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

Trabalho de conclusão de curso
IME-USP

Aluno: André Casimiro
Orientador: João Eduardo Ferreira

Contextualização

Projeto internacional REDS-II: segurança em transfusões de sangue.

Data-IME fez um Data Warehouse para integrar dados de 3 hemocentros (SP, MG, PE).

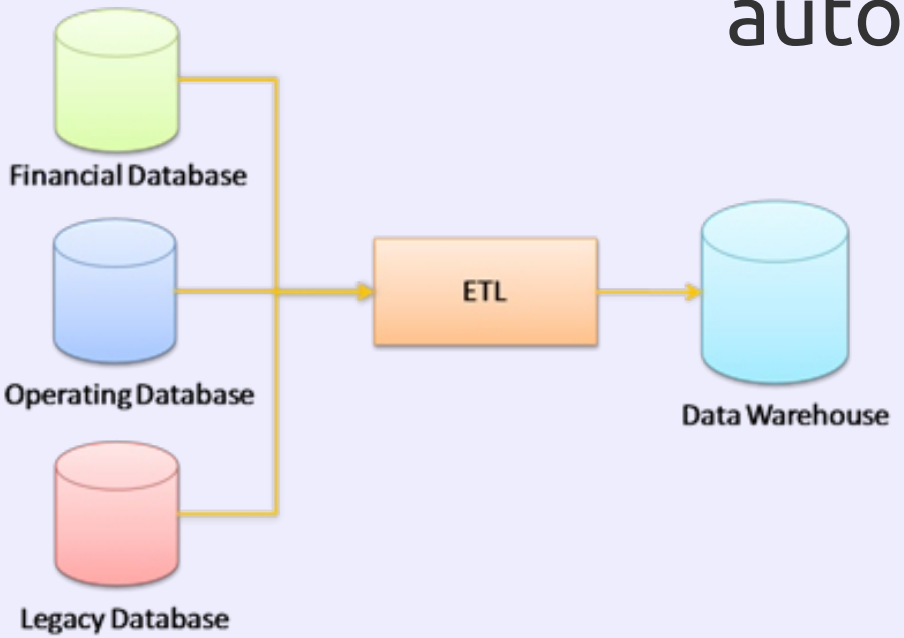
Estudar a relação entre doações de sangue e anemia.

Data Warehouse
(depósito de dados)

