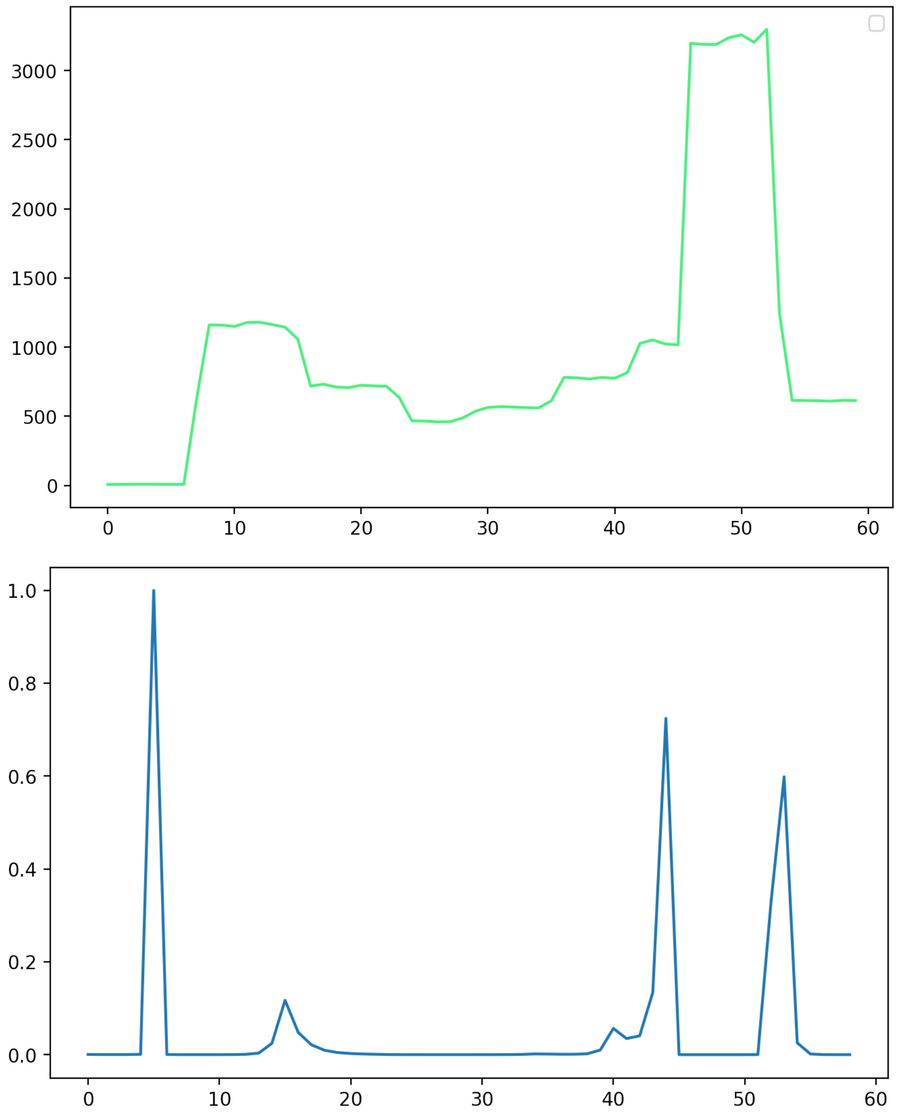

Figura 8: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.1.4 Caso 4

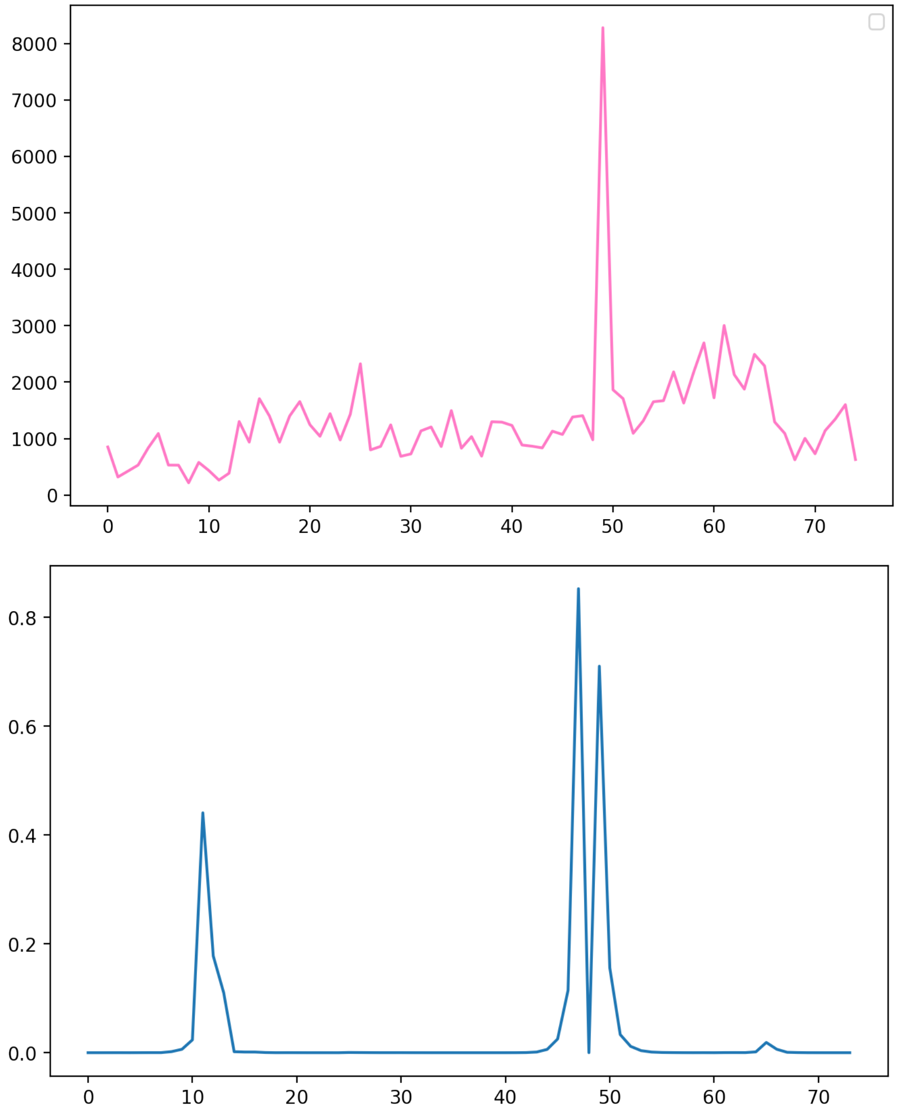

