UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Física
Bacharelado em Engenharia Física
Tópicos especiais de engenharia física

Deomar Santos da Silva Junior / 00260682

Análise da distribuição por *kernel density estimation* e da densidade espectral de séries temporais

Introdução

As ferramentas matemáticas utilizadas para a análise de séries temporais, modelos preditivos e tomadas de decisão baseadas em testes de hipóteses utilizados em grandes empresas são fundamentados na suposição de que os dados analisados têm determinada distribuição. Entretanto, caso a distribuição suposta esteja incorreta, as tomadas de decisão baseadas em dados podem levar a empresa a uma estratégia errônea [1]. Pode-se inferir a distribuição de determinada série de dados a partir de uma análise não paramétrica de uma amostra por estimação de densidade por kernel, ou seja, a aplicação de uma função de janela que atribui a cada ponto amostral discreto uma função contínua.

Para este trabalho, utilizamos a função de Epanechnikov $K(u)$ como kernel. A função é dada por:

$K(u) = (3/4)(1 - u^2)$ se $|u| \leq 1$ e $K(u) = 0$ para outros valores

Sendo $u = (x - x_n)/h$

De forma que x é o ponto avaliado no eixo x e x_n é o ponto da série temporal de número n.

$K(u)$ possui a seguinte forma:

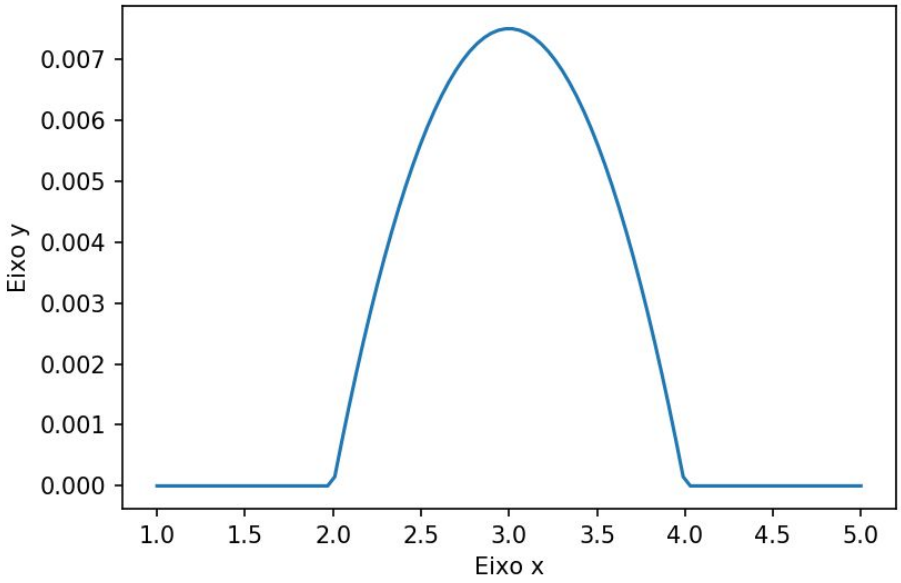

Figura 1. Função Kernel de Epanechnikov.

Para obter a densidade pela aproximação do kernel de Epanechnikov, calcula-se o kernel para cada ponto da série e, então, soma-se a contribuição de cada ponto. Por exemplo, para uma série com os pontos x = 3,4,5, e janela = 2, obtém-se:

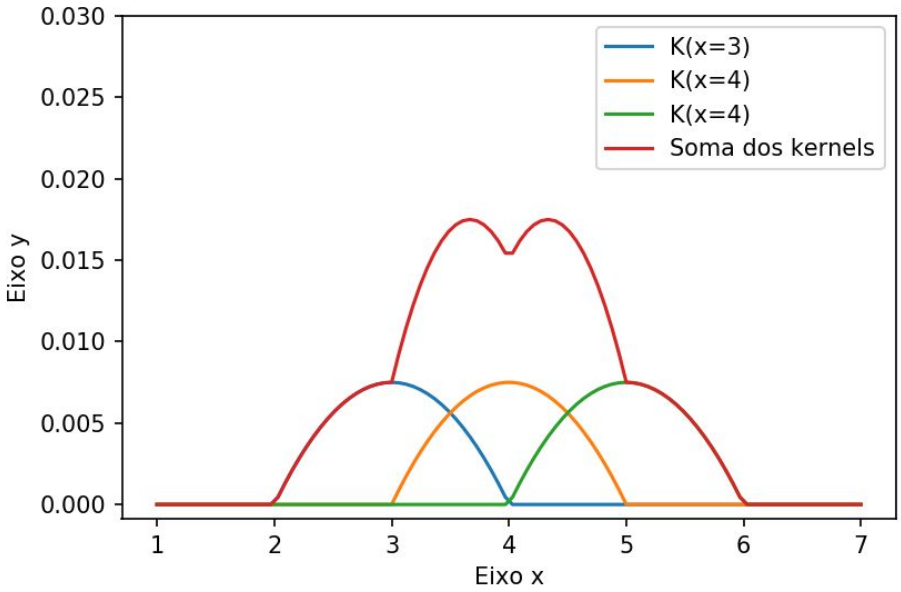

Figura 2. Funções de kernel para 3 pontos e a sua soma.

