DESENVOLVIMENTO DE UM PROGRAMA ESTATÍSTICO PARA ANÁLISE DA CORRELAÇÃO DE KARL PEARSON

DEVELOPMENT OF A STATISTICAL PROGRAM FOR PEARSON CORRELATION ANALYSIS

Alfredo BONINI NETO[1]*

Carolina dos Santos Batista BONINI[2]

RESUMO

Neste trabalho é apresentado o desenvolvimento de uma interface gráfica, ou seja, um programa iterativo, para o estudo da análise de correlação de Karl Pearson entre duas variáveis. O objetivo principal é fazer um programa simples e de fácil entendimento e que possa ser utilizado em diversas áreas de aplicação, inclusive na área de engenharia de Biossistemas. A análise correlacional indica a relação entre 2 variáveis lineares (x, y) e os valores sempre serão entre -1 e +1. O sinal indica a direção, se a correlação é positiva (regressão linear crescente) ou negativa (regressão linear decrescente), e o tamanho da variável indica a força da correlação entre as variáveis x e y. O programa foi desenvolvido no ambiente Matlab que fornece todos os comandos possíveis para o desenvolvimento.

Palavras-chave: Correlação, Programa Iterativo, Interface Gráfica.

ABSTRACT

This paper presents the development of a graphical interface, i.e., an iterative program for the study of the Pearson correlation analysis between two variables. The main goal is make a simple program and easy to understand and that can be [1]used in various application areas, including in the area of Biosystems Engineering. The correlation analysis shows the linear relationship between two variables (x, y) and the values will always be between -1 and +1. The sign indicates the direction, if the correlation is positive (growing linear regression) or negative (decreasing linear regression), and the size of the variable indicates the strength of the correlation between the variables x and y. The program was

[1] Departamento de Matemática - Faculdades de Dracena/UNIFADRA

* alfredobonininetO@hotmail.com

[2] Faculdade de Engenharia de Ilha Solteira - UNESP

developed in Matlab environment that provides all possible commands for development.

Keywords: Correlation, Iterative Program, Graphic Interface

INTRODUÇÃO

Karl Pearson (Britânico) foi um grande contribuidor para o desenvolvimento da estatística como uma disciplina científica séria e independente. Foi o fundador do Departamento de Estatística Aplicada na University College London em 1991; foi o primeiro departamento universitário dedicado à estatística em todo o mundo. O coeficiente de correlação de Pearson é uma medida do grau de relação linear entre duas variáveis quantitativas x e y (STANTON, 2001). Este coeficiente varia entre os valores -1 e 1. O valor de 0 indica que não há relação linear, o valor 1 indica uma relação linear perfeita e o valor -1 também indica uma relação linear perfeita mas inversa, ou seja, quanto mais próximo estiver de 1 ou -1, mais forte é a associação linear entre as duas variáveis x e y. Um fato que atrai pesquisadores aplicados das mais diversas áreas é a possibilidade de obter uma análise de ligação entre as variáveis x_i e y_i em um conjunto de dados, ou seja, (x_1, y_1) (x_2, y_2) (x_3, y_3).....(x_{n-1}, y_{n-1}) (x_n, y_n). A colocação destes pares ordenados num plano cartesiano, depende dos valores de x_i e y_i, (i=1..n) e pode fornecer um gráfico de dispersão, figura 1.

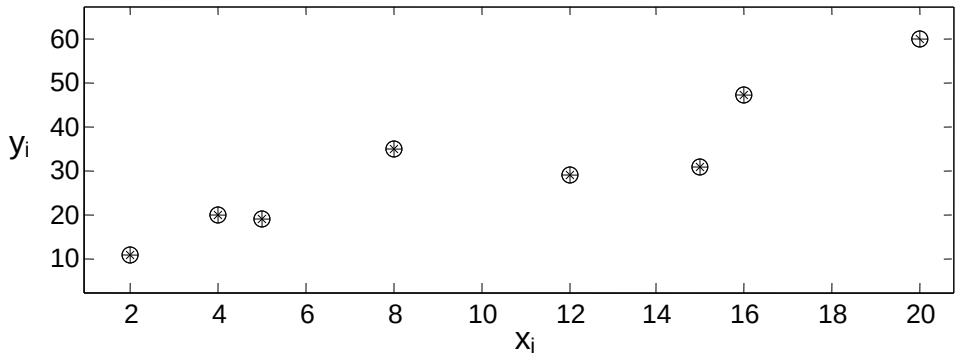

Figura 1 - Diagrama de dispersão

Adotou-se neste trabalho, como forma de análise de possíveis correlações entre as variáveis tratadas os parâmetros propostos por Karl Pearson (STANTON, 2001). O coeficiente de correlação de Karl Pearson é normalmente representado pela letra r e sua fórmula de cálculo é:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{\sqrt{(n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2)(n\sum_{i=1}^{n} y_i^{\,2} - (\sum_{i=1}^{n} y_i)^2)}} \quad (1)$$

