Instituto de Matemática e Estatística - USP

MAC499 - Trabalho de Formatura Supervisionado

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

---

Aluno: André Henrique Serafim Casimiro

Orientador: João Eduardo Ferreira

Colaboradora: Dra. Ester Sabino

1 de dezembro de 2011

Resumo

O grupo de banco de dados do IME (DATA-IME) 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. Além disso, tem prestado apoio aos pesquisadores na mineração dessa base de dados. Um tema preocupante sobre doações 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 explora uma forma de visualizar os dados das doações para tentar prever possíveis casos de anemia nos doadores.

Palavras-chave: Banco de dados, Data Warehouse, Mineração de Dados, Doação de Sangue, Anemia.

Abstract

The IME's database group (DATA-IME) works with 3 brazilian blood centers helping them to manage and integrate the huge amount of data from the blood donations in each one of them. Besides it, have been providing support to the researches on the mining of this database. A concerning topic in the blood donation process is about the risk that donors might be taking of evolving into anemia due to donations. This correlation is still not well understood by medical specialists and varies a lot from donor to donor. This work explores a method for viewing the donations data as an attempt for prevising potential anemic cases on donors.

Keywords: Database, Data Warehouse, Data Mining, Blood Donation, Anemia.

Sumário

1 Introdução

1.1 Contextualização

O Grupo de Banco de Dados do IME-USP (Data-IME) participa desde 2006 do REDS-II, um projeto americano sobre segurança em transfusões de sangue. O Data-IME trabalha com 3 hemocentros brasileiros: o de São Paulo (Fundação Pró-Sangue), o de Minas Gerais (Hemominas) e o de Pernambuco (Hemope). Para trabalhar com todas as informações agregadas em um único lugar foi necessário criar um Data Warehouse, trabalho este realizado pelo Data-IME em 2009. [1] Agora é possível fazer análises sobre os dados de forma a tirar conclusões dos mesmos. Este trabalho tem como objetivo encontrar métricas que permitam classificar os doadores com relação ao desenvolvimento de anemia, a saber utilizando níveis de hematócrito.

1.2 Organização do trabalho

Seção 1: breve contextualização e apresentação deste trabalho.

Seção 2: bases conceituais usadas na concepção do projeto.

Seção 3: aplicação dos conceitos no domínio de doações de sangue.

Seção 4: exposição dos experimentos de mineração de dados mais relevantes.

Seção 5: resultados derivados dos experimentos e proposta de continuidade.

Seção 6: conclusão do trabalho e propostas para sua continuação.

Seção 7: impressões do autor sobre sua experiência durante o trabalho.

2 Fundamentos

Nesta seção são apresentados os conceitos de banco de dados multidimensional, modelo multidimensional, hematócrito, anemia, série temporal e redução dimensional. Também apresentamos um pouco do processo de doação de sangue. São estes os conceitos que formam a base teórica deste trabalho.

2.1 Banco de dados multidimensional

Bancos de dados estão por toda parte, das pequenas às grandes empresas e instituições. Eles são parte fundamental da maioria dos sistemas de gerenciamento de informação. Este trabalho, no entanto, lida com um tipo específico de banco de dados, os bancos de dados multidimensionais. Também conhecidos como data warehouses (depósito de dados), os bancos de dados multidimensionais são construídos para integrar várias bases de dados e permitem que se faça uma análise global dos dados sob diversos aspectos (as dimensões).

Segundo [2], Warehousing é uma técnica utilizada para recuperação e integração de dados a partir de fontes distribuídas, autônomas e, possivelmente, heterogêneas.

A principal diferença entre um data warehouse e os bancos de dados convencionais presentes na maioria dos sistemas de informação é o propósito de uso. Estes sistemas de informação tem como principal objetivo o armazenamento e recuperação de dados de forma pontual. Esse tipo de comportamento é conhecido como OLTP (On-Line Transaction Processing) e é responsável pelo processamento de operações críticas para o bom funcionamento de qualquer empresa, por exemplo. Os data warehouses são bases de dados construídas a partir de bases OLTP, e são projetadas para prover aplicações OLAP (On-Line Analytical Processing). Tais aplicações proveem uma interface de consultas analíticas a uma base de dados unificada e tem como função auxiliar a tomada de decisões a nível gerencial.

Assim, enquanto aplicações OLTP são boas para realizar tarefas sobre dados específicos, tais como: inserir uma nova venda, atualizar o cadastro de um determinado cliente, remover uma cobrança, etc. As aplicações OLAP existem para responder perguntas como: qual o valor arrecadado com a venda de televisores no primeiro trimestre de 2011? ou, qual o valor médio do nível de hematócrito da população de mulheres da região metropolitana de SP?

Integrar bases de dados grandes e diferentes não é uma tarefa fácil. Os dados provenientes de diversos bancos transacionais podem ser completamente diferentes, da tecnologia à semântica dos campos e tabelas. Assim, é necessário projetar o processo de inserção dos dados no data warehouse, é o que chamamos de *processo de carga*. Trata-se de um conjunto de operações pelas quais os dados devem passar, de forma a obter-se uma base de dados integrada, limpa e confiável. Seguindo este padrão, todas as cargas são feitas de forma incremental ao longo do tempo.

2.1.1 Modelo multidimensional

A implementação de data warehouses utiliza a mesma tecnologia que os bancos OLTP (e.g. PostgreSQL, MySQL, Oracle) mas possui uma modelagem diferente. Os modelos multidimensionais são construídos a partir de um fato (ou mais) e as dimensões sobre as quais se deseja analisar aquele determinado fato. Ele representa os interesses dos especialistas de negócio sobre os fatos. Segundo [2][4], segue uma definição para modelo multidimensional.

Definição 1. *Uma* ***base de dados multidimensional*** *é uma coleção de relações* $D_1, \ldots, D_n, F$*, onde:*

- *Cada* D_i *é uma* ***tabela dimensão****, isto é, uma relação caracterizada por um identificador* d_i *que identifica unicamente cada entrada (di é a chave primária de Di).*
- *F é uma* ***tabela fato****, isto é, uma relação que conecta todas as tabelas* $D_1, \ldots, D_n$*; o identificador de F é composto pelas chaves estrangeiras* $d_1, \ldots, d_n$ *de todas as tabelas dimensões conectadas. O esquema de F contém, ainda, um conjunto de atributos adicionais que são as métricas sobre as quais se pode aplicar funções de agregação.*

Note que as relações em um modelo multidimensional não precisam estar normalizadas. Isto evita buscas e *joins* em tabelas secundárias e aumenta a eficiência quando a base de dados é muito extensa (o que geralmente é o caso).

Exemplo A figura 1 mostra um exemplo de um modelo multidimensional simples de uma empresa de vendas qualquer. Ele contém o fato *venda (F)* e as dimensões *produto* (D_1)*, vendedor* (D_2)*, loja* (D_3) e *tempo* (D_4)*.*

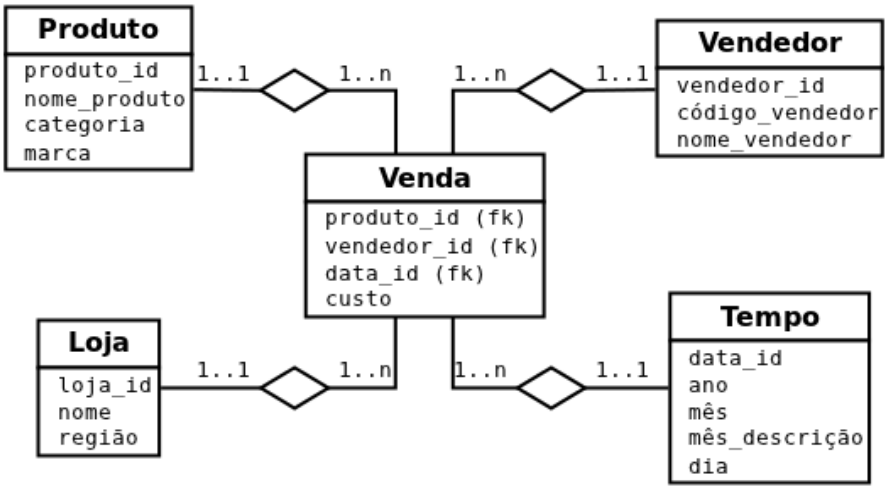


Figura 1: Exemplo de modelo multidimensional com o fato *venda (F)* e as dimensões *produto* (D_1)*, vendedor* (D_2)*, loja* (D_3) e *tempo* (D_4)*.*

Atributos presentes na tabela fato são chamados de métricas, e podem ser classificadas em *aditivas*, *não aditivas* e *semi-aditivas*, de acordo com a semântica que possuem em relação às dimensões do modelo. [2] No exemplo acima temos *custo* como uma métrica aditiva em *venda*, pois ele pode ser sumariado (somado) ao se fazer qualquer projeção (agregação de vendas) de uma dimensão nas outras.

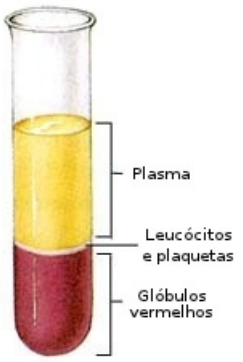
O modelo, então, permite responder rapidamente a perguntas como:

- Quantas vendas de suco de laranja (produto), os vendedores João e Eduardo efetuaram na região sul de São Paulo, no mês de setembro?
- Quais os produtos mais vendidos no ano passado?
- Qual loja é a que menos vendeu chocolates e balas no segundo trimestre?

Em praticamente todos os domínios de aplicação de data warehouses há interesse em fazer análise temporal dos dados armazenados. O que se reflete no próprio processo incremental de carga dos dados. Assim, a dimensão tempo é praticamente obrigatória em qualquer modelo multidimensional.

2.2 Hematócrito e anemia

Hematócrito é 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 entre de 42% a 52% nos homens e de 36% a 48% nas mulheres. Caso o valor seja inferior à média significa que existe pouca quantidade de glóbulos vermelhos e se for superior existe uma maior quantidade de glóbulos vermelhos no volume de sangue. [6]



Hemoglobina é uma metaloproteína presente nos glóbulos vermelhos que contém ferro e viabiliza o transporte de oxigênio pelo sistema circulatório. [6]

Anemia é a doença caracterizada pela capacidade diminuída de transporte de oxigênio devido à diminuição da contagem de glóbulos vermelhos ou à concentração de hemoglobina nestas células. [7]

Neste trabalho vamos trabalhar com dados de doações de sangue para determinar classes de doadores com relação ao desenvolvimento de anemia. Como as fontes de dados são heterogêneas, para alguns registros da tabela fato teremos valores de hematócrito *(HT)* e para outros de hemoglobina *(HB)*. Para todos os fins, quando necessário, utilizaremos o seguinte mapeamento entre os dois índices.

$$HT = 3 * HB$$

2.3 Doação de sangue

A ciência avançou muito e fez várias descobertas, mas ainda não foi encontrado um substituto para o sangue humano. Por isso, sempre que alguém precisa de uma transfusão de sangue só pode contar com a solidariedade de outros. [8]

O sangue recolhido nas doações é ministrado em pacientes com os mais variados problemas. Nesse processo segurança é um requisito fundamental. A figura 2 mostra as etapas de uma doação de sangue.

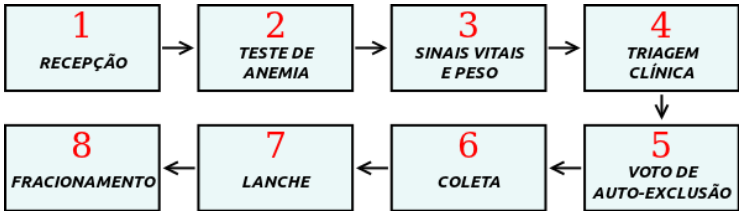


Figura 2: Etapas do processo de doação de sangue

Para o escopo deste trabalho, não necessitamos explicar cada uma cada uma das etapas do processo de doação. Vamos nos limitar a ressaltar que doadores que são diagnosticados com anemia são impedidos de doar, e que a etapa de fracionamento separa o sangue em 4 partes: concentrado de hemácias, concentrado de plaquetas, plasma e crioprecipitado.

2.3.1 Intervalos entre doações

Homens: 60 dias (até 4 doações por ano)

Mulheres: 90 dias (até 3 doações por ano)

2.3.2 Aférese (doação de plaquetas)

Aférese é o nome que se dá ao processo que permite a separação e coleta específica de plaquetas. Através de uma agulha colocada na veia do braço do doador, o sangue é bombeado para o interior de um equipamento que separa o sangue nos seus constituintes e retém parte das plaquetas, devolvendo para o organismo do doador as células restantes. [8]

Aféreses podem ser realizada a cada semana pois a reposição de plaquetas no organismo é rápida, sendo feita em apenas 72 horas. [8] Além disso, como não são retirados glóbulos vermelhos elas não influenciam no desenvolvimento de anemia.

