

Envirocar visualization

exploring an environmental and traffic data set

Rodrigo Claro Zembruski
Programa de Pós Graduação em Ciência da Computação
Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre, Brasil
Web page: <http://inf.ufrgs.br/~rczembruski>

Abstract—Envirocar é um projeto que coleta e armazena dados ambientais e dados de tráfego de automóveis. São dados de sensores implantados dentro dos veículos que trazem informações precisas sobre navegação dos carros e dados relativos à conservação do meio-ambiente, como emissão de gás-carbonico e consumo do veículo.

Esse artigo tem por objetivo caracterizar os dados e apresentar uma serie de visualizações para análise exploratória desse banco de dados. Com isso, será obtida uma intuição sobre o tipo de dados apresentado. Além disso, serão obtidas informações acuradas sobre a viabilidade do uso desse banco de dados para modelagem de algoritmos de aprendizado de máquina.

Keywords—envirocar; smartcars; kibana; geolocation; data visualization;

I. INTRODUCTION

Envirocar é um projeto da Universidade de Münster. O objetivo do projeto é oferecer uma plataforma simples para coleta, armazenamento e distribuição de dados ambientais e dados de logística de tráfego.

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

A. Motivacao

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

Uma base de dados consistente é crucial para que se confirmem – ou refutem – hipóteses relativas a comportamento no trânsito em diferentes partes do globo. Além disso, a base consistente pode ser muito importante para a inferência de padrões comportamento e de pilotagem dos motoristas de automóvel.

Uma análise exploratória é um passo necessário para que se verifique se a base realmente pode ser utilizada com os propósitos citados ou se ela deve ser descartada.

B. Objetivos

O objetivo desse trabalho é fazer uma exploração básica pela base dados para ganhar intuição sobre ela. Serão observados

padrões básicos de comportamento em diferentes partes do globo e também padrões individuais de pilotagem para determinados motoristas.

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

Contribuições: O trabalho dará uma sugestão sobre a possibilidade de uso da base de dados envirocar em projetos de aprendizagem de máquina. Também será feita uma análise comparativa entre diferentes técnicas de visualização dos dados coletados.

II. CARACTERIZAÇÃO DOS DADOS

Os dados são obtidos através de sensores comuns em diversos automóveis (OBD-II) e enriquecidos com informações de geolocalização que estão presentes em smartphones. Para preservar a identidade dos usuários, suas informações pessoais não são disponibilizadas e também são retiradas as informações dos primeiros e últimos 200m de deslocamento, com objetivo que o usuário não seja identificado através das informações de origem e de destino.

Uma vez que os automóveis estão em deslocamento, suas informações mudam rapidamente. Desse modo, informações populadas no dataset a cada 2 segundos.

Os dados originais do servidor são persistidos num banco de dados NoSQL – MongoDB – e oferecidos abertos ao público através de um RESTful web service. Os dados são oferecidos em um formato JSON, que é simples de entender e processar.

Para realização desse trabalho, foram acessadas as APIs REST e persistidas as informações de cada deslocamento no banco de dados Elasticsearch, por motivos que serão esclarecidos em seguida.

A. Caracterização geral dos dados

Cada item do dataset é composto por dois atributos principais, que são subdivididos em diversas partes e serão mais detalhados na sequência: 'properties' e 'features'.

- **properties:** possui as características gerais do veículo em questão. Esta subdividido nos itens apresentados na tabela 1. Veja um exemplo de propriedades extraído do banco de dados.

```
"properties": {
```

TABLE I
PROPERTIES: PROPRIEDADES GERAIS DE CADA VEÍCULO

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

TABLE II
FEATURES: INFORMAÇÕES 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 características do deslocamento – ou da 'viagem' em si. A viagem é 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. Questões a serem respondidas

A ideia básica é responder questões relativas a (a) padrões de motoristas em diferentes partes do globo, (b) padrões de comportamento individual no volante e (c) possibilidade de uso da base em modelos de aprendizagem de máquina.

- Existem muitos registros nesse banco de dados?
- Em que período de tempo esse banco de dados foi utilizado? Ainda hoje ele é bastante utilizado?
- Em que regiões do globo estão os usuários desse sistema?
- Os modelos e fabricantes mais comuns na Europa são os mesmos do Brasil?
- Quais marcas e modelos consomem mais combustível?
- Quais os modelos e fabricantes mais comuns?
- Quais são as marcas e modelos que mais poluem? quais são as marcas e modelos que menos poluem?
- Podemos descobrir quais são as regiões mais poluídas e regiões menos poluídas?
- É possível utilizar este dataset para extrair padrões de comportamento de um determinado usuário?

