Analisando um *dataset*: Acidentes em POA

Gustavo J. Feller; *William Lopes*

INF/EA

Novembro de 2015

1. Objetivo de negócio
2. Objetivo de mineração
3. Dataset escolhido
4. Pré-processamento
5. Mineração de dados

Objetivo de negócio

Objetivo de negócio

- Posicionar-se como gestor na área de trânsito na cidade de Porto Alegre
- Através de técnicas de Descoberta de Conhecimento em Base de Dados extrair informações para auxiliar na gestão do trânsito
- Utilizar a decisão baseada em fatos para recomendar modificações ou açõ
- Buscar a melhoria na segurança do trânsito em Porto Alegre
- Analisar observações de comportamento de acidentes de trânsito para evitar futuros acidentes
- Com os resultados aplicar modificações no trânsito (Sinalização, conscientização, intervenção, etc...)

Objetivo de mineração

Objetivo de mineração

- Compreender a situação dos acidentes de trânsito na cidade de Porto Alegre
- Fornecer uma base para a tomada de decisões do objetivo de negócio
- Evidenciar informações relevantes para auxiliar na tomada de decisão

Dataset escolhido

Dataset escolhido

- A base de dados escolhida foi o *dataset* fornecido pelo governo municipal em relação aos acidentes de trânsito registrados em Porto Alegre
- A base escolhida conta com 14 *datasets* sendo que cada *dataset* representa um ano de registros, compreendendo o período de 2000 a 2014.
- Os dados são públicos e estão disponíveis para consulta juntamente com vários conjuntos de *datasets* de outros setores do governo.
- Outros dois datasets com a localização dos pardais e a localização das lombadas eletrônicas foram utilizados.

Dataset escolhido

Tabela: Estrutura do *dataset*

Campo	Descritivo
Id	Sequencial
Log1	Logradouro
Log2	Logradouro 2
Predial1	Número próximo
Local	Tipo de logradouro
Tipo_Acid	Tipo de acidente
Local_Via	Endereço acidente
Data_hora	Data e hora do ac.
Dia_sem	Dia da semana
Feridos	Num. feridos
Morte	Num. Mortes
Morte_post	Num. Mortes posterior
Fatais	Acidentes fatais
Auto	Automóveis envolvidos
Taxi	Taxis envolvidos
Lotacao	Lotações envolvidas
Onibus_urb	Ônibus urb. envolvidos
Onibus_int	Ônibus int. envolvidos
Caminhão	Caminhões envolvidos

Campo	Descritivo
Moto	Motos envolvidas
Carroca	Carroças envolvidas
Bicicleta	Bicicletas envolvidas
Outro	Outros veículos envolvidos
Tempo	Cond. Meteorológica
Noite_Dia	Momento do acidente
Fonte	Fonte da ocorrencia
Boletim	Num. Boletim ocorrencia
Regiao	Região da cidade
Dia	Dia do acidente
Mes	Mês do acidente
Ano	Ano do acidente
Fx_hora	Faixa de hora do acidente
Cont_acid	Núm. de carros envolvidos
Cont_Vit	Núm. de vítimas no momento
UPS	Núm. de atendidos por UPS
Latitude	Latitude
Longitude	Longitude

Pré-processamento

Pré-processamento

- Retirada de campos não considerados:
  - **Id:** Sequencial do arquivo, considerado como não relevante, visto informações de data e hora do acidente.
  - **Boletim:** Boletim de ocorrência, sequencial sem relação com outros campos.
  - **Data_Hora:** Data e hora do acidente, campo duplicado, considerado que o campo "FX_hora" é mais relevante para classificação.
  - **Data:** Duplicado de "Data_Hora" e campos "Dia", "Mes" e "Ano" contém as informações a um nível mais detalhado
  - **Fonte:** O campo fonte do registro do acidente não se enquadrou nos objetivos do presente trabalho

Pré-processamento

- Análise de campos vazios (Ex: Log2 , Hora $< 1\%$) - Linhas com colunas relevantes que estavam vazias foram retiradas (3 linhas no total)
- Coluna de fatalidade (S para sim ou N para não), existe a coluna com quantidade de fatalidades e de mortes pós acidente.
- Criação de um dataset auxiliar apenas com as linhas ao qual houveram acidentes fatais (LOG1, LOCAL, TIPO ACID, DIA SEM, TEMPO, NOITE DIA e REGIAO).
- Datasets auxiliares (lombadas e pardais) foram designados latitude e longitude pois tinham apenas uma coluna, o endereço, sendo retirado dados duplicados.
- Ajustes necessários em latitude e longitude, mesmo sendo um dataset preparado para kdd

Mineração de dados

Mineração de dados - Análise dos dados

- Analisados os últimos 4 anos, sendo que foram trabalhados:
  - Histogramas para conhecimento da estrutura (horários acidentes, dias acidentes, meses acidentes)
  - "Ranges"por dataset
  - Algoritmo *apriori* em características de acidentes fatais
  - Tentativa de k-nn
  - Visualização de densidade de acidentes (*heatmap*) em mapa (latitude/longitude) com localização de lombadas e pardais