2.4 Séries temporais

Enunciaremos a seguir uma definição não muito apurada de **série temporal**. A razão disso é que, como veremos em 3, não é possível a utilização das mesmas para análises no domínio de doações de sangue.

Uma *série temporal* pode ser definida como um conjunto de observações de uma variável ordenadas ao longo do tempo.[5][3] Assim, se X(t) é a tal variável observada, então podemos representar a série temporal como o vetor:

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

São exemplos de séries temporais:

- valores diários de poluição na cidade de São Paulo;
- valores mensais de temperatura na cidade de Cananéia-SP;
- índices diários da Bolsa de Valores de São Paulo;
- número médio anual de manchas solares;
- temperatura horária na Avenida Paulista;

2.5 Redução dimensional

A análise dimensional é feita por especialistas de domínio (e.g. gerentes) que utilizam as informações para auxiliar a tomada de decisões e a criação estratégias de negócio. Tais pessoas geralmente analisam os dados em ferramentas OLAP alimentadas pelo data warehouse em estruturas chamadas de *cubos de dados*, que contém dados e algumas consultas pré-computadas. As aplicações OLAP proveem aos especialistas de domínio um ambiente completo de análise, permitindo qualquer manipulação e combinação das dimensões. É neste ponto que, *aparentemente*, acabam as tarefas computacionais do processo de análise.

Exemplo Vamos voltar ao exemplo da empresa de vendas de produtos usado em 2.1.1 e digamos que ela possui em sua base de dados:

40 produtos: Suco, Chocolate, Pão, ...

50 vendedores: João, Eduardo, André, ...

10 lojas: Butantã, Vila Indiana, Paulista, ...

Observação: *o termo* ***cardinalidade da dimensão*** *se refere a quantidade de elementos presentes em uma dada dimensão do modelo. Assim, a cardinalidade da dimensão produto no exemplo acima é 40.*

Vamos então levar este quadro para que um gerente da empresa possa analisar o comportamento de seus vendedores dia a dia ao longo da última semana. Para cada escolha de produto, vendedor e loja que ele fizer, obterá uma série temporal de vendas diferente.

$$Z('Suco','Eduardo','Paulista') = (0, 20, 12, 0, 15, \ldots)$$

O número de séries temporais que podemos extrair vem do produto cartesiano entre as dimensões *produto*, *vendedor* e *loja*, ou seja, $40 * 50 * 10 = 20000$ séries temporais diferentes. Este exemplo foi baseado em um caso real, onde havia, inclusive, outras dimensões envolvidas. Ou seja, a análise individualizada é, muitas vezes, totalmente impraticável.

No dia a dia de data mining, entretanto, o especialista utiliza seu conhecimento implícito do domínio e analisa apenas algumas séries temporais. Ele escolhe alguns *representantes* que julga adequados para mostrar padrões nos dados e assume que eles indicam o comportamento de um grupo todo, ou seja, a seleção de representantes depende da classe de elementos que se deseja representar. Voltando ao nosso exemplo, onde o gerente queria analisar as vendas de seus vendedores no último mês, a escolha dos representantes (vendedores) poderia ser baseada nas seguintes perguntas:

1. Quem foi um vendedor regular? (não importando o produto nem a loja)
2. Quem foi um bom vendedor de suco?
3. Quem foi um vendedor ruim na loja da Vila Indiana?

2.5.1 Técnicas de redução dimensional

Como o número de séries temporais pode ser muito grande, deseja-se reduzir a cardinalidade das dimensões. Existem existem duas abordagens possíveis:

Eliminação de variáveis Em [3], podemos encontrar um estudo de redução de dimensionalidade baseado na determinação de variáveis que não influenciem no comportamento do conjunto de outras. Isso seria o equivalente a, por exemplo, encontrar um produto que tivesse um padrão de venda igual para todos os vendedores, em qualquer loja. O que nos permitiria reduzir a cardinalidade da dimensão produto ignorando o mesmo.

Agrupamento Neste trabalho, vamos usar uma abordagem diferente da de eliminação de variáveis. Queremos reduzir drasticamente a quantidade de elementos em uma dimensão através da detecção de padrões de comportamento (possivelmente obtidos por análise de séries temporais). Vamos substituir os elementos antigos por novos que foram construídos para representar as classes encontradas. Reduzindo os 50 vendedores ('João', 'Eduardo', 'André', ...) para, digamos, 3 classes de vendedores ('Bom', 'Regular', 'Ruim').

***Nota**: para simplificar o trabalho, iremos levar em conta apenas 1 dimensão ao fazer os agrupamentos, ou seja, não será uma análise estatística multivariada.*

3 Descrição do problema e solução

Nesta seção vamos utilizar os fundamentos vistos em 2 para definir o problema e o escopo de estudo a que se propôs este trabalho. Vamos apresentar o modelo de dados a ser utilizado e a mudança de metodologia necessária para dar continuidade a exploração dos dados. Também descreveremos brevemente alguns tópicos sobre a implementação.

3.1 Modelo

O banco de dados que vamos utilizar, é uma pequena fração do original desenvolvido em [1]. O modelo, como ilustrado na figura 3, contém a tabela fato (doações de sangue) e apenas 2 dimensões, sendo uma delas o tempo.

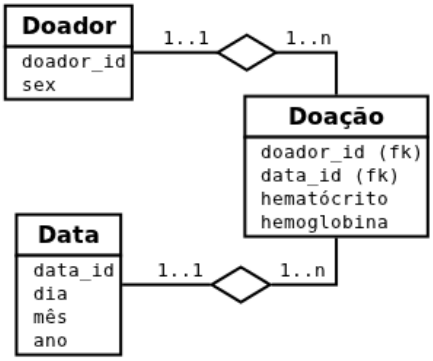


Figura 3: Modelo multidimensional do domínio de doações de sangue utilizado. Temos o fato *doação* e as dimensões *doador* e *tempo*.

Os atributos *hematócrito* e *hemoglobina* da relação fato são métricas não aditivas, pois a medida é única por doação e não há significado em somá-las. No contexto desse trabalho, como dito em 2.2, eles representam a mesma informação. Assim, para uma dada instância armazenada no banco, somente um dos campos representará uma informação válida, enquanto que o outro conterá um valor inválido, e se necessário poderemos mapear um valor no outro.

3.2 Metodologia de análise

O objetivo principal deste trabalho é encontrar classes de doadores de sangue de acordo com sua suscetibilidade ao desenvolvimento de anemia; valendo-se, para isso, da série temporal que pode ser extraída dos índices de hematócrito por doador e analisando seu comportamento ao longo do tempo.

3.2.1 Problema: desnormalização

A proposta inicial deste trabalho era utilizar análise de séries temporais para classificar os padrões de doadores mas, como veremos, isso não foi possível.

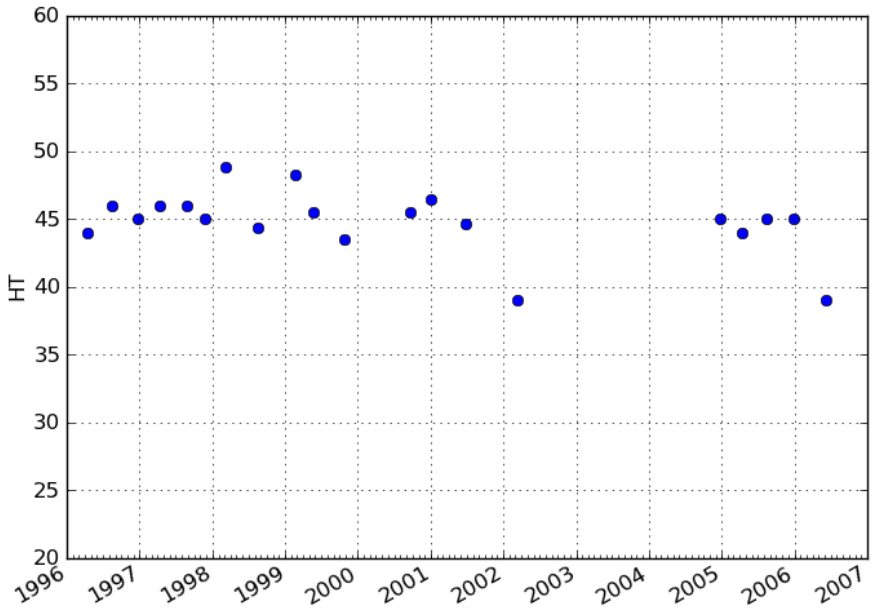


Figura 4: Série temporal do nível de hematócrito de um doador

A figura 4 ilustra bem a situação real encontrada no domínio de doações de sangue, que é: *as doações não ocorrem em períodos regulares de tempo*. A figura 5 mostra a distribuição dos intervalos de tempo entre doações consecutivas.

Em outras palavras, a série temporal de doações de sangue por doador é muito **desnormalizada**. Esta característica intrínseca ao domínio constitui uma grande barreira ao uso de análise de séries temporais (vale notar que todos os exemplos dados em 2.4 possuem um intervalo regular de medição). Os modelos de análise fazem uso dessa propriedade e, por conta disso, usaremos uma abordagem mais simples (mas não ingênua) para tentar encontrar as classes de doadores.

3.2.2 Solução: regressão linear

Dado um conjunto de valores de hematócrito aferidos ao longo do tempo em cada doação feita pelo doador, construímos uma reta de aproximação pelo método dos mínimos quadrados. Assim, cada série temporal:

$$HT(t) = (ht_1, ht_2, ..., ht_n)$$

dá origem a uma reta:

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

da qual utilizaremos o coeficiente angular (α) como base para a classificação dos doadores com relação ao seu desenvolvimento de anemia.

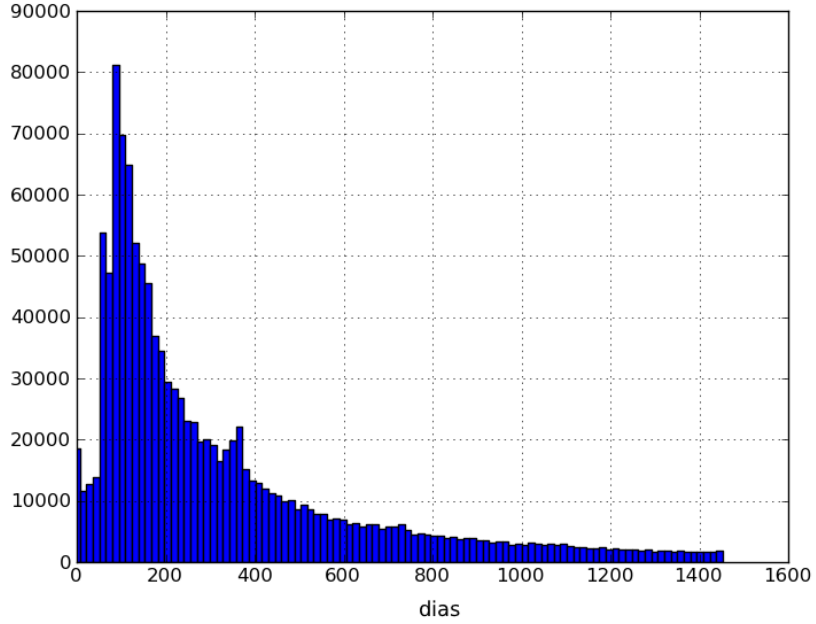


Figura 5: Histograma do intervalo entre doações consecutivas de um mesmo doador. Contagem de 1 dia até 4 anos.

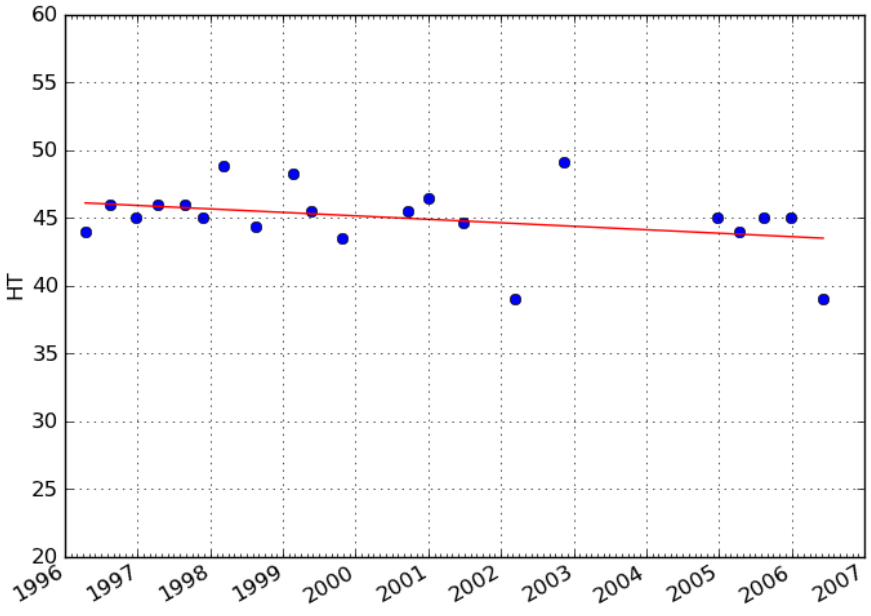


Figura 6: Série temporal do nível de hematócrito de um doador e sua aproximação por uma reta.