III. PREPARAÇÃO DOS DADOS PARA VISUALIZAÇÃO

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

Para que se realizasse a obtenção da base inteira para que pudessem ser feitas as visualizações, utilizou-se um script capaz de varrer a base de maneira sequencial. Esse mesmo script é responsável pela persistência dos dados no *Elasticsearch*.

A base de dados possui 15 mil viagens. O número bruto não é tão grande quanto se supunha, mas, para cada viagem, é enviado ao banco de dados uma série de informações a cada dois segundos. Dessa maneira, embora o número de viagens seja relativamente pequeno, o tamanho de cada viagem é grande. A indexada no *Elasticsearch* ocupou 6 Gb de memória em disco.

IV. TÉCNICA DE VISUALIZAÇÃO DESENVOLVIDA

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

Kibana é uma ferramenta utilizada para análise por inspeção manual e visualização de informações que funciona de maneira natural com o *ElasticSearch*. Dessa forma, é ela que irá apresentar os dados armazenados no *ElasticSearch*, em uma interface, via browser, altamente customizável com histogramas, mapas e outros painéis que propiciam uma visão geral sobre os dados. O Kibana possibilita transformar os logs em informações úteis (valor) através de Dashboards, pois permite realizar correlação de eventos, filtrar logs por origem, hospedeiros, entre outras combinações.

Foi utilizado o *ElasticSearch* como base primária dos dados, pois Além do suporte à persistência, vem junto com uma série

de mecanismos e algoritmos de recuperação de informação. A ferramenta permite combinar geolocalização com outras técnicas de recuperação de informações baseadas em texto.

Outras ferramentas para visualização de informações utilizadas foram a linguagem R. A linguagem também possui uma variada gama de técnicas para visualização de dados. Além de possuir um suporte muito grande a questões de estatística e de probabilidade.

O R não é uma ferramenta de tão alto nível de abstração quanto o Kibana em que, uma vez que os dados estão persistidos, se montam visualizações de maneira simplificada. Em vez disso, é necessário que se escreva código para que as visualizações apareçam. Os gráficos surgem da maneira mais crua do que no Kibana, mas possui muito mais flexibilidade, maleabilidade e extensibilidade. Além disso, existe um conjunto muito mais vasto de visualizações que o Kibana, com vieses mais estatísticos, como *boxplots* e matrizes de correlações.

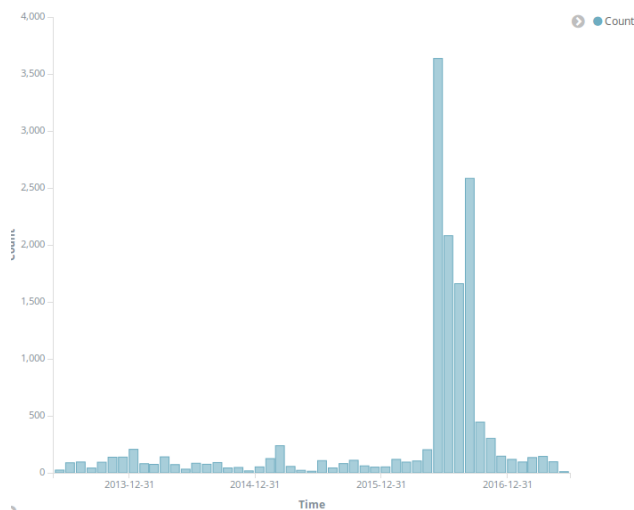
Por fim, para detectar padrões de comportamento individuais, foi implementada utilizando a API do Google Maps uma ferramenta simples para visualizar por quais lugares um determinado motorista frequenta com maior frequência.

A. Histograma

Com o intuito de descobrir se existem muitos registros no banco de dados e, sobretudo, quais são os períodos em que a ferramenta *envirocar* são mais utilizados, foi desenvolvido um histograma em que o eixo x corresponde à variável temporal e o eixo y corresponde à quantidade de viagens naquele período de tempo. Com isso, poderá ser inferido se bastante gente vem utilizando a ferramenta e, principalmente, se seu uso vem crescendo ou decrescendo com o passar do tempo. É um estudo de tendências, portanto.

