Relatório de Atividades
Preparação de Datasets para Análises

Jefferson Colares

[1]CEFET - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Rio de Janeiro - RJ - Brasil

jefferson.colares@eic.cefet-rj.br

Resumo. *Esse relatório apresenta informações sobre os métodos e processos utilizados para aprimorar os dados nos datasets VRA, WU e apresenta o novo dataset ASOS com dados de sensores meteorológicos instalados em aeroportos ao redor do mundo.*

1. Introdução

Os Datasets VRA e WU contém dados de partidas e chegadas de voos nacionais e dados meteorológicos, respectivamente. Ambos são utilizados como base para diversas análises de dados que suportam pesquisas e artigos científicos desenvolvidos por alunos e professores do PPCIC.

Durante os meses de dezembro/2018 a fevereiro/2019, varias modificações foram realizadas nesses dois conjuntos de dados com o objetivo de:

1. Melhorar a qualidade dos dados
2. Tornar os datasets mais fáceis de manipular.

No decorrer do trabalho, surgiu a possibilidade de utilizar um novo dataset, mais completo, com dados sobre o clima em aeroportos no Brasil e no Exterior.

Esse novo conjunto de dados é o ASOS e esse documento serve de guia introdutório para os que desejarem utilizá-lo em suas pesquisas.

As seções a seguir descrevem brevemente cada um desses três conjuntos de dados.

1.1. VRA

Os dados neste dataset são registrados pelas empresas aéreas e consolidados pela ANAC na base de dados VRA (Voo Regular Ativo). NO endereço `https://sistemas.anac.gov.br/sas/siros/(S(ky34tqfgeulexxodp54cqdcp))/view/vra/frmConsultaVRA` é possível consultar e obter dados do período de 01/01/2000 a 31/12/2018.

Consultas e download de dados a respeito de voos ocorridos a partir de 01/01/2019 podem ser realizados no endereço `https://sistemas.anac.gov.br/sas/siros/(S(ky34tqfgeulexxodp54cqdcp))/view/registro/frmConsultaVoosHistorico`, entretanto, alguns campos importantes para análises, como o horário real de chegada e o de partida não estão mais disponíveis.

Os dados do VRA tratados nesse trabalho foram obtidos em `https://github.com/eogasawara/flight-data`

1.2. WU

Os dados do conjunto WU são originários do Weather Underground (https://www.wunderground.com/), um website que disponibiliza informações obtidas a partir de milhares de estações meteorológicas ao redor do globo.

O Weather Underground, que pertence à empresa Weather Company, fornecia chaves gratuitas de acesso a sua API, para download de dados. Em 2016 a Weather Company foi adquirida pela IBM e, a partir de 2017, não foram mais disponibilizadas chaves gratuitas. A partir de 2019 a API foi totalmente desativada. A empresa disponibiliza a opção de adquirir uma chave para acesso às APIs pagas da Weather Company.[WUA]

Ainda é possível fazer download dos dados diretamente através da página web do Weather, mas este método consome muito tempo e esforço quando são necessários dados meteorológicos de um período grande. O endereço a seguir exemplifica a forma como os dados são disponibilizados para a estação meteorológica do aeroporto Santos Dumont, no Rio: https://www.wunderground.com/history/daily/br/rio-de-janeiro/SBRJ/date/2019-2-16?cm_ven=localwx_history.

O dataset WU tratado neste trabalho foi obtido no endereço https://github.com/eogasawara/flight-data.

1.3. ASOS

ASOS (Automated Surface Observing Systems) é um programa que envolve diversas agências governamentais norte-americanas, criado com o objetivo de formar uma rede oficial de informações meteorológicas para suportar primariamente as entidades da área de aviação, mas também aquelas voltadas a pesquisas meteorológicas, climatológicas e hidrológicas. [ASO]

O departamento de Agronomia da Iowa State University, nos Estados Unidos, compila diariamente não apenas as informações do sistema ASOS dos EUA, mas também de diversas entidades ligadas a aviação civil e militar de todo o planeta.