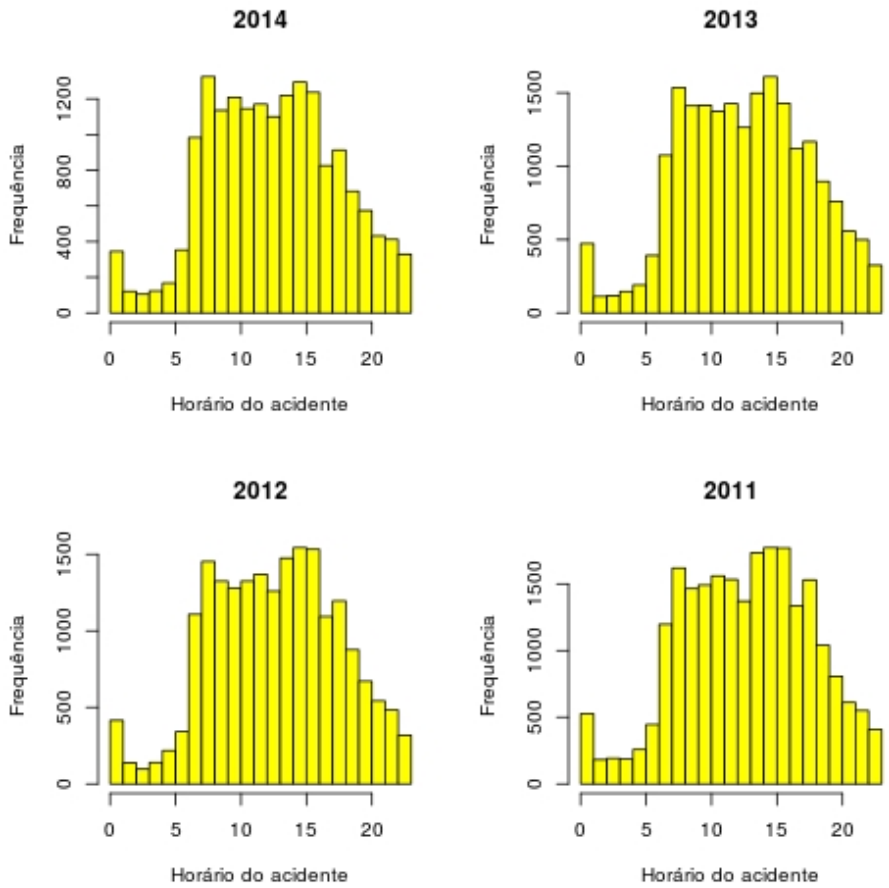
Mineração de dados - Análise dos dados

Tabela: Número de acidentes por ano e número de acidentes fatais por ano

Ano	Acid.	Fatais	% Fatais	†	Máx. †	Máx. † pós	Máx tot. †
2011	23.579	135	0,57	141	3	2	3
2012	20.202	97	0,48	100	1	2	2
2013	20.799	117	0,56	124	4	1	5
2014	17.203	135	0,78	135	1	1	1
Total	81.703	484	0,59	500	-	-	-

Mineração de dados - Histogramas

Figura: Histogramas dos horários dos acidentes entre 2011 e 2014



Mineração de dados - Correlação

Tabela: Correlação entre acidentes com tipo de condução em 2014

Tipo	Auto	Taxi	Lot.	O.U.	O.M.	O.I.	Cam.	Mot.	Car.	Bic.	Out.
†	-0.054	-0.018	-0.001	0.022	0.011	0.003	0.018	0.025	-0.001	0.037	-0.003
F.G.	-0.163	-0.031	-0.017	0.005	0.010	-0.004	-0.038	0.187	0.014	0.024	0.001

Algoritmo *apriori*

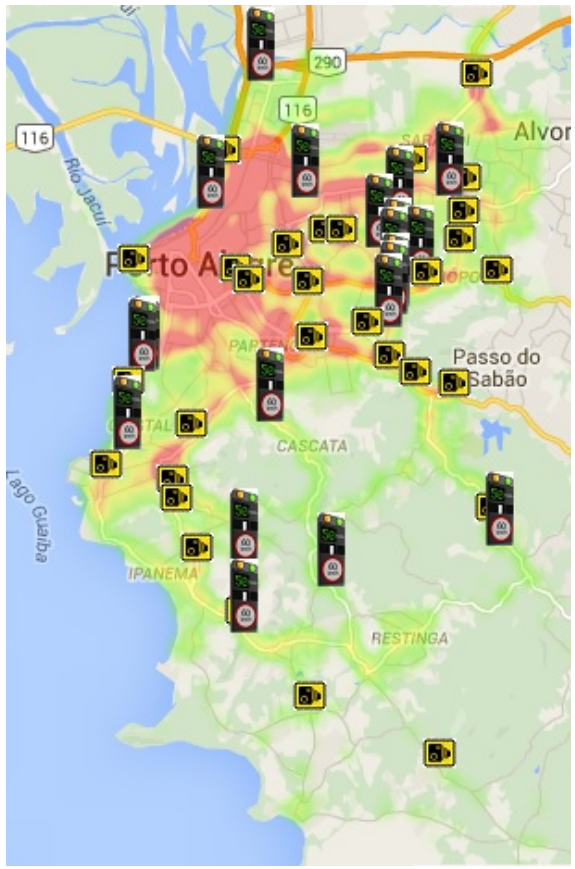
- Aplicado ao dataset completo, sem relevância para acidentes do tipo fatal
- Dataset auxiliar apenas com os acidentes fatais foi utilizado
  - Atropelamentos ocorrem principalmente no centro e de dia
  - Av. Baltazar de Oliveira Garcia tem forte relação com acidentes fatais à noite

