Envirocar visualization exploring an environmental and traffic data set

Rodrigo Claro Zembrzuski
Programa de Pos Graduacao em Ciencia da Computacao
Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre, Brasil
Web page: http://inf.ufrgs.br/∼rczembrzuski

***Abstract*—Envirocar eh um projeto que coleta e armazena dados ambientais e dados de trafego de automoveis. Sao dados de sensores implantados dentro dos veiculos que trazem informacoes precisas sobre navegacao dos carros e dados relativos a conservacao do meio-ambiente, como emissao de gas-carbonico e consumo do veiculo.**

Esse artigo tem por objetivo caracterizar os dados e apresentar uma serie de visualizacoes para analise exploratoria desse banco de dados. Com isso, sera obtida uma intuicao sobre o tipo de dados apresentado. Alem disso, serao obtidas informacoes acuradas sobre a viabilidade do uso desse banco de dados para modelagem de algoritmos de aprendizado de maquina.

***Keywords*-envirocar; smartcars; kibana; geolocation; data visualization;**

I. Introduction

Envirocar eh um projeto da Universidade de Munster. O objetivo do projeto eh oferecer uma plataforma simples para coleta, armazenamento e distribuicao de dados ambientais e dados de logistica de trafego.

A plataforma une dados gerados por sensores comuns, que vem na maioria dos automoveis atuais com sensores de geolocalizacao, presente na maioria dos smartphones. Com isso, cria-se uma base de dados supostamente relevante para que seja medido o fluxo de automoveis nas avenidas e quantidade de emissao de poluentes nas cidades.

A. Motivacao

Um dos grandes desafios da sociedade nos nossos dias eh preservar e desenvolver a mobilidade urbana de maneira sustentavel. Isto eh, mitigar o impacto do transporte de passageiros meio-ambiente e no bem-estar geral das pessoas [ref].

Uma base de dados consistente eh crucial para que se confirme – ou refute – hipoteses relativas a comportamento no transito em diferentes partes do globo. Alem disso, a base consistente pode ser muito importante para a inferencia de padroes comportamento e de pilotagem dos motoristas de automovel.

Uma analise exploratoria eh um passo necessario para que se verifique se a base realmente pode ser utilizada com os propositos citados ou se ela deve ser descartada.

B. Objetivos

O objetivo desse trabalho eh fazer uma exploracao basica pela base dados para ganhar intuicao sobre ela. Serao observados padroes basicos de comportamento em diferentes partes do globo e tambem padroes individuais de pilotagem para determinados motorias.

Com isso, sera determinado se o uso dessa base pode ser aconselhavel em experimentos mais complexos, como treinamento de modelos de aprendizagem .

Contribuicoes: O trabalho dara uma sugestao sobre a possibilidade de uso da base de dados envirocar em projetos de aprendizagem de maquina. Tambem sera feita uma analise comparativa entre diferentes tecnicas de visualizacao dos dados coletados.

II. Caracterizacao dos dados

Os dados sao obtidos atraves de sensores comuns em diversos automoveis (OBD-II) e enriquecidos com informacoes de geolocalizacao que estao presentes em smartphones. Para preservar a identidade dos usuarios, suas informacoes pessoais nao sao disponibilizadas e tambem sao retiradas as informacoes dos primeiros e ultimos 200m de deslocamento, com objetivo que o usuario nao seja identificado atraves das informacoes de origem e de destino.

Uma vez que os automoveis estao em deslocamento, suas informacoes mudam rapidamente. Desse modo, informacoes populadas no dataset a cada 2 segundos.

Os dados originais do servidor sao persistidos num banco de dados NoSQL – MongoDB – e oferecidos abertos ao publico atraves de um RESTful web service. Os dados sao oferecidos em um formato JSON, que eh simples de entender e processar.

Para realizacao desse trabalho, foram acessadas as apis rest e persistidas as informacoes de cada deslocamento no banco de dados ElasticSearch, por motivos que serao esclarecidos em seguida.

A. Caracterizao geral dos dados

Cada item do dataset eh composto por dois atributos principais, que sao subdivididos em diversas partes e serao mais detalhados na sequencia: 'properties' e 'features'.

- properties: possui as caracteristicas gerais do veiculo em questao. Esta subdividido nos itens apresentados na tabela 1. Veja um exemplo de propriedades extraido do banco de dados.

```
"properties": {
```

TABLE I
PROPERTIES: PROPRIEDADES GERAIS DE CADA VEICULO

Variavel	Descricao
type	categorica
constructionYear	discreta
model	categorica
fuelType	categorica
engineDisplacement	discreta
manufactuerer	categorica

TABLE II
FEATURES: INFORMACOES DE CADA TIMESTAMP DO DESLOCAMENTO

Variavel	Unidade de medida
coordinates	geoespacial
speed	continuo
rpm	continuo
gps accuracy	continuo
maf	continuo
engine load	continuo
gps pdop	continuo
o2 lambda voltage	continuo
throttle position	continuo
consumption	continuo
gps vdop	continuo
gps speed	continuo
gps bearing	continuo
intake pressure	continuo
co2	continuo
time	temporal


```
  "sensor": {
    "type": "car",
    "properties": {
      "constructionYear": 2011,
      "model": "Avensis",
      "fuelType": "gasoline",
      "id": "574e78cbe4b09078f97bbb4a",
      "engineDisplacement": 1800,
      "manufacturer": "Toyota"
    }
}
```