Pode-se observar o gráfico abaixo, que não existe uma tendência de aumento ou de diminuição ao longo do tempo. O que se observa é um 'boom' de uso no período em volta do ano de 2016, mas que não se mantém no restante do tempo. No período de novembro de 2016, por exemplo, tivemos 3600 viagens naquele mês. Mas no início de 2017, não passavam de 100 viagens por mês.

Não se pode observar, portanto, um uso massivo da plataforma. Entretanto, temos uma amostra que parece razoável 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.

B. Heatmaps

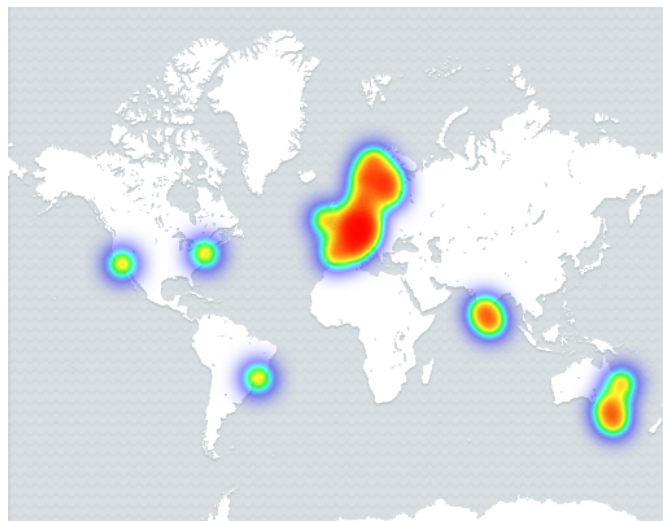
Uma vez visualizada a distribuição temporal da utilização da plataforma, irá ser observada a distribuição geográfica dos usuários. Como já se sabe que é uma plataforma desenvolvida na Alemanha, pode-se supor que muita gente na Europa irá utilizar a ferramenta, em relação às outras regiões do globo.

A visualização por calor, utilizando como métrica a quantidade de viagens dispostas no globo terrestre, dará um senso de quais regiões são mais utilizadas a ferramenta.

Observando o gráfico a seguir, observamos aquilo que já era suposto de antemão: de fato, na Europa, a plataforma é mais utilizada. Além disso, conseguimos obter alguns dados que não eram imaginados. Vemos pontos isolados de uso nos Estados Unidos, Brasil, Índia e Austrália.

Com as informações geoespaciais, ganha-se confiança de que poderão ser respondidas outras questões, como o comportamento dos motoristas em diferentes regiões do globo. Além disso, pode-se utilizar esse gráfico para clusterizar de maneira macro o globo terrestre de acordo com essa base de dados: América do Norte, América do Sul, Europa, Índia e Austrália.

A visualização de heatmap 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 está sendo utilizada, com ênfase na Europa. Traz confiança de que poderemos observar padrões de comportamento em diversos pontos do globo. Ela traz uma ideia básica, mas não definitiva, da quantidade de pessoas que utiliza a ferramenta ao longo do globo.



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

C. Gráfico em pizza

O heatmap deu uma ideia básica sobre a quantidade de pessoas que utilizam a plataforma ao redor do globo. Entretanto, não trouxe uma ideia definitiva sobre as proporções de cada região.

Para isso, será utilizado um gráfico em pizza, que traz claramente a ideia de proporções ao usuário, ao dispor em fatias cada segmento analisado.

Nesse gráfico, será utilizado o cluster inferido visualmente do heatmap com as seguintes regiões: Europa, América do Norte, América do Sul, Índia e Austrália.

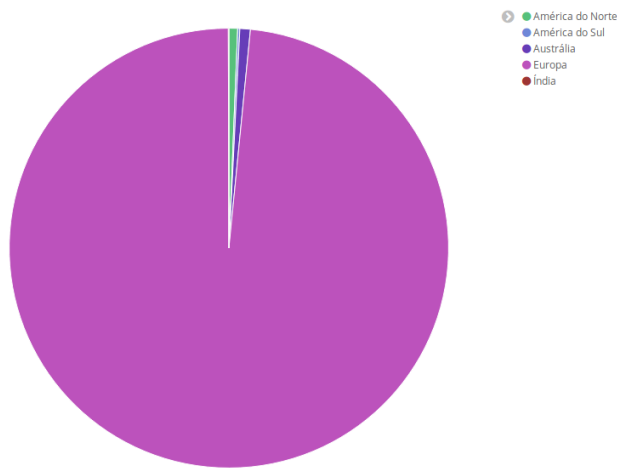
Aqui, pode-se observar a maioria, de fato, utilizando na Europa: 97% dos usuários estão lá. Temos alguma representatividade, com 1% dos usuários na América do Norte e 1% na Austrália.