3.3 Implementação

Tecnologias Utilizadas

A implementação deste projeto e de todas as etapas da mineração de dados foi feita utilizando a linguagem de programação Python e diversas tecnologias relacionadas. Eis a lista das tecnologias utilizadas:

- **PostgreSQL**: sistema de gerenciador de bancos de dados relacionais. [9]
- **Python**: linguagem de programação multiparadigma. [10]
- **psicopg2**: adaptador Python para o PostgreSQL. [11]
- **NumPy**: pacote Python para cálculos matemáticos. [12]
- **matplotlib**: biblioteca Python para plotagem de gráficos. [13]

Processamento

Na mineração de dados, o processamento é feito em 3 etapas, como mostrado na figura 7.

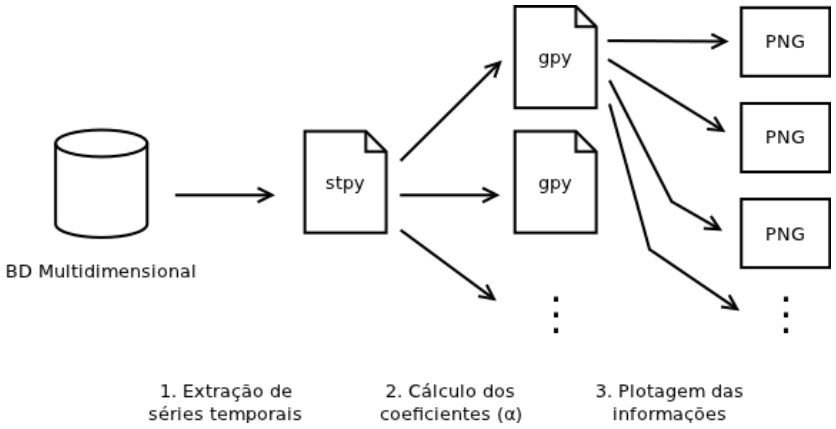


Figura 7: Etapas de processamento na mineração de dados.

Estatísticas gerais do BD

Número de doadores	1.366.197
Número de doações	2.650.499
Período de doações	11 anos

Arquivos intermediários

O volume de dados é muito grande e, para evitar retrabalho, estados intermediários de processamento são armazenados em arquivos textos com extensão *stpy* ou *gpy*. São extensões sem muito rigor de definição baseadas nas estruturas de dados da linguagem Python, de modo a facilitar a construção de seus respectivos *parsers*.

- **stpy**: armazena séries temporais de doadores.
- **gpy**: armazena grupos de doações (cf. 4) e seus respectivos coeficientes angulares.

Distribuição do código

Buscou-se separar o código por suas funcionalidades. Segue breve descrição da função de cada um dos scripts presentes no processamento dos dados:

- **settings.py**: arquivo de configurações globais.
- **db.py**: define as classes *Donor* e *SingletonDB*, que serve para a interação com o banco de dados e foi implementada seguindo o padrão de desenho Singleton. [14]
- **extractor.py**: responsável pela extração das séries temporais de doações.
- **calculator.py**: recebe as séries temporais geradas pelo extractor.py e faz o cálculo dos coeficientes angulares (α).
- **parser.py**: define classes auxiliares que iteram sobre os tipos de arquivos intermediários criados por *extractor.py* e *calculator.py*.
- **drawer**: pacote que define várias funções que desenham gráficos para os coeficientes calculados pelo *calculator.py*.
- **experiment.py**: definição da classe *Experiment* que expressa o processamento de um experimento, coordenando-o.
- **lab.py**: define um banco de experimentos. Quando invocado pela linha de comando espera como argumentos uma lista de números de experimentos a serem processados. (e.g. `./lab.py 3 1`)

4 Mineração de dados

4.1 Introdução

A utilização de séries temporais, como explanado anteriormente, encontrou um limite teórico no domínio de doações de sangue, contudo ainda podemos fazer análises estatísticas das mais diversas em busca de padrões. A este processo dá-se genericamente o nome de mineração de dados. Nesta seção iremos expor os processos e resultados obtidos seguindo o aprofundamento natural da mineração de dados que fizemos.

4.2 Definições básicas

Em todas as definições desta seção considere que, dado um doador e a série temporal do nível de hematócrito do mesmo, seja $D = \{d_1, d_2, ..., d_n\}$ o conjunto das doações que o mesmo realizou ao longo do tempo.

Definição 2 (ordenação de doações)**.** *Dados $d, d' \in D$, diremos que $d < d'$ se d ocorreu antes de d' ao longo do tempo.*

Definição 3 (distância entre doações)**.** *Dados $d, d' \in D$, definiremos $d - d'$ como o número de dias entre as datas das duas doações. (note que, pela natureza dos dados, nenhum doador pode doar 2 vezes em um mesmo dia, ou seja, $d - d' \neq 0$ vale sempre.*

4.3 Experimento 1: Análise geral

Motivação: como primeiro experimento que fizemos com os dados, quisemos analisá-los da forma mais geral possível.

Descrição: para cada doador com no mínimo 2 doações aproximamos sua série temporal para uma única reta, como na figura 8.

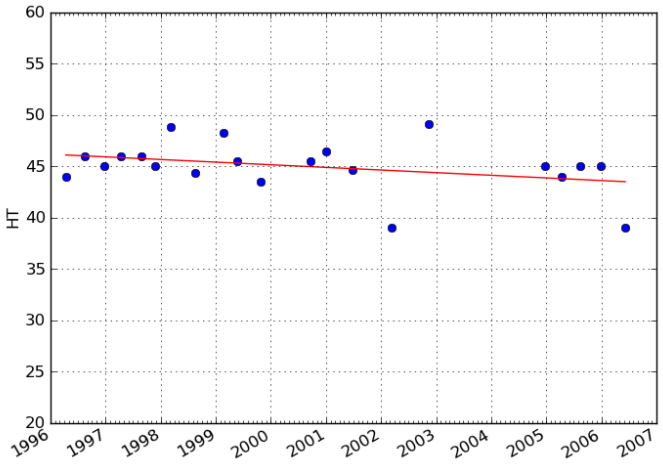


Figura 8: Método de aproximação usado na análise geral do experimento 1.

Resultados

Número de doações	Número de doadores	Média ($*10^3$)	Desvio Padrão ($*10^3$)
2+	459211	-0.280	130.161

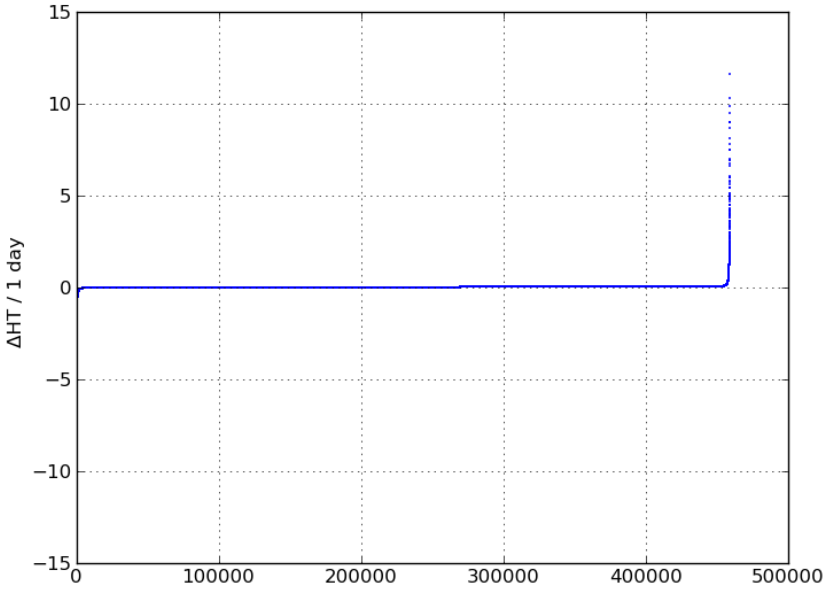


Figura 9: Coeficientes lineares obtidos no experimento 1 em ordem crescente.

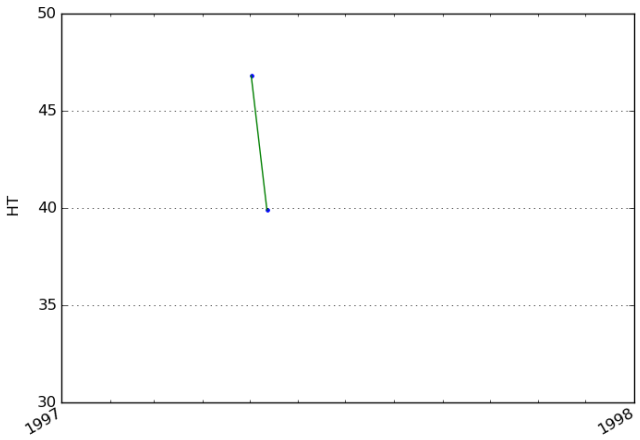


Figura 10: Série temporal com coeficiente angular muito baixo devido a doações muito próximas.

Considerações e conclusões

- Os coeficientes angulares são da ordem de grandeza de 10^{-1} a 10^{-4}. Eles medem variação de HT *por dia*, por isso são tão próximos de zero. Um valor de $\alpha = -0.05$, por exemplo, representa uma perda de 4.5% de HT em 3 meses.
- Valores exorbitantes ($|\alpha| > 0.5$) de coeficientes angulares são poucos e provenientes de casos degenerados na coleta (confira em 2.3) que podem ser ignorados.

4.4 Experimento 2: Intervalo de retorno máximo (1 ano)

Motivação: quando um doador fica um longo período sem doar, uma resposta biológica natural é a recuperação do nível de HT. Isso faz com que a reta de aproximação tenha seu coeficiente angular atenuado, levando α para mais perto de 0. Este experimento divide o conjunto de doações usados para fazer o cálculo das retas sempre que a distância entre doações consecutivas é maior do que 1 ano.

Descrição: deste experimento em diante, cada doador pode gerar 0, 1, ou vários valores de α. Sempre que um doador fica mais de 1 ano sem doar separamos o conjunto de doações e calculamos uma reta para cada um. **Obs:** para tentar reduzir o número de casos degenerados de α, ignoramos doações com distância menor do que 7 dias da anterior.

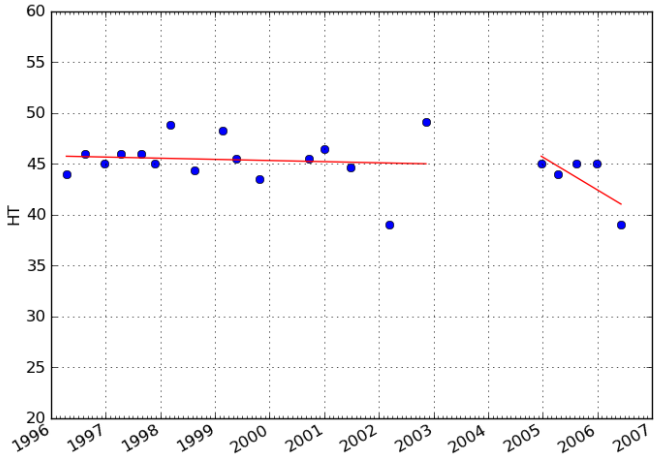


Figura 11: Exemplo de agrupamento de doações usado no experimento 2.