Figura 9: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.1.5 Caso 5

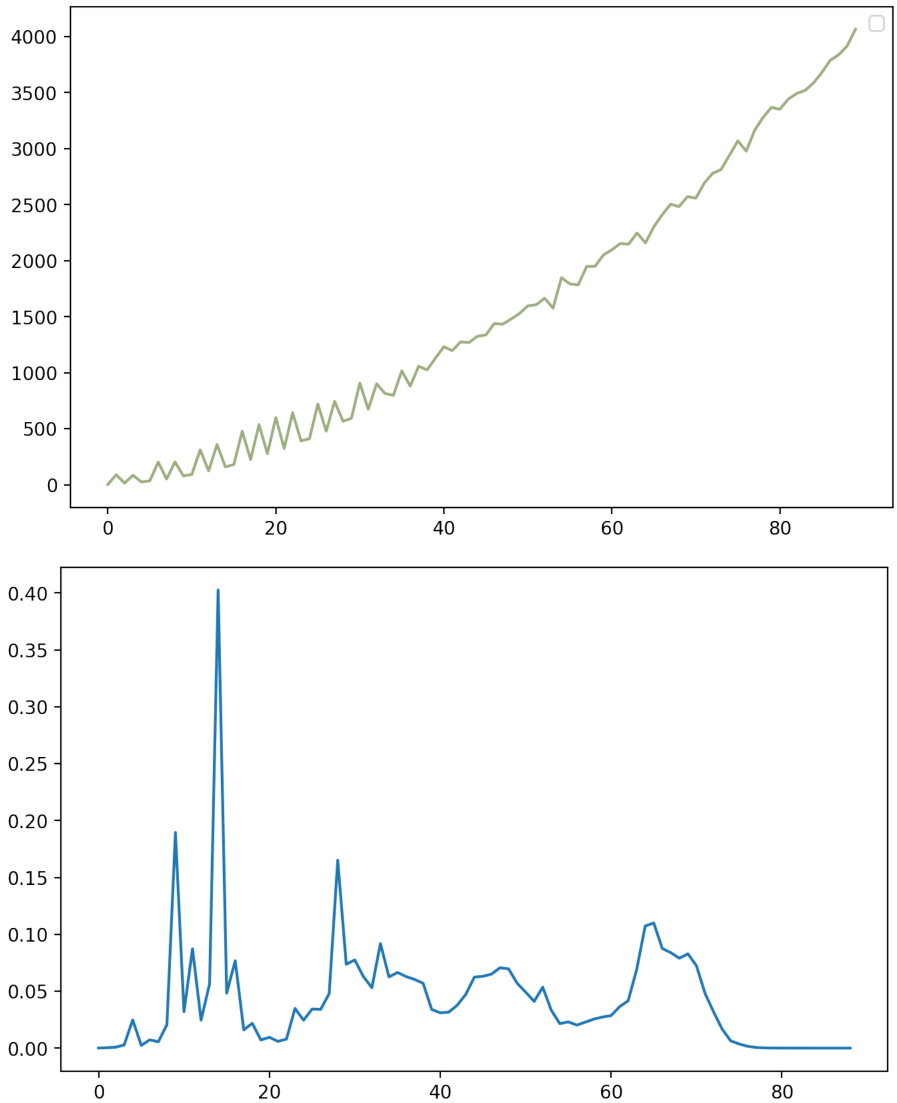

Figura 10: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.1.6 