Grandes quantidades de dados

Integra bases de dados distribuídas, autônomas e heterogêneas

É OLAP, e não OLTP



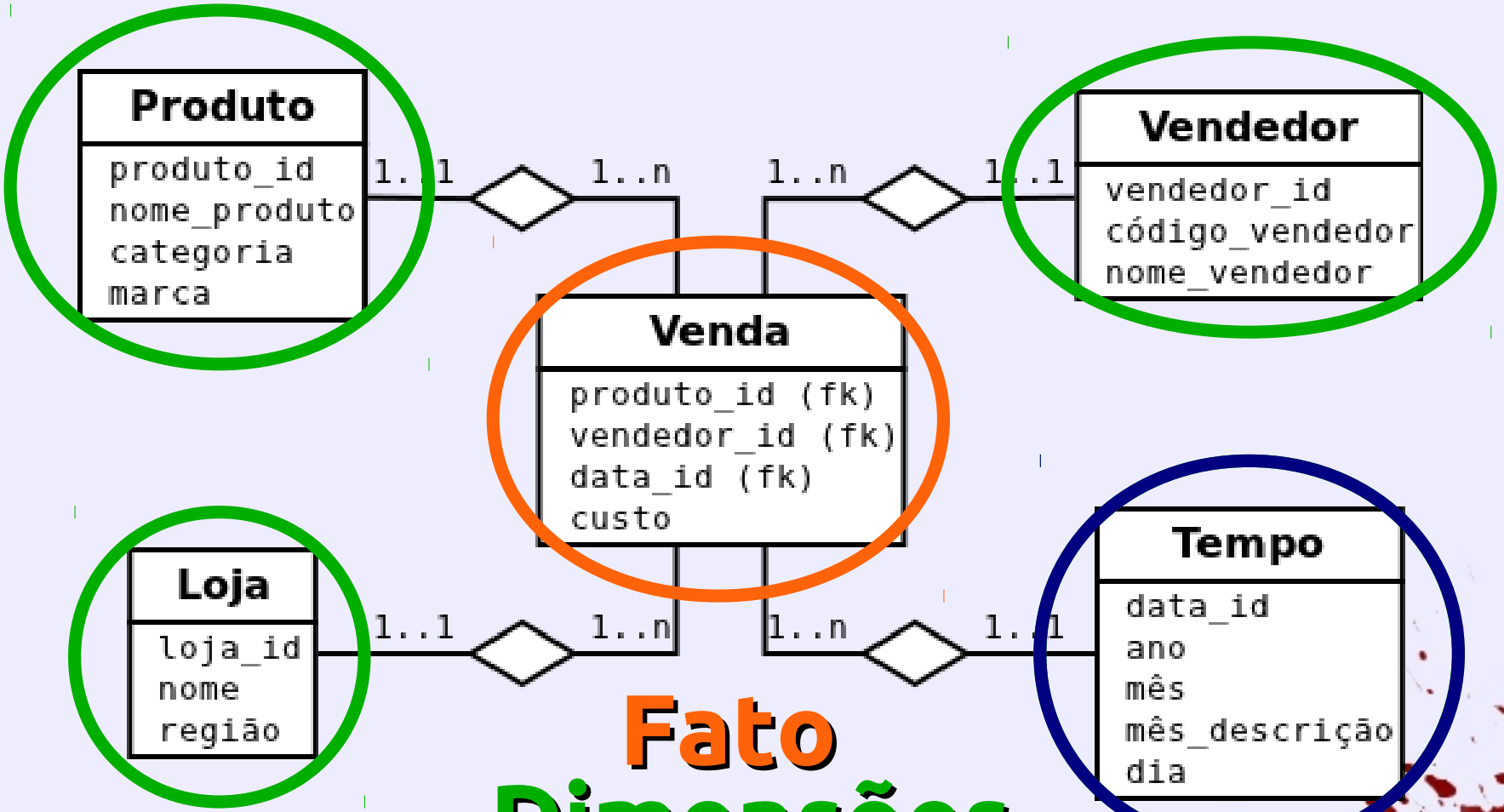
Modelo Multidimensional

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 que identifica unicamente cada tupla (*d_i *é a chave primária de* D_i*).*
- 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 um conjunto de atributos adicionais* V *(que representam os valores sobre os quais serão aplicadas as funções de agregação)*

Não tem exemplo???

Modelo Multidimensional



Redução Dimensional

No exemplo anterior, imagine que há:

- 40 vendedores (João, José, Pedro...)
- 10 lojas (Butantã, Av. Paulista, SBC...)
- 50 produtos (refrigerante, bala, suco...)

O número de séries temporais distintas será:

$$40 \times 10 \times 50 = \mathbf{20000}$$

Análise manual inviável.
Como reduzir esse número?

Redução Dimensional

Classificação dos elementos das dimensões pelo seu **comportamento ao longo do tempo**:

- Vendedores: {bom, médio, ruim}
- Lojas: {nobre, periferia}
- Produtos: {sazonais, permanentes}

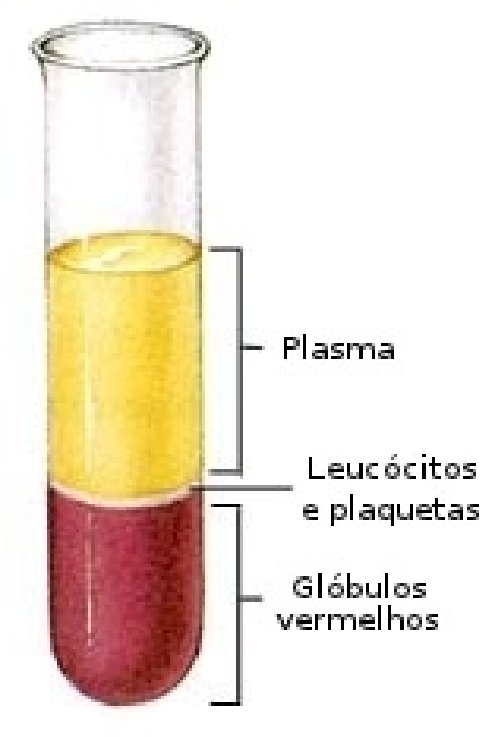
E o número de séries temporais...

3 x 2 x 2 = **12 :)**

Hematócrito e Anemia

Hematócrito (HT) é à **porcentagem ocupada pelos glóbulos vermelhos** no volume total de sangue.

Varia de 36% a 52% conforme sexo e idade.



