Mineração de Dados Complexos
Curso de Aperfeiçoamento
INF-0611
Dados Complexos e Recuperação de Informação

Tarefa 2
INF-0611 - Dados Complexos e Recuperação de Informação

Felipe Wolff Ramos
Lucas Aoki Heredia

Prof. Adín Ramírez Rivera

Abril de 2018

Sumário

Dados e procedimentos

Séries temporais

Foram fornecidas as seguintes séries temporais que representam a variação da temperatura (a cada 10 minutos) em Campinas, correspondentes a um dia de verão e inverno do mesmo ano.

```
A <- c(21.7, 21.7, 21.6, 21.6, 21.7, 21.7, 21.7, 21.6, 21.5, 21.5, 21.4, 21.2, 21.2, 21.1, 21.0, 20.9,
20.9, 21.0, 20.9, 20.9, 20.8, 20.7, 20.6, 20.6, 20.5, 20.5, 20.5, 20.5, 20.5, 20.4, 20.3, 20.2, 20.1,
20.0, 20.0, 20.0, 20.0, 19.9, 19.8, 19.8, 19.8, 20.0, 20.3, 20.8, 21.1, 21.7, 22.3, 22.6, 23.0, 23.8,
24.4, 24.8, 24.7, 25.1, 25.8, 26.3, 26.6, 26.5, 27.0, 27.2, 27.6, 27.6, 27.9, 28.1, 28.2, 28.2, 28.6,
29.0, 29.0, 29.1, 29.4, 29.4, 29.5, 29.5, 29.6, 30.1, 30.1, 30.4, 30.2, 30.5, 30.6, 30.4, 30.6, 30.2,
30.4, 30.6, 30.1, 30.2, 30.3, 30.2, 30.3, 30.5, 30.1, 30.0, 30.3, 31.1, 31.2, 31.1, 31.2, 31.3, 31.6,
31.3, 30.8, 30.0, 30.5, 29.9, 29.7, 29.9, 29.2, 28.7, 28.4, 28.2, 26.4, 25.0, 24.4, 23.9, 23.7, 23.7,
23.8, 23.9, 23.9, 23.8, 24.0, 24.1, 24.2, 24.2, 24.1, 24.1, 24.0, 24.0, 24.0, 24.0, 23.9, 23.6, 23.4,
23.4, 23.4, 23.3, 23.2, 23.1, 23.0, 22.9, 22.9, 22.8)
```

```
B <- c(21.4, 21.3, 21.3, 20.9, 20.4, 20.0, 19.8, 19.9, 19.9, 19.7, 20.0, 19.8, 19.7, 20.1, 20.1, 19.9,
19.7, 18.8, 19.0, 18.3, 18.0, 17.5, 17.4, 17.5, 17.7, 18.0, 18.0, 17.5, 17.5, 17.7, 18.1, 18.0, 17.9,
17.6, 17.2, 17.3, 17.5, 17.1, 17.2, 17.5, 17.4, 17.7, 18.0, 18.0, 17.8, 17.7, 17.6, 17.9, 19.3, 20.2,
20.6, 21.6, 22.3, 21.7, 21.5, 21.7, 22.2, 22.4, 22.6, 23.1, 23.4, 24.0, 24.1, 24.5, 24.8, 25.0, 25.7,
25.8, 25.8, 26.4, 26.6, 27.0, 26.8, 26.9, 27.0, 27.3, 27.1, 27.8, 28.0, 28.2, 28.2, 27.9, 27.4, 27.2,
27.2, 27.3, 27.2, 27.1, 27.4, 27.7, 27.4, 27.3, 27.2, 27.7, 27.8, 28.2, 28.0, 27.8, 27.7, 27.7, 27.7,
27.8, 27.5, 26.6, 25.7, 25.0, 24.2, 23.5, 23.2, 22.9, 22.5, 22.3, 22.0, 21.6, 21.3, 21.0, 20.8, 20.4,
20.3, 20.0, 19.7, 19.5, 19.3, 19.1, 19.0, 18.9, 18.7, 18.6, 18.5, 18.4, 18.4, 18.4, 18.4, 18.3, 18.3,
18.4, 18.4, 18.4, 18.4, 18.3, 18.3, 18.3, 18.4, 18.3)
```

Procedimentos

Um programa em R foi desenvolvido para determinar a representação SAX (*Symbolic Aggregate approXimation*), realizando uma redução de dimensionalidade para 24 dimensões mediante o método *Piecewise Aggregate Approximation* ($n = 144$, $w = 24$ e portanto $n/w = 6$).

