MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Conselho Nacional de Desenvolvimento Científico e Tecnológico

AMBIENTE COMPUTACIONAL PARA MONITORAMENTO E ANÁLISES DE DADOS METEOROLÓGICOS

RELATÓRIO FINAL DE PROJETO DE INICIAÇÃO CIENTÍFICA
(PIBIC/CNPq/INPE)

Almir de Oliveira Giornes (FATEC Cruzeiro, Bolsista PIBIC/CNPq)
Email: almirgiornes@gmail.com

Eduardo Batista de Moraes Barbosa (CPTEC/INPE, Orientador)
Email: eduardo.barbosa@cptec.inpe.br

Julho de 2018

Sumário

Resumo.....1
1. Introdução.....2
2. Objetivo.....2
   2.1. Objetivos Específicos.....2
3. Fundamentação Teórica.....2
   3.1. Sistema de Monitoramento Terrestre.....3
   3.2. Linguagem de Programação Python.....3
   3.3. Linguagem de Programação R.....4
   3.4. Distribuições de Probabilidade.....4
4. Metodologia.....4
5. Resumo das Realizações.....5
6. Análises e Resultados.....6
7. Conclusões.....9
8. Referências Bibliográficas.....10

Lista de Ilustrações

Tabela 1 - Exemplo de tabela SYNOP do dia 01 de Janeiro de 2017, às 12:00…….....…...5
Figura 1 - Cobertura de dados de superfície - Synop em azul/Metar em vermelho/Ema em amarelo……...6
Figura 2 - Histogramas dos dados SYNOP nos horários das 20:00, 21:00, 22:00 e 23:00….....7
Figura 3 - Comparação gráfica dos ajustes das distribuições de Cauchy, Normal, Logística e Exponencial……...8
Tabela 2 - Testes estatísticos elaborados nos dados de observação com relação aos ajustes de distribuição de probabilidade.....8

Resumo

Nas últimas duas décadas, o número de redes automáticas de estações meteorológicas aumentou consideravelmente como consequência da necessidade de dados meteorológicos em tempo quase real e da grande evolução de sistemas automáticos de aquisição de dados. No Brasil, o Centro de Previsão de Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais adquire diariamente um amplo conjunto de dados meteorológicos, que em grande parte são utilizados como insumo de rotinas operacionais, bem como no desenvolvimento de estudos científicos. A partir deste projeto pretende-se desenvolver um sistema para monitoramento de dados composto por mapas e informações estatísticas. Para este estudo, foram adquiridos três anos de dados meteorológicos do tipo SYNOP, originados de estações meteorológicas que reportam condições da superfície terrestre em horários sinóticos, que são às 12 horas, às 18 e à meia noite, por meio do Sistema de Telecomunicação Global. Inicialmente, foram confeccionados mapas para avaliar a distribuição espacial dos dados. A partir de estudos estatísticos foi possível conhecer alguns parâmetros, em que um deles é o valor da média das observações por hora, a partir dos quais pode-se observar que as maiores quantidades de observações se encontram nos horários sinóticos, com uma média de 6710 observações, e nos horários intermediários aos sinóticos, que são de três em três horas, com uma média de 5965 observações, os demais horários apresentam uma média menor, apresentando 1974 observações. Outro parâmetro observado foi a média diária, que no qual apresenta cerca de 3397 observações. Identificou-se um crescimento ao longo dos anos em relação ao total de observações por dia, em que a média do total de observações diárias registrado em 2015 foi de 67817, enquanto que em 2017 foi uma média de 90049, apresentando um crescimento de 32,78%. A partir de distribuições de frequência foram realizados ajustes da distribuição de probabilidade visando conhecer o padrão dos dados e prever possíveis erros. Os resultados preliminares revelam semelhanças entre horários de chegada dos dados, principalmente nos horários sinóticos, que apresentam quantidades de observações similares e com valores mais elevados. Para chegar aos ajustes, foram realizados testes através de programas em R, um ambiente estatístico que disponibiliza ferramentas para análises estatísticas aprimoradas. Foram realizados diversos testes com diferentes distribuições de probabilidade, e a melhor foi a distribuição de Cauchy. Essa distribuição se assemelha à distribuição Normal, porém com os testes realizados, esta não foi adotada como a melhor por conta da dispersão dos dados. A partir dos ajustes espera-se que os dados correntes sigam o mesmo padrão identificados em anos anteriores.