- features: possui as caracteristicas do deslocamento – ou da 'viagem' em si. A viagem eh caracterizada pela composicao, a cada dois segundos, do seguinte conjunto de dados:


```
{
      "geometry": {
        "coordinates": [
          6.443663779195363,
          51.20348336793408
        ],
        "type": "Point"
      },
      "type": "Feature",
      "properties": {
        "phenomenons": {
          "Speed": {
            "value": 34.647194623947144,
            "unit": "km/h"
          },
          "Rpm": {
            "value": 1584.3050694465637,
            "unit": "u/min"
          },
          "GPS Accuracy": {
            "value": 2.999999910593033,
            "unit": "%"
          },
          "MAF": {
            "value": 9.437765815629945,
            "unit": "l/s"
          },
          "Engine Load": {
            "value": 43.96216858346017,
            "unit": "%"
          },
          "GPS PDOP": {
            "value": 1.4999999776482582,
            "unit": "precision"
          },
          "O2 Lambda Voltage": {
            "value": 3.2762881521175586,
            "unit": "V"
          },
          "Throttle Position": {
            "value": 21,
            "unit": "%"
          },
          "Consumption": {
            "value": 3.102402130874109,
            "unit": "l/h"
          },
          "GPS VDOP": {
            "value": 1.1779116287827491,
            "unit": "precision"
          },
          "GPS Speed": {
            "value": 33.47147411240462,
            "unit": "km/h"
          },
          "GPS HDOP": {
            "value": 0.899999986588955,
            "unit": "precision"
          },
          "Intake Pressure": {
            "value": 44.14216932654381,
            "unit": "kPa"
          },
          "GPS Bearing": {
            "value": 304.7174273121018,
            "unit": "deg"
          },
          "Intake Temperature": {
            "value": 13.000000387430191,
            "unit": "c"
          },
          "CO2": {
            "value": 7.290645007554156,
            "unit": "kg/h"
          },
          "O2 Lambda Voltage ER": {
            "value": 0.9988191702782387,
            "unit": "ratio"
          },
          "GPS Altitude": {
            "value": 104.90115790988267,
            "unit": "m"
          }
        },
        "id": "590ad752268d1b08a47f18d4",
        "time": "2017-03-27T04:51:05Z"
```

```
    }
  }
```

B. Questoes a serem respondidas

A ideia basica eh responder questoes relativas a (a) padroes de motoristas em diferentes partes do globo, (b) padroes de comportamento individual no volante e (c) possibilidade de uso da base em modelos de aprendizagem de maquina.

- Existem muitos registros nesse banco de dados?
- Em quer periodo de tempo esse banco de dados foi utilizado? Ainda hoje ele eh bastante utilizado?
- Em que regioes do globo estao os usuarios desse sistema?
- Os modelos e fabricantes mais comum na Europa sao os mesmos do Brasil?
- Quais marcas e modelos consomem mais combustivel?
- Quais os modelos e fabricantes mais comuns?
- Quais sao as marcas e modelos que mais poluem? quais sao as marcas e modelos que menos poluem?
- Podemos descobrir quais sao as regioes mais poluidas e regioes menos poluidas?
- Eh possivel utilizar este dataset para extrair padroes de comportamento de um determinado usuario?

III. Preparacao dos dados para visualicao

O envirocar eh uma base aberta de dados. Entretanto, nao existe uma forma de simples para fazer download da base inteira. A entrega dos dados eh feita via servicos REST que possibilitam acesso a determinadas partes do banco de dados.

Para que se realizasse a obtencao da base inteira para que pudessem ser feitas as visualizacoes, utilizou-se um script capaz de varrer a base de maneira sequencial. Esse mesmo script eh responsavel pela persistencia dos dados no elasticsearch.

A base de dados possui 15 mil viagens. O numero bruto nao eh tao grande quanto se supunha, mas, para cada vaigem, eh enviado ao banco de dados uma serie de informacoes a cada dois segundos. Dessa maneira, embora o numero de viagens seja relativamente pequeno, o tamanho de cada viagem eh grande. A indexada no elasticsearch ocupou 6 Gb de memoria em disco.

IV. Tecnica de visualizao desenvolvida

A exploracao dos dados foi feita com tres abordagens. Uso de Kibana, linguagem R e API de mapas do Google.

Kibana eh uma ferramenta utilizada para analise por inspecao manual e visualizacao de informacoes que funciona de maneira natural com o ElasticSearch. Dessa forma, eh ela que ira apresentar os dados armazenados no ElasticSearch, em uma interface, via browser, altamente customizvel com histogramas, mapas e outros painis que propiciam uma viso geral sobre os dados. O Kibana possibilita transformar os logs em informaes teis (valor) atravs de Dashboards, pois permite realizar correlao de eventos, filtrar logs por origem, hospedeiros, entre outras combinaes.

Foi utilizado o ElasticSearch como base primaria dos dados, pois alem do suporte a persistencia, vem junto com uma serie de mecanismos e algoritmos de recuperacao de informacao. A ferramenta permite combinar geolocalizacao com outras tecnicas de recuperacao de informacoes baseadas em texto.