Hematócrito e Anemia

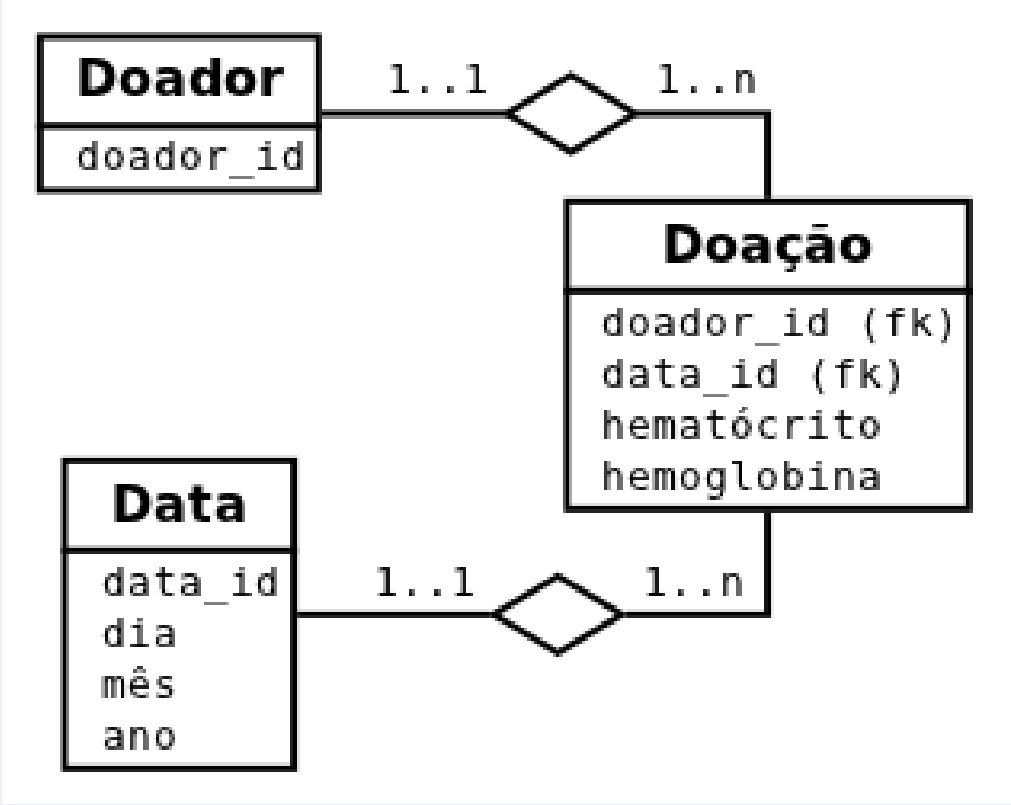
Anemia é a doença caracterizada pela capacidade diminuída de transporte de oxigênio devido a diminuição da contagem de glóbulos vermelhos.

Uma pessoa está anêmica se:

♂ HT < 39%

♀ HT < 38%

O Modelo de Doações de Sangue



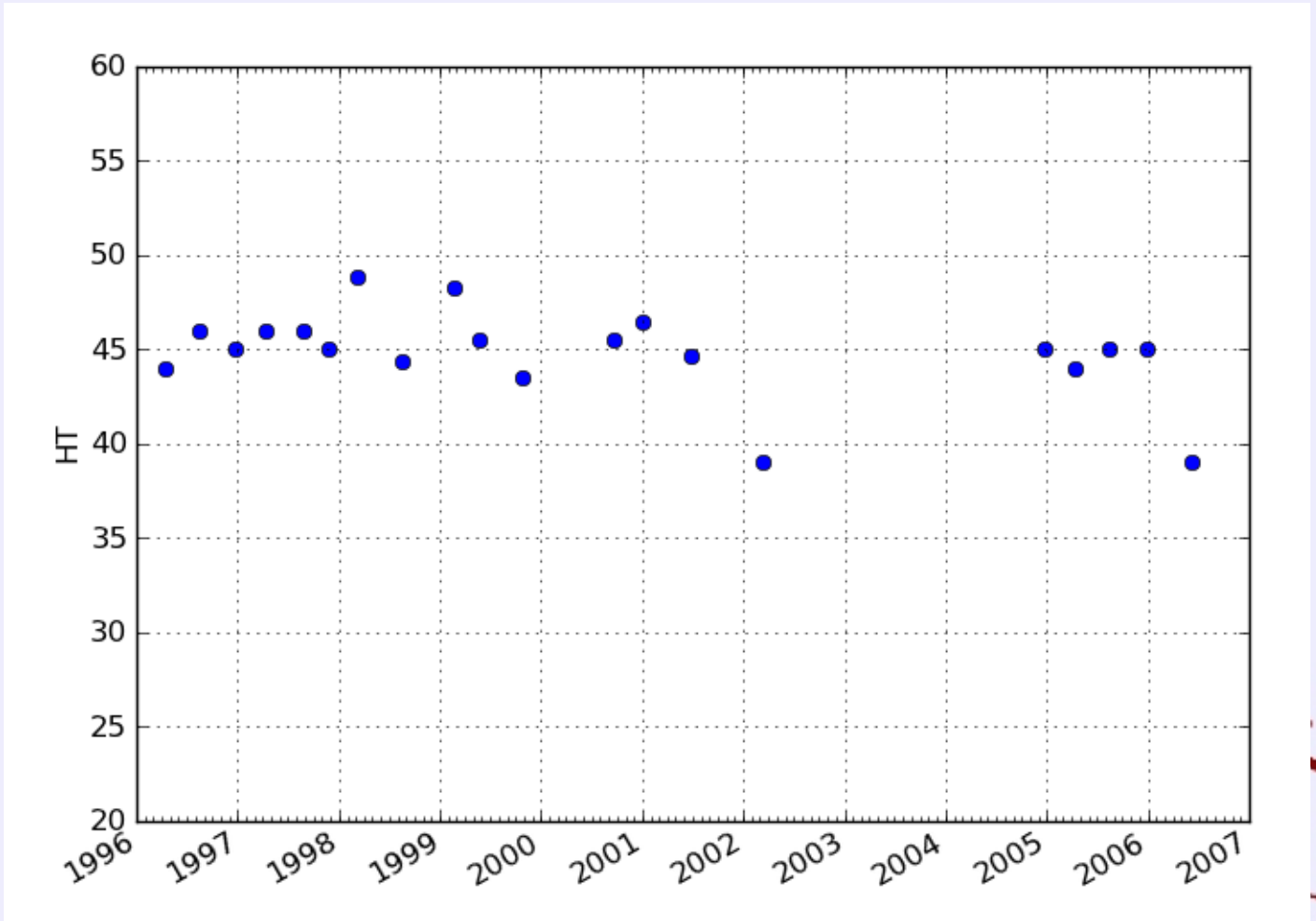
Estatísticas:

- 1.366.197 doadores
- 2.650.499 doações
- 4013 dias (01/01/96 ~ 31/12/06)

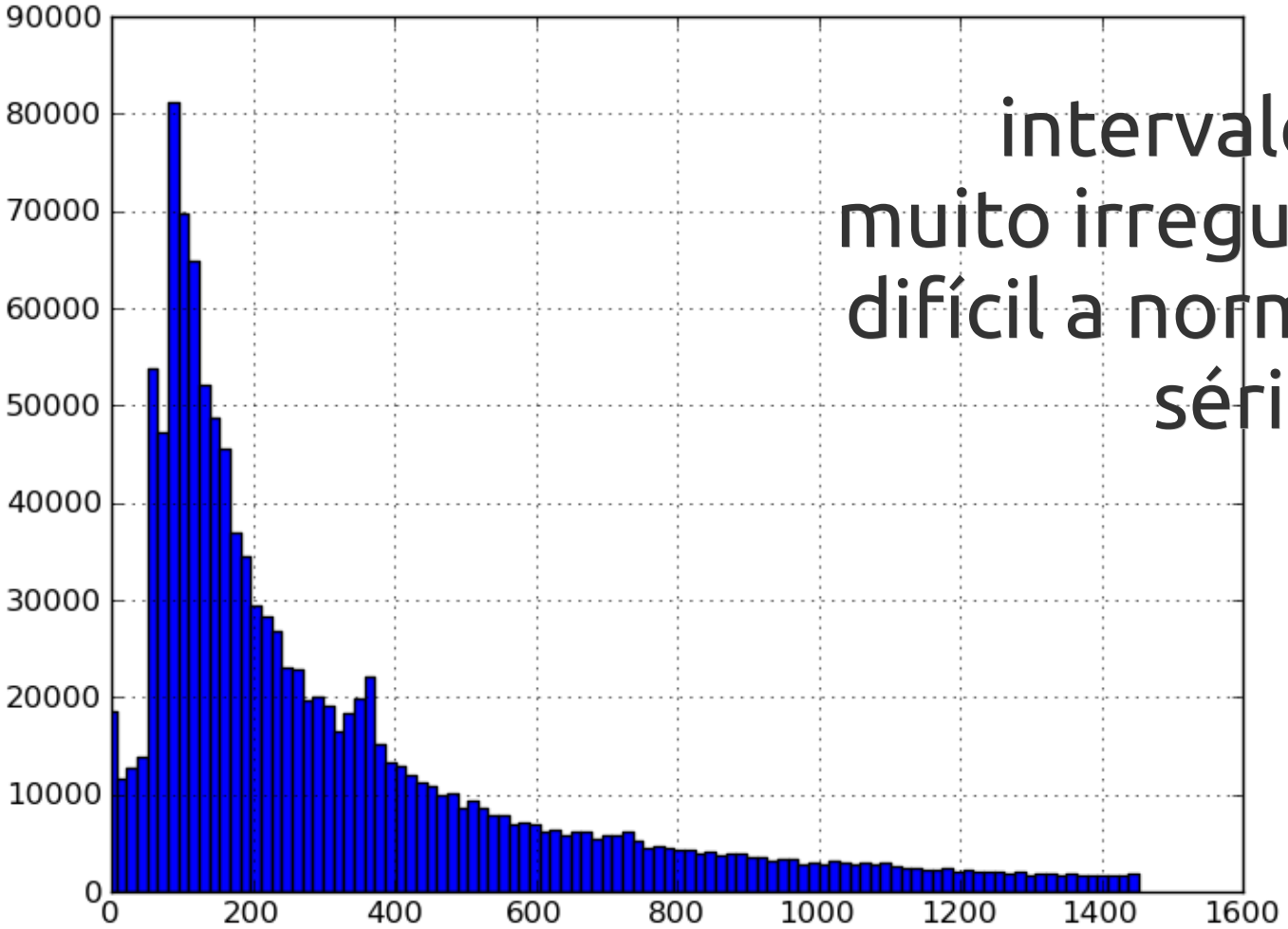
HT = 3 * HB

Série Temporal de HT

Em cada doação colhe-se o nível de HT



Série Temporal de HT



PROBLEMA

intervalos de doação muito irregulares tornam difícil a normalização das séries temporais