Mineração dos dados - Densidade de acidentes

- Adicionado densidade ao mapa por tipo de acidente, pela latitude e longitude
- Interatividade via Google Maps
- Utilizado os datasets das lombadas eletronicas e pardais

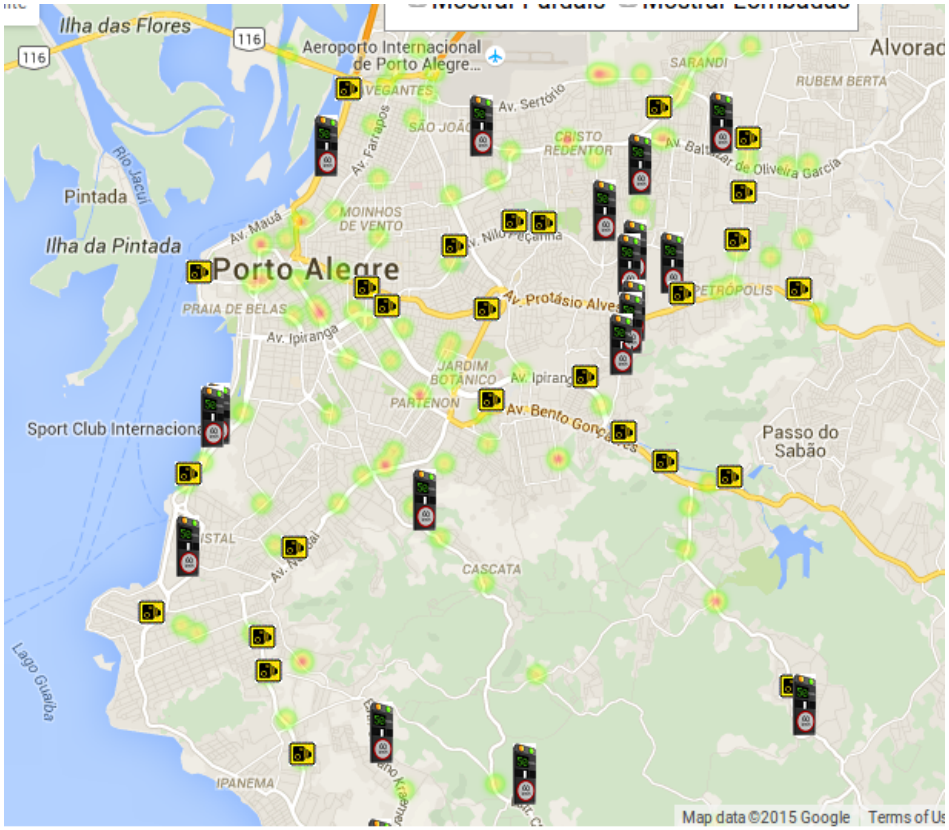
Mineração dos dados - Densidade de acidentes

Figura: Heatmap com os acidentes no ano de 2014



Mineração dos dados - Densidade de acidentes

Figura: Heatmap com os acidentes fatais no ano de 2014



Conclusões

- Os histogramas de densidade no horário de acidente inicia-se às 6 horas da manhã e tem seu pico às 3 horas da tarde, diminuindo consideravelmente após esse horário.
- Também, o mapa com a densidade dos acidentes comprovou que locais com controladores de velocidade e pardais os acidentes fatais são bem raros, mostrando-se um sistema eficiente para diminuição dos acidentes. Outro ponto importante que pode ser visto no mapa são os locais com maiores quantidades de acidentes.
- Também notou-se que os acidentes mais graves ocorrem em entradas de vias principais, como, por exemplo, a Av. Bento Gonçalves. Nas próprias avenidas não há registros de acidentes graves, porém, logo após as entradas secundárias destas avenidas pode-se verificar registros com densidades consideráveis de serem avaliadas.

Sugestões

- Melhorar sinalização nas esquinas, principalmente nas vias de entrada das principais avenidas por conterem o maior número de acidentes do tipo fatais (no heatmap fica claro).
- Em vias com um grande número de acidentes fatais diminuir a velocidade máxima, como, por exemplo, a Estrada João de Oliveira Remião.
- Aumentar a sinalização onde as densidades dos acidentes ocorrem mais
- Refazer o KDD anualmente, na medida que as informações dos datasets estão disponíveis para verificar o comportamento do trânsito e inferir novas ações;
- Sugestão de estudos futuros: *Dataset* de blitz X acidentes fatais à noite.

Obrigado!

Gustavo J. Feller; *William Lopes*
INF/EA