Outras ferramenta para visualizacao de informacoes utilizada foi a linguagem R. A linguagem tambem possui uma variada gama de tecnicas para visualizacao de dados, alem de possuir um suporte muito grande a questoes de estatistica e de probabilidade.

O R nao eh uma ferramenta de tao alto nivel de abstracao quanto o kibana em que, uma vez que os dados estao persistidos, se montam visualizacoes de maneira simplificada. Em vez disso, eh necessario que se escreva codigo para que as visualizacoes aparecam. Os graficos surgem da maneira mais crua do que no Kibana, mas possui muito mais flexibilidade, maleabilidade e estensibilidade. Alem disso, existe um conjunto muito mais vasto de visualizacoes que o kibana, com vieses mais estatisticos, como boxplots e matrizes de correlacoes.

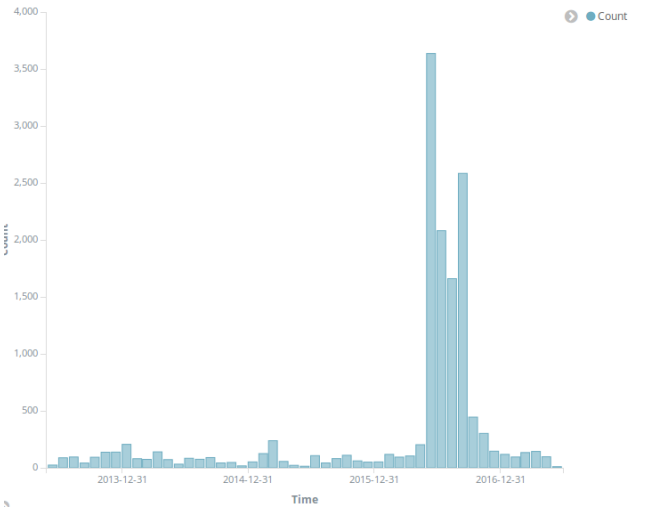
Por fim, para detectar padroes de comportamente individuais, foi implementada utilizando a API do Google Maps uma ferramenta simples para visualizar por quais lugares um determinado motorista frequenta com maior frequencia.

A. Histograma

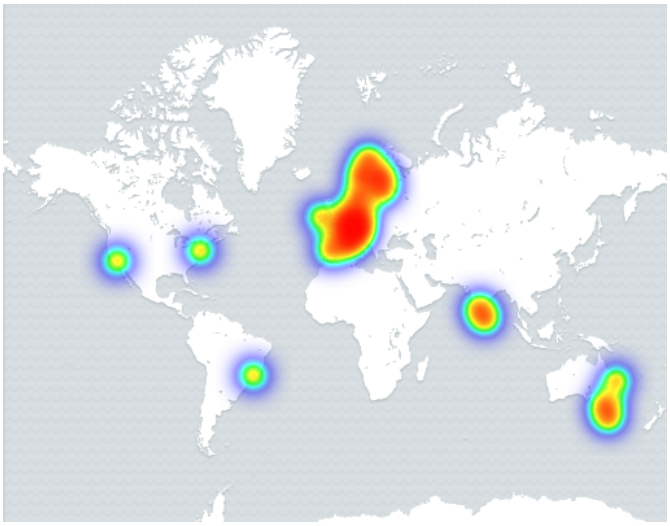
Com o intuito de descobrir se existem muitos registros no banco de dados e, sobretudo, quais sao os periodos em que a ferramente envirocarm sao mais utilizados, foi desenvolvido um histograma em que o eixo x corresponde a variavel temporal e o eixo y corresponde a quantidade de viagens naquele periodo de tempo. Com isso, podera ser inferido se bastante gente vem utilizando a ferramenta e, principalmente, se se uso vem crescendo ou decrescendo com o passar do tempo. Eh um estudo de tendencias, portanto.

Pode-se observar o grafico abaixo, que nao existe uma tendencia de aumento ou de diminuicao ao longo do tempo. O que se observa eh um 'boom' de uso no periodo em volta do ano de 2016, mas que nao se mantem no restante do tempo. No periodo de novembro de 2016, por exemplo, tivemos 3600 viagens naquele mes. Mas no inicio de 2017, nao passavam de 100 viagens por mes.

Nao se pode observar, portanto, um uso massivo da plataforma. Entretanto, temos uma amostra que parece razoavel para que se possa tentar responder as outras perguntas do artigo.



No eixo x, observamos a quantidade de uso da plataforma ao longo dos meses.



Aqui, podemos observar que a maioria dos usuarios, de fato, se encontra, na Europa. Contudo, existem pessoas utilizando em diversos outros pontos do globo tambm.

B. *Heatmaps*

Uma vez visualizada a distribuicao temporal da utilizacao da plataforma, ira ser observada a distribuicao geografica dos usuarios. Como ja se sabe que eh uma plataforma desenvolvida na Alemanha, pode-se supor que muita gente na Europa ira utilizar a ferramenta, em relacao as outras regioes do globo.

A visualizacao por calor, utilizando como metrica a quantidade de viagens dispostas no globo terrestre, dara um senso de quais regioes sao mais utilizadas a ferramenta.

Observando o grafico a seguir, observamos aquilo que ja era suposto de antemao: de fato, na Europa, a plataforma eh mais utilizada. Alem disso, conseguimos obter alguns dados que nao eram imaginados. Vemos pontos isolados de uso nos Estados Unidos, Brasil, India e Australia.