Sendo que o resultado obtido pode ser interpretado por meio dos seguintes parâmetros:
0,7 ≤ r < 1 correlação linear fortemente positiva.
0,3 ≤ r < 0,7 correlação linear moderada positiva.
0 ≤ r < 0,3 correlação linear fraca positiva.
r = 0 não existe correlação linear.
-1 ≤ r < -0,7 correlação linear fortemente negativa.
-0,7 ≤ r < -0,3 correlação linear moderada negativa.
-0,3 ≤ r < 0 correlação linear fraca negativa.

Na Matemática existe a Teoria de Interpolação que é a área que estuda tais processos para obter funções que passam exatamente pelos pontos dados, enquanto que a Teoria de Aproximação estuda processos para obter funções que passem o mais próximo possível dos pontos dados.

É óbvio que se pudermos obter funções que passem próximas dos pontos dados e que tenham uma expressão fácil de ser manipulada, teremos obtido algo positivo e de valor científico.

Dentre os processos matemáticos que resolvem tal problema, com certeza, um dos mais utilizados é o método dos Mínimos Quadrados, que serve para gerar o que se chama em Estatística: Regressão Linear ou Ajuste Linear. A curva mai comum utilizada pelos estatísticos é a função do primeiro grau ($y = a_0 + a_1 x$), Aguiar e Moreira, (2009).

O sistema a seguir é formado para obtenção da função do primeiro grau $y = a_0 + a_1 x$:

$$\begin{cases} na_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^{\,2}\right) a_1 = \sum_{i=1}^{n} x_i y_i \end{cases} \quad (2)$$

onde n é o número de ponto dados, e os coeficientes são:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n \quad (3)$$

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 + ... + y_n \quad (4)$$

$$\sum_{i=1}^{n} x_i^{\,2} = x_1^{\,2} + x_2^{\,2} + x_3^{\,2} + ... + x_n^{\,2} \quad (5)$$

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + ... + x_n y_n \quad (6)$$

Existem vários métodos na literatura que resolvem sistemas lineares. Neste trabalho será aplicado o método $\mathbf{A}.\mathbf{X} = \mathbf{B}$, pois a matriz dos coeficientes (A) sempre possui o mesmo numero de linhas e colunas (quadrada), permitindo assim um modo rápido de resolução $\mathbf{X} = \mathbf{A}^{-1}.\mathbf{B}$. Portanto, para resolver o sistema de equações (2) será utilizado o

programa Matlab, através da forma matricial:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^{\,2} \end{bmatrix} . \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

ou ainda,

$$\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^{\,2} \end{bmatrix}^{-1} . \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \quad (7)$$

O SOFTWARE MATLAB

O Matlab é uma linguagem de programação apropriada ao desenvolvimento de aplicativos de natureza técnica. Para isso, possui facilidades de computação, programação e baixo custo, dentro de um ambiente amigável e de fácil aprendizado (Huang, Zhang, 2000). Com o Matlab é possível resolver problemas computacionais mais rápido do que com linguagens de programação tradicionais, como C, C + + e Fortran (Mathworks, 2009). O Matlab foi desenvolvido no início da década de 80 por Cleve Moler, no Departamento de Ciência da Computação da Universidade do Novo México, EUA. As versões posteriores ao Matlab 4.0, foram desenvolvidas na firma comercial MathWorks Inc., que detêm os direitos de autores destas implementações. O Matlab foi originalmente desenvolvido para prover um acesso amigável ao tratamento de vetores e matrizes. Atualmente o Matlab dispõe de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados que auxiliam muito o trabalho do programador. E ainda possui uma vasta coleção de bibliotecas denominadas toolboxes para áreas específicas como: equações diferencias ordinárias, estatística, processamento de imagens, processamento de sinais, finanças, entre outras.