Resultados

Número de doações	Número grupos de doações	Média ($*10^3$)	Desvio Padrão ($*10^3$)
2+	358021	-1.310	43.006

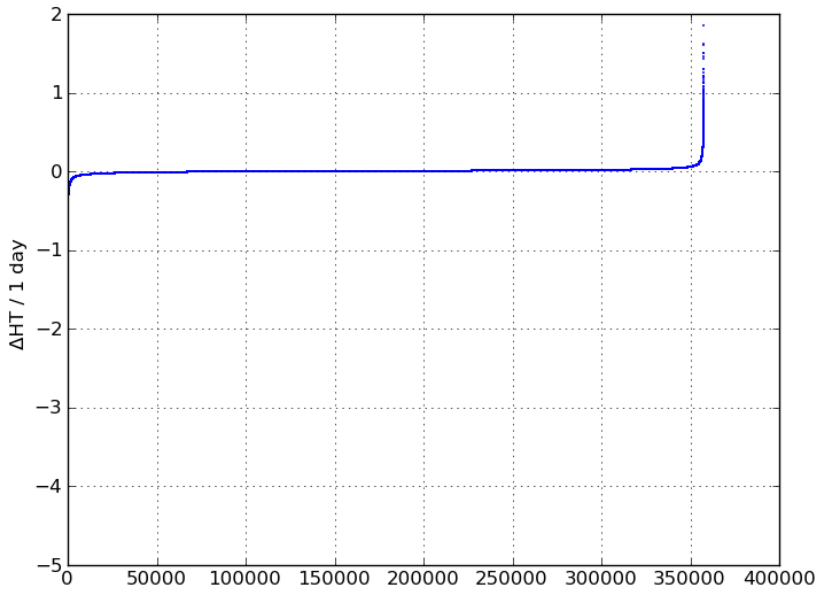


Figura 12: Coeficientes lineares obtidos no experimento 2 em ordem crescente.

Considerações e conclusões

- Mesmo tendo a possibilidade de que um doador gerasse mais de um α a quantidade deles total dos mesmos caiu pois muitos desses doadores tinham doações muito espaçadas no tempo.
- Os valores de α ficaram mais homogêneos: a variância caiu para menos da metade e a curva tornou-se mais suave. Entretanto, ainda é necessário um maior refinamento dos dados para aprofundarmos a análise e tirar conclusões mais fortes.

4.5 Experimento 3: Grupos de doações (1 ano)

Motivação: da mesma forma que o nível de HT se recupera quando um doador fica muito tempo sem doar, ele deve sofrer uma queda tanto mais acentuada quanto mais frequentemente doar. Deste experimento em diante estaremos assumindo que o comportamento de um mesmo doador pode variar ao longo do tempo, de modo que cada α tem uma localidade temporal de no máximo 1 ano.

Descrição: deste experimento em diante iremos restringir o intervalo temporal de cada conjunto de doações. Para descrever como serão construídos tais conjuntos em cada experimento vamos introduzir a definição de **grupo de doações**.

Definição 4. *Seja $S = \{G_1, G_2, ..., G_k\}$ uma partição de D (como em 4.2) tal que:*

$$d < d', \quad \forall d \in G_i, d' \in G_{i+1} \quad i = 1, 2, ..., k-1$$

diremos que cada um dos G_i é um ***grupo de doações*** *se, dados dois inteiros positivos Δ e δ, G_i é o maior conjunto tal que:*

- *o elemento máximo de G_i não está a uma distância maior do que Δ do elemento mínimo. (Δ define um intervalo máximo para as doações de um grupo)*
- *nenhum elemento de G_i está a uma distância maior do que δ do elemento imediatamente anterior. (δ define uma distância máxima entre doações consecutivas num grupo)*

O intuito de particionar as doações da forma descrita acima é permitir uma análise temporalmente mais localizada.

Neste experimento utilizaremos $\Delta = 1$ ano e $\delta = 1$ ano. A figura 13 mostra um exemplo desse particionamento.

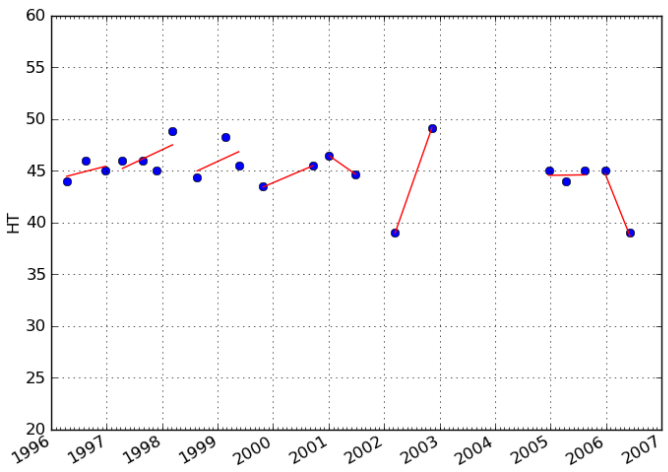


Figura 13: Exemplo de agrupamento de doações usado no experimento 3 ($\Delta = 1$ ano e $\delta = 1$ ano.

Resultados

Doações por grupo	Número de grupos (%)	Média ($*10^3$)	Desvio Padrão ($*10^3$)
0	1216 (0.28%)	-0.267	56.690
1	6955 (1.59%)	-12.121	78.878
2	297116 (67.82%)	-2.424	43.095
3	96369 (22.00%)	-1.903	18.946
4	28010 (6.39%)	-1.866	12.359
5	6419 (1.47%)	-2.534	11.451
6	1639 (0.37%)	-3.244	11.191
7	285 (0.07%)	-2.834	9.955
8	49 (0.01%)	-4.165	12.527
9	4 (0.00%)	-4.613	11.327
10	6 (0.00%)	6.051	14.606
11	2 (0.00%)	-14.779	9.529
12+	4 (0.00%)	-26.353	44.231

Observação: aféreses (confira 2.3.2) são utilizadas para o cálculo de α, mas não são contabilizadas na coluna "Doações por grupo".

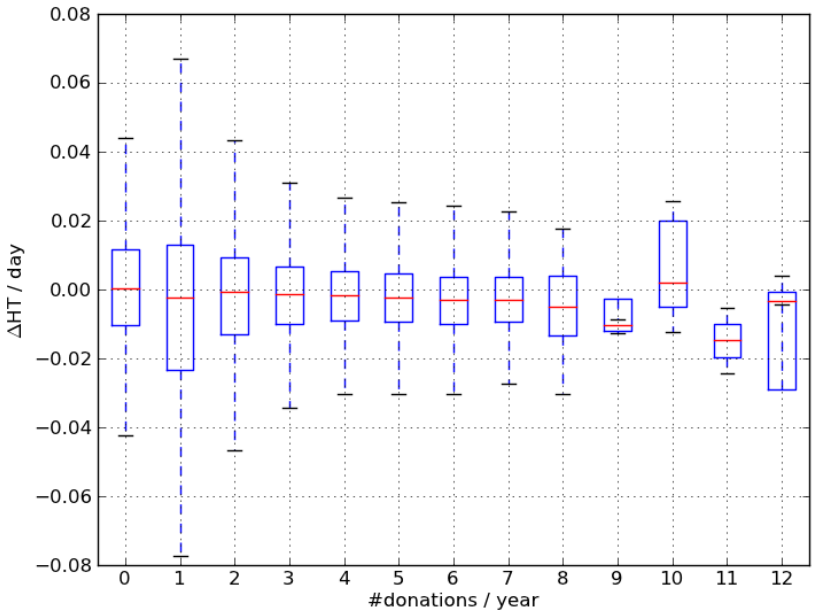


Figura 14: Comparação entre os boxplots de cada uma das linhas da tabela 4.5.

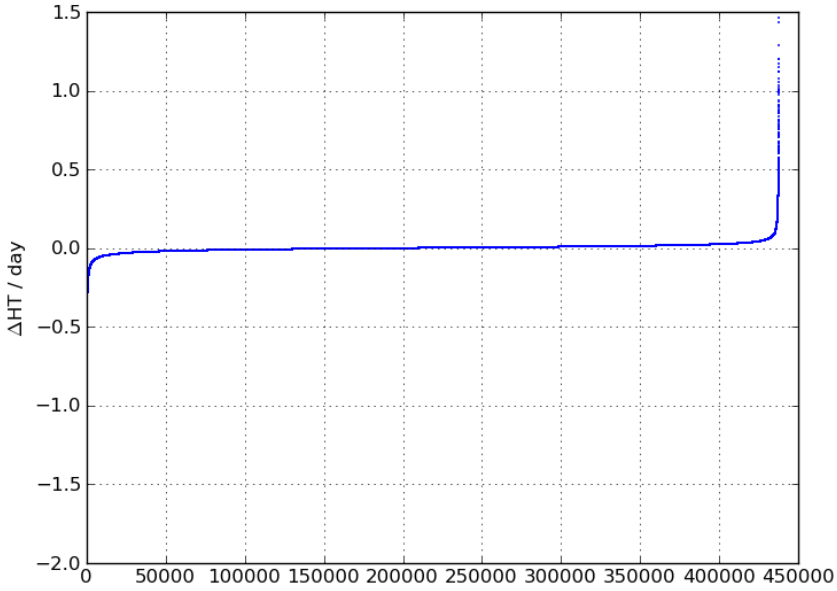


Figura 15: Coeficientes angulares de todos os grupos obtidos no experimento 3 em ordem crescente.

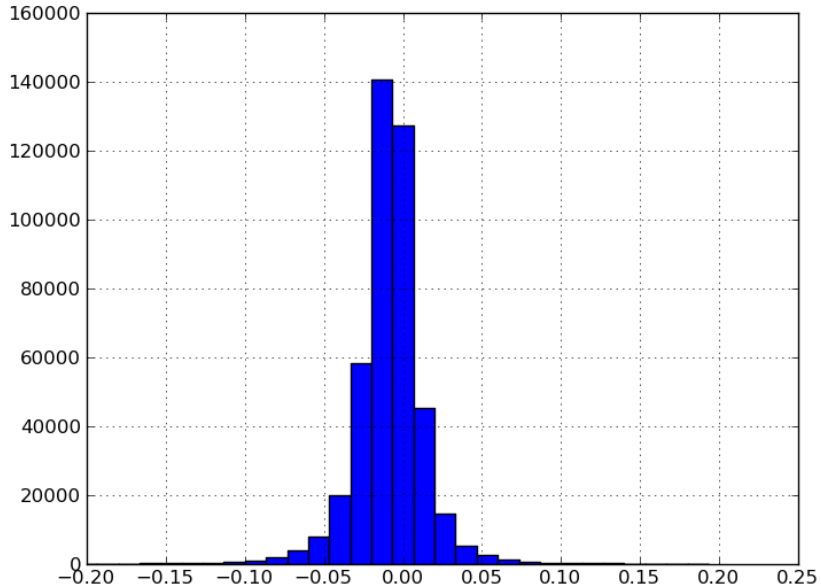


Figura 16: Histograma dos coeficientes angulares de todos os grupos do experimento 3 (mesmos dados que os da figura 15).

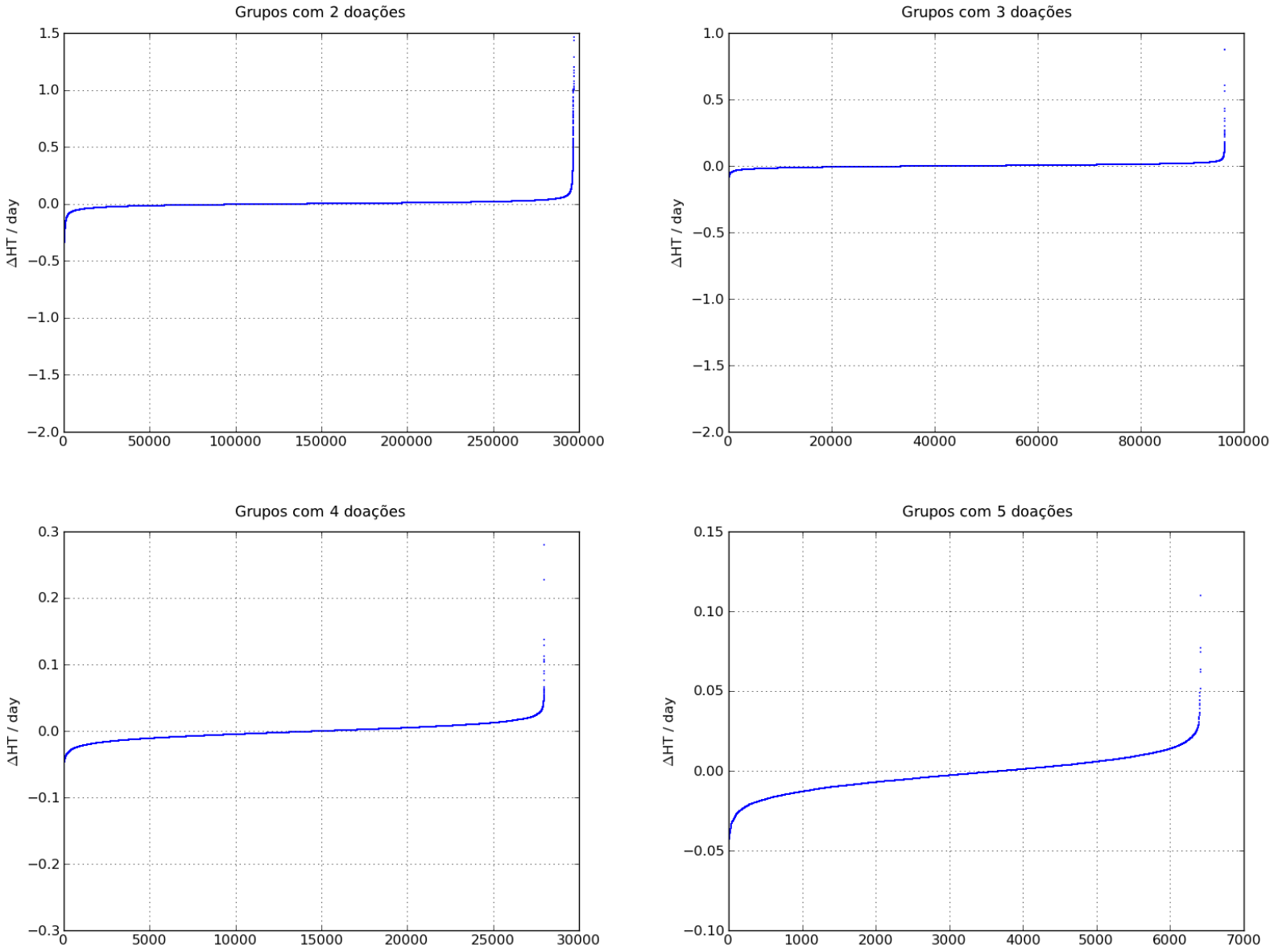


Figura 17: Coeficientes angulares em ordem crescente separados pelo número de doações no grupo. Apenas os gráfico mais relevantes para o experimento 3.