1. Introdução

Nas últimas décadas, o número de estações automáticas meteorológicas aumentou consideravelmente. Isso é consequência da necessidade de dados meteorológicos em tempo quase real e da evolução dos sistemas automáticos de aquisição de dados (MILLER E BARTH, 2003).

A alta necessidade na precisão desses dados também aumentou, e por conta disso, é essencial efetuar melhorias constantes e fortalecer as redes de observações meteorológicas em escala global, para que seja possível subsidiar aplicações operacionais e o desenvolvimento de pesquisas aplicadas (WMO, 2008).

No Brasil, o Centro de Previsão do Tempo e Estudos Climáticos do Inpe adquire diariamente uma grande quantidade de dados meteorológicos, que em grande parte são utilizados como matéria-prima de produtos operacionais e em estudos científicos. No entanto, a existência de erros pode influenciar a qualidade dos trabalhos.

Este projeto se concentra em produzir informações estatísticas em relação aos dados meteorológicos, focando no tipo SYNOP, com o propósito de informar sobre a qualidade e disponibilidade de dados essenciais para trabalhos rotineiros e pesquisas. Os resultados das análises estatísticas serão produzidos para apoiar, principalmente, a tomada de decisão quanto ao uso desses dados.

2. Objetivo

Este projeto tem como objetivo desenvolver análises e visualização de conjuntos de dados meteorológicos.

2.1 Objetivos Específicos

- Preparar informações estatísticas a respeito dos dados (como por exemplo, mapas da distribuição espacial dos dados);
- Identificar parâmetros estatísticos (como a média por dia ou por hora, desvio padrão, total, etc) dos dados;
- Facilitar o diagnóstico de erros e a tomada de decisão quanto ao uso de dados; e
- Contribuir para a formação e inserção de estudantes de graduação em atividades de pesquisa.

3. Fundamentação Teórica

O conhecimento adquirido a partir dos estudos realizados durante o período da bolsa são apresentados nas Seções 3.1, 3.2, 3.3 e 3.4. A Seção 3.1 descreve o sistema de monitoramento terrestre, de onde são adquiridos os conjuntos de dados objeto deste estudo. Na Seção 3.2 é apresentada a linguagem de programação Python, utilizada no desenvolvimento de programas para confecção de mapas de distribuição de dados. A Seção 3.3 irá tratar sobre a linguagem de programação R, utilizada para as análises e identificação de parâmetros estatísticos, onde será apresentado, tanto graficamente quanto numericamente, a justificativa da escolha da distribuição de probabilidade. Na Seção 3.4 será explicado o que é uma distribuição de probabilidade.

3.1 Sistema de Monitoramento Terrestre

O Sistema de Telecomunicação Global da Organização Meteorológica Mundial é um componente de comunicação e gerenciamento de dados que permite ao Observatório Meteorológico Mundial (do inglês, *World Weather Watch* - WWW) operar através da coleta e distribuição de informações críticas para seus processos. O GTS é implementado e operado pelos Serviços Meteorológicos Nacionais de Membros da Organização Meteorológica Mundial e Organizações Internacionais (WMO-No. 386, 2015).

Além do GTS, existe o Sistema de Observações Global (do inglês, *Global Observing System* - GOS) e o Sistema de Dados-Processados e de Previsão Global (do inglês, *Global Data-Processing and Forecasting System* - GDPFS), ambos também são componentes do WWW. O Sistema de Observações Global providencia os dados de observação para agrometeorologia, meteorologia aeronáutica, e climatologia, inclusive o estudo de clima e mudança global. Esses dados trafegam pelo GOS e GTS, de forma automática, para a distribuição de dados de observação. O GDPFS é composto por centros de meteorologia que fornecem os dados processados, análises e produtos para previsão do tempo. Os centros meteorológicos são organizados em três níveis: mundiais, especializados regionais e nacionais. O CPTEC está inserido no contexto da WWW como Centro Meteorológico Nacional (do inglês, *National Meteorological Center* - NMC) do Sistema de Dados-Processados e de Previsão (CINTRA, 2005).