Com as informacoes geoespaciais, ganha-se confianca de que poderao ser respondidas outras questoes, como o comportamento dos motoristas em diferentes regioes do globo. Alem disso, pode-se utilizar esse grafico para clusterizar de maneira macro o globo terrestre de acordo com essa base de dados: America do Norte, America do Sul, Europa, India e Australia.

A visualizacao de heatmat traz uma ideia muito clara sobre os pontos que mais utilizam a ferramenta. Traz-nos a ideia de que em diversos pontos do globo ela esta sendo utilizada, com enfase na Europa. Traz confianca de que poderemos observar padroes de comportamento em diversos pontos do globo. Ela traz uma ideia basica, mas nao definitiva, da quantidade de pessoas que utiliza a ferramenta ao longo do globo.

C. *Grafico em pizza*

O heatmap deu uma ideia basica sobre a quantidade de pessoas que utilizam a plataforma ao redor do globo. Entretanto, nao trouxe uma ideia definitiva sobre as proporcoes de cada regiao.

Para isso, sera utilizado um grafico em pizza, que traz claramente a ideia de proporcoes ao usuario, ao dispor em fatias cada segmento analizado.

Nesse grafico, sera utilizado o cluster inferido visualmente do heatmatp com as seguintes regioes: Europa, America do Norte, America do Sul, India e Australia.

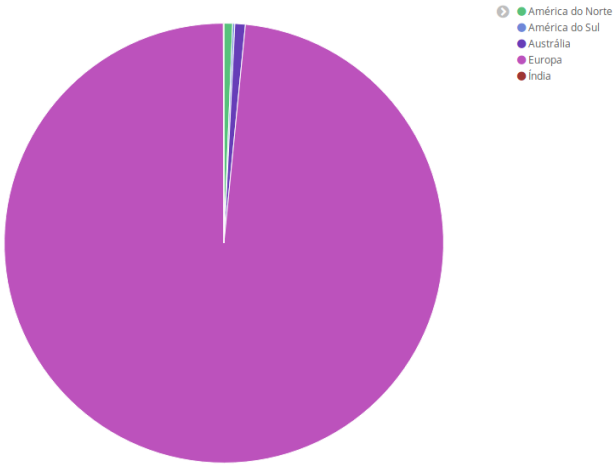
Aqui, pode-se observar a maioria, de fato, utilizando na Europa: 97% dos usuarios esta la. Temos alguma representatividade, com 1% dos Usuarios na America do Norte e 1% na Australia.

Na America do Sul e na India, nao chega a 1% de usuarios. Parece ter sido realmente um usuario curioso que comecou a utilizar a plataforma em sua terra natal.

Essa informacao podera ser utilizada da seguinte maneira: para comparacao de dados globais, como estatisticas e tentativas de deteccao de comportamento em massa, devera ser dado foco nos usuarios da Europa.

Entretanto, essa falta de representatividade em outras partes do globo podera favorecer um olhar mais individualista. Por exemplo, pode-se acreditar que na America do Sul e na India, eh sempre a mesma pessoa que utiliza a ferramenta. Desse modo, poderemos observar algum comportamento individual. Do mesmo modo pode ser feito na America do Norte e na Australia, com um pouco mais de criterio, pois la houve muito mais uso.

Entretanto, na Europa, nao podemos fazer isso, pois eh muita gente usando. Na Europa, serao feitas analises mais genericas.



98% dos deslocamentos esta na Europa. 1% esta na America do Norte e 1% esta na Australia. Outras regioes sao sao insignificantes, que nem visualizamos no grafico .

D. Graficos em pizza

Observadas algumas questoes temporais e espaciais, irao ser analisados alguns padroes de comportamento em diferentes regioes do globo. Sera utilizado aquele cluster inferido do primeiro heatmap para fazer uma analise sobre quais sao os fabricantes mais utilizados em diferentes partes do globo.

Sobre padroes de comportamento, espera-se verificar se supostos padroes culturais podem de uma determinada regiao do globo pode ser inferida atraves deste banco de dados. Por exemplo, eh verdade que americanos gostam de carros espacosos? Eh verdade que na Europa e Estados Unidos os carros sao muito melhores do que os da America do Sul?

Para responder essa pergunta, algumas visualizacoes foram feitas algumas suposicoes de tecnicas de visualizacao: sunbsurst e nugget.

A tecnica de sunbsurst nao pareceu muito adequada devido a quantidade de informacoes – muito grande para um espaco pequeno – e o biscoito nao transpareceu de maneira clara o que estava sendo perguntado.

Para isso, foi utilizada novamente a tecnica de visualizacao de graficos em pizza. Entretanto, agora, em vez de plotar somente um grafico, o foram plotados 5 graficos em pizza, em que cada pizza corresponde a uma regiao do cluster. Essa visualizacao foi escolhida em detrimento do sunburst pois a segmentacao trouxe mais evidencia para respoder aquilo que estava sendo perguntado. O sunburst trouxe muita informacao em muito pouco espaco fisico, o que pareceu meio confuso.