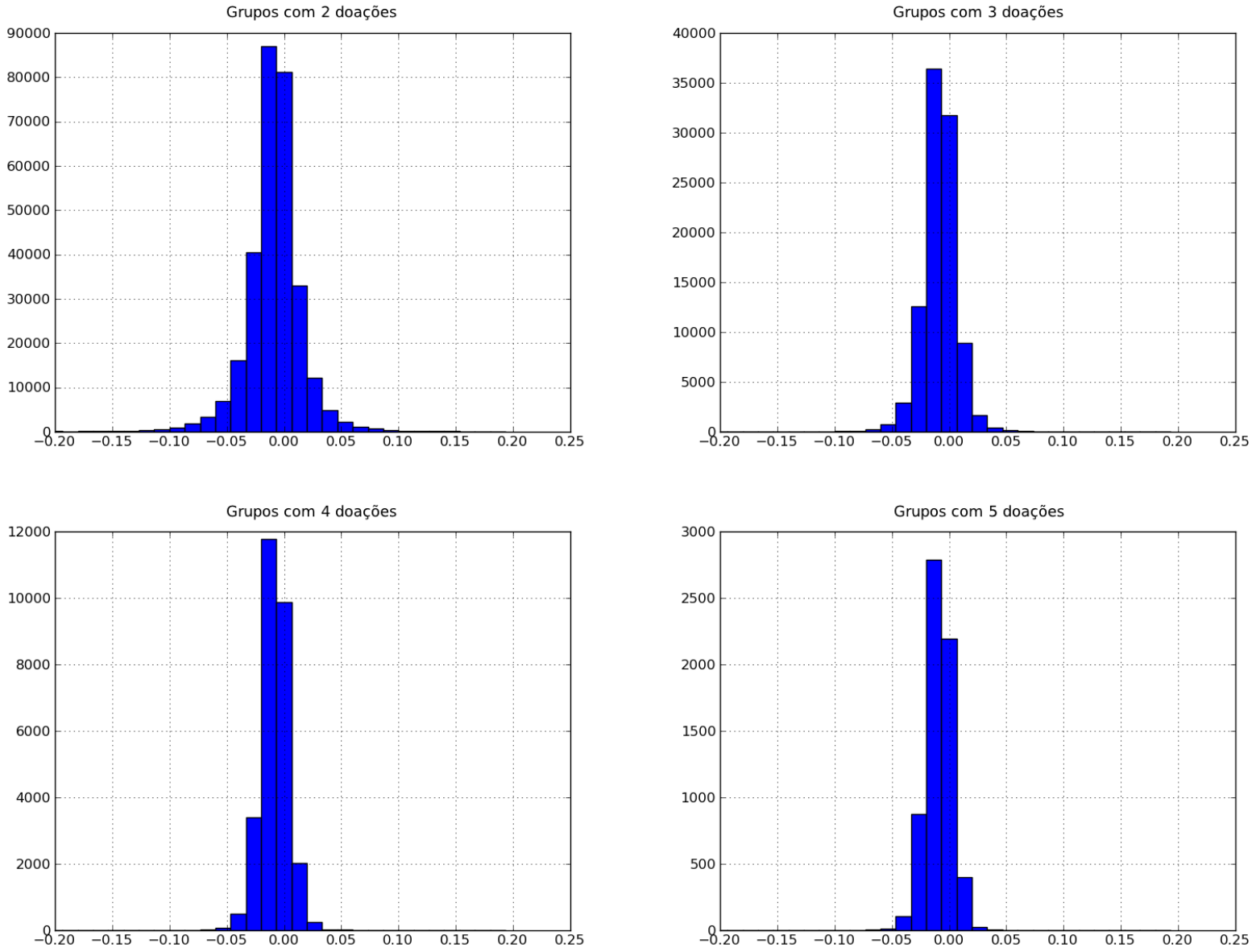


Figura 18: Histogramas dos respectivos gráficos figura 17.

Considerações e conclusões

- Confirmou-se a hipótese de que o número de doações influencia na perda de HT. Na figura 14 vemos que quanto mais doações dentro de 1 ano, mais o doador perde HT. Isso pode ser também notado nos gráficos da figura 17, que mostram como os coeficientes angulares se espalham conforme o número de doações no grupo.
- A distribuição dos grupos quanto ao número de doações condiz com os limites estabelecidos em 5.3. Os grupos com mais de 6 doações representam menos de 0.1% da população e, portanto, vamos ignorar tal comportamento degenerado.
- Olhando para os histogramas nas figuras 16 e 18 vemos que a distribuição dos coeficientes angulares assemelha-se com uma distribuição normal de probabilidades.

4.6 Experimento 4: Grupos de doações (2 anos)

Motivação: verificar se o tamanho de janela para o cálculo de α utilizado no experimento 3 é bom aumentando o tamanho do intervalo de cada grupo.

Descrição: Neste experimento utilizaremos $\Delta = 2$ anos e $\delta = 1$ ano, conforme a definição 4. A figura 19 mostra um exemplo desse particionamento.

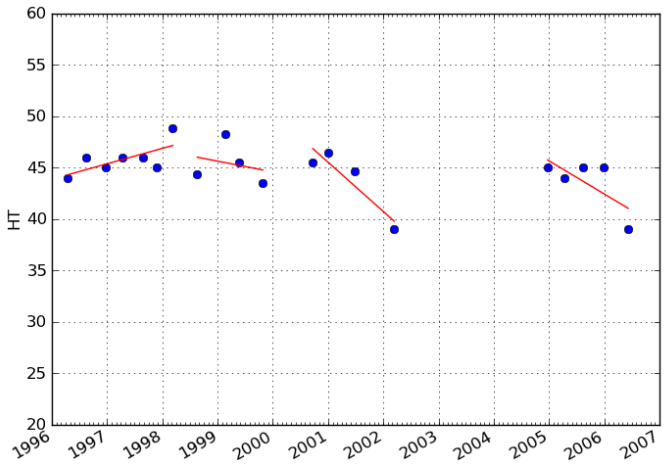


Figura 19: Exemplo de agrupamento de doações usado no experimento 4 ($\Delta = 2$ anos e $\delta = 1$ ano.

Observação: Note que sendo o mesmo doador da figura 13 apresentou uma doação a menos no final de 2002. Esta doação não aparece pois para $\Delta = 2$ anos e $\delta = 1$ ano ela ficou sozinha em um grupo, situação em que não é possível construir uma reta.

Resultados

Doações por grupo	Número de grupos (%)	Média ($*10^3$)	Desvio Padrão ($*10^3$)
0	541 (0.11%)	-1.562	78.000
1	4087 (0.86%)	-13.127	78.261
2	267738 (56.51%)	-2.068	39.506
3	106727 (22.53%)	-1.376	14.329
4	48934 (10.33%)	-1.157	8.374
5	24025 (5.07%)	-1.192	6.349
6	11676 (2.46%)	-1.139	5.665
7	5462 (1.15%)	-1.276	5.346
8	2614 (0.55%)	-1.556	5.177
9	1090 (0.23%)	-1.907	5.249
10	540 (0.11%)	-2.391	4.917
11	220 (0.05%)	-2.255	4.947
12+	155 (0.03%)	-3.369	9.950

Observação: aféreses (confira 2.3.2) são utilizadas para o cálculo de α, mas não são contabilizadas na coluna "Doações por grupo".

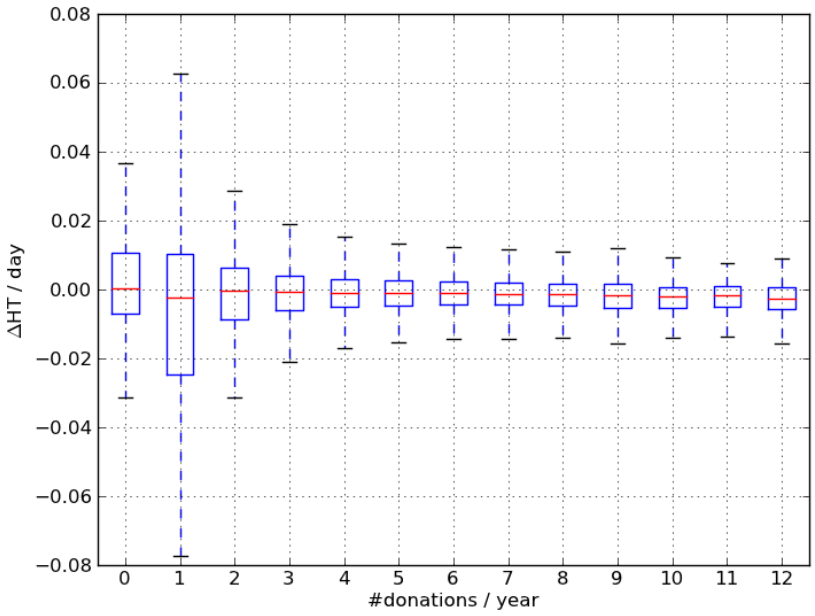


Figura 20: Comparação entre os boxplots de cada uma das linhas da tabela 4.6.

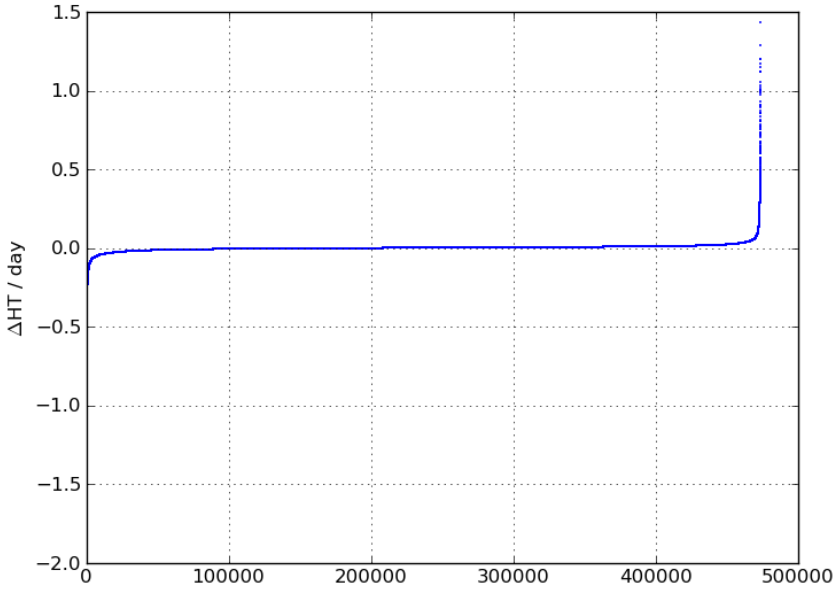


Figura 21: Coeficientes angulares de todos os grupos obtidos no experimento 4 em ordem crescente.