O GOS providência diversos tipos de dados observacionais, dentre eles, os principais são:

- SYNOP, usados para reportar observações meteorológicas de superfície do continente, feitas em horários periódicos, geralmente em intervalos de três horas e seis horas (BLUESTEIN, 1993);
- METAR, rotinas originárias de registros em aeroportos, que informa a condição do tempo no local e regiões subjacentes (WMO-No. 782, 2014);
- BUOY, que são dados observacionais originários de boias à deriva rastreadas e localizadas via satélite, com a finalidade de informar as condições climáticas no mar, bem como estudos da circulação oceânica (STEVENSON e KAMPEL, 1997);
- SHIP, que são registros originários de navios em alto mar, que informam as condições em alto mar, como o estado do mar (em relação às ondas), temperatura da água, temperatura local. (WMO-No. 471, 2017);
- AIREP, observações reportadas por aeronaves em voo, informando valores das condições climáticas (WMO-No. 1200, 2017);
- TEMP, observações de ar superior medidas por radiossondas (balões meteorológicos), obtidas duas a quatro vezes ao dia, reportando temperatura, umidade e velocidade do vento; e
- PILOT, observações de vento obtidas por meio de balões piloto. A precisão desta observação é um pouco mais baixa se comparada às observações feitas por meio de radiossonda, por conta da velocidade de subida dos balões (CINTRA, 2005).

3.2 Linguagem de Programação Python

Com a linguagem de programação Python, em conjunto com bibliotecas externas Pandas e Matplotlib, é possível elaborar programas de computador para criar uma prévia da visualização dos dados utilizando uma tabela de dados meteorológicos. Resumidamente, esses programas realizam a leitura das tabelas (Tabela 1, Seção 5), identificando as colunas necessárias para elaborar a visualização.

A biblioteca Pandas foi desenvolvida com base na biblioteca NumPy, uma biblioteca padrão do Python, muito utilizada na realização de cálculos científicos. Ela fornece um alto desempenho e uma maior facilidade em usar as ferramentas de análise de dados e uma implementação eficiente para as camadas de dados (em Inglês, *Data Frames*) (VANDERPLAS, 2017). Uma camada de dados é um registro dos dados, semelhante à uma matriz, cujas colunas podem possuir nomes e diferentes tipos de dados. Cada linha de um Data Frame é um registro, funcionando assim, como uma tabela.

Com a biblioteca Matplotlib é possível criar a visualização dos dados através de mapas e gráficos 2D (bidimensional). Possui suporte para visualização interativa, em que é possível utilizar o zoom, alternar entre mapas/gráficos, sendo também possível salvar as figuras geradas em diversos formatos de saída (JPEG, PNG, PS, dentre outras) (TOSI, 2009).

3.3 Linguagem de Programação R

A linguagem R é amplamente empregada em ambientes onde há a necessidade de analisar, visualizar e obter resultados dos dados. Esta linguagem de programação disponibiliza inúmeras ferramentas que fornecem um melhor suporte à essas necessidades (MATLOFF, 2011). Uma das ferramentas utilizadas no projeto foi a biblioteca “fitdistrplus”, que de maneira bem simples, auxilia no ajuste dos dados à uma distribuição de probabilidade. Esta biblioteca, além de possuir funções que retornam valores numéricos, ela também possui funções que geram figuras, como comparações de densidade das distribuições de probabilidade.

3.4 Distribuições de Probabilidade

Segundo Grimmett e Stirzaker (1992), grande parte da vida é baseada na crença de que o futuro é em grande parte imprevisível. Expressamos essa crença no comportamento casual através do uso de palavras como “aleatório” ou “probabilidade”. A probabilidade é definida como a chance de um determinado evento ocorrer, baseado em análises prévias de outras ocorrências e uma margem de erro.

Uma distribuição de probabilidade nos permite identificar padrões nos dados em análise, onde esses dados podem se ajustar e assim, ser representado por uma distribuição de probabilidade, assumindo-a como um padrão para dados posteriores.