Na América do Sul e na Índia, não chega a 1% de usuários. Parece ter sido realmente um usuário curioso que começou a utilizar a plataforma em sua terra natal.

Essa informação poderá ser utilizada da seguinte maneira: para comparação de dados globais, como estatísticas e tentativas de detecção de comportamento em massa, deverá ser dado foco nos usuários da Europa.

Entretanto, essa falta de representatividade em outras partes do globo poderá favorecer um olhar mais individualista. Por exemplo, pode-se acreditar que na América do Sul e na Índia, é sempre a mesma pessoa que utiliza a ferramenta. Desse modo, poderemos observar algum comportamento individual. Do mesmo modo pode ser feito na América do Norte e na Austrália, com um pouco mais de critério, pois lá houve muito mais uso.

Entretanto, na Europa, não podemos fazer isso, pois é muita gente usando. Na Europa, serão feitas análises mais genéricas.



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

D. Graficos em pizza

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

Sobre padrões de comportamento, espera-se verificar se supostos padrões culturais podem de uma determinada regioao do globo pode ser inferida através deste banco de dados. Por exemplo, é verdade que americanos gostam de carros espacosos? é verdade que na Europa e Estados Unidos os carros são muito melhores do que os da America do Sul?

Para responder essa pergunta, algumas visualizações foram feitas algumas suposicoes de técnicas de visualização: sunburst e nugget.

A tecnica de sunburst nao pareceu muito adequada devido a quantidade de informações – 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 visualização 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 regioao do cluster. Essa visualização 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.

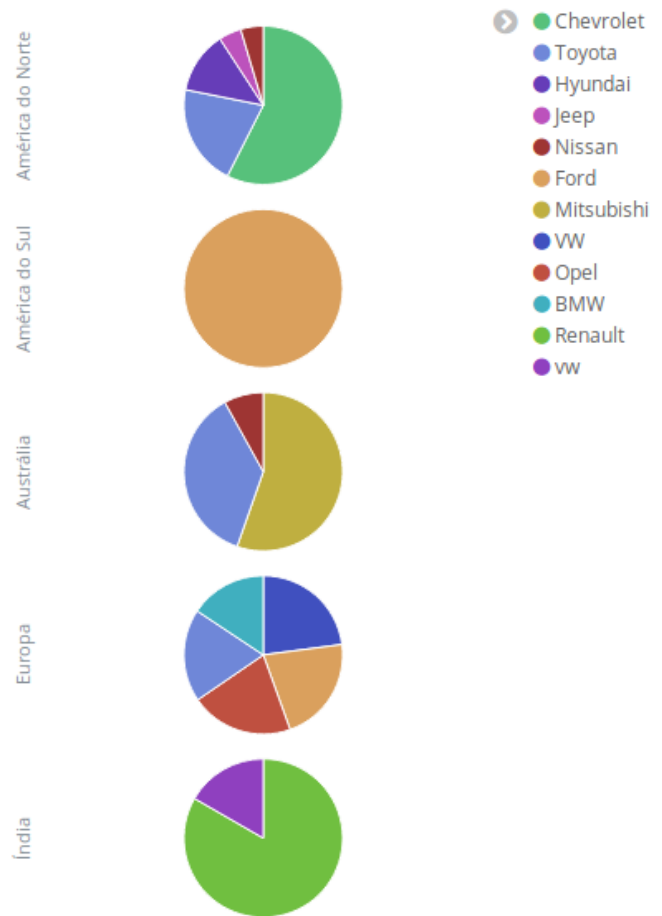
Observando os fabricantes, vemos que as principais marcas na America do Norte, Australia e Europa são as mesmas.

- 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 conclusão definiva para as questoes levantadas. A distribuição 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 conclusão 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.