Os recursos instalados também podem ser estendidos pelo usuário através da implementação de funções Matlab (M-files) ou de rotinas escritas em linguagem C ou Fortran.

A necessidade de um programa em língua portuguesa para atender as necessidades das aulas e permitir ao aluno de graduação um primeiro contato com os computadores e programas de computadores relacionados à pesquisa científica. Além de permitir o uso do programa por pesquisadores (alunos e professores) dos cursos de pós-graduação.

Características dos Recursos Gráficos do Matlab

Existem muitos comandos para criação da interface gráfica no Matlab, citaremos alguns dos comandos. Podemos criar uma janela através da função Figure e formatar essa janela através de seus parâmetros (alguns destes parâmetros serão mostrados a seguir). A figura 2 mostra uma janela feita com a função Figure e seus parâmetros devidamente configurados.

```
dx=0.2850;
dy=0.2200;
pos = [(1-dx)*0.5, (1-dy)*0.5, dx, dy];
h0 = figure('Color',[ 0.800 0.800 0.800], ...
        'Units','normalized', ...
        'MenuBar','none', ...
        'NumberTitle','off', ...
        'Position',pos, ...
        'Resize','off', ...
        'name','');
```

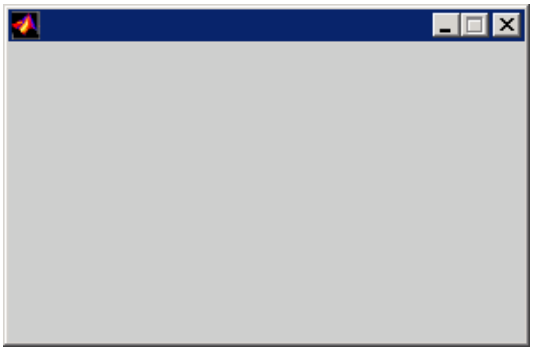

Figura 2: Exemplo da função F*igure.*

Parâmetros da Função Figure:

Color: Representa a cor de fundo da janela. É um vetor com os componentes RGB. Exemplo: a seqüência Color [0 0 0] equivale a cor preta e a seqüência Color[1 1 1] equivale a cor branca.

Units: É uma unidade usada para posicionar o controle. A posição e tamanho de um controle dentro da janela, que são feitos através de coordenadas como: Normalized (máximo e mínimo da janela correspondendo a 0 e 1) e Pixels(pontos gráficos).

MenuBar: Se o valor dessa propriedade for ‘none’ nenhum menu é mostrado na janela. Se for ‘Figure’ a janela terá o menu padrão de figuras.

NumberTitle: Se o valor dessa propriedade for ‘on’ aparecerá o nome e o número da janela. Se for ‘off’ a barra de título aparece em branco.

Position: Especifica a posição e tamanho da janela através das propriedades: [left, bottom, width, height].

Resize: Se estiver em ‘on’ a janela pode ter seu tamanho alterado. Se tiver em ‘off’ o tamanho da janela não pode ser alterado.

Name: Dá um nome para a janela. O valor desta propriedade deve ser uma string.

Controlando os Controles

No *Matlab* existe uma maneira muito prática de se programar a resposta de um controle ao usuário. Por exemplo, ao apertar-se um botão queremos que seja plotado um gráfico, ou fechar a janela que se está operando. Os controles também servem para retornar algum valor para o usuário de maneira mais amigável.

O Comando Uicontrol

O *Uicontrol* é um comando de controle para a janela que está ativa. Para criar os controles, deve-se configurar apropriadamente seus parâmetros. A figura 3 apresenta o exemplo de *BackgroundColor*

```
h0 = figure('Color',[ 0.800 0.800 0.800], ...
            'Units','normalized', ...
            'MenuBar','none', ...
            'NumberTitle','off', ...
            'Position',pos, ...
            'Resize','off', ...
            'name','');
h1 = uicontrol('Parent',h0, ...
           'Units','normalized', ...
           'BackgroundColor',[ 1 1 1], ...
           'ForegroundColor',[0.000 0.000 0.502],...
           'HorizontalAlignment','center', ...
           'Position',[0.0787  0.7678  0.8287  0.1547],
...
           'String','Sistemas:', ...
           'FontSize',17,...
           'Fontname','Arial',...
           'Style','text', ...
           'Tag','StaticText1');
```