Simplificação do Problema

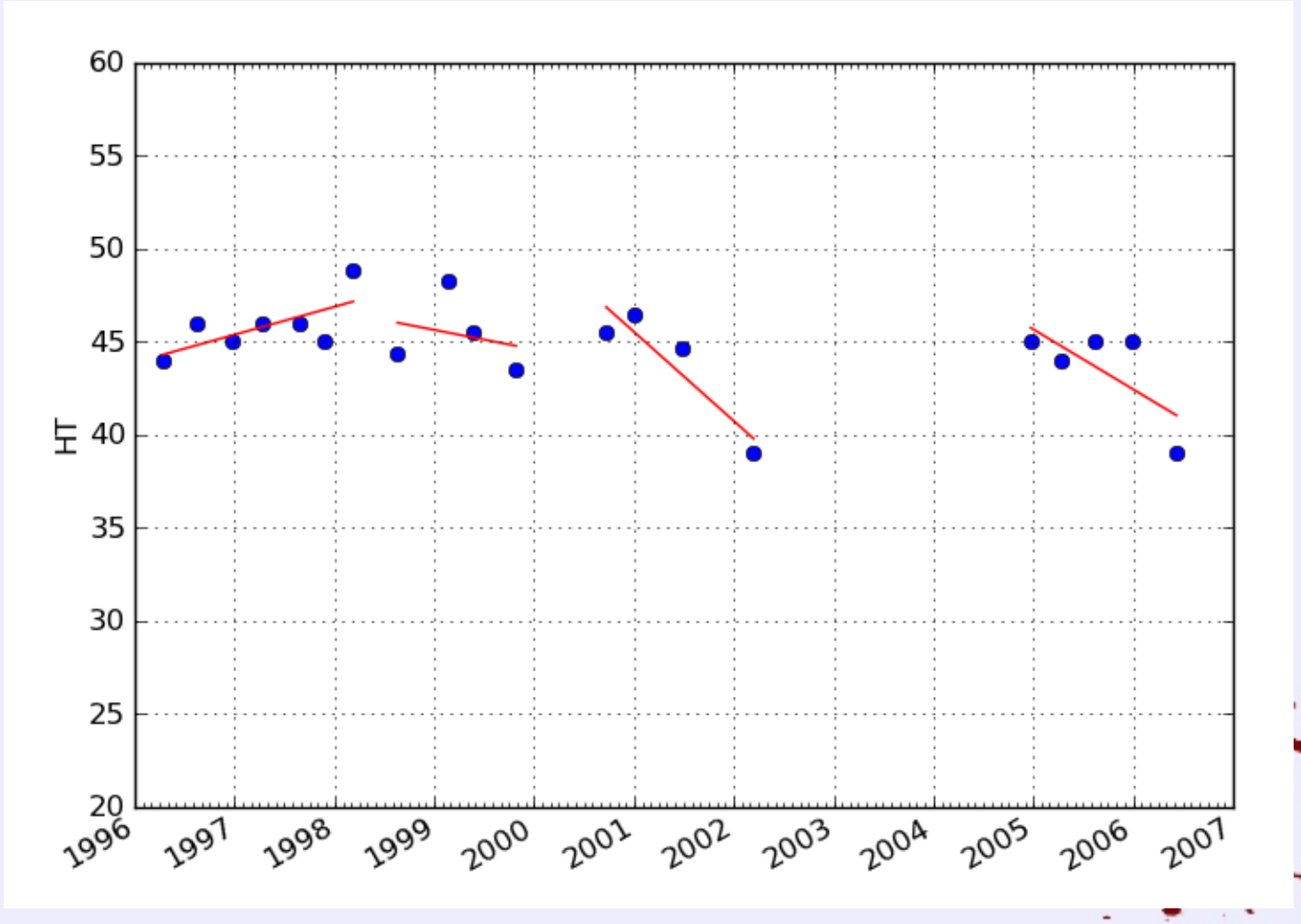
Para poder continuar o estudo resolvemos adotar uma abordagem simples e geral, mas que ainda não foi explorada.

A partir das séries temporais, separamos as doações em grupos temporalmente localizados, aproximamos uma reta e utilizamos o **coeficiente angular** (α) como representante daquele grupo.