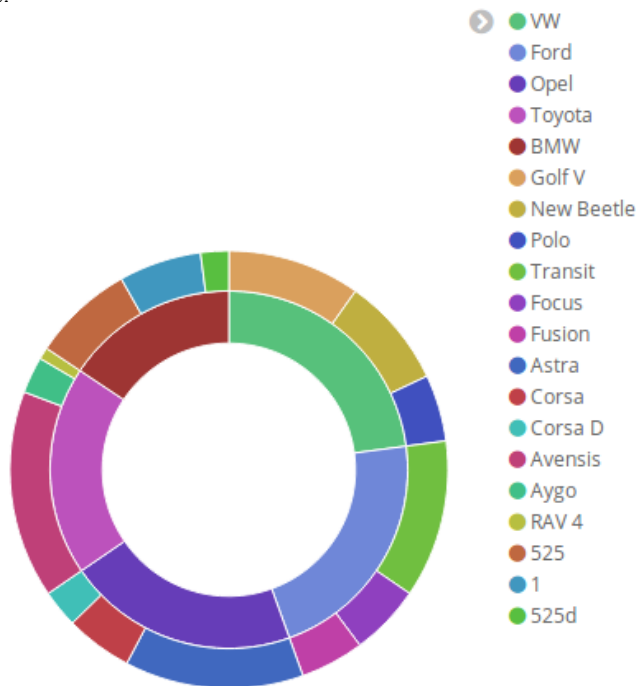
E. Sunburst

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

Como na Europa existe uma amostra muito mais significativa do que em outras regioes do globo, a pergunta fica limitada a regioao da Europa. Além disso, para que a visualização 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 intuição de que na Europa o nível dos automóveis é maior do que no Brasil. Observamos que os fabricantes mais comuns lá são 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 são os mais presentes por lá.

A diferença de padrão de veículos se observa nos modelos mais comuns, uma vez que os fabricantes são os mesmos. Por exemplo, os modelos mais comuns da VW são New Beetle, Golf e Polo. Notavelmente, são veículos mais sofisticados que os modelos comuns brasileiros. O mesmo se observa nos veículos da Ford, cujos modelos mais populares são Transit, Fusion e Focus. Aqui, esses automóveis são automóveis de elite.



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

F. Radar

Para que se possa identificar quais são as marcas e modelos que mais poluem e que menos poluem de uma forma bastante rica, optou-se por fazer uma visualização em forma de radar. Dessa forma, foi possível ver, para cada fabricante, uma série de atributos que parecem ser correlacionados. São eles: rotação do motor (rpm), consumo, velocidade, duração da viagem, tempo de viagem e quantidade de gás carbônico emitido.

Para que isso pudesse ser sintetizado adequadamente na forma de um radar, foram feitas uma série de computações 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 média para cada um desses atributos em cada viagem. Posteriormente, foi feita uma normalização desses dados, pois a ferramenta de visualização que o Kibana oferece não permite que se coloque diferentes escalas em cada um dos eixos.

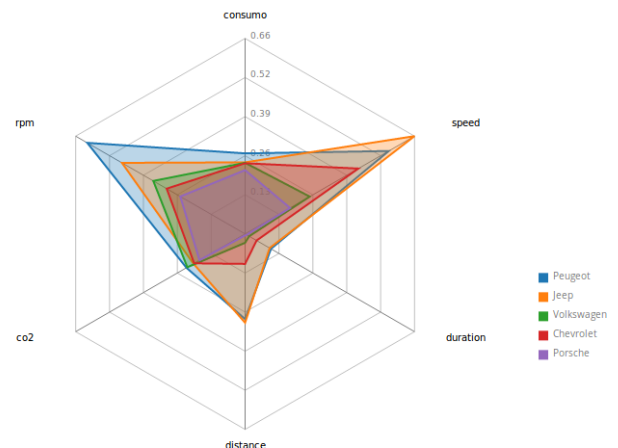
Computadas essas médias e normalizados os dados, podemos observar o gráfico de radar. E ele nos leva a algumas informações bastante curiosas:

Peugeot é o fabricante que normalmente se anda com a rpm mais alta e é o cara que tem maior consumo, mas não é o fabricante que anda mais rápido.