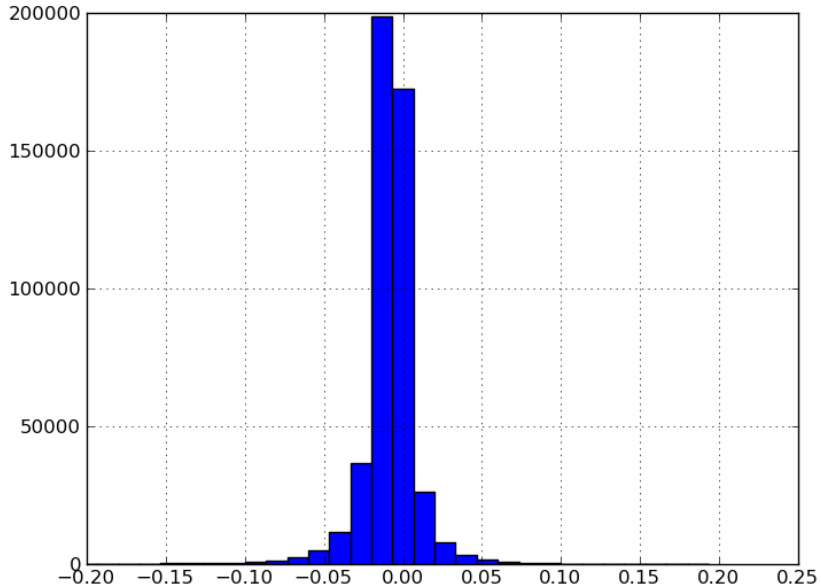


Figura 22: Histograma dos coeficientes angulares de todos os grupos do experimento 4 (mesmos dados que os da figura 21).

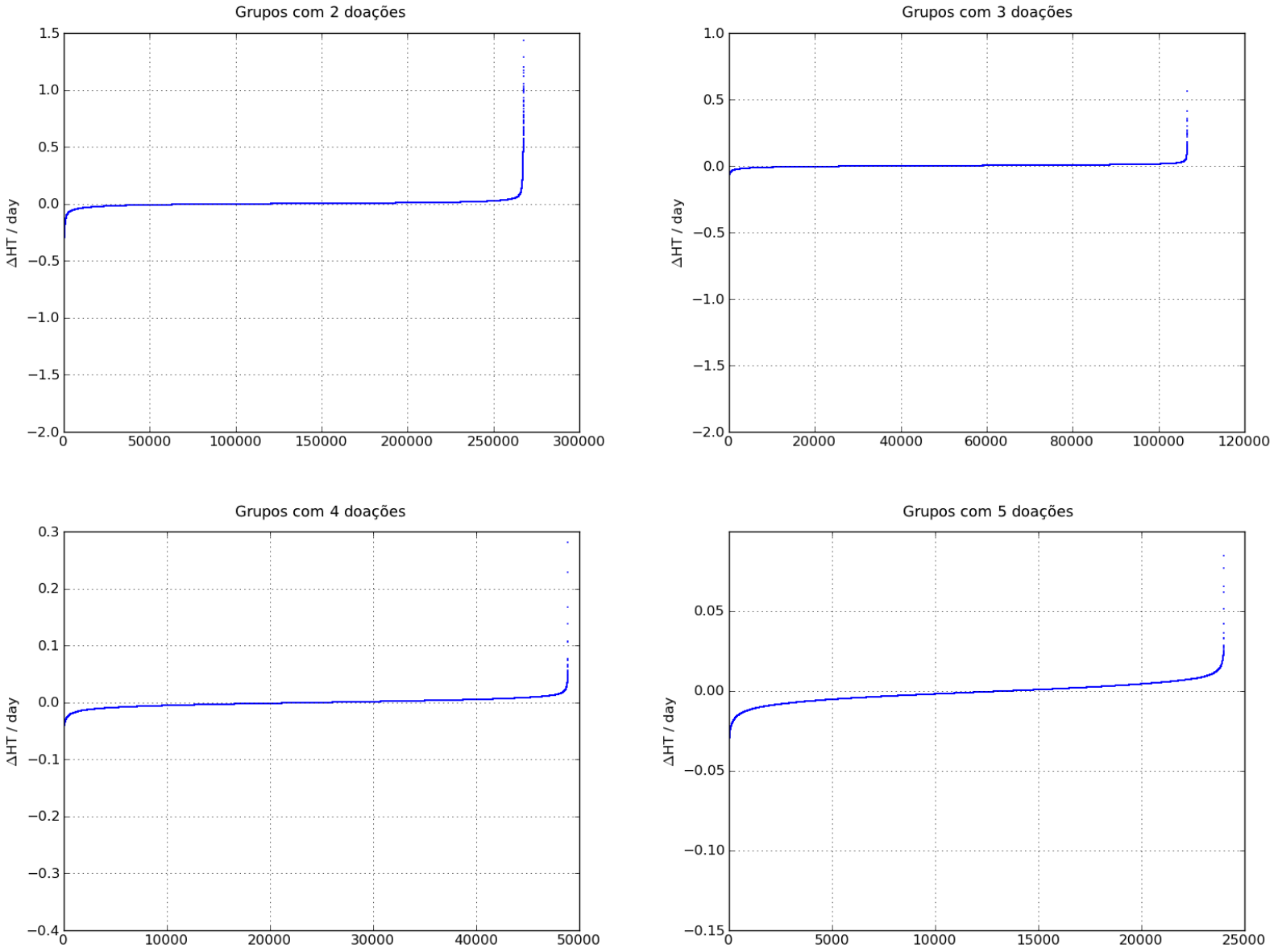


Figura 23: Coeficientes angulares em ordem crescente separados pelo número de doações no grupo. Apenas os gráfico mais relevantes para o experimento 4.

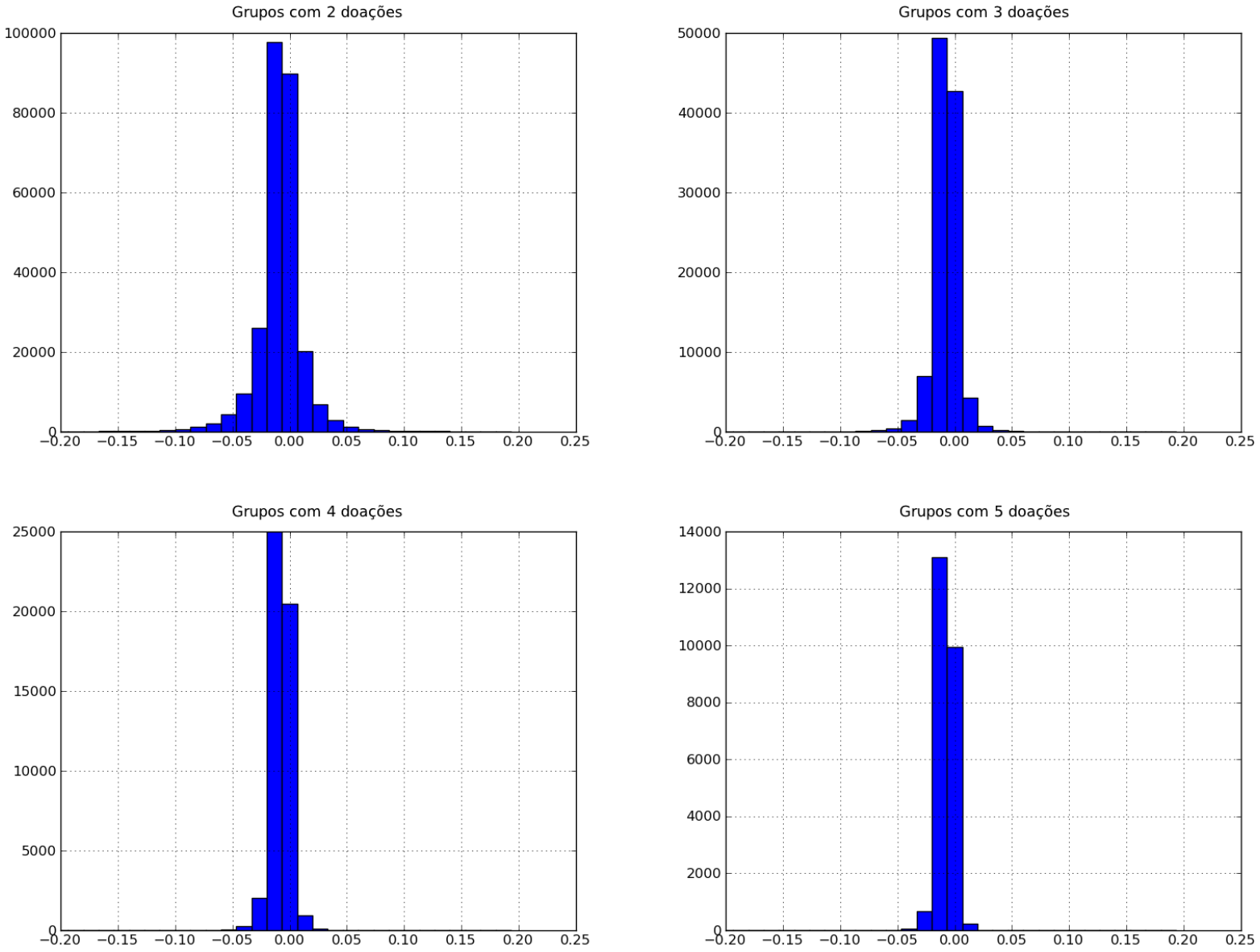


Figura 24: Histogramas dos respectivos gráficos figura 23.

Considerações e conclusões

- A figura 20 mostra que a distribuição dos coeficientes angulares ficou mais bem distribuída com relação ao número de doações de um grupo em comparação com o experimento 3 (confira 14). Com isso, perde-se um pouco da expressividade na variação de α.
- Repete-se o fato de que quanto maior o número de doações em um determinado período de tempo, maior é a perda no nível de HT.

4.7 Conclusão

Alguns outros experimentos foram feitos, e muitos outros gráficos gerados. Colocamos nesta seção aqueles que ajudavam a explicar a evolução do trabalho de mineração. Na próxima detalharemos os resultados obtidos que foram mais significativos.

5 Resultados

Nesta seção apresentaremos os principais resultados obtidos neste trabalho, quase todos obtidos no experimento 3, onde o particionamento dos grupos gerou coeficientes mais expressivos. Os resultados mostrarão como a regressão linear permitiu-nos encontrar uma boa métrica para determinação de comportamento anêmico nos doadores de sangue.

5.1 Coeficientes angulares dos grupos de doações formam uma métrica útil para previsão de anemia

No início deste trabalho, não se sabia se o uso da técnica de regressão linear em substituição à analise de séries temporais seria útil na busca de padrões que levassem um dado doador à anemia. Entretanto, conforme a mineração dos dados evoluiu, os valores de α mostraram-se uma métrica válida na detecção de padrões de anemia.

A Dra. Ester Sabino acompanhou estudos sobre o desenvolvimento de anemia em doadores de sangue que utilizavam técnicas de análise de sobrevivência estatística para compreender o caminho do nível de hematócrito de um dado doador até que chegue pela primeira vez em um estado anêmico.

As figuras 5.1, 26 e 27 mostram que, para pessoas que desenvolveram anemia, os valores de α são mais negativos até o primeiro estado anêmico, onde, dali em diante, equipara-se a índices de doadores não anêmicos.

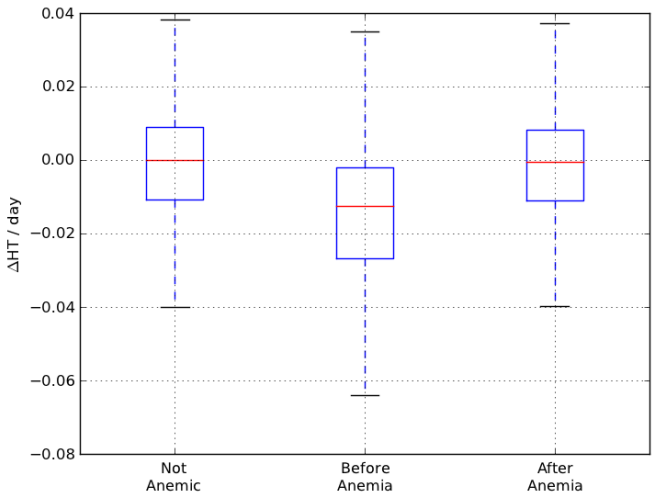


Figura 25: Comparação entre boxplots de doadores anêmico e não anêmicos.

Embora não seja possível classificar um doador como anêmico ou não, os coeficientes angulares funcionam como uma métrica válida para o estudo de desenvolvimento de anemia em relação à doações de sangue.