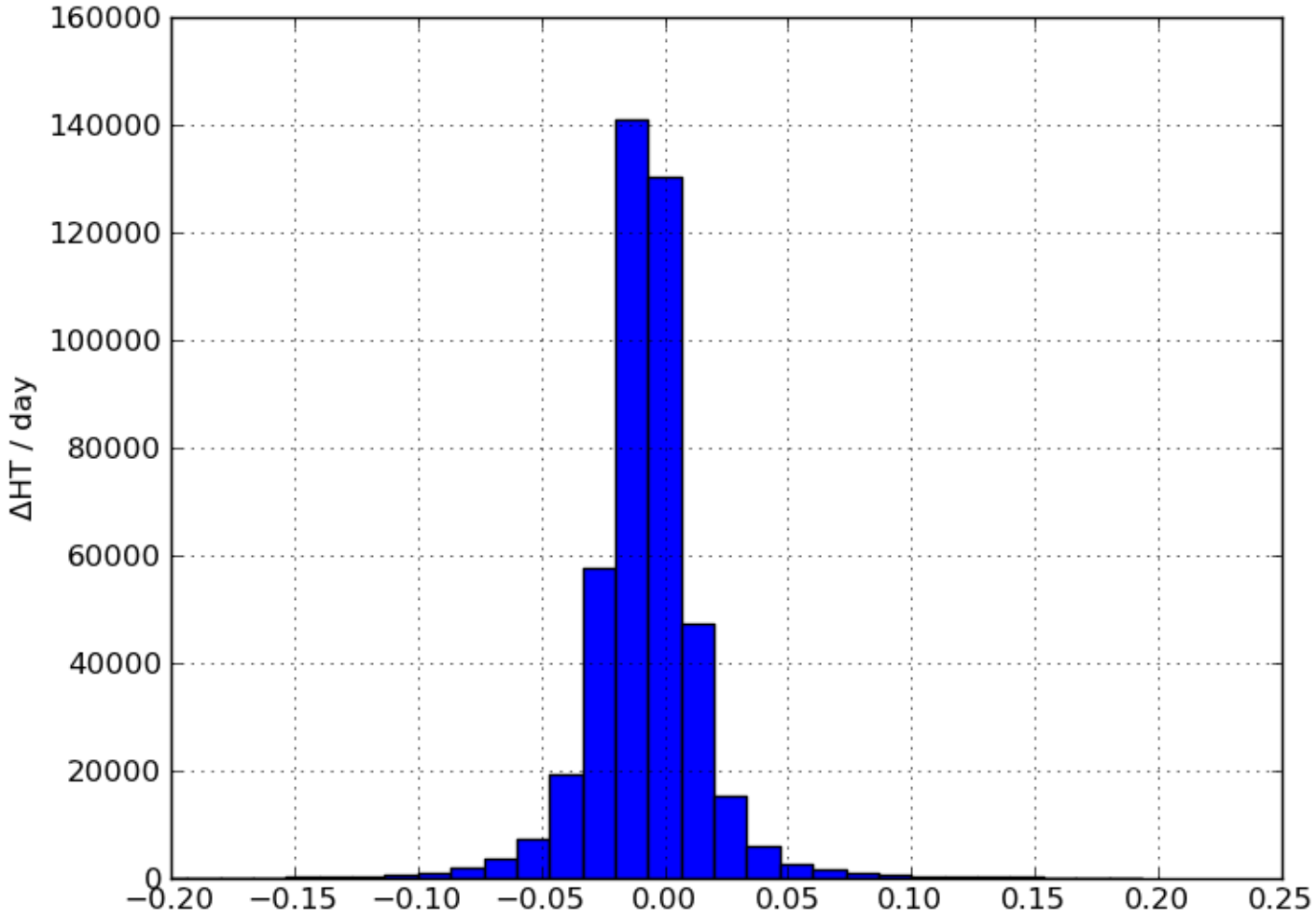
Não tem figura???

Simplificação do Problema

Grupos de doações e suas respectivas retas de aproximação (cada uma tem um α)