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

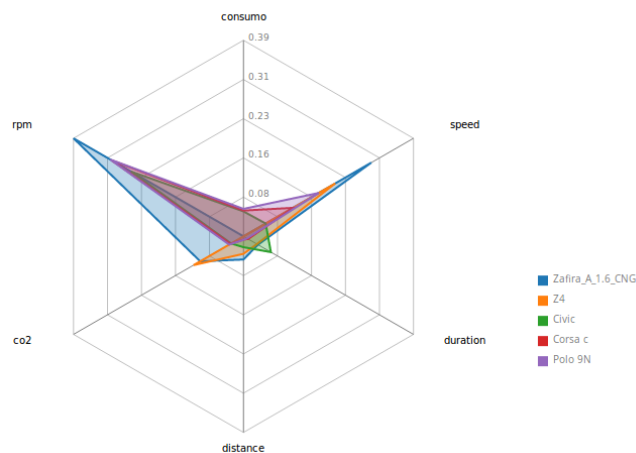
Curioso observar também, que o Porsche é a marca que se anda com rpm mais baixa – ao contrário do que normalmente se imagina – e cuja velocidade é mais baixa também. Acredito que isso se deve porque o Porsche deve ter uma representatividade muito baixa na amostra. Portanto, se você quiser andar rápido, não compre um Porsche!

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

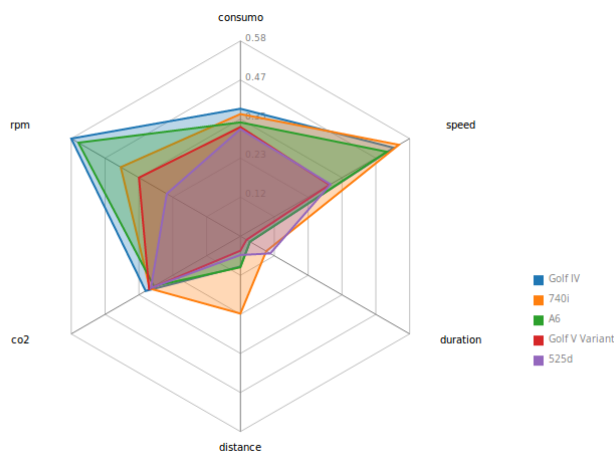


Diversos atributos seccionado por fabricante.

O mesmo tipo de visualização pode ser aplicado, não aos fabricantes, mas aos modelos dos automóveis em si. Na figura X, observam-se os modelos que menos poluem – observe que dos 5 apresentados, 2 são Chevrolet, o que está de acordo com o gráfico de radar apresentado por fabricantes – e observe que os que mais poluem são modelos esportivos, cuja rotação do motor normalmente é alta.



Modelos que menos poluem.



Modelos que mais poluem.

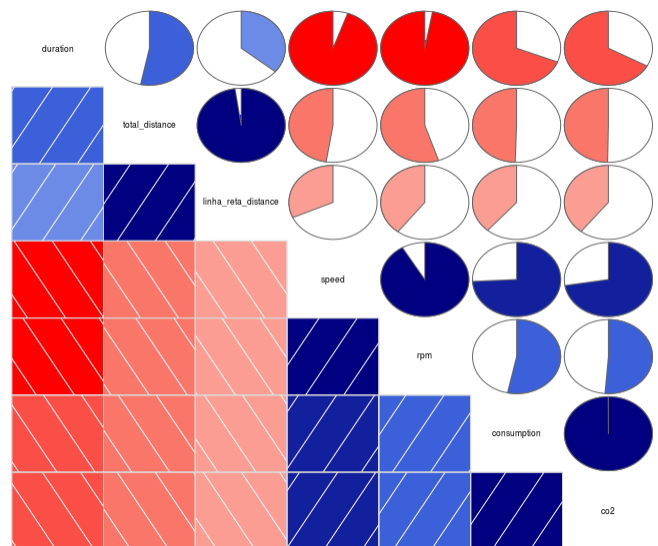
G. Matriz de correlacao

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

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

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

Um ponto a ser considerado é 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

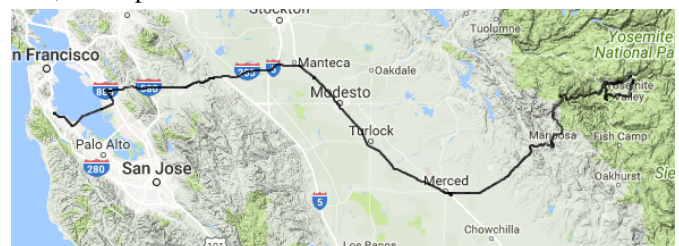
Para uma análise individual de determinados motoristas, será utilizada novamente a tecnica de heatmaps. Infereriu-se que registros de uma regioa isolada – como Brasil ou India –, em que existem poucos registros na base, correspondem ao mesmo motorista.

