Mineração de dados para classificação de Comportamento anêmico em doadores de sangue

IME - Instituto de Matemática e Estatística

Aluno: André Henrique Serafim Casimiro
Orientador: João Eduardo Ferreira

FUNDAÇÃO PRÓ-SANGUE HEMOCENTRO DE SÃO PAULO

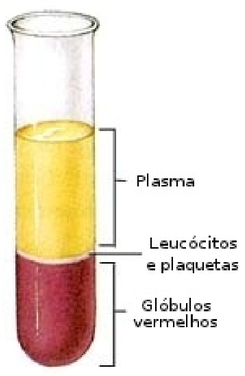
1. INTRODUÇÃO

O grupo de banco de dados do IME (DATA-IME, http://www.data.ime.usp.br) atua junto a 3 hemocentros brasileiros (SP, MG e PE) auxiliando-os a gerenciar e integrar a enorme base de dados provenientes das doações de sangue em cada um deles.

Um tema preocupante no processo de doação de sangue diz respeito ao risco que os doadores correm de **desenvolver anemia por causa das doações**. Essa correlação ainda não é muito bem entendida pelos médicos e varia muito de doador para doador. Este trabalho expõe uma forma de visualizar os dados das doações para tentar prever possíveis casos de anemia nos doadores.

2. HEMATÓCRITO E ANEMIA

Hematócrito (abrevia-se HT) é a medida utilizada na identificação de anemia. Corresponde a **porcentagem ocupada pelos glóbulos vermelhos**, ou hemácias, no volume total de sangue. Os valores médios são diferentes segundo o sexo e idade, e variam de 42% a 52% nos homens e de 36% a 48% nas mulheres. [1]



Hemoglobina é uma metaloproteína presente nas hemácias que contém ferro e permite o transporte de oxigênio pelo sistema circulatório. [1]

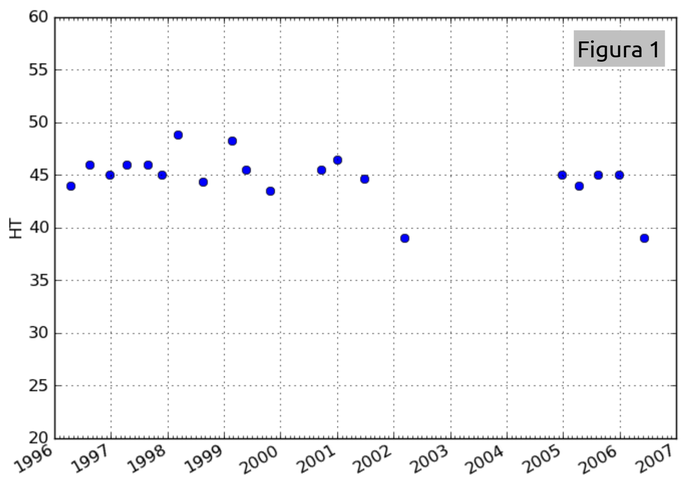
Anemia é a doença caracterizada pela capacidade diminuída de transporte de oxigênio devido à **diminuição da contagem de glóbulos vermelhos** no sangue. Uma pessoa é considerada anêmica se o nível de hematócrito é menor do que 39% em homens, ou 38% em mulheres.

3. SÉRIES TEMPORAIS

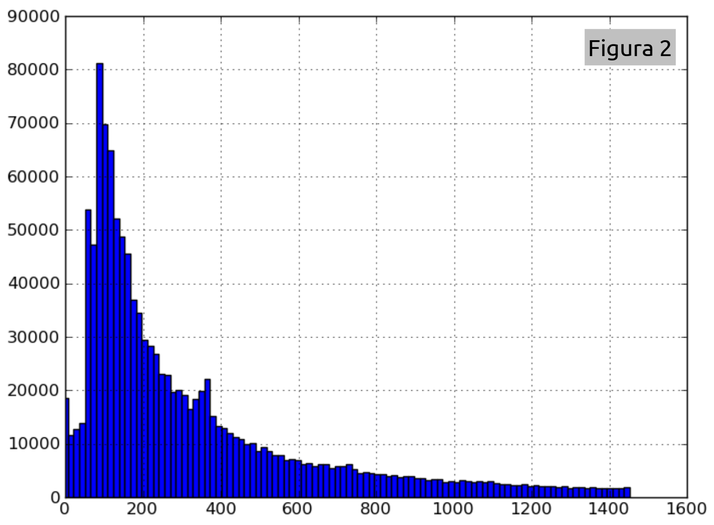
Uma série temporal pode ser definida como uma **sequência de observações de uma variável** ordenadas ao longo do tempo. [2][3] Assim, se X(t) é essa tal variável observada, podemos representar a série como o vetor:

$$X(t) = (x_1 , x_2 , ..., x_n)$$

O processo de doação de sangue inclui a coleta de dados para qualificação dos doadores. Entre eles obtém-se o nível de hematócrito. Assim, podemos olhar para a sequência de obervações de HT de um doador como uma série temporal.



A proposta inicial deste trabalho era utilizar o conceito de distância entre séries temporais para classificar os padrões de doadores. [3] Entretanto, uma característica intrínseca ao domínio constituiu uma enorme barreira ao uso desta técnica: *as doações não ocorrem em intervalos regulares de tempo*. Em outras palavras, é muito complicado normalizar as séries temporais. A figura 2 mostra a distribuição dos intervalos entre doações.



Embora seja possível notar uma certa concentração na distribuição dos intervalos de doação, é difícil escolher um valor padrão que seja significativo para normalizar as séries temporais.

4. MINERAÇÃO DE DADOS

A mineração de dados tem como principal objetivo explorar grandes volumes de dados à procura de padrões de informação. [4] Foi o que fizemos como alternativa à analise de séries temporais.

Para tanto, aproximamos as séries temporais para retas e utilizamos o **coeficiente angular** como representação daquele conjunto de doações. Dessa forma, cada série temporal

$$HT(t) = (ht_1 , ht_2 , \dots , ht_n)$$

dá origem, pelo método dos mínimos quadrados, a uma reta

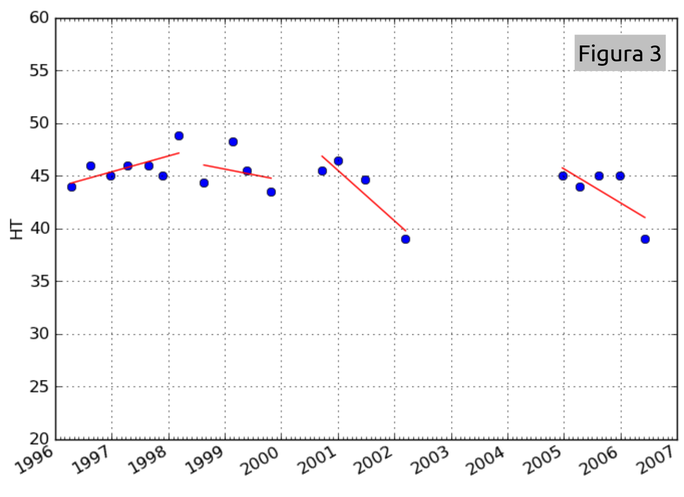
$$f(t) = \alpha * t + h$$

onde α expressa a **variação de hematócrito por unidade de tempo** (dias ou anos).

Conforme a evolução do trabalho de mineração, mostrou-se importante fazer o cálculo das retas de forma mais localizada no tempo. O motivo para isso é que uma resposta biológica natural é a recuperação do nível de HT após longos períodos sem doação. Decidimos, então, **particionar temporalmente as doações em grupos** pelos seguintes critérios:

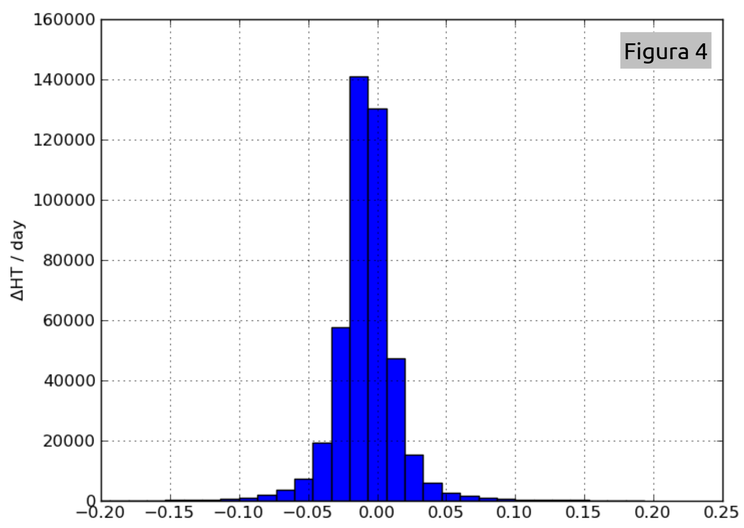
- Doador ficou muito tempo (δ) sem doar.
- Doa regularmente, mas já está doando há um certo tempo (Δ). Assim obtivemos α tendo ano como intervalo de tempo.