Resultados



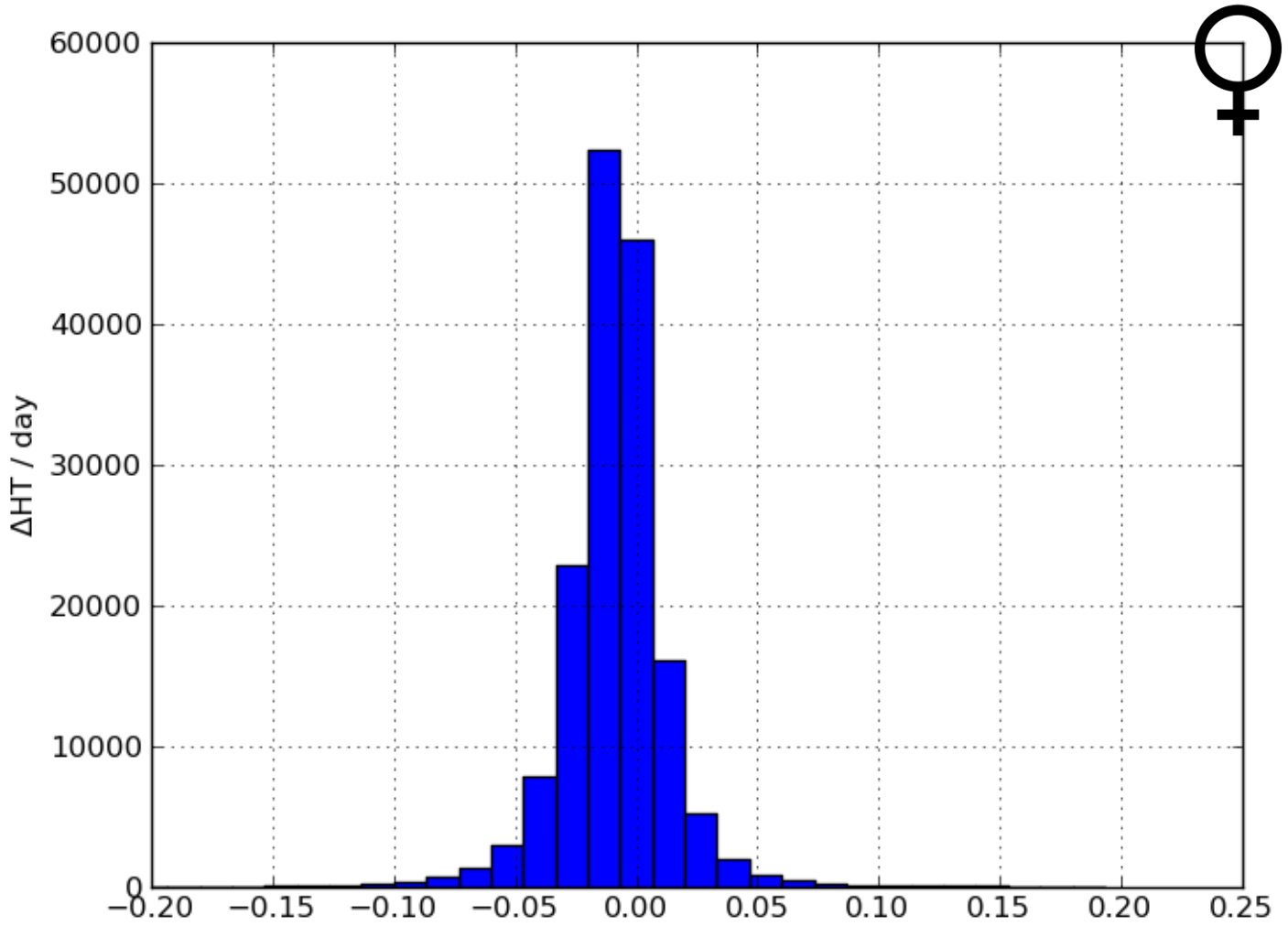
Coeficientes seguem sempre um mesmo padrão

Resultados

Doações por grupo	Número de grupos (%)	Média ($*10^3$)	Desvio Padrão ($*10^3$)
2	313237 (68.44%)	-1.21082926554	48.4493775925
3	101041 (22.08%)	-1.51009191388	15.8126076667
4	31698 (6.93%)	-1.82756254255	12.3719006218
5	8296 (1.81%)	-2.22804637993	11.6145182359
6	2452 (0.54%)	-2.96952312405	11.1580344505
7	617 (0.13%)	-2.33875606638	10.4688665234
8	193 (0.04%)	-4.47660616743	13.0620859077
9	61 (0.01%)	-2.75941067998	9.18043146082
10	35 (0.01%)	-6.02224674979	10.380046483
11	32 (0.01%)	-2.85877985861	9.33093668725
12+	13 (0.00%)	-3.08502923848	6.58982678328