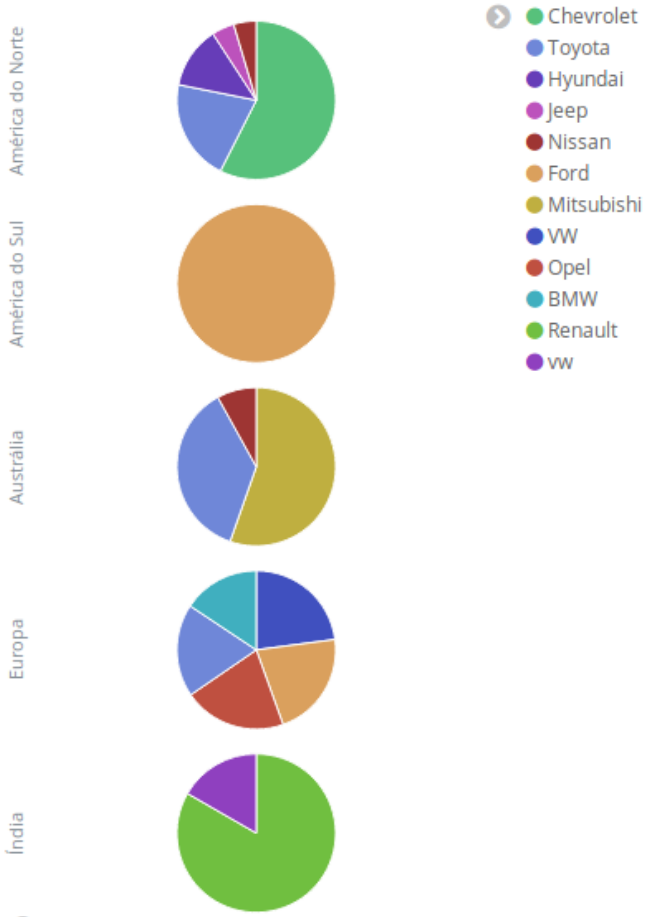
Observando os fabricantes, vemos que as principais marcas na America do Norte, Australia e Europa sao as mesmas.

- America do Norte: Chevrolet, Toyota e Hiunday
- Australia: Ford, Toyota e Opel
- Europa: Toyota, Ford, Opel, VW, BMW

Ja na America do Sul e India, parecem ser regioes mais humildes tambem, pois aqui nao observamos Hiunday e BMW.

- America do Sul: Ford
- India: Renault e VW

Baseado nessa amostra de dados, nao se pode chegar a uma conclusao definiva para as questoes levantadas. A distribuicao entre Europa, Estados Unidos e Australia ficaram razoavelmente parecidas. Ja na America do Sul e India, realmente parecem regioes menos desenvolvidas. Entretanto, a amostra dos dados nao parece ser significativa o suficiente para uma conclusao assertiva. A observacao do grafico parece corroborar a ideia inicial, utilizada para fazer as pergutnas, mas nao ha evidencias que, de fato, elas acontecem.



Estados Unidos e Europa apresentam resultados parecidos. America do Sul supoe que somente um cara utilizou a plataforma.

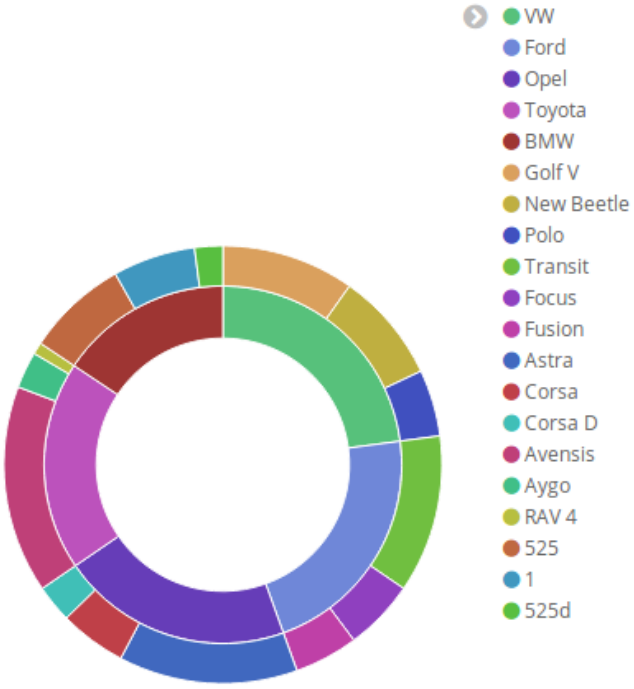
E. Sunburst

A visualizacao por sunburst eh uma tecnica de visualizacao radial apra exibir estruturas hierarquicas como arvores. Essa tecnica mostra-se muito adequada para condensar informacoes hierarquicas, como para responder a pergunta: para cada fabricante, qual eh o modelo de automovel mais utilizado.

Como na Europa existe uma amostra muito mais significicativa do que em outras regioes do globo, a pergunta fica limitada a regiao da Europa. Alem disso, para que a visualizacao nao acontecesse de maneira muito confusa, para que ela ficasse clara, foram escolhidos os 5 fabricantes mais utilizados e os 3 modelos mais utilizados por cada fabricante. Se nao houvesse essa limitacao, o grafico ficaria muito poluido, o que gera um desconforto ao usuario.