1. Cálculo da densidade dos dados de uma série temporal

Utilizando-se a série temporal do trabalho anterior do número de passageiros mensais das linhas aéreas dos EUA de 1949 até 1960 após tornar a função estacionária [2] e aplicando-se a função kernel para cada ponto da série, obtém-se:

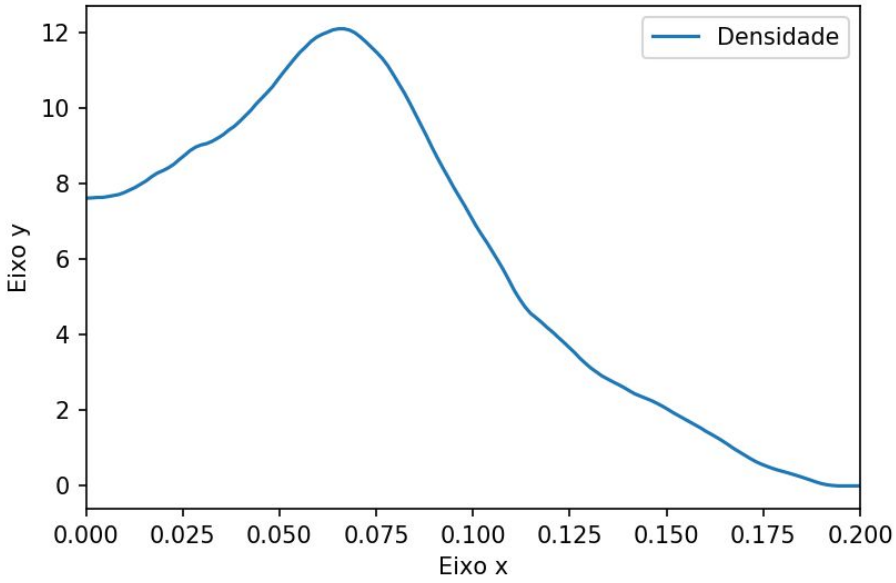

Figura 3. Distribuição da série temporal.

Como pode se observar, a distribuição da série temporal não tem um comportamento gaussiano, como assumido na maioria dos tratamentos matemáticos. Uma simples consequência disso é que a média não é igual à moda e a mediana, como pode ser facilmente assumido erroneamente para a modelagem futura, dentre outras premissas. O ponto máximo da distribuição está em torno de 0.07, o que é condizente com a série temporal que parece ter a maior concentração de pontos em torno do mesmo valor na figura 4.

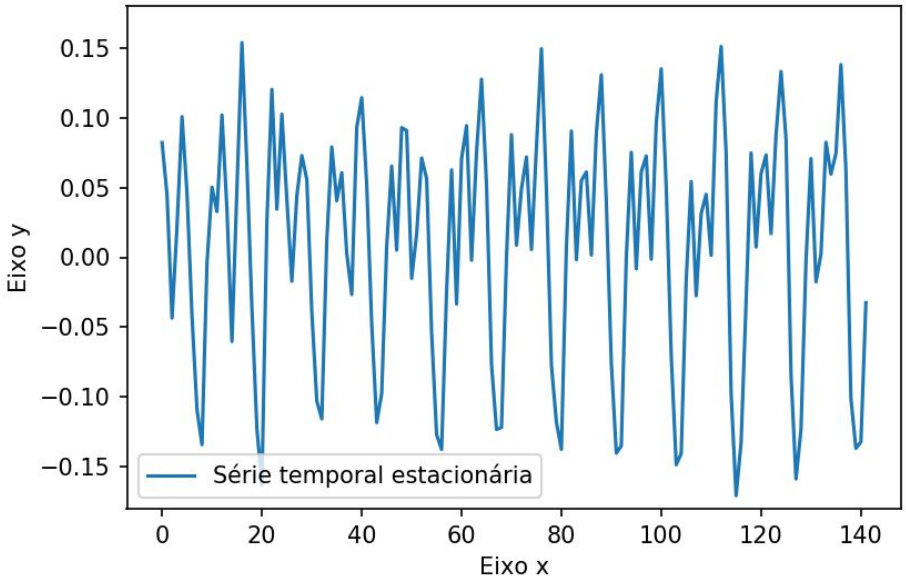

Figura 4. Série temporal.