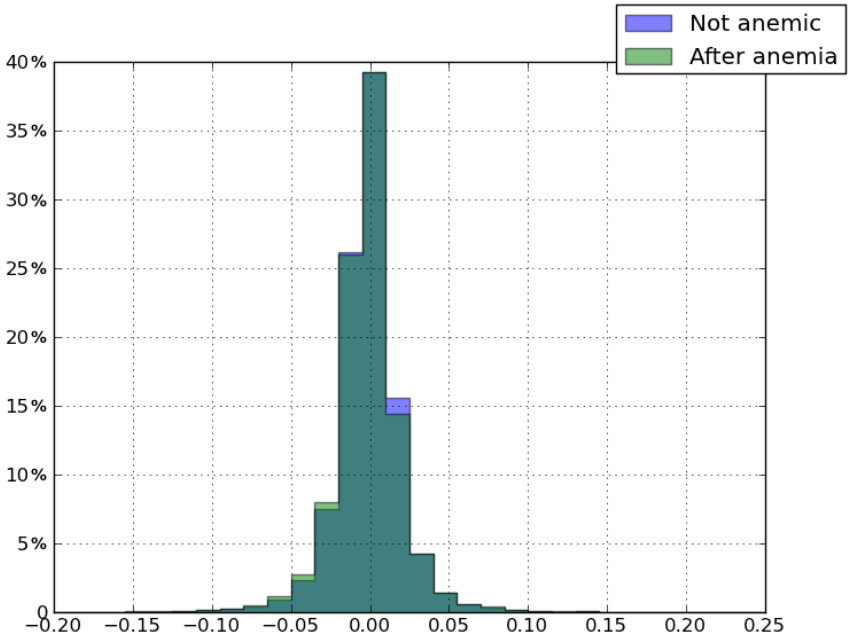


Figura 26: Comparação grupos de doações posteriores ao primeiro estado anêmico de doadores que nunca desenvolveram anemia

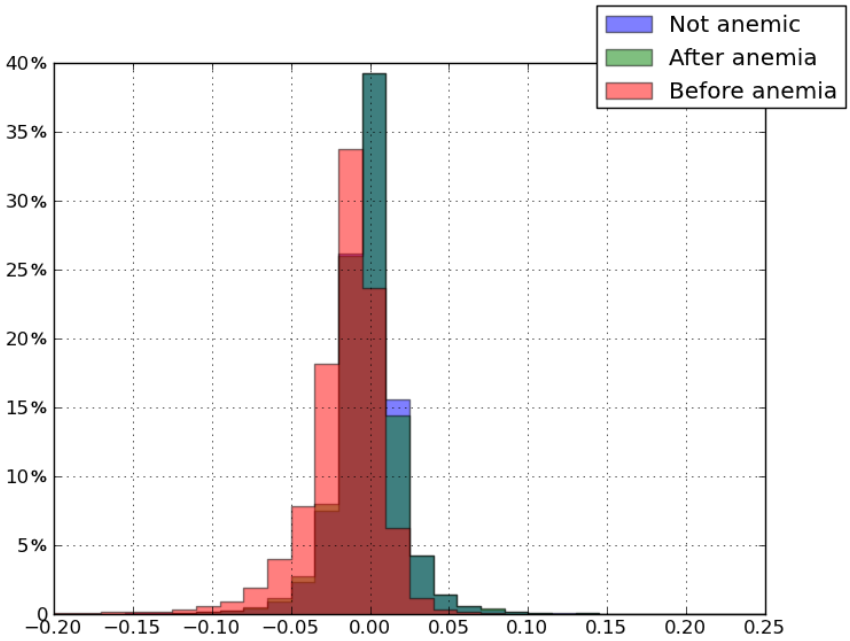


Figura 27: Em contraste com a figura 26, o histograma em vermelho representa valores de α que antecedem a primeira anemia de um doador

5.2 Doar sangue provoca leve diminuição de HT

Em todos os experimentos realizados, o valor médio dos coeficientes angulares é sempre um valor negativo e muito próximo de zero (confira as tabelas em 4.3, 4.4, 4.5, 4.6. Isso indica uma perda de HT na população de doadores, mas que, por serem valores tão próximos de 0 (da ordem de 10^{-3}), são imperceptíveis na grande maioria dos casos.

Assim, conforme esperado, concluímos que doar sangue não provoca anemia na grande maioria dos doadores. O risco de evoluir para um quadro anêmico é tanto menor quanto menos frequentes forem as doações.

5.3 Mulheres são mais suscetíveis a decaimento do nível de HT do que homens

É sabido que o organismo feminino é mais suscetível a anemia do que o masculino, o que é corroborado pelas recomendações vistas em . De forma geral, os coeficientes angulares dos grupos de doações foram capazes de evidenciar este fato.

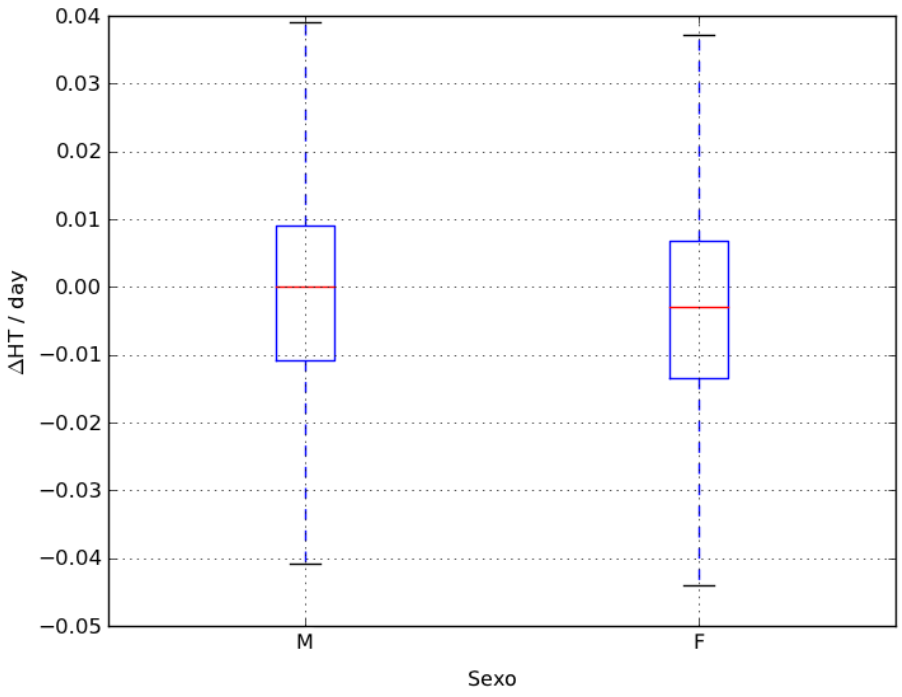


Figura 28: Comparação entre os boxplots de cada sexo.

Note como na figura 29 o histograma feminino sobressai-se ao masculino do lado negativo, enquanto que com o masculino ocorre o oposto. Este é o motivo da diferença entre as médias na figura .

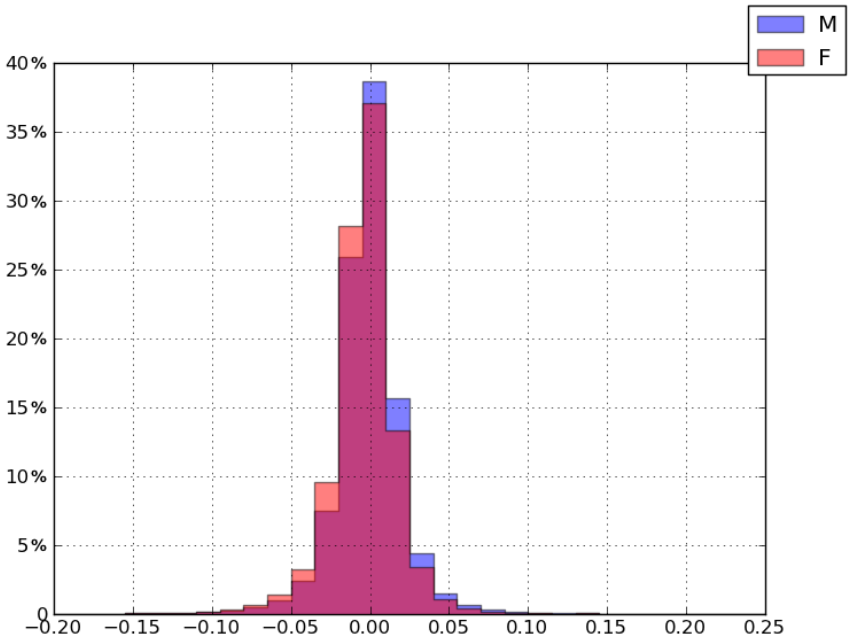


Figura 29: Comparação entre os histogramas de cada sexo.

5.4 Classificação *temporalmente localizada* de um doador

Um dos principais objetivos deste trabalho é o de poder classificar doadores com relação a sua propensão ao desenvolvimento de anemia. Gostaríamos de ser capazes de prever as chances de um doador evoluir para um estado anêmico no futuro e de determinar um intervalo de retorno seguro para que ele não venha a ficar anêmico.

Podemos olhar para os coeficientes angulares calculados e usá-los para classificar os doadores determinando intervalos relevantes que separem doadores em classes de risco. Seria o equivalente a encontrar bons representante para $v_1, v_2, ...v_N$ que dessem significado às classes de 0 a N.

Classe	**Intervalo**
0	$\alpha < v_1$
1	$v_1 \leq \alpha < v_2$
...	...
N-1	$v_{N-1} \leq \alpha < v_N$
N	$v_N \leq \alpha$

Há, entretanto, um problema com esta abordagem. Um mesmo doador possui vários grupos de doações (cf. 4) e, assim, pode pertencer a mais de uma classe. Portanto, os valores de α só podem fornecer uma classificação para o período determinado pelo grupo de doações, ou seja, é uma classificação temporalmente localizada.

5.4.1 Proposta: classificação pela distribuição normal

Definir intervalos e classes como descritos anteriormente é uma tarefa que exige conhecimento mais aprofundado dos processos biológicos envolvidos. Nos limitaremos em propor uma classificação baseado na semelhança percebida nos histogramas de valores de α com a distribuição normal de probabilidades.

A figura 30 mostra o gráfico da função densidade de probabilidade da distribuição normal [15] e destaca, partindo da média (μ) as probabilidades acumuladas entre 1, 2, e 3 desvios padrões (σ); propriedade de qualquer curva normal.

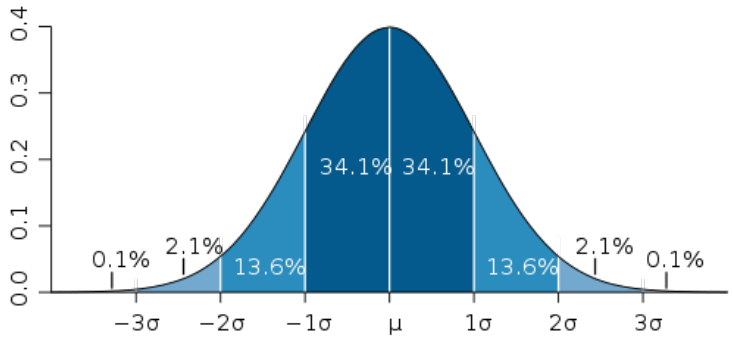


Figura 30: Distribuição normal e suas probabilidades acumuladas.

Os gráficos da figura 32 ilustram como os valores de α adequam-se satisfatoriamente a uma distribuição normal e os da figura 33 exibem um teste visual de aderência a distribuição normal conhecido como *Normal Probability Plot*. [16]

O padrão da figura 31 encontrado nos testes indicam uma distanciação da normal devido a uma cauda longa, ou seja, a variável aleatória em questão apresenta mais variância do que o comum em uma distribuição normal.

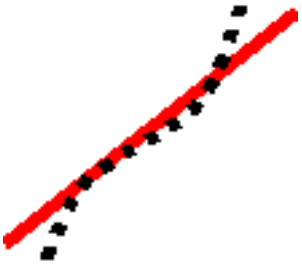


Figura 31: Padrão que indica cauda longa no testes de normalidade.