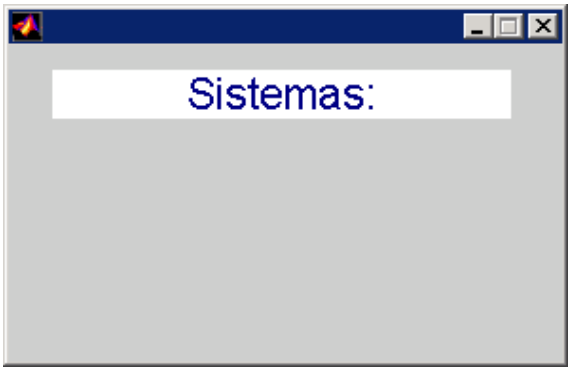

Figura 3 - Exemplo do *BackgroundColor*

O PROGRAMA DESENVOLVIDO E RESULTADOS

Utilizaremos o programa desenvolvido para o cálculo da correlação de Pearson nos dados apresentados na tabela 1 a seguir. A tabela 1 apresenta os dados de precipitação de chuva no ano de 2010 na cidade de Selvíria/ MS. Foram computados os valores em milímetros (mm) durante os 12 meses, de janeiro a dezembro (EMIS, 2010).

Tabeal 1 - Precipitação em 2010 na cidade de Selvíria/MS (em mm)

Ano de 2009	**mm de chuva mensal**
Janeiro	240
Fevereiro	180
Março	210
Abril	50
Maio	25
Junho	10
Julho	0
Agosto	0
Setembro	100
Outubro	140
Novembro	150
Dezembro	180

Para calcular a correlação de Pearson foi utilizada a fórmula da equação (1), onde x_i é o vetor representando os meses, variando de 1 (janeiro) até 12 (dezembro) e o vetor y_i representa o volume de chuva em mm. O objetivo deste trabalho é verificar o grau de correlação entre as variáveis x e y, ou seja, se há um grau de correlação forte entre as variáveis. A figura 4 apresenta a tela inicial do programa. Ao clicar no botão "CÁLCULO DA CORRELAÇÃO" outra tela é aberta (figura 5), esta tela representa a figura de entrada de dados do exercício.

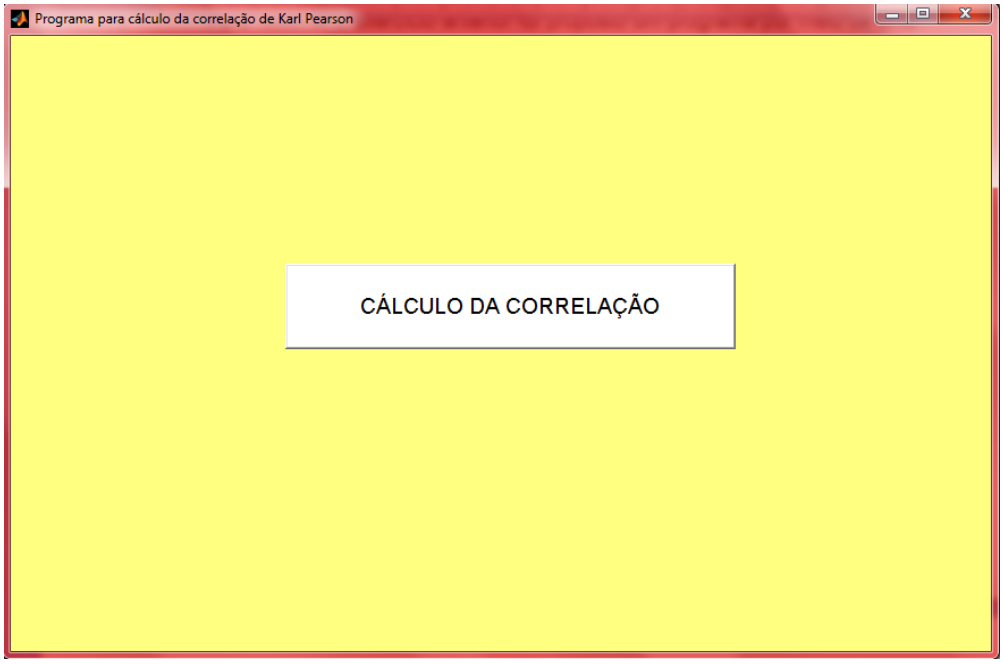

Figura 4 - Tela inicial do programa.