Caso 6

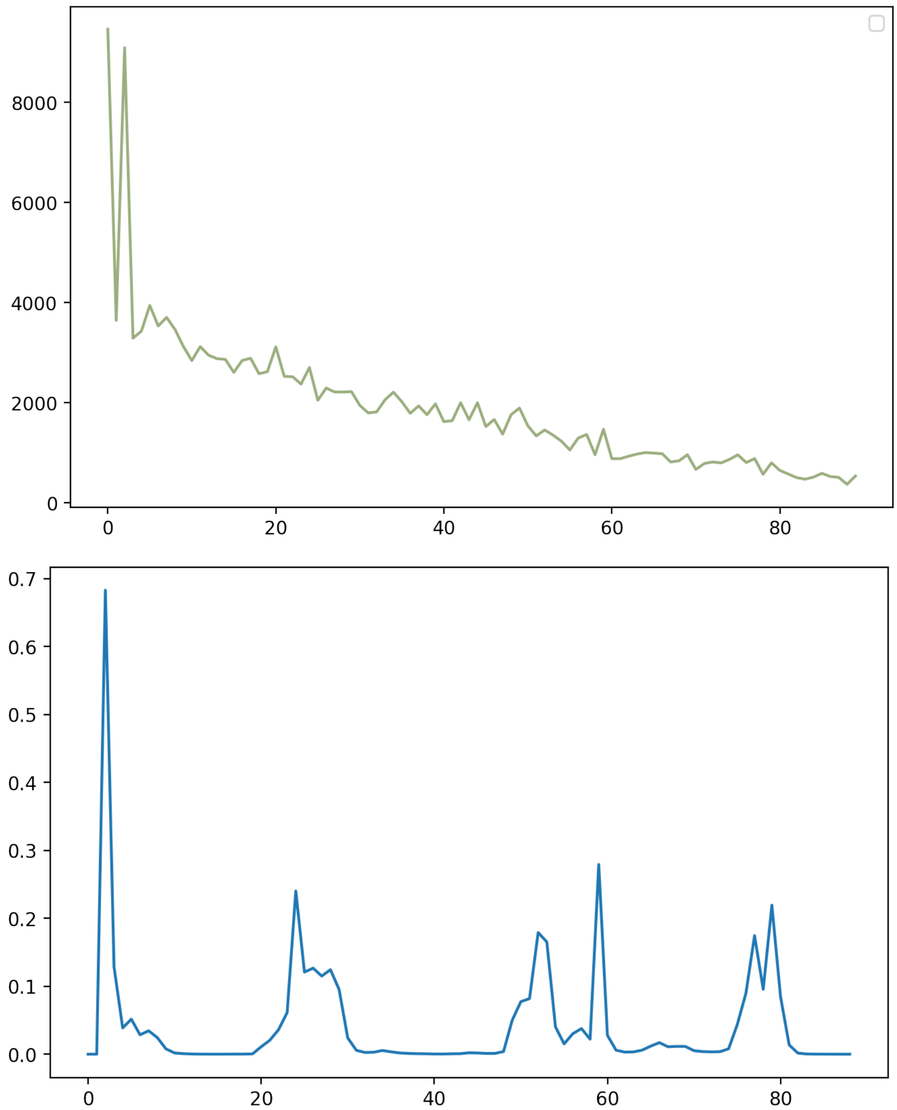

Figura 11: Resultado da probabilidade de um ponto de mudança (azul) em comparação aos dados.