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

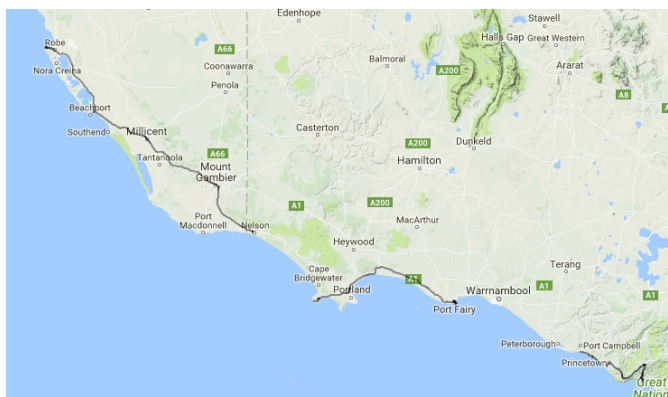
Espera-se dessa maneira, verificar se os caminhos de um determinado motorista são recorrentes ou nao. Além disso, espera-se verificar se essa tecnica é adequada para vislumbrar essa hipotese.

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.

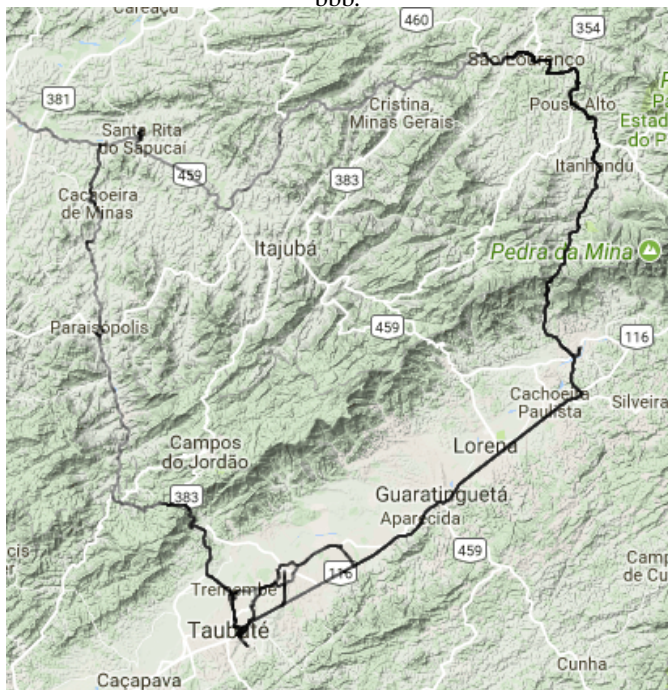
Esse terceiro mapa apoia a ideia de que a tecnica do heatmap pode ser utilizada com muita eficiencia par verificar rotas recorrentes, pois claramente se observa pontos escuros que, de fato, correspondem as rotas mais utilizadas.



aaa.



bbb.



ccc.

V. RESULTADOS

As visualizações deram uma visão 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. Além 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 padrões de comportamento utilizando esse dataset: na Europa, os fabricantes mais comuns são praticamente mesmos fabricantes do Brasil. Entretanto, os carros comercializados por esses fabricantes por la são superiores aos brasileiros.
- observamos de maneira clara, através da visualização de radar, quais são as marcas e fabricantes cujos veículos poluem mais e quais poluem menos. Vimos, nesse mesmo grafico de radar, quais são 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 máquinas, 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 visualização em pizza.
- Os graficos de radar ajudarām a ter claramente uma ideia de muitos atributos sintetizados numa pequena regioao espacial.

VI. CONCLUSÃO

Foi atingido objetivo de se fazer uma exploração básica na base de dados oferecida pelo envirocar. Obteve-se intuição sobre essa base de dados e padrões de comportamento dos motoristas de automóvel.

padrões culturais, como o tipo de veículos e fabricantes utilizados na Europa em comparacao com o Brasil foram identificados. Além disso, padrões individuais de pilotagem de automóveis foram identificados.

Além disso, foi observadas as distribuicoes temporais e espaciais dos registros de veículos.

A base de dados nao é 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.

Além disso, pode-se fazer uma análise básica sobre o uso de ferramentas como Kibana e como R para visualização de informações. 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 detecção de intrusão usando técnicas de Big Data para a análise de logs
- 5) VAARANDI, R.; NIZINSKI, P. Comparative analysis of open-source log management solutions for security monitoring and network forensics