4. Metodologia

O projeto propõe produzir informações estatísticas de dados meteorológicos, tendo um foco maior nos dados do tipo SYNOP. Para que estas informações sejam visualizadas, utilizou-se da linguagem Python (URL: www.python.org), uma linguagem de programação interpretada de alto nível, amplamente empregada no processamento de dados científicos e criação de páginas dinâmicas para Internet (MATTHES, 2015), e para as análises estatísticas, foi usada a linguagem de programação R (URL: www.r-project.org), uma linguagem que fornece ferramentas que facilitam essa tarefa, sendo possível obter resultados rápidos e precisos.

Para a produção de análises a respeito dos conjuntos de dados, foram desenvolvidos programas em linguagem Python para criação de mapas com a distribuição espacial dos dados, e com a linguagem R, foram desenvolvidos programas para criar a distribuição de frequências e gráficos a respeito dos dados (MORETTIN e BUSSAB, 2014). Nesta etapa também foi feita a investigação das características dos dados por meio de parâmetros

estatísticos em termos de qualidade e disponibilidade, em que serão analisados o número de observações, média e desvio-padrão de observações com relação ao valor esperado.

5. Resumo das Realizações

O Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) adquire e processa em tempo quase-real os dados meteorológicos de todo o mundo. Esses dados são adquiridos pelo Sistema de Telecomunicação Global (do inglês, *Global Telecommunication System* - GTS), coordenado pela Organização Meteorológica Mundial (do inglês, *World Meteorological Organization* - WMO), e, também, pelo Sistema de Distribuição de Dados pela Internet (do inglês, *Internet Data Distribution* - IDD), do programa UNIDATA. Os sistemas possuem uma variedade de dados de observações e derivados de satélite em diversos formatos.

Durante o período vigente da bolsa, foram realizados estudos sobre dados meteorológicos utilizados pelo CPTEC, com foco nos dados SYNOP, para o conhecimento de seus tipos e suas finalidades. Uma das atividades programadas na primeira etapa do projeto consiste na confecção de mapas com a distribuição espacial dos dados. Para realização dessa atividade, foi necessário preparar programas para coletar os dados por meio do software GEMPAK, um sistema responsável em produzir pacotes de dados meteorológicos em formato binário. Para agilizar o processo foram desenvolvidos programas utilizando a linguagem de comando interativa Korn Shell (ROSENBLATT, 1993) do Linux, para converter arquivos binários em arquivos ASCII em forma de uma tabela (Tabela 1).

Tabela 1. Exemplo de tabela SYNOP do dia 01 de Janeiro de 2017, às 12:00

Código	Data	Latitude	Longitude	Hora (GMT)
835760	170101/1200	-19.75	-47.97	12h00
835790	170101/1200	-19.60	-46.93	12h00
835820	170101/1200	-20.00	-45.98	12h00
...	...	...	...	...

Através dos programas, também, foram adquiridos três anos de observações meteorológicas compreendidos no período 2015 e 2017.

Para a identificação de parâmetros e ajustes à uma distribuição de probabilidade, foram feitos programas de computador na linguagem de programação R, uma linguagem que disponibiliza um ambiente para visualização e análises de dados. A escolha do R se deve ao fato da facilidade em identificar a distribuição de probabilidade dos dados, bem como compará-las em uma única imagem, e em realizar testes estatísticos.

6. Análises e Resultados

Como resultado da primeira etapa do projeto, pode-se destacar o mapa de distribuição espacial dos dados meteorológicos (Figura 1). Na Figura 1, são destacados dados do tipo SYNOP, METAR e de Estações Meteorológicas de Automáticas (EMA), usados para reportar observações meteorológicas de superfície do continente automaticamente a cada hora (URL: www.inmet.gov.br).