5.2 Aplicação do Algoritmo *Online*

Para a aplicação *Online*, dois gráficos são gerados. Um deles mostra apenas a probabilidade de um ponto de mudança ter ocorrido (gráfico de linha em azul), e outro que apresenta, em escala logarítmica de cor, a probabilidade posterior do *run length* atual e verticalmente o tamanho do r_t.

5.2.1 Caso 1

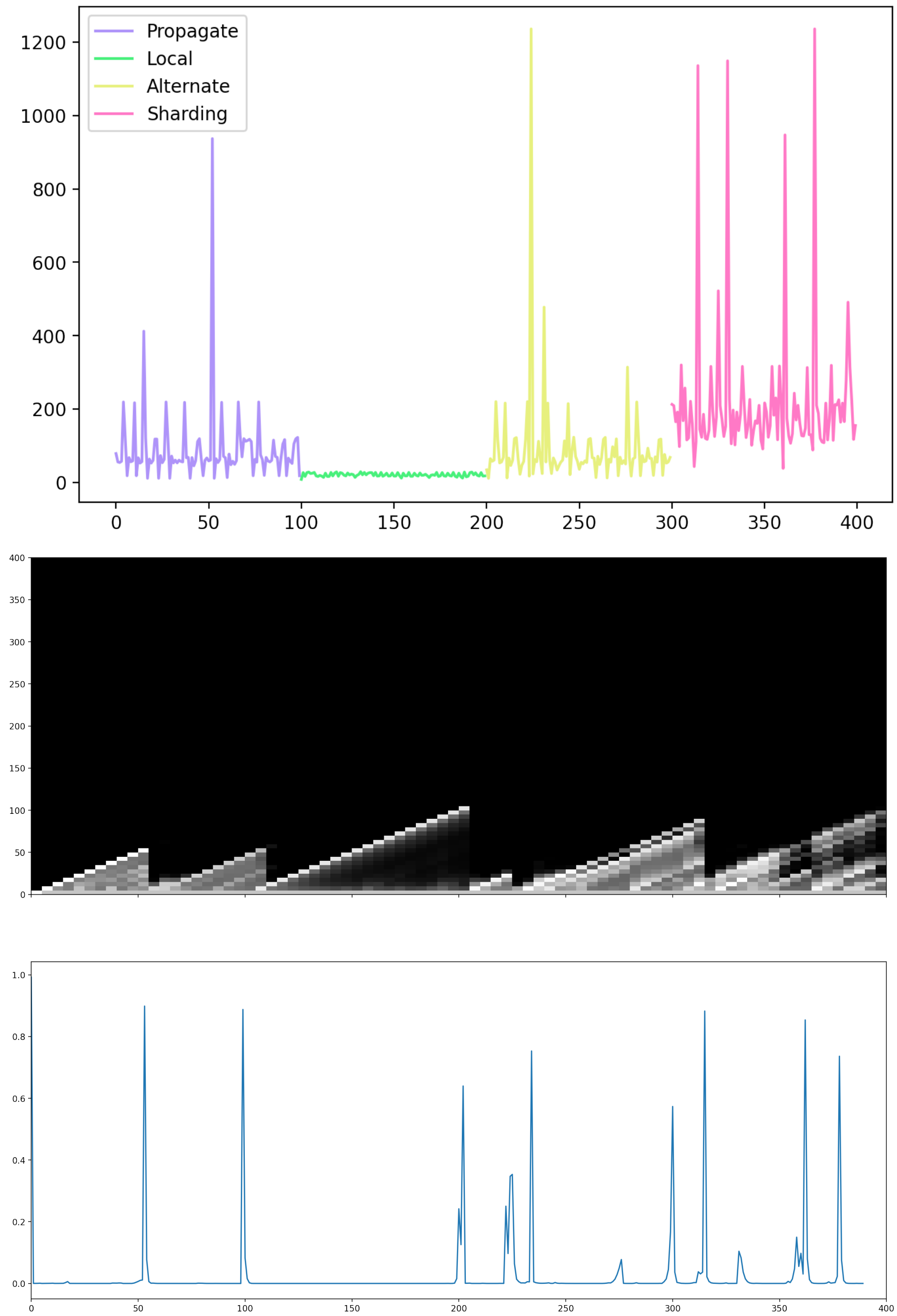

Figura 12: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.