Foi realizado também a normalização (pela média e desvio padrão) das séries com o objetivo de auxiliar na comparação entre as mesmas. Fórmula de normalização utilizada:

$$x_i = \frac{x_i - \mu}{\sigma},$$

sendo μ e σ a média e desvio padrão das observações, respectivamente.

Após a aplicação do método de redução de dimensionalidade, foram feitas as discretizações das séries utilizando-se da tabela pré-definida (abaixo), a qual indica os valores de quebra das séries, dado um número de símbolos definidos, para que estes sejam equiprováveis.

β_i \ a	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Tabela pré-definida para a quantidade de símbolos e os valores de quebra das séries, correspondentes

Para esta tarefa, foram utilizadas as seguintes quantidades de símbolos: 4, 5, 6 e 7. Os resultados obtidos serão apresentados na próxima seção deste relatório.

Feitas as discretizações das séries para as diferentes quantidades de símbolos, as distâncias entre as séries foram calculadas utilizando-se da seguinte função (MINDIST):

$$MINDIST(\hat{Q},\hat{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} \left(dist(\hat{q}_i,\hat{c}_i)\right)^2}$$

$$cell_{r,c} = \begin{cases} 0, & if \ |r-c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)}, & otherwise \end{cases}$$

Resultados obtidos

Inicialmente, calculamos o MINDIST entre as duas séries dadas, variando a quantidade de símbolos utilizados para discretizar as amostras já normalizadas e reduzidas a 24 dimensões. Os resultados seguem:

- 4 símbolos -> MINDIST = 0
- 5 símbolos -> MINDIST = 1.445199
- 6 símbolos -> MINDIST = 0
- 7 símbolos -> MINDIST = 0.955301

Quanto mais próximo de 0 é o valor de MINDIST, mais próximas são as duas séries consideradas. Os resultados acima mostram que, dependendo do número de símbolos escolhidos, temos como resultado curvas consideradas exatamente iguais, ou não. Para analisar estes resultados de forma mais aprofundada, geramos os gráficos para cada quantidade de símbolos escolhida.

Discretização com 4 símbolos

Começando com a discretização feita com 4 símbolos, logo abaixo, notamos que as curvas diferem mais em suas bordas, tanto no início quanto no fim, sendo que a região central tem alto grau de similaridade. Ao utilizarmos 4 símbolos, foi possível reduzir o impacto dessas diferenças no cálculo da distância, atingindo como resultado o valor 0, o que significa que as curvas foram classificadas como iguais.