Essa universidade disponibiliza esses dados gratuitamente para download através do website https://mesonet.agron.iastate.edu/request/download.phtml. Os dados tratados e disponibilizados através deste trabalho foram obtidos nesse endereço.

2. Desenvolvimento

As atividades foram desenvolvidas individualmente para cada dataset e são apresentadas neste relatório da mesma forma.

2.1. VRA

O dataset VRA utilizado neste trabalho é formado de 18 arquivos .ZIP contendo dados de partidas e chegadas de voos em aeroportos brasileiros, nos anos 2000 a 2017.

Cada .ZIP ontém 12 arquivos .CSV, correspondentes a cada mês do ano (exceto junho e julho de 2014, quando a ANAC não fez a coleta dos dados), compreendendo um total de 214 arquivos que totalizam 2.2GB.

A seguir são descritos os problemas encontrados na análise do dataset e as soluções aplicadas a cada um.

2.1.1. Formatação de arquivos

Os zip files correspondentes aos anos 2009 em diante, quando descompactados, colocam os CSVs dentro de pastas. (isso não ocorre com os zips dos anos 2000-2008). Como consequência disso, a leitura do conteúdo dos CSVs em uma única operação, o que dificulta a manipulação dos dados pelos usuários do dataset.

A solução aplicada para esse problema foi reempacotar os zips de 2009 em diante. Desta forma, agora, ao desempacotar os zips, todos os arquivos ficam na mesma pasta.

2.1.2. Formatação de dados

Alguns dos arquivos têm os nomes de colunas diferentes dos demais. Como consequência, encontramos erros ao tentar ler ou concatenar o conteúdo de múltiplos arquivos. A solução aplicada nesse caso foi padronizar os nomes de colunas.

Ainda com respeito a formatação, a coluna Número.do.Voo apresentava um espaço adicional à esquerda que provocava erros nas análises dos dados (o espaço adicional modifica o resultado das buscas por strings). A solução aplicada foi remover o caractere adicional indesejado.

Outro problema encontrado diz respeito ao formato das colunas com horários dos voos, que era ”dd/mm/aaaa h:mm”. Esse formato dificulta as operaçõe envolvendo cálculos de duração e atrasos de voos (em especial, as conversões de tipos de dados para POSIX). A solução aplicada foi alterar o formato dos campos de datas/horas para ”aaaa-mm-dd h:mm”

2.1.3. Qualidade de dados

Com relação à qualidade dos dados, observamos alguns voos com data/hora de chegada menor que a data/hora de partida. Embora isso possa ocorrer em alguns voos longos internacionais é improvável que isso ocorra em voos nacionais.

Foram imaginadas algumas possíveis soluções para esse caso, como:

1. Remover do dataset as linhas com duração de voo negativa.
2. Criar uma coluna ”Status.Registro” contendo 0 para registros bons e 1 para registros ruins.
3. Criar campos adicionais indicando a duração do voo (Duração.Real, Duração.Prevista).

Ao final, foi aplicada a solução número 3, pois dessa forma, fica a cargo do usuário decidir quais observações ele deseja invalidar.

2.2. WU

O dataset WU é composto de 518 arquivos do tipo .Rdata, contendo dados meteorológicos do período de 2009 a 2017, em 62 aeródromos do Brasil e exterior.

Diversos problemas foram observados nesse dataset, todos envolvendo qualidade dos dados. Eles são apresentados a seguir, juntamente com suas respectivas soluções.