5.2.2 Caso 2

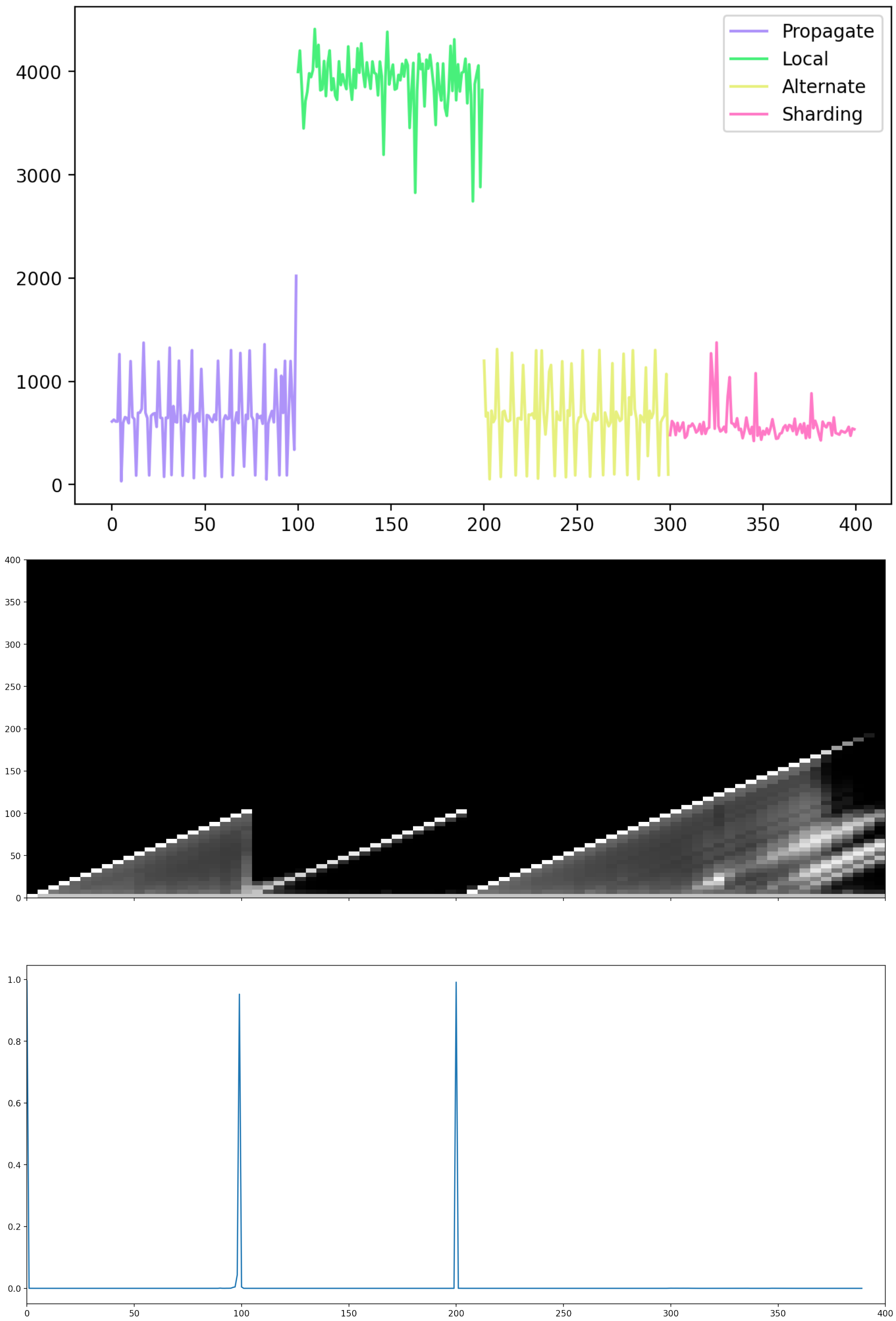

Figura 13: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.