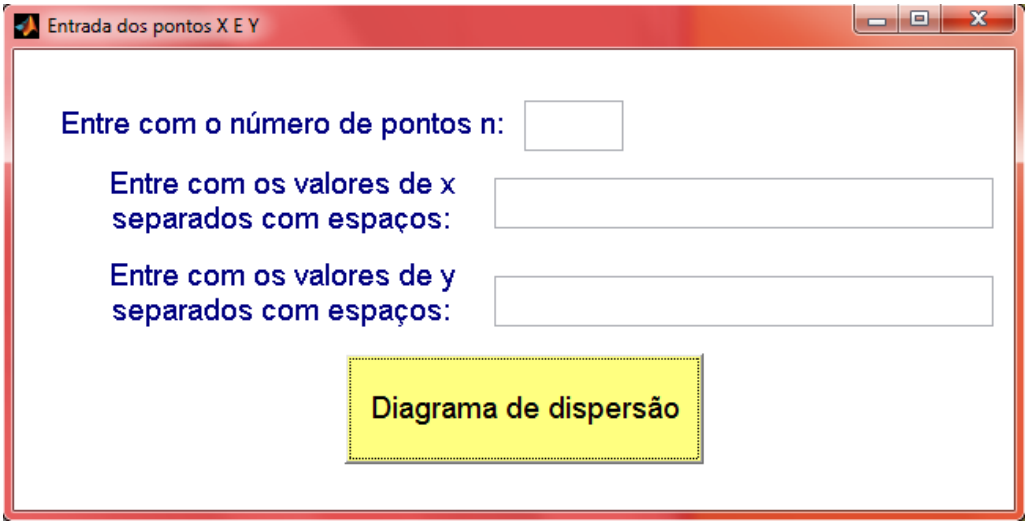

Figura 5 - tela de entrada de dados.

Apresenta-se na figura 6, a mesma figura 5 com os valores de entrada de dados para a aplicação apresentada, foram 12 pontos (n=12), com 12 valores para x representando os meses e 12 valores para y representando o volume da precipitação de chuva nesses 12 meses em milímetros. Para montar o diagrama de dispersão basta o usuário clicar no botão “DIAGRAMA DE DISPERSÃO”. A figura 7 representa o diagrama de dispersão para os pontos da aplicação.

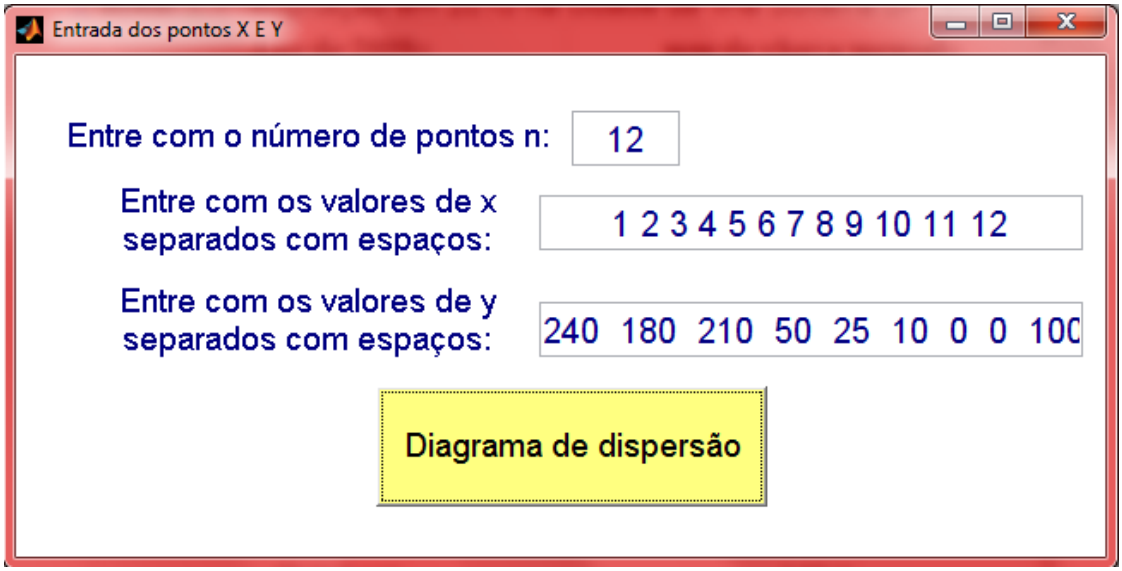

Figura 6 - tela de entrada de dados.