2.2.1. Colunas deslocadas

Possivelmente representando o maior problema do dataset, 243.089 (4,6%) das observações no dataset estão com as colunas deslocadas, ou seja, os dados apresentados em uma determinada coluna pertencem, na realidade, à coluna ao lado. Como consequência, a acurácia de todos os resultados de análises feitas no dataset é comprometida. Especialmente aquelas que são mais sensíveis à existência de outliers.

```
> summary(all_wu)
  data.airport     data.datetime                data.temperature data.dew_point
 LPPT   : 157966   Min.   :2009-01-01 00:00:00   Min.   : 0.00    Min.   : 0.00
 SKBO   : 109220   1st Qu.:2011-04-26 03:00:00   1st Qu.:20.00    1st Qu.:15.00
 SBPV   : 106996   Median :2013-08-03 20:53:00   Median :24.00    Median :19.00
 SAEZ   : 106603   Mean   :2013-07-14 22:45:47   Mean   :23.12    Mean   :18.09
 SBSV   : 105595   3rd Qu.:2015-09-27 02:00:00   3rd Qu.:27.00    3rd Qu.:22.00
 SBNT   : 105421   Max.   :2017-12-31 23:58:00   Max.   :90.00    Max.   :81.00
 (Other):4529220                                 NA's   :60994    NA's   :229530
 data.humidity    data.pressure    data.visibility                 data.events
 Min.   :  0.00   Min.   :   0.0   Min.   :    0                          :3469276
 1st Qu.: 59.00   1st Qu.:1011.0   1st Qu.:   10     \xc2                 :1047973
 Median : 74.00   Median :1014.0   Median :   10     Rain                 : 269182
 Mean   : 70.27   Mean   : 967.6   Mean   :   69     Rain , Thunderstorm:  78224
 3rd Qu.: 88.00   3rd Qu.:1017.0   3rd Qu.:   10     Thunderstorm         :  68220
 Max.   :999.00   Max.   :1100.0   Max.   :14481     Fog                  :  43782
 NA's   :65254    NA's   :137135   NA's   :1205432   (Other)              : 244364
        data.conditions
 Clear           :1214580
 Scattered Clouds:1049769
 Mostly Cloudy   :1015122
 Partly Cloudy   : 848644
 Light Rain      : 216500
 (Other)         : 807462
 NA's            :  68944
```

Figure 1. Valores máximos fora das faixas esperadas

A solução aplicada para esse problema foi utilizar os dados nas colunas data.visibility e data.pressure, que tem valores limites bem definidos, como referência

para identificar as linhas que tiveram suas colunas deslocadas. Após a correção, o número de observações com valores fora dos parâmetros esperados caiu de 241.878 para apenas 44.

2.2.2. Valores negativos de temperatura ausentes

Os campos contendo temperaturas (data.temperature e data.dew_point) apresentavam como positivos valores que deveriam ser representados como negativos. Como consequência, os resultados de análises envolvendo dados de aeródromos de regiões frias e do exterior poderiam sair distorcidos.

A solução adotada para esse problema foi obter os dados corretos a partir da base de dados ASOS e substituir os valores errados.

Apenas os aeródromos de regiões frias do país e do exterior foram selecionados para que tivessem seus dados de temperatura substituídos.

Sigla OACI	Descrição	Cidade	UF	País
SCEL	COMODORO ARTURO MERINO BENÃTEZ INTERNATIONAL	SANTIAGO DO CHILE		CHILE
KJFK	JOHN F. KENNEDY INTERNATIONAL AIRPORT	NEW YORK		ESTADOS UNIDOS DA AMERICA
LPPT	LISBOA	LISBOA		PORTUGAL
SUMU	CARRASCO GRAL. CESÃREO L. BERISSO INTERNATIONA	MONTEVIDEU		URUGUAI
SABE	JORGE NEWBERY AIRPORT	BUENOS AIRES/AEROPARQUE		ARGENTINA
SAEZ	MINISTRO PISTARINI INTERNATIONAL AIRPORT (EZEIZA I	BUENOS AIRES		ARGENTINA
SBCA	ADALBERTO MENDES DA SILVA	CASCAVEL	PR	BRASIL
SBCT	AFONSO PENA	SÃO JOSÉ DOS PINHAIS	PR	BRASIL
SBFI	CATARATAS	FOZ DO IGUAÇU	PR	BRASIL
SBLO	GOVERNADOR JOSÉ RICHA	LONDRINA	PR	BRASIL
SBMG	REGIONAL DE MARINGÁ SÍLVIO NAME JÚNIOR	MARINGÁ	PR	BRASIL
SBPA	SALGADO FILHO	PORTO ALEGRE	RS	BRASIL
SBCH	SERAFIN ENOSS BERTASO	CHAPECÓ	SC	BRASIL
SBFL	HERCÍLIO LUZ	FLORIANÓPOLIS	SC	BRASIL
SBJV	LAURO CARNEIRO DE LOYOLA	JOINVILLE	SC	BRASIL
SBNF	MINISTRO VICTOR KONDER	NAVEGANTES	SC	BRASIL