A figura 3 mostra o agrupamento e o cálculo das retas do exemplo da seção 3 para Δ=2 anos e δ=1 ano.



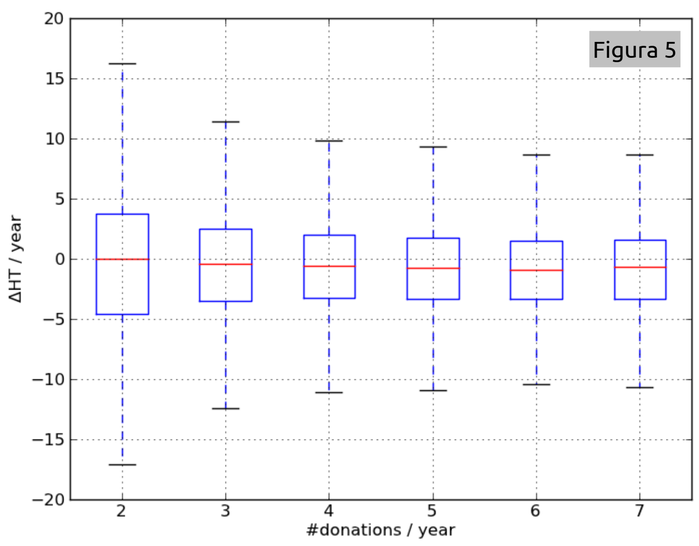
5. RESULTADOS

Em praticamente todas as análises feitas com os coeficientes angulares (α) encontramos uma distribuição muito semelhante. A figura 4 mostra esse padrão em um histograma. (note que α=0.05/dia representa uma perda de aproximadamente 4,5% de HT em 3 meses)



Das muitas análises, eis alguns fatos que pudemos constatar sobre os coeficientes (α):

- A média é sempre negativa, mas muito próxima de zero. **Doar sangue leva a diminuição do nível de HT, mas não faz mal a maioria das pessoas**.
- O sexo do doador não mostrou diferenças notáveis na distribuição dos coeficientes. Isso é contra intuitivo pois **esperava-se uma média menor nas mulheres**.
- A média cai conforme o número de doações por ano aumenta, ou seja, **quanto mais doações em um ano, maior a perda de HT**.



Dos estudos realizados, concluímos que existem 4 classes distintas que um doador pode assumir em um dado instante, tendo em vista o intervalo de α. O conhecimento delas é de fundamental importância para nortear os procedimentos médicos dos especialistas na área hematologia:

Sensível [α < -0.04] - doando regularmente, corre o risco de ficar anêmico em alguns meses.

Negativo [-0.04 < α < 0] – caso comum em que HT decai mas não chega a ser prejudicial.

Positivo [0 < α < 0.02] – a maior parte dos doadores flutua entre negativo e positivo.

Insensível [0.02 < α] – não importa o quanto doe seu organismo não sofre o impacto.

6. REFERÊNCIAS

[1] Skikne B, Lynch S, Borek D, Cook J. Iron and blood donation. Clin Haematol. 1984;13:271-87.

[2] P.A. Morettin e C.M.C. Toloi: Análise de séries temporais, Edgard Blücher, 2004.

[3] E. P. Kameda: Redução de dimensionalidade em modelos de bancos de dados multidimensionais, 2005. Dissertação (Mestrado em Ciências) – IME-USP, São Paulo. 2001.

[4] J. E. Ferreira: Introdução a banco de dados, 2005. http://www.ime.usp.br/~jef/apostila.pdf Acessado Set/2011.