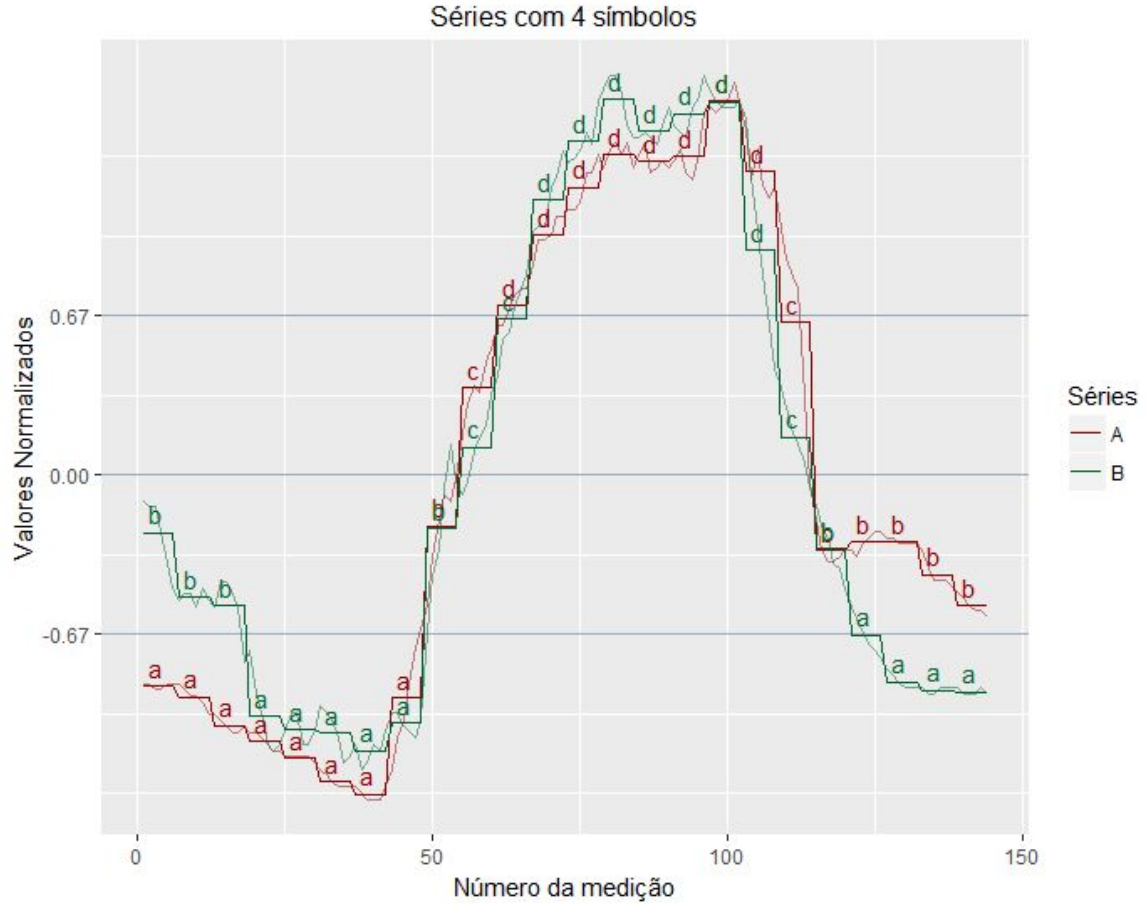

Gráfico comparativo entre as séries A e B considerando discretização com 4 símbolos

Discretização com 5 símbolos

Ao aumentarmos o número de símbolos para 5, obtivemos um resultado diferente de 0. Ele se deveu basicamente a um único ponto, o inicial, que ficou em região limítrofe nesta classificação. O aumento do número de símbolos, neste caso, deu mais relevância as distâncias encontradas nos extremos inicial e final das curvas. É possível ver isso também em alguns pontos no final da curva, que ficaram próximos dos limites e que poderiam ter causado um aumento de distância, caso seus valores fossem ligeiramente diferentes.