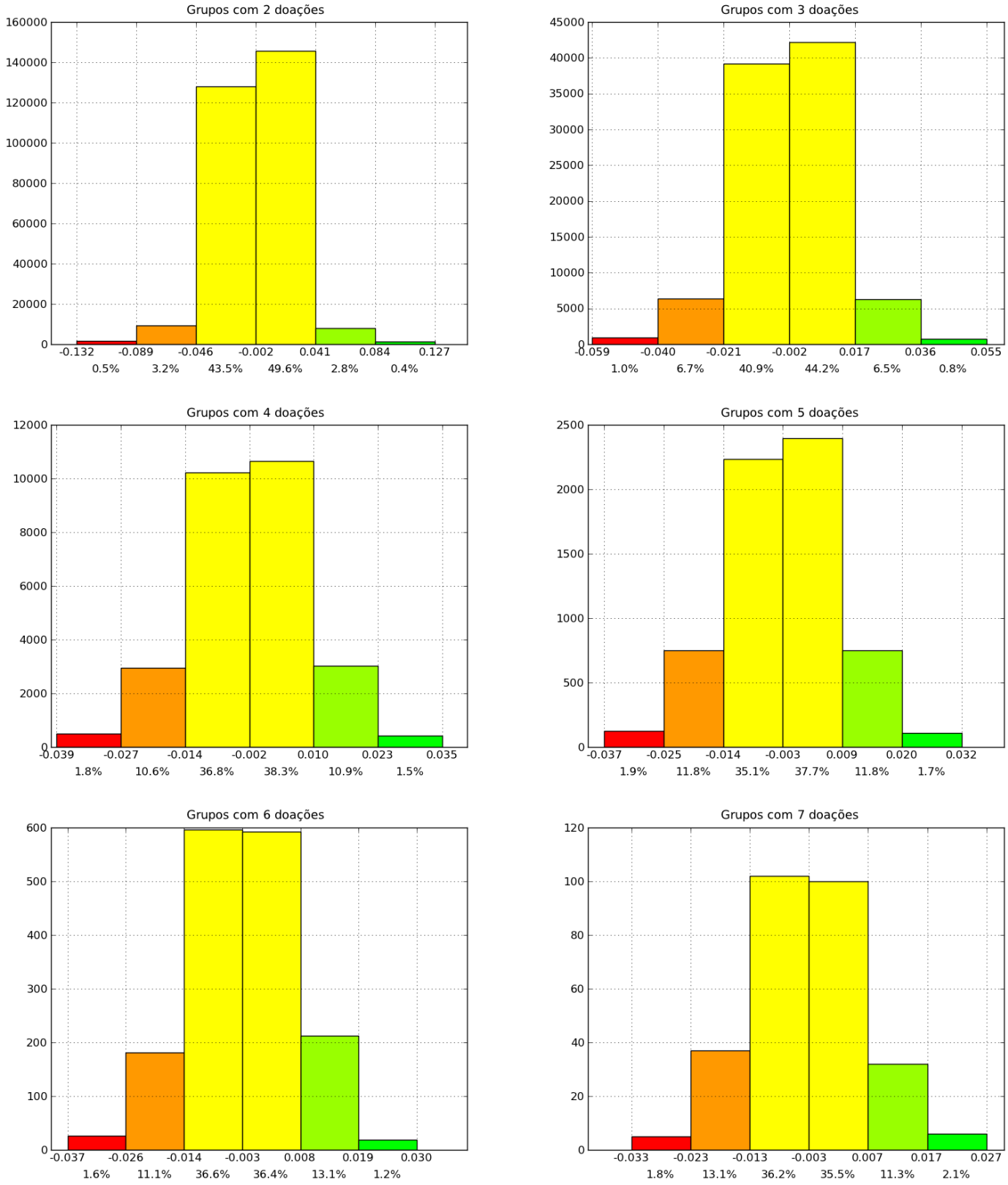


Figura 32: Proposta de classificação de α baseado na distribuição normal.

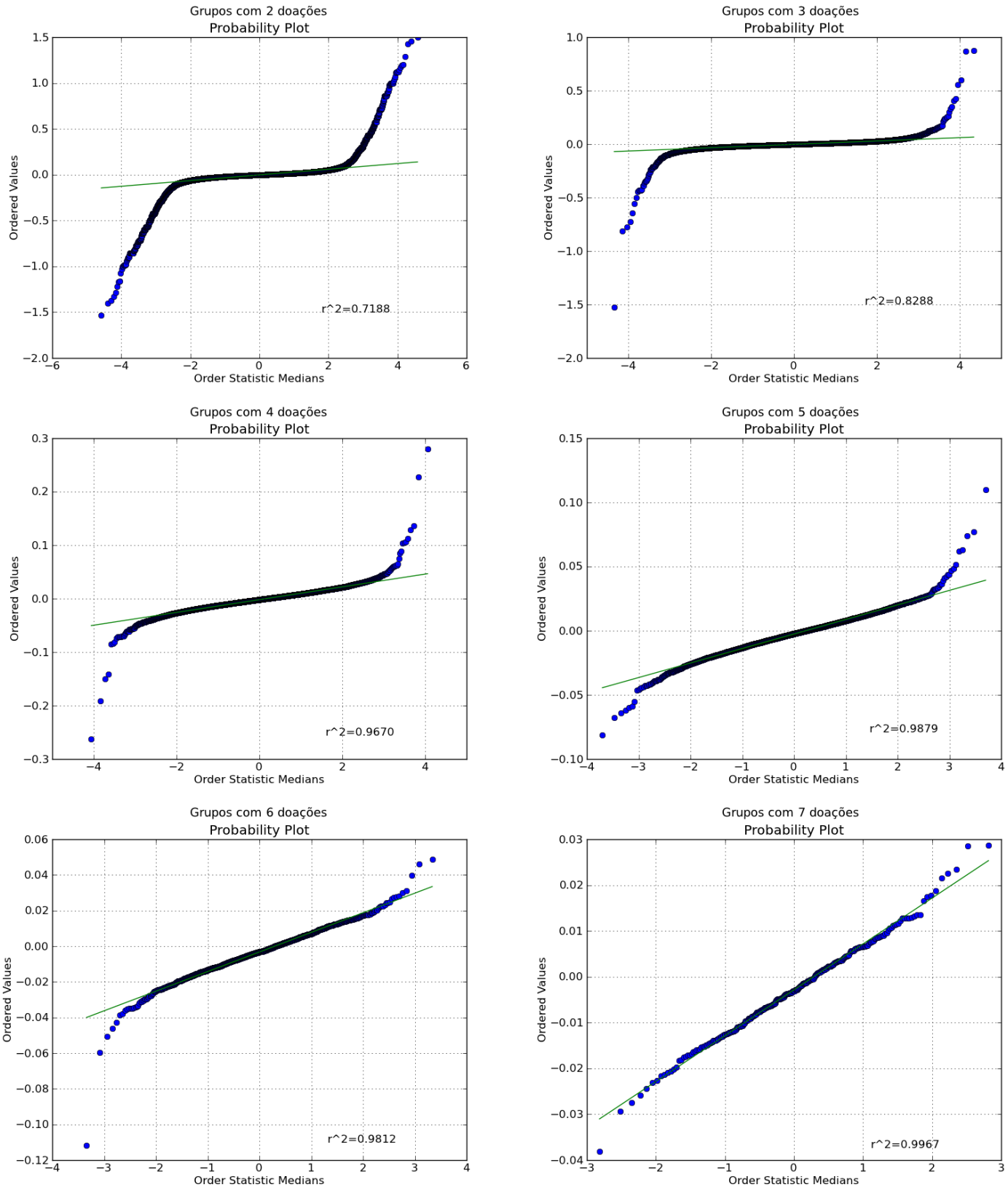


Figura 33: Testes de adequação a curva normal dos histogramas da figura 32.

Assim, para ilustrar uma possível construção de classes. Construímos uma tabela de classificação de doadores baseada na média entre os extremos dos intervalos encontrados na figura 32.

Classe	**Intervalo**
Prejudicado	$\alpha < -0.04$
Sensível	$-0.04 \leq \alpha < -0.02$
Indiferente	$-0.02 \leq \alpha < 0.015$
Insensível	$0.0015 \leq \alpha < 0.033$
Beneficiado	$0.033 \leq \alpha$

6 Conclusão

6.1 Coeficientes angulares

A construção dos coeficientes angulares baseado em grupos de doações localizados no tempo é uma abordagem nova no estudo de anemia em doações de sangue. Este trabalho mostrou que tais coeficientes são métricas válidas para identificação de possíveis casos de anemia.

6.2 Estudos futuros

Há muito o que se explorar e aprimorar no uso dos grupos de doações e sua regressão linear. Seguem algumas possibilidades de estudo:

- Definir, por inferência estatística, as distribuições das seguintes variáveis aleatórias: tempo de retorno de um doador, índice de HT, coeficiente angular de grupo de doações (assemelha-se a uma normal).
- Estudar o comportamento dos sexos pelo número de doações de seus grupos.
- Para tentar encontrar uma classificação única por doador, talvez seja possível construir os caminhos que os valores de α traçam em um autômato de classes locais. A partir disso, poderíamos ponderar a probabilidade de se alcançar um estado anêmico.

7 Parte Subjetiva

7.1 Desafios e frustrações

O projeto que desenvolvi neste trabalho foi desafiador e motivante em sua natureza. Desafiador pois tive de lidar com assuntos que, como cientista da computação, não domino (biologia e estatística). Motivante por saber que meus estudos poderiam ser utilizados na elucidação de problemas reais da medicina, ajudando a evitar anemia em pessoas reais.

Do ponto de vista acadêmico, este foi o trabalho de caráter mais formal que desenvolvi. Foi bastante interessante ter de estudar e escolher bases teóricas para fundamentar o trabalho, eu gostei bastante. Confirmei minha desconfiança de que estatística é uma área muito importante para resolver problemas da área de computação. Sinto por não ter mais conhecimento nessa área.

No primeiro semestre desse ano, o Prof. João Eduardo Ferreira estava nos EUA e, à parte de algumas reuniões que fizemos por Skype, passei a maior parte do tempo estudando os fundamentos teóricos deste trabalho. A programação de fato, por assim dizer, começou somente no segundo semestre. Eu sempre fui uma pessoa bastante ativa, cheia de atividades. Tive bastante dificuldade em conciliar estudos, estágio e as atividades que tenho na igreja em que frequento. Assim, no segundo semestre tomei uma atitude drástica e saí de São Bernardo do Campo e vim morar perto da USP junto com um amigo de turma (valeu, Chico!). Foi bem desgastante, mas graças a Deus deu tudo certo!

7.2 Disciplinas relevantes

MAE0121 - Introdução a Probabilidade e a Estatística I e MAE0212 - Introdução à Probabilidade e à Estatística II

Conheci estimadores estatísticos básicos (como média e variância) e aprendi a construir e interpretar gráfciso, histogramas e boxplots. Também conheci a famosa distribuição normal e fiquei surpreso ao vê-la aparecer nesse trabalho.

MAC0426 - Sistemas de Bancos de Dados

Foi onde aprendi conceitos básicos de modelagem de dados e transações em bancos de dados relacionais. O grande interesse que tenho na área hoje começou nesta matéria.

MAC0439 - Laboratório de Bancos de Dados

Conheci conceitos avançados de bancos de dados, inclusive os bancos de dados multidimensionais. Desenvolvi um projeto prático de construir um bom modelo de dados para um sistema real. Foi nesta matéria, também, onde conheci o Prof. João Eduardo Ferreira, meu orientador.

MAC0441 - Programação Orientada a Objetos

Aprendi a reconhecer erros de modelagem e a fazer bons desenhos orientados a objeto. Conheci o padrão de projeto chamado *Singleton* e utilizei-o neste trabalho na classe que interagia com o banco de dados.

7.3 Continuação dos estudos

Como já disse, gostei bastante de desenvolver este trabalho e isso me serve de grande incentivo para ingressar em um mestrado. Ainda tenho o 1º semestre do ano que vem para terminar a graduação e, talvez eu possa continuar este trabalho sob a forma de uma iniciação científica. Se isso for acontecer, vou precisar estudar mais estatítica. Confesso, entretanto que estou cansado do tema e gostaria de trablhar com outra coisa.

Referências

[1] P. P. S. B. Silva: Atualização Incremental De Data Warehouses, 2009. http://www.ime.usp.br/~cef/mac499-09/monografias/pedro-paulo/ Acessado Setembro/2011.

[2] J. E. Ferreira, I. C. Italiano, O. K. Takai: Introdução a banco de dados, 2005. http://www.ime.usp.br/~jef/apostila.pdf. Acessado Setembro/2011.

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

[4] E. Baralis, S. Paraboschi, E. Teniente: Materialized view selection in a Multidimensional Database. Proceedings of the 23rd VLDB Conference, Atenas, Grécia, 1997.

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

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

[7] Wikipedia: Hemácia. Doenças e ferramentas de diagnóstico. http://pt.wikipedia.org/wiki/Hemácia#Doenças_e_ferramentas_de_diagnóstico Acessado Setembro/2011.

[8] Fundação PRÓ-SANGUE. http://www.prosangue.sp.gov.br Acessado Novembro/2011.

[9] PostgreSQL, a powerful, open source object-relational database system. http://www.postgresql.org/

[10] Python. http://python.org

[11] psycopg2, the most popular PostgreSQL adapter for Python. http://initd.org/psycopg/

[12] NumPy, scientific computing with Python. http://numpy.scipy.org/

[13] matplotlib, a python 2D plotting library. http://matplotlib.sourceforge.net/

[14] E. Gamma, R. Helm, R. Johnson e J. Vlissides: Design Pattern. Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995; 127-134.

[15] M.N. Magalhães e A.C.P. de Lima: Noções de Probabilidade e Estatística, Edusp, 2008; 6:183-187

[16] Wikipedia: Normal Probability Plot. http://en.wikipedia.org/wiki/Normal_probability_plot Acessado Novembro/2011.