Esse grafico, novamente, corrobora a intuicao de que na Europa o nivel dos automoveis eh maior do que no Brasil. Observamos que os fabricantes mais comuns la sao os mesmos que os presentes no Brasil. Exceto pela BMW, que muito pouco se observa por aqui, vemos que VW, Ford, Opel (GM) e Toyota sao os mais presentes por la.

A difenca de padrao de veiculos se observa nos modelos mais comuns, uma vez que os fabricantes sao os mesmos. Por exemplo, os modelos mais comuns da VW sao New Beetle, Golf e Polo. Notavelmente, sao veiculos mais sofisticados que os modelos comuns brasileiros. O mesmo se observa nos veiculos da Ford, cujos modelos mais populares sao Transit, Fusion e Focus. Aqui, esses automoveis sao automoveis de elite.



Observamos que as marcas mais comuns na Europa sao as mesmas do brasil, exceto pelo BMW. Nunca vemos BMWs pelo Brasil.

F. Radar

Para que se possa identificar quais sao as marcas e modelos que mais poluem e que menos poluem de uma forma bastante rica, optou-se por fazer uma visualizacao em forma de radar. Dessa forma, foi possivel ver, para cada fabricante, uma serie de atributos que parecem ser correlacionados. Sao eles: rotacao do motor (rpm), consumo, velocidade, duracao da viagem, tempo de viagem e quantidade de gas carbonico emitido.

Para que isso pudesse ser sintetizado adequadamente na forma de um radar, foram feitas uma serie de computacoes em cada um dos itens. Lembrando que diversos desses atributos mencionados, como co2, rpm, consumo e velocidade mudam periodicamente – a cada dois segundos –, foi feita uma media para cada um desses atributos em cada viagem. Posteriormente, foi feita uma normalizacao desses dados, pois a ferramenta de visualizacao que o Kibana oferece nao permite que se coloque diferentes escalas em cada um dos eixos.

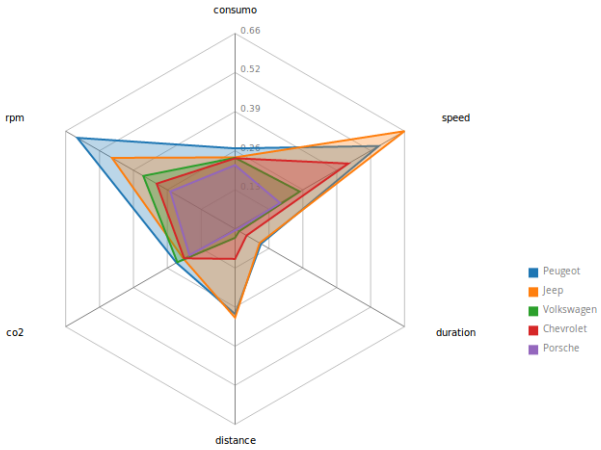
Computadas essas medias e normalizados os dados, podemos observar o grafico de radar. E ele nos leva a algumas informacoes bastante curiosas:

Peugeot eh o fabricante que normalmente se anda com a rpm mais alta e eh o cara que tem maior consumo, mas nao eh o fabricante que anda mais rapido.

O fabricante que anda mais rapido eh o Jeep, e ele consome menos que os Peugeot, normalmente.

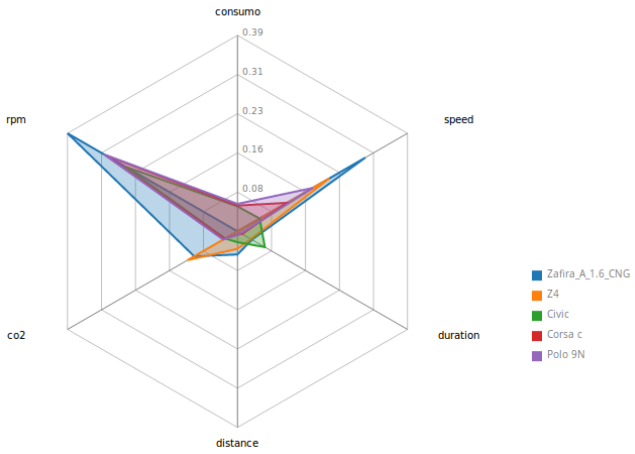
Curioso observar tambem, que o Porshe eh a marca que se anda com rpm mais baixa – ao contrario do que normalmente se imagina – e cuja velocidade eh mais baixa tambem. Acredito que isso se de porque o Porshe deve ter uma representatividade muito baixa na amostra. Portanto, se voce quiser andar rapido, nao compre um Porshe!

Para quem deseja comprar um automovel que consuma pouco e que ande com a rotacao do motor baixa, recomenda-se, de acordo com esse dataset, o uso de um Chevrolet.

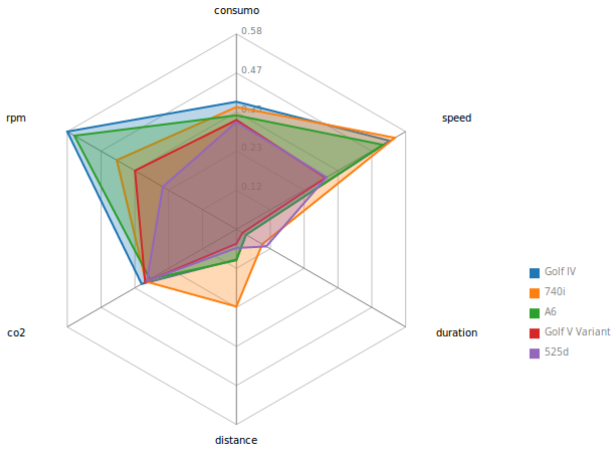


Diversos atributos seccionado por fabricante.