Figure 2. Relação de aeródromos que tiveram os dados de temperatura substituídos

2.3. ASOS

2.3.1. Aquisição de dados

Os dados que compõem o dataset ASOS foram baixados do departamento de Agronomia da Iowa State University, através do website mencionado na introdução desse relatório.

O download foi feito da seguinte forma:

Todos os aeródromos do Brasil foram incluídos nas consultas. No entanto, foi necessário fazer uma consulta separada para cada ano, de forma que o volume de dados manipulado de uma única vez não ultrapassasse a capacidade da máquina.

Cada um desses arquivos foi colocado em um diretório separado, nomeado de acordo com o ano dos dados no arquivo.

Foram baixados 19 arquivos com os anos completos de 2000 a 2018.

No caso dos aeroportos do exterior, não foi possível baixá-los todos em um único arquivo, como ocorreu com os aeroportos brasileiros, devido a limitações da interface do site. Em compensação, por se tratarem de poucos aeroportos, foi possível baixar todo o período de 2000 a 2018 em um único arquivo para cada país.

Esse download foi feito em um diretório de nome ASOS_EXT, onde cada arquivo representa um ou mais aeroportos de cada país.

2.3.2. Tratamento de arquivos

Depois de baixados os arquivos, o script abaixo foi utilizado para juntar todos os aeroportos em um só dataframe e salvar os registros de cada ano em um arquivo na pasta correspondente

```
# Juntar os arquivos de aeroportos estrangeiros e
# gerar arquivos separados para cada ano na pasta do ano correspondente
setwd(paste(base_dir, "EXT_ASOS", sep = "/"))
file_list = list.files(path = ".")
EXT_ASOS <- read.csv(file=file_list[1], stringsAsFactors = FALSE)
for ( i in 2:length(file_list)) {
  tmp <- read.csv(file=file_list[i], quote = "\"", stringsAsFactors = F
  EXT_ASOS <- rbind(EXT_ASOS, tmp)
  }
EXT_ASOS$ano = substr(EXT_ASOS$valid, 1,4)
anos = EXT_ASOS %>% distinct(ano)
for(i in 1:length(anos$ano)) {
  filename = paste("../", anos$ano[i], "/ASOS_EXT_", anos$ano[i], ".txt
  tmp_asos <- EXT_ASOS %>% filter(ano == anos$ano[i])
  tmp_asos$ano <- NULL
  write.csv(tmp_asos, file = filename,row.names = FALSE)
}
```