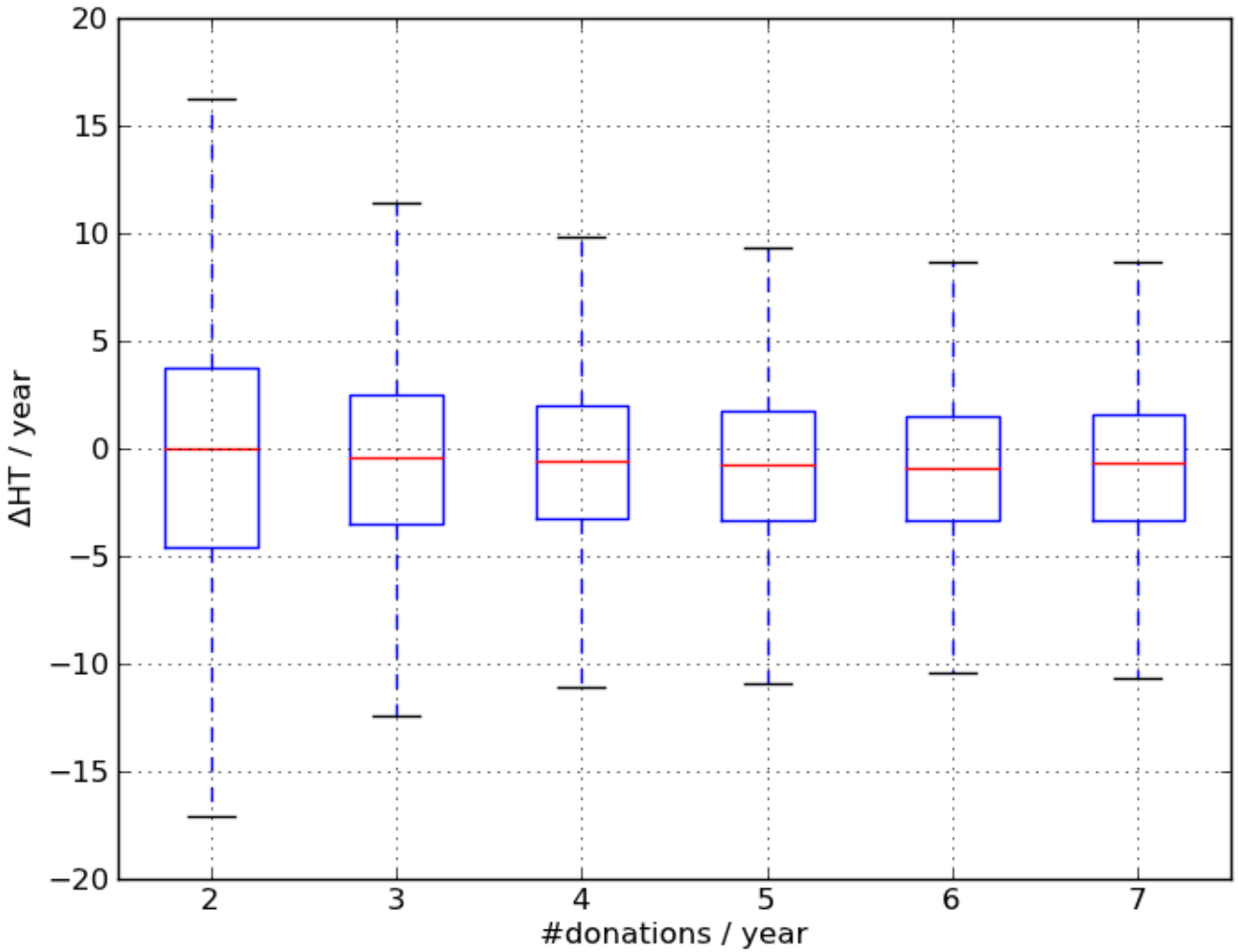
A média dos (α) é sempre negativa e muito próxima de 0

Resultados



Sexo não influi na distribuição dos (α)

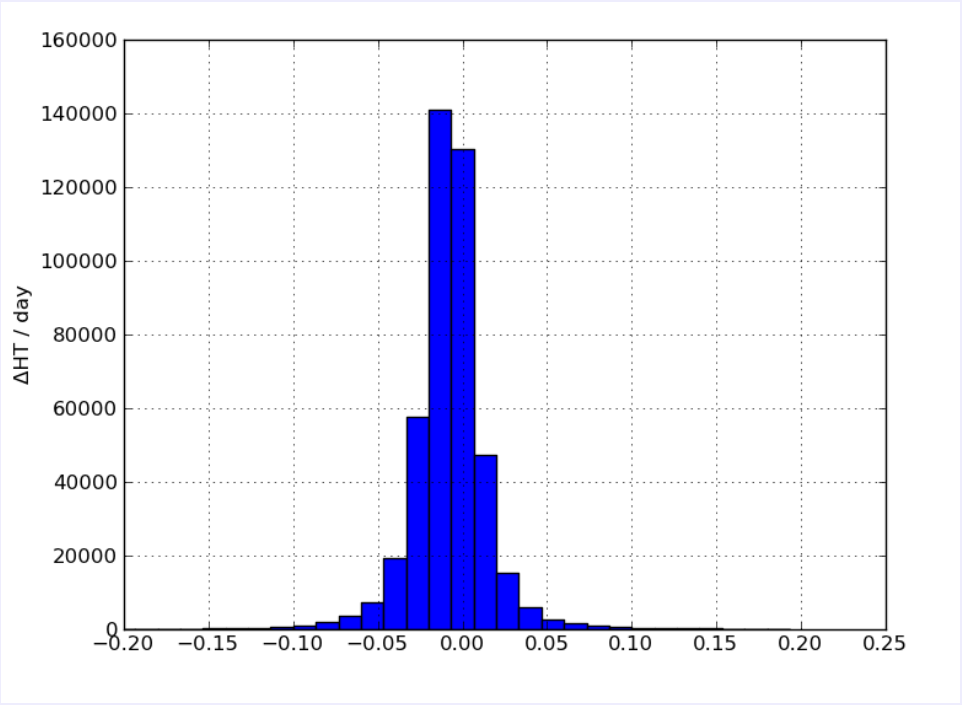
Resultados



Número de doações por ano faz a média cair

Resultados

Com base neste comportamento dos coeficientes (α), decidimos dividí-los em 4 classes distintas:



Sensível
[α < -0.04]

Negativo
[-0.04 < α < 0]

Positivo
[0 < α < 0.02]

Insensível
[0.02 < α]

Perguntas

?

Obrigado! :)