5.2.3 Caso 3

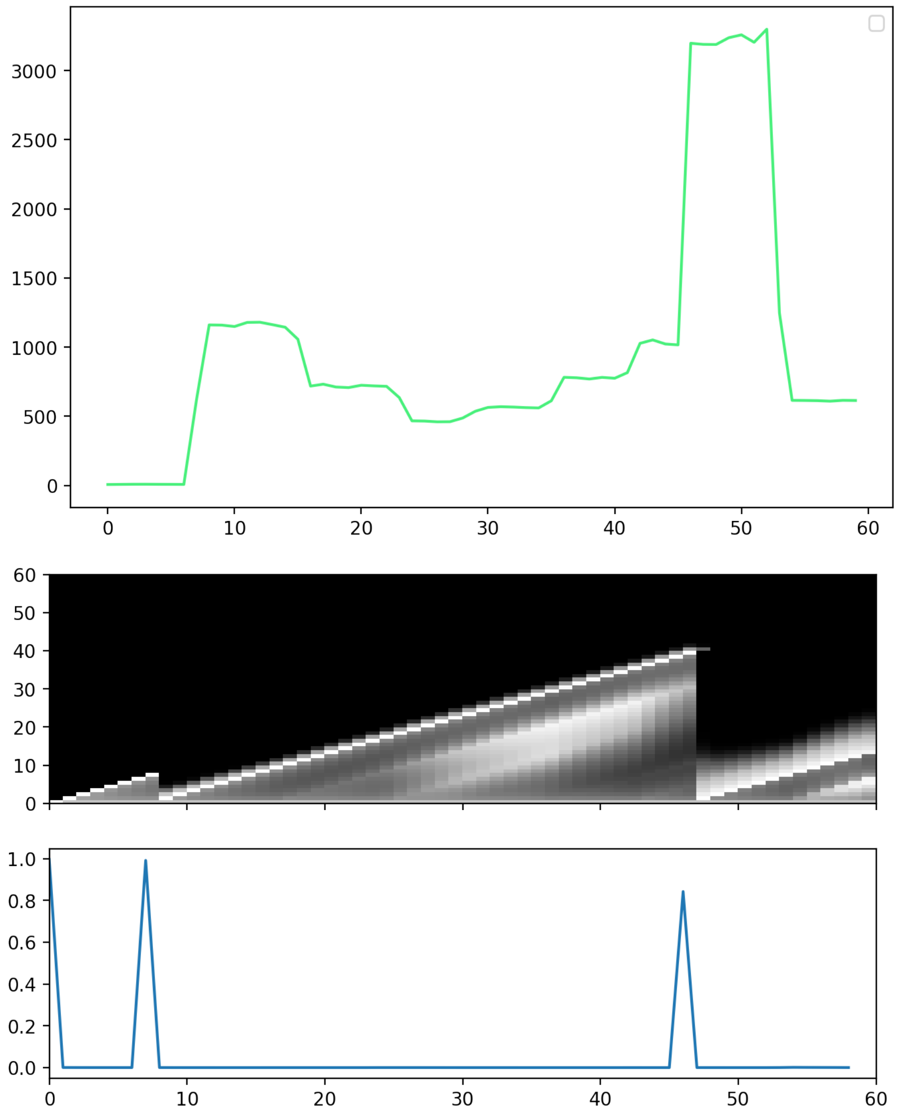

Figura 14: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.