Nesse ponto, cada uma das 19 pastas nomeadas 2000 a 2018, contém dois arquivos: Um com os dados de aeroportos brasileiros, outro com os de aeroportos estrangeiros. O passo seguinte é juntar ambos em um único arquivo para cada ano. Isso é feito com o trecho de código abaixo:

```
# Juntar os aeroportos brasleiros e estrangeiros em um único arquivo pa
# no mesmo diretório.
dir.create(paste(base_dir, "consolidado", sep = "/"))
for(i in 1:length(anos$ano)) {
  dir <- paste(base_dir, anos$ano[i], sep="/")
  setwd(dir)
  file_list = list.files(path = ".")
  tmp <- read.csv(file=file_list[1], quote = "\"", stringsAsFactors = F
  for(j in 2:length(file_list)) {
    tmp1 <- read.csv(file=file_list[j], quote = "\"", stringsAsFactors
    tmp <- rbind(tmp, tmp1)
  }
  filename <- paste("../consolidado/ASOS", anos$ano[i], ".txt",sep = ""
  write.csv(tmp, file = filename, row.names = FALSE)
  # file.remove(file_list)
```

}

Ao final do tratamento, ficamos com os 19 arquivos com dados de todos os aeroportos do Brasil e alguns do exterior, sendo um para cada ano completo entre 2000 e 2018

2.3.3. Tratamento de dados

Com o objetivo de facilitar a utilização dos dados por parte de nossos pesquisadores, algumas etapas de tratamento de dados foram realizadas:

1. Conversão de NULLS em NAs.
2. Correção dos códigos dos aeroportos americanos para o padrão IATA.
3. Atribuição de tipos de dados apropriados para as colunas numéricas.
4. Criação de novas colunas com temperaturas em graus Celsius.

3. Resultados

O resultado final do trabalho é composto pelos três datasets, cujos formatos e conteúdos são descritos abaixo.

3.1. VRA

3.1.1. Formato

O VRA está disponível em 6 arquivos do tipo RData. Cada um deles contém os seguintes períodos:

- **vra1.RDAta**: 2000 a 2002
- **vra2.RDAta**: 2003 a 2005
- **vra3.RDAta**: 2006 a 2008
- **vra4.RDAta**: 2009 a 2011
- **vra5.RDAta**: 2012 a 2014
- **vra6.RDAta**: 2015 a 2017

3.1.2. Conteúdo

Todos os arquvios contem apenas um objeto: o data frame **vra**. As colunas desse dataframe são:

- **Sigla.da.Empresa** - sigla IATA da companhia aérea. Ex: GLO, TAM, ONE
- **Número.do.Voo** - Identificador do voo ANAC
- **D.I** - Digito identificador do tipo do voo 0: Vôo Regular, 1: Vôo extra com HOTRAN, 2: Vôo extra sem HOTRAN, 3: Vôo de retorno, 4: Inclusão de etapa em um vôo previsto em HOTRAN, 5: Vôo Cargueiro, 6: Vôo de Serviço, 7: Vôo de fretamento, 9: Vôo charter, A: Vôo de instrução, B: Vôo de experiência)
- **Tipo.de.Linha** - Dígito identificador do tipo de linha do voo. (C: Cargueiro, E: Especial, G: Cargueiro internacional, H: Sub-Regional, I: Internacional, L: Rede Postal, N: Nacional, R: Regional)
- **Aeroporto.Origem** - Código ICAO do aeroporto de origem do voo
- **Aeroporto.Destino** - Código ICAO do aeroporto de destino do voo
- **Partida.Prevista** - Data e hora previstas da partida do voo
- **Partida.Real** - Data e hora reais da partida.
- **Chegada.Prevista** - Data e hora previstas da chegada.
- **Chegada.Real** - Data e hora reais da chegada do voo.
- **Situação** - Situação do voo (Realizado ou Cancelado)
- **Justificativa** - Código para o atraso ou cancelamento do voo, quando houver, de acordo com a normativa IAC 1504 da ANAC [IAC]. (Exemplos: XN: Cancelamento por motivos técnico-operacionais, XO: Cancelamento - aeroporto de origem abaixo dos limites, XT: Cancelamento - aeroporto de destino abaixo dos limites)
- **Duração.Real** - Duração real do voo, em minutos.
- **Duração.Prevista** - Duração prevista do voo, em minutos.