O mesmo tipo de visualizacao pode ser aplicado, nao aos fabricantes, mas aos modelos dos automoveis em si. Na figura X, observam-se os modelos que menos poluem – observe que dos 5 apresentados, 2 sao chevrolet, o que esta de acordo com o grafico de radar apresentado por fabricantes – e observe que os que mais poluem sao modelos esportivos, cuja rotacao de motor normalmente eh alta.



Modelos que menos poluem.



Modelos que mais poluem.

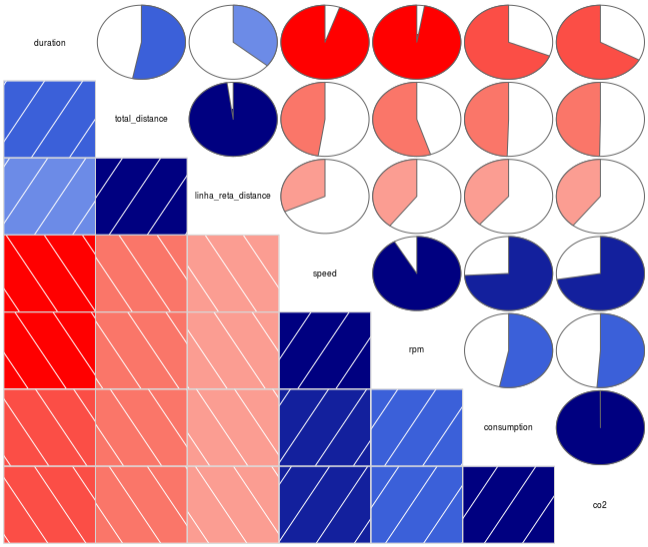
G. Matriz de correlacao

Como foi observado no grafico de radar que a rotacao do motor esta intimamente ligada ao consumo dos automoveis, decidi por fazer uma apresentacao de uma matriz de correlacao. O intuito, com essa visualizacao eh verificar quais atributos estao mais ligados um ao outros. Por exemplo, para ter certeza de que, se o automovel esta com rotacao maior, que ele esta poluindo mais. Alem disso, para poder observar se faz sentido associar a velocidade ah quantidade de emissao de poluentes.

Nessa visualizacao, ao contrario das anteriores, foi utilizada linguagem R. Ela possui mais suporte a visualizacoes e trabalhos estatisticos do que o Kibana.

Nessa representacao, quanto mais azul e quanto mais cheio esta o circulo superior, maior eh a correlacao entre os atributos. Observamos que quanto maior o consumo de combustivel, maior eh a quantidade de gas carbonico emitido. Observamos tambem que existe grande correlacao entre rpm e velocidade.

Um ponto a ser considerado eh que parece nao haver correlacao entre velocidade e duracao de uma viagem. Inicialmente pode-se supor que em viagens mais longas se tende a andar mais rapidamente. Aqui, entrentanto, esse fato nao pode ser observado.



Diversos atributos seccionado por fabricante.

H. Heatmap com Google Maps

Para uma analise individual de determinados motoristas, sera utilizada novamente a tecnica de heatmaps. Infereriu-se que registros de uma regiao isolada – como Brasil ou India –, em que existem poucos registros na base, correspondem ao mesmo motorista.

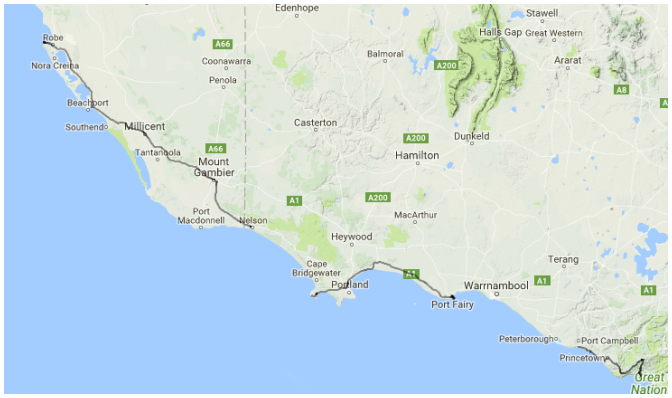
Com essa hipotese, sao plotadas as viagens de um determinado motorista. As regioes mais escuras no mapa indicam que ele passou por aquele lugar muitas vezes. Regiose claras indicam lugares que ele frequentou pouco.

Espera-se dessa maneira, verificar se os caminhos de um determinado motorista sao recorrentes ou nao. Alem disso, espera-se verificar se essa tecnica eh adequada para vislumbrar essa hipotese.