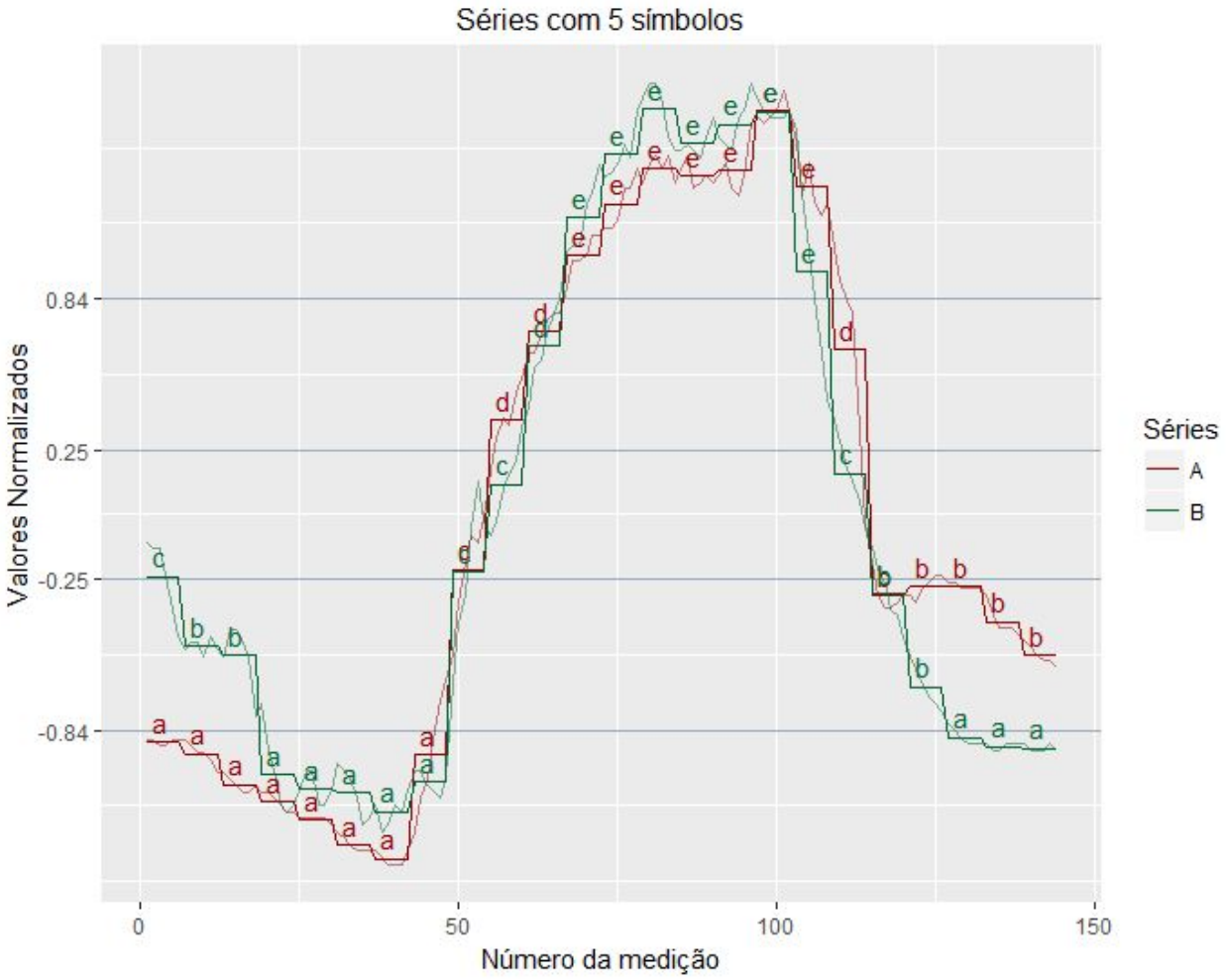

Gráfico comparativo entre as séries A e B considerando discretização com 5 símbolos

Discretização com 6 símbolos

Ao aumentarmos o número de símbolos para 6, voltamos a ter como resultado de distância o valor 0. Neste caso, podemos ver que a nova divisão de regiões favoreceu a diminuição das distâncias entre os pontos. Comparando-se ponto a ponto a classificação obtida no gráfico em questão com o gráfico que utiliza 4 símbolos, vemos algumas diferenças, com alguns pares caindo em regiões distintas em relação a comparação com menor número de símbolos, mas não relevantes o suficiente para considerarmos esta classificação como melhor que a de 4 símbolos em termos de precisão no cálculo da distância.