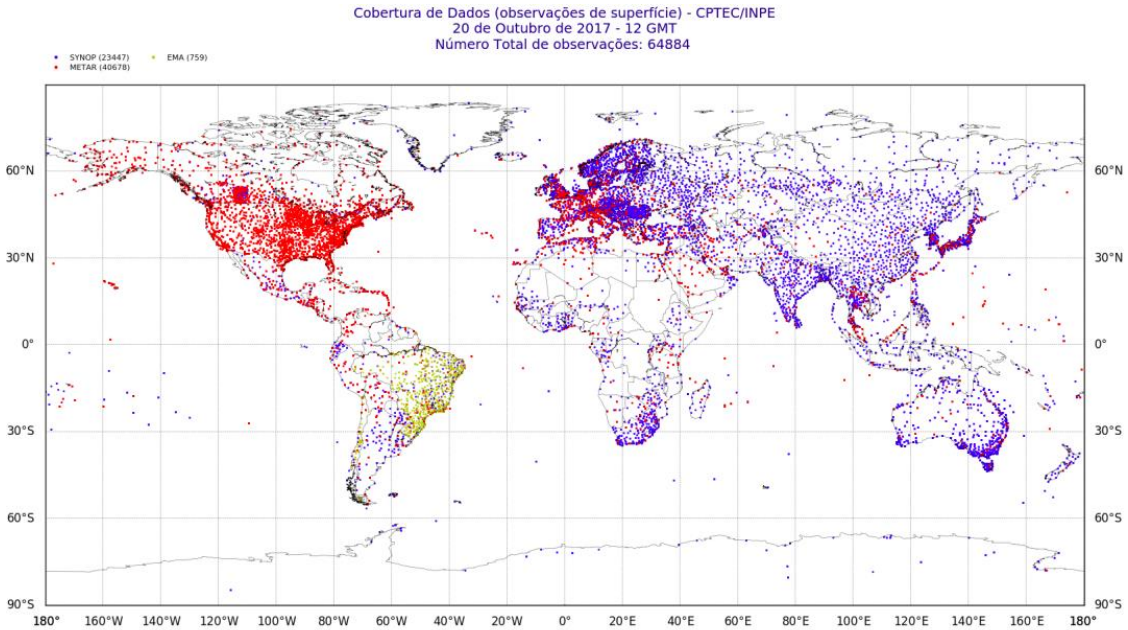

Figura 1. Cobertura de dados de superfície - Synop em azul/Metar em vermelho/Ema em amarelo

Analisando o mapa, observa-se uma grande concentração de dados METAR (em vermelho) sobre a América do Norte, principalmente, na região dos Estados Unidos. Os dados EMA (em amarelo) estão concentrados sobre a América do Sul, na maior parte do Brasil. Os dados SYNOP (em azul) estão na maior parte no continente Europeu.

Como resultado da continuação da primeira etapa, foi efetuado o levantamento de parâmetros estatísticos dos dados SYNOP e elaborada as análises estatísticas da distribuição de probabilidade, embasadas numa população total de três anos (2015, 2016, 2017), onde foi necessário obter uma amostra, ou seja, uma parte da população, para efetuar análises e obter resultados para tomada de decisão. Os parâmetros estatísticos levantados em relação aos três anos de observações são as médias por dia observado e a média por hora. No ano de 2015, tomando como alvo os dados SYNOP, foi obtido uma média total de 65651 observações por dia, cerca de 2735 observações por hora, já no ano de 2017 esse número cresceu, em que eram obtidas aproximadamente 90056 observações por dia, e 3752 observações por hora, isso representa e comprova que houve um aumento significativo no número de estações meteorológicas de dados SYNOP ao redor do mundo, cerca de 37,17%, em que cada observação é originada de uma estação meteorológica. Com a contagem das observações, foi elaborada a tabela de frequência, em que era possível identificar a quantidade de vezes que uma amostra se apresentava em um intervalo de dados. A amostra para esse caso são os horários, onde elabora-se uma representação gráfica dos intervalos através de histogramas, como no exemplo apresentado na Figura 2. Cada histograma possui valores diferentes nos intervalos e na frequência por conta dos horários em que foram utilizados para confeccioná-los.