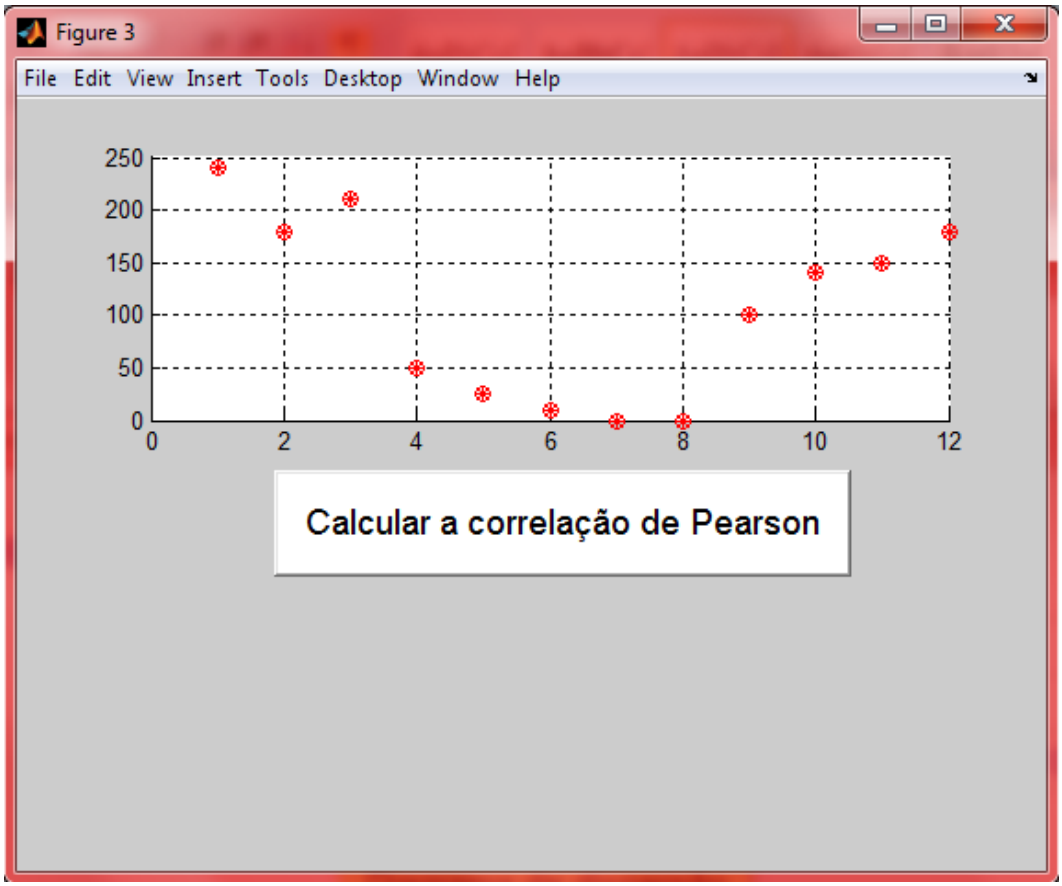

Figura 7 - Diagrama de dispersão da aplicação (precipitação VS meses).

Ao clicar no botão “CALCULAR A CORRELAÇÃO DE PEARSON” aparecerá no gráfico de dispersão a reta de regressão e abaixo do gráfico aparecerá o valor da correlação de Karl Pearson e em que faixa de

correlação se encontra estes valores, como visto os parâmetros no início do artigo, figura 8. O valor da correlação foi de -0.1808, uma correlação linear fraca negativa.