3.2. WU

3.2.1. Formato

O dataset WU está armazenado em um único arquivo, de nome **wu.RData**, contendo dados dos anos 2009 a 2017.

3.2.2. Conteúdo

O arquivo contém apenas um objeto, o dataset **wu**, que apresenta as seguintes colunas:

- **data.airport** - codigo ICAO do aeroporto onde os dados foram coletados
- **data.datetime** - data e hora da coleta
- **data.temperature** - temperatura, em graus Celsius
- **data.dew_point** - ponto de orvalho em Celsius
- **data.humidity** - percentual de umidade do ar
- **data.pressure** - pressão atmosférica, em mbar
- **data.visibility** - visibilidade em kms
- **data.events** - Eventos atmosféricos no momento da medição.
- **data.conditions** - Descrição textual das condições atmosféricas

3.3. ASOS

3.3.1. Formato

O dataset ASOS está disponível em dois formatos: .csv ou .RData. Em ambos os casos, são 19 arquivos, contendo os dados de cada ano entre 2000 e 2018. O conteúdo dos dois conjuntos é o mesmo e a ideia de oferecê-lo em dois formatos visa apenas dar opções aos usuários.

3.3.2. Conteúdo

Todos os arquivos do conjunto no formato RData contém um único dataframe de nome **asos**. As colunas nesse dataframe são as mesmas do conjunto no formato csv:

- **station** - Código OACI do aeroporto
- **valid** - data/hora de coleta dos dados
- **tmpf** - temperatura em Fahrenheits
- **dwpf** - ponto de orvalho em Fahrenheits
- **relh** - percentual de umidade relativa do ar
- **drct** - direção do vento em graus, a partir do Norte
- **sknt** - velocidade do vento, em nós
- **p01i** - precipitação na hora anterior à medição, em polegadas.
- **alti** - altímetro de pressão, em polegadas
- **mslp** - pressão atmosférica, em mbar
- **vsby** - visibilidade, em milhas
- **gust** - rajadas de vento, em nós
- **skyc1** - nebulosidade, no aeroporto, na hora da coleta de dados.
- **skyc2** - nebulosidade, na região do aeroporto, nas 2 horas antecedentes

- **skyc3** - nebulosidade, no país, nas 2 horas antecedentes
- **skyc4** - nebulosidade
- **skyl1** - teto (altitude das nuvens) no aeroporto, real-time, em pés
- **skyl2** - teto (altitude das nuvens) na região do aeroporto, nas duas horas anteriores, em pés
- **skyl3** - teto (altitude das nuvens) no país, nas duas horas anteriores, em pés
- **skyl4** - teto (altitude das nuvens), em pés
- **wxcodes** - códigos de condições atmosféricas (extraído do metar) [MET]
- **ice_accretion_1hr** - gelo acumulado na última hora, em polegadas
- **ice_accretion_3hr** - gelo acumulado nas últimas 3 horas, em polegadas
- **ice_accretion_6hr** - gelo acumulado nas últimas 6 horas, em polegadas
- **feel** - sensação teérmica, en Fahrenheits
- **metar** - texto completo do METAR [MET]
- **tmpc** - temperatura, em graus Celsius
- **dwpc** - ponto de orvalho, em graus Celsius
- **feelc** - sensação térmica, em graus Celsius

References

[WUA] End of service for the weather underground api. https://apicommunity.wunderground.com/weatherapi/topics/end-of-service-for-the-weather-underground-api. Accessed: 2019-02-17.

[MET] Metar help. https://weather.cod.edu/notes/metar.html. Accessed: 2019-02-17.

[IAC] Procedimentos para o registro de alteração em vôos de empresas de transporte aéreo regular. http://www.anac.gov.br/assuntos/legislacao/legislacao-1/iac-e-is/iac/iac-1504. Accessed: 2019-02-17.

[ASO] Us national weather service - automated surface observing systems. https://www.weather.gov/asos/. Accessed: 2019-02-17.