5.2.4 Caso 4

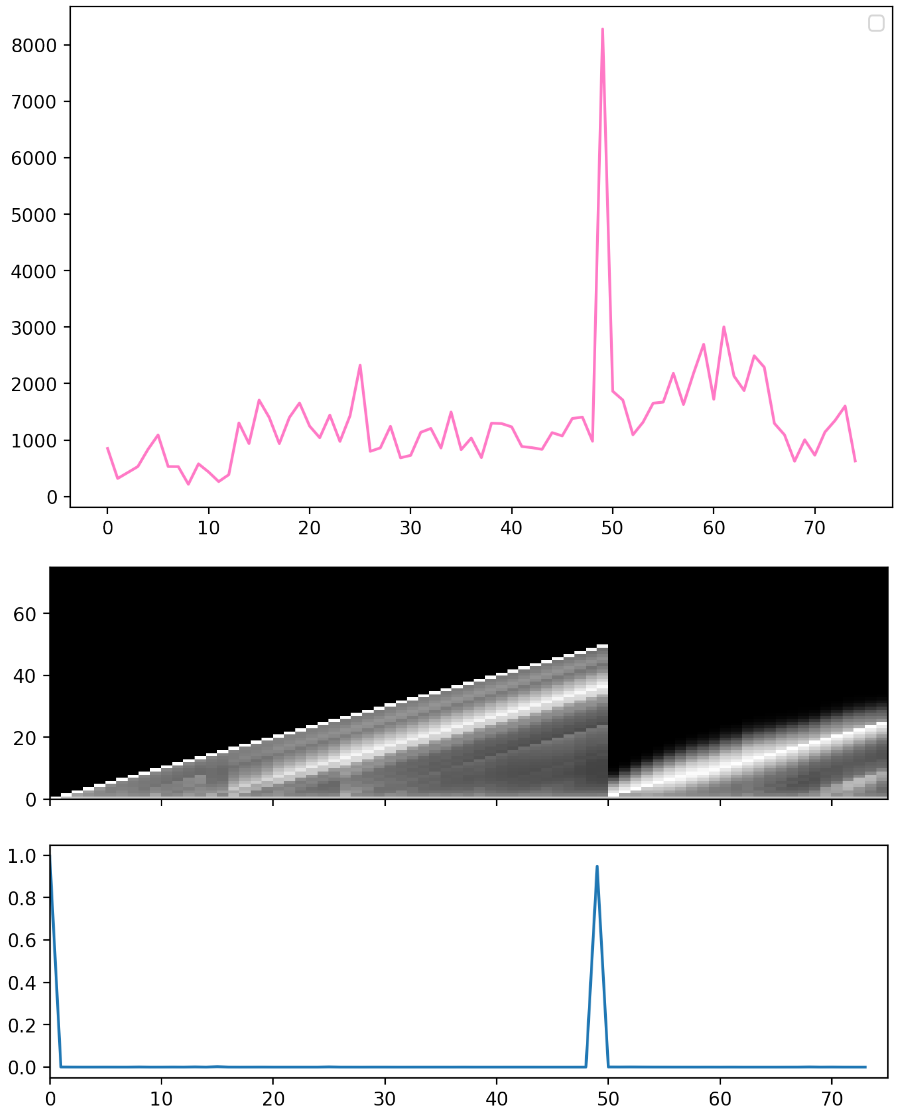

Figura 15: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.