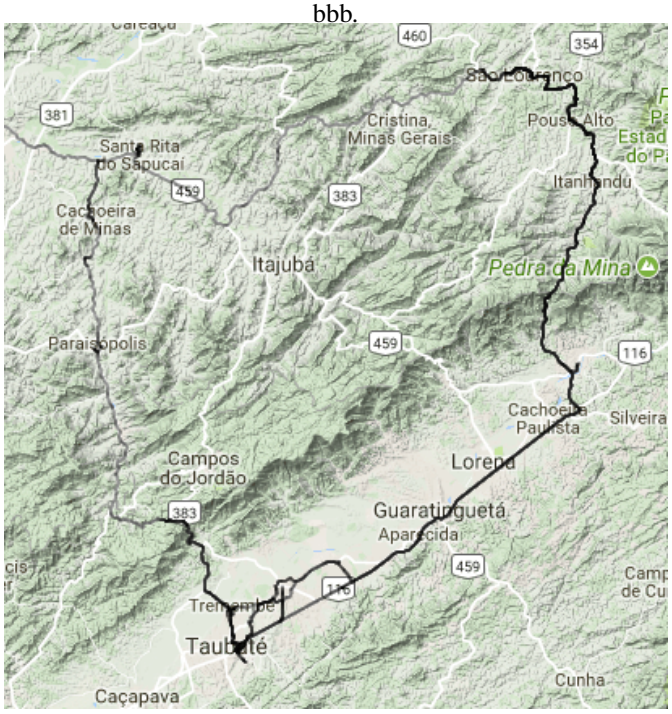
Na primeira figura, observamos que o motorista passou muitas vezes pela mesma rota, dada a linha escura. Na segunda figura, observamos que o motorista nao repete muitas trajetorias. Na terceira figura, observa-se que em algumas rotas o motorista usa muitas vezes e em algumas rotas ele frequenta pouco.

Esse terceiro mapa apoia a ideia de que a tecnica do heatmap pode ser utilizada com muita eficiencia par verificar rotas recorrentes, pois claramente se observa pontos escuros que, de fato, correspondem as rotas mais utilizadas.

aaa.



bbb.



ccc.

V. Resultados

As visualizacoes deram uma visao geral sobre a base de dados, e praticamente todas as perguntas que puderam ser respondidas.

- observamos que, de fato, a maioria dos usuarios esta na Europa. Alem disso, observamos que por um curto periodo de tempo, houve muita gente utilizando a plataforma, mas que essa quantidade de gente a utilizando nao se sustentou por muito tempo.
- pode-se observar alguns padroes de comportamento utilizando esse dataset: na Europa, os fabricantes mais comuns sao praticamente mesmos fabricantes do Brasil. Entretanto, os carros comercializados por esses fabricantes por la sao superiores aos brasileiros.
- observamos de maneira clara, atraves da visualizacao de radar, quais sao as marcas e fabricantes cujos veiculos poluem mais e quais poluem menos. Vimos, nesse mesmo grafico de radar, quais sao os que gastam mais e os que gastam menos combustivel.
- observamos ainda que esse dataset, embora nao seja tao rico quanto se supunha no inicio. Para treinamento de modelos de aprendizagem de maquinas, supoe-se que essa base nao seja suficiente
- A tecnica de heatmap se mostra bastante eficiente par visualizar rotas repetidas, pois deu claramente essa nocao ao usuario
- Heatmaps realmente ajudam a se ter uma ideia sobre as regioes que possuem maior uso da ferramenta. Entretanto, para se ter uma nocao mais exata de proporcoes, foi necessaria uma visualizacao em pizza.
- Os graficos de radar ajudaram a ter claramente uma ideia de muitos atributos sintetizados numa pequena regiao espacial.

VI. Conclusao

Foi atingido objetivo de se fazer uma exploracao basica na base de dados oferecida pelo envirocar. Obteve-se intuicao sobre essa base de dados e padroes de comportamento dos motoristas de automovel.

Padroes culturais, como o tipo de veiculos e fabricantes utilizados na Europa em comparacao com o Brasil foram identificados. Alem disso, padroes individuais de pilotagem de automoveis foram identificados.

Alem disso, foi observadas as distribuicoes temporais e espaciais dos registros de veiculos.

A base de dados nao eh tao utilizada quanto se supunha – o autor imaginava que mais gente utilizasse a plataforma –, mas possui-se uma amostra significativa de pessoas para que se possa levar o trabalho mais adiante e utilizar essa base para treinar algum modelo de machine learning em trabalhos futuros.

Alem disso, pode-se fazer uma analise basica sobre o uso de ferramentas como Kibana e como R para visualizacao de informacoes. O Kibana, mais simples, apresenta graficos mais bonitos. O R, entretanto, apresenta graficos mais cruz e uma linguagem de baixo nivel. O R, portanto, vai ser mais aproveitado por usuarios mais experientes – como programadores –, enquanto o Kibana pode ser utilizado por usuarios mais leigos.

References

1) JIRKA, S, RAMKE, A, BRORING, A: enviroCar Crowd Sourced Traffic and Environment Data for Sustainable Mobility
2) BRORING, A; STASCH, C: enviroCar: A Citizen Science Platform for Analyzing and Mapping Crowd-Sourced Car Sensor Data
3) LIN, MAIO; JING WEN: Mining GPS data for mobility patterns: A survey
4) VIDAL, WILLIAN. Estudo de viabilidade para deteco de intruso usando tcnicas de Big Data para a anlise de logs
5) VAARANDI, R.; NIZINSKI, P. Comparative analysis of open-source log management solutions for security monitoring and network forensics