O significado do eixo y não será explicado, pois a série original sofreu um processo de transformação para se tornar estacionária através da aplicação de médias e funções logarítmicas como mencionado no trabalho 1. Assim como o eixo x que poderia ser entendido como um período entre anos, mas não será mencionado, pois não auxilia na interpretação do presente trabalho.

2. Densidade espectral

Outra técnica utilizada para analisar a série estacionária é a transformada de fourier [3] que transforma o sinal original no domínio do tempo para o domínio de frequência, assim é possível detectar as frequências que compõem o sinal original e tirar conclusões acerca de comportamentos sazonais dos valores no tempo que, através de uma análise mais elaborada, podem ser úteis para a construção de um modelo de previsão adequado.
A transformada de Fourier do tempo 't' para a frequência 'w' é dada por:

$$F(w) = \int_{-\infty}^{\infty} f(t)\, exp(-2\pi i w t) dt$$

Em que i é o número imaginário $\sqrt{-1}$ e w é um número real.

2.1. Transformada Discreta de Fourier (DFT)

Não é necessário calcular a transformada de fourier para todos os ‘w’ possíveis, além de ser custoso e, de fato, impossível computacionalmente. Então calcula-se a transformada de fourier discreta em que a frequência ‘w’ é discretizada da seguinte forma:

$$w = 2\pi ki/N$$

Onde k é a frequência discretizada, n o dado de número n de série e N o tamanho da amostra.

Dependendo da frequência de amostragem e a frequência discretizada ‘k’, pode-se calcular a transformada para um sinal periódico defasado no tempo t, o que ocasiona a introdução de ruído na transformada. Para evitar esse problema, aplica-se uma função janela no sinal original no tempo que, basicamente, seleciona períodos igualmente amostrados no tempo t para diminuir o ruído esse ruído. Para este cálculo utilizou-se a janela de Hann que é dada por:

$$w(n) = sen(\pi n/N - 1)$$

Com n e N dados acima.

Calculando a DFT para a série temporal, convertendo a frequência para Hz e calculando a amplitude do sinal complexo, obtém-se:

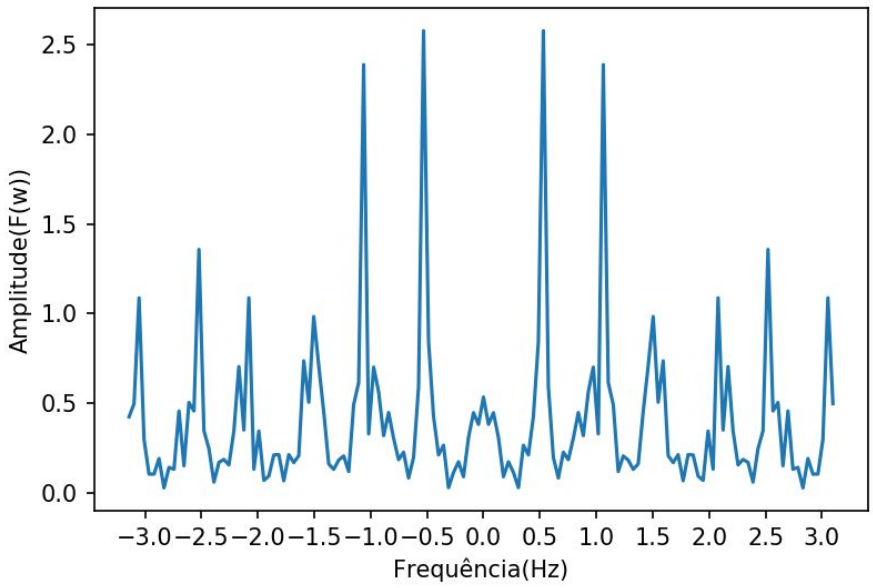

Figura 5. DFT da série temporal.

Percebe-se que o cálculo da DFT tem, no mínimo, 6 picos bem visíveis, ou seja, a série pode ser composta por 6 sinais periódicos de frequências diferentes, contando os seus harmônicos.

Conclusões

Pode-se aplicar duas técnicas de análise para a série temporal que demonstra duas características diferentes e complementares; sendo a distribuição dos valores que identifica o tipo de curva e a densidade espectral que decompõe a série em suas componentes fundamentais periódicas.

Referências:

[1] https://www.cursospm3.com.br/ - Curso online de *Product Manager*. Seção "Como usar dados para tomar decisões." <Disponível em 07/06/2020>

[2] https://www.kaggle.com/chirag19/air-passengers <Disponível em 07/06/2020>

[3] https://en.wikipedia.org/wiki/Fourier_transform <Disponível em 07/06/2020>