5.2.5 Caso 5

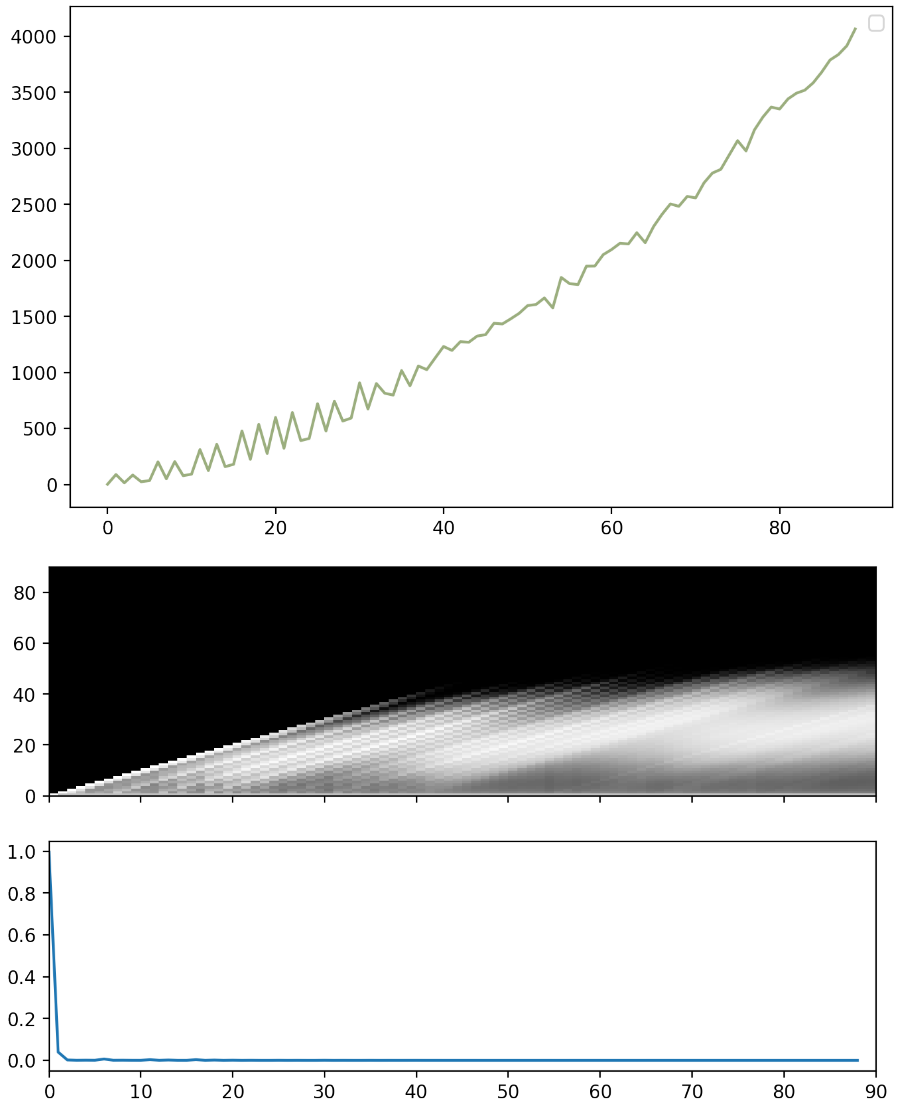

Figura 16: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.

5.2.6 Caso 6

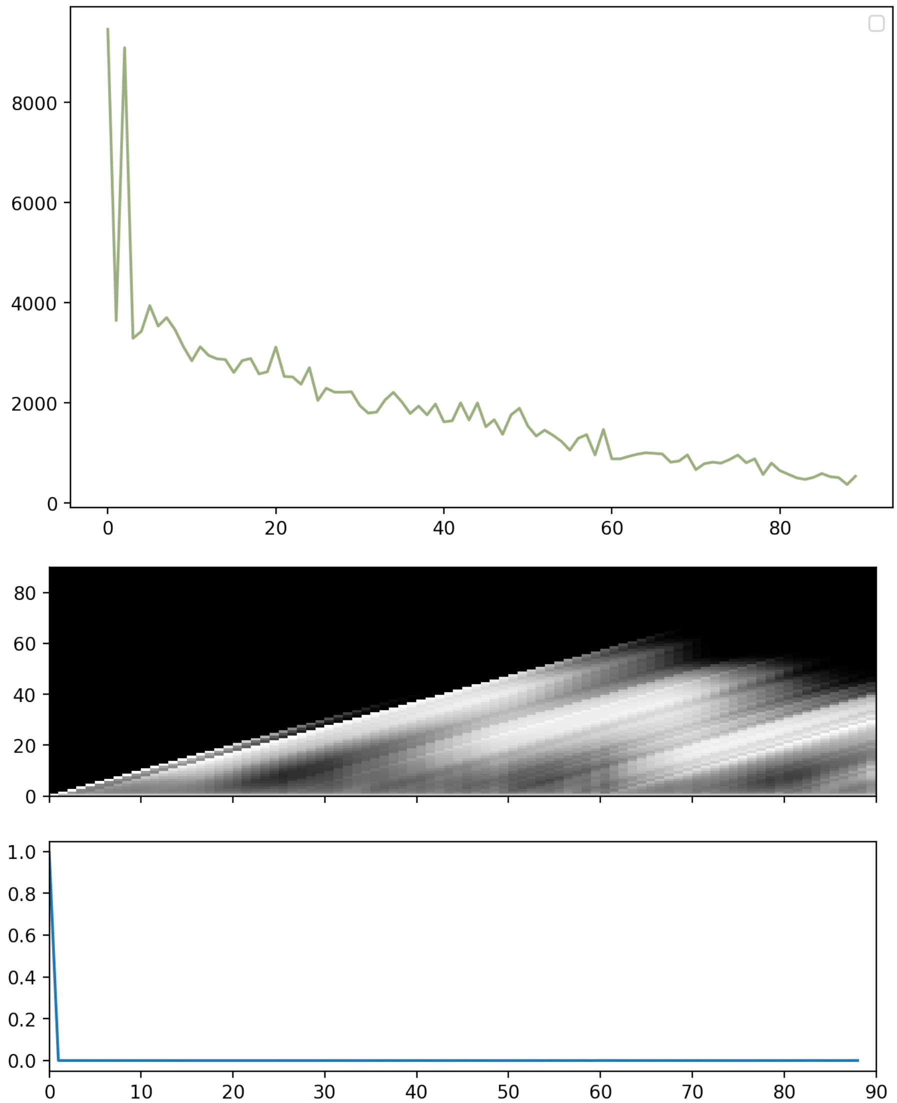

Figura 17: Resultado da probabilidade de um ponto de mudança (azul) e probabilidade posterior do *run length* em comparação aos dados originais.