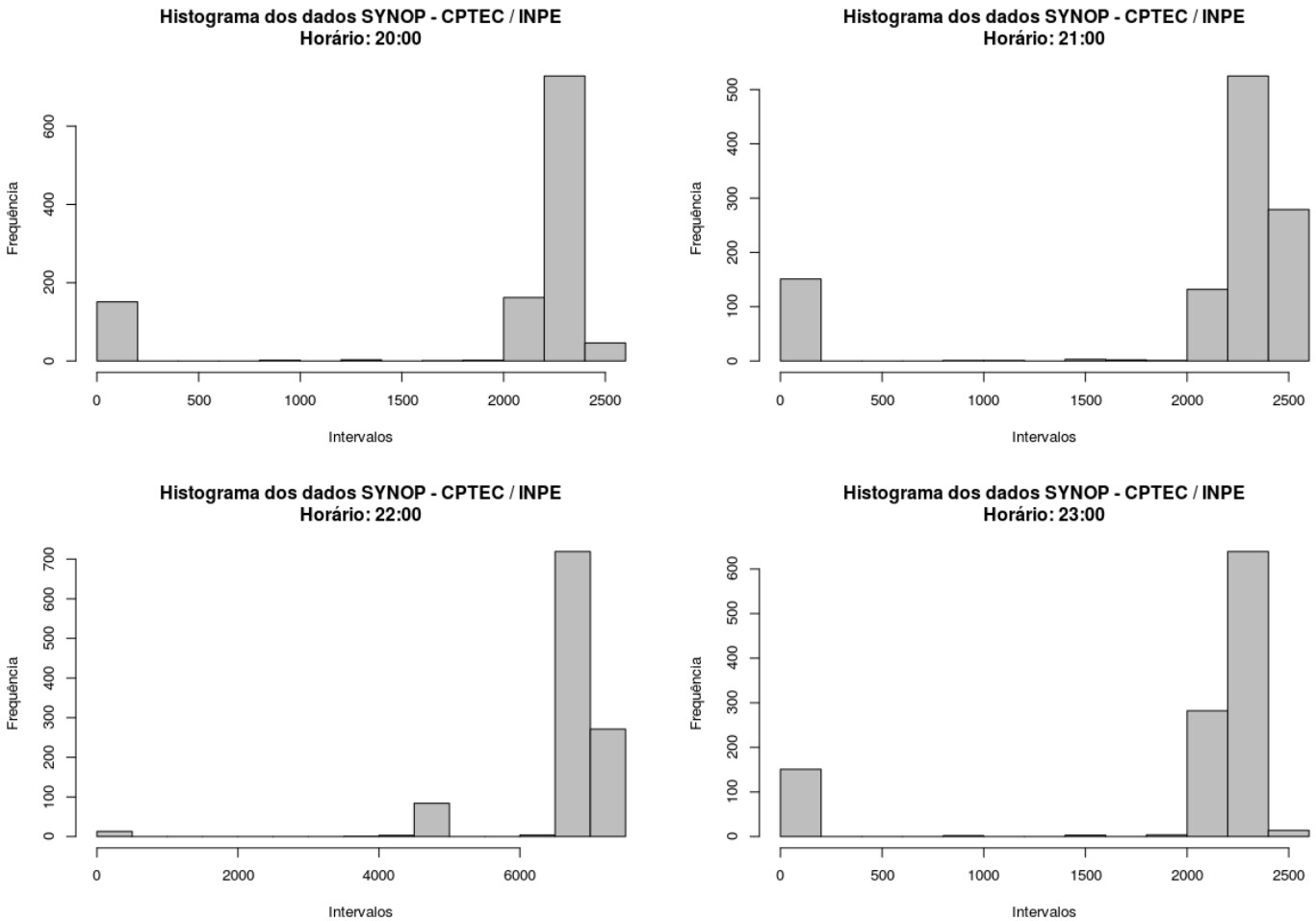

Figura 2. Histogramas dos dados SYNOP nos horários das 20:00, 21:00, 22:00 e 23:00

Com os valores da tabela de frequência e com a representação gráfica em forma de histograma de cada horário, foi realizado testes para apontar qual distribuição de probabilidade se ajusta melhor aos dados.

Para este projeto, foram testadas as distribuições de probabilidade de Cauchy, Gaussiana, Logística, Exponencial, Poisson, Gamma, Beta e Weibull. As distribuições apresentadas a seguir na Figura 3 são o comparativo dos ajustes dos dados em relação às distribuições que apresentaram melhores ajustes, que são a de Cauchy, Gaussiana (também conhecida na estatística como distribuição Normal), Logística e Exponencial.