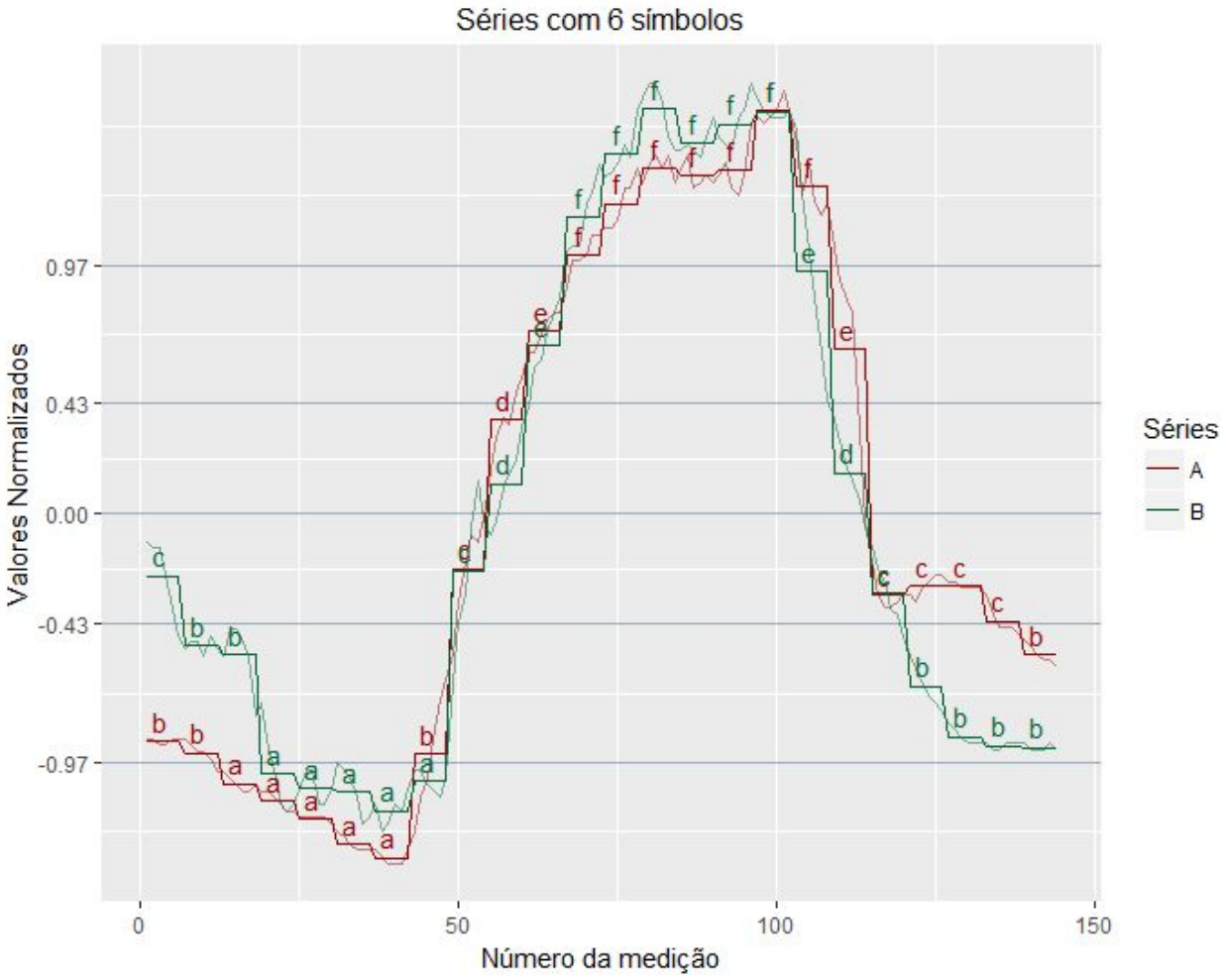

Gráfico comparativo entre as séries A e B considerando discretização com 6 símbolos

Discretização com 7 símbolos

Por último, geramos o gráfico considerando 7 símbolos, que também gerou uma distância diferente de 0. Neste caso, a distância resultante se deveu basicamente à maior sensibilidade em regiões de variações mais acentuadas no eixo y, quando as curvas estão deslocadas entre si no eixo x. Mais uma vez, apenas 1 ponto causou o valor diferente de 0, sendo que o valor da série B estava situado muito próximo a linha de transição para classificação de símbolos.

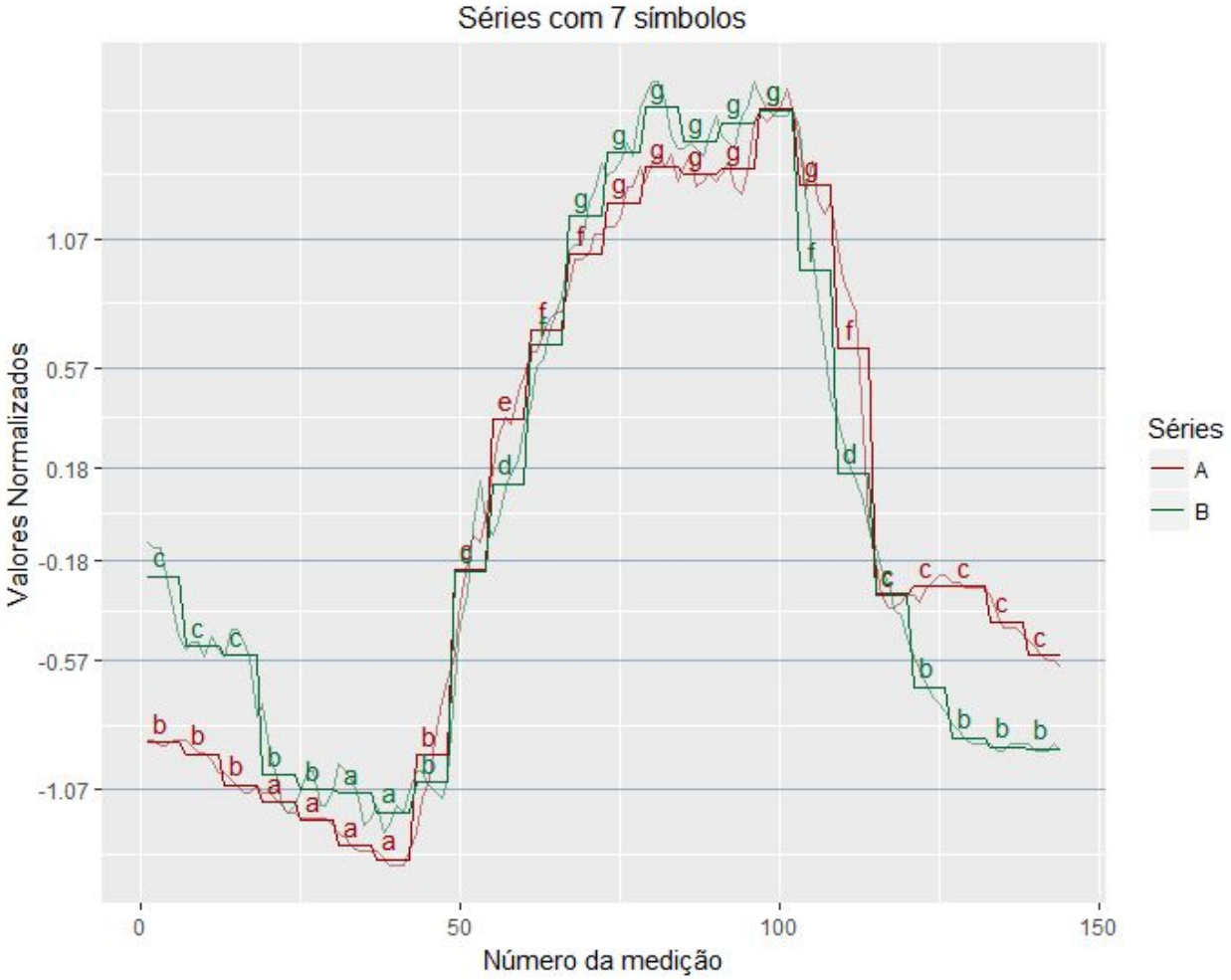

Gráfico comparativo entre as séries A e B considerando discretização com 7 símbolos

Conclusão

Com a análise feita através da utilização de diferentes níveis de discretização, foi possível observar a influência que o aumento no número de símbolos pode ter no resultado final de comparação de distância. Dependendo dos valores a serem comparados, nem sempre o aumento de símbolos trás uma melhor comparação entre duas séries. Ao aumentarmos o número de símbolos, começamos dar uma maior ênfase em variações abruptas e deslocadas entre as duas séries, mesmo que ambas tenham uma forma muito similar, como foi visto no gráfico de 7 símbolos.

Mesmo obtendo-se resultados diferentes de MINDIST para certos números de símbolos, os resultados foram todos muito próximos ou iguais a 0, demonstrando que as duas curvas dadas são muito similares entre si. Isso foi corroborado através da construção dos gráficos, que ajudaram a visualizar o que o MINDIST demonstrou através de um único valor.