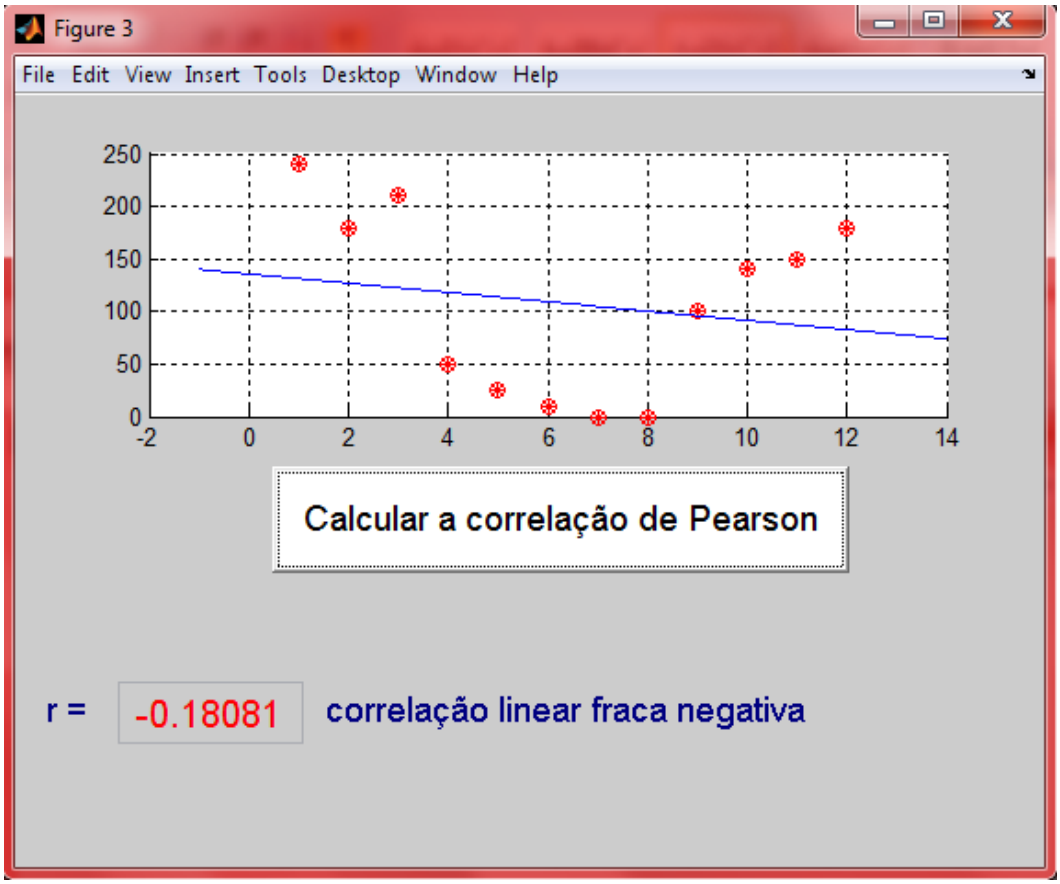

Figura 8 - Diagrama de dispersão e valor da correlação (precipitação VS tempo).

CONCLUSÕES

Neste trabalho foi apresentado o método do cálculo de correlação de Karl Pearson desenvolvido a partir de uma interface gráfica. Foi utilizado o software Matlab para criação da interface gráfica tornando o programa mais iterativo para o usuário. O programa também auxiliou na resolução do sistema de equações lineares do método dos Mínimos Quadrados para obtenção da reta de regressão. Por fim, foi apresentado ao trabalho uma aplicação da correlação de Karl Pearson para dados de precipitação de chuva na região de Selvíria MS no ano de 2010. Dos resultados obtidos, verificou-se que houve uma correlação fraca entre os dados, o que é esperado, pois há meses que apresentam uma precipitação em torno de 200 mm e outros meses com 0,0 mm devido à estiagem, ou seja, alta dispersão de dados. Neste trabalho também foi apresentado o diagrama de dispersão.

REFERÊNCIAS BIBLIOGRÁFICAS

STANTON, J. M. (2001), Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. Journal of Statistical Education, 9,3. Disponível em: <http://www.amstat.org/publications/JSE/v9n3/stanton.html>.

AGUIAR, F. L. E MOREIRA, W. I. Ajuste de curvas por quadrados mínimos lineares. Disponível em: < http://www.mat.ufmg.br/gaal/aplicacoes/quadrados_minimos.pdf>. Acesso em: 06 out. 2009.

HUANG, G.M.; ZHANG, H. A. New Education Matlab Software for Teaching Power Analysis that Involves the Slack Bus Concept and Allocation Issues. Power Engineering Society Winter Meeting. IEEE, v. 2, 23-27, p. 1150- 1158, Jan. 2000.

MATHWORKS. Disponível em: < http://www.mathworks.com>. Acesso em: 15 mar. 2009.

EMIS – Estação Metereológica de Ilha Solteira, 2010. Disponível em: <www.agr.feis.unesp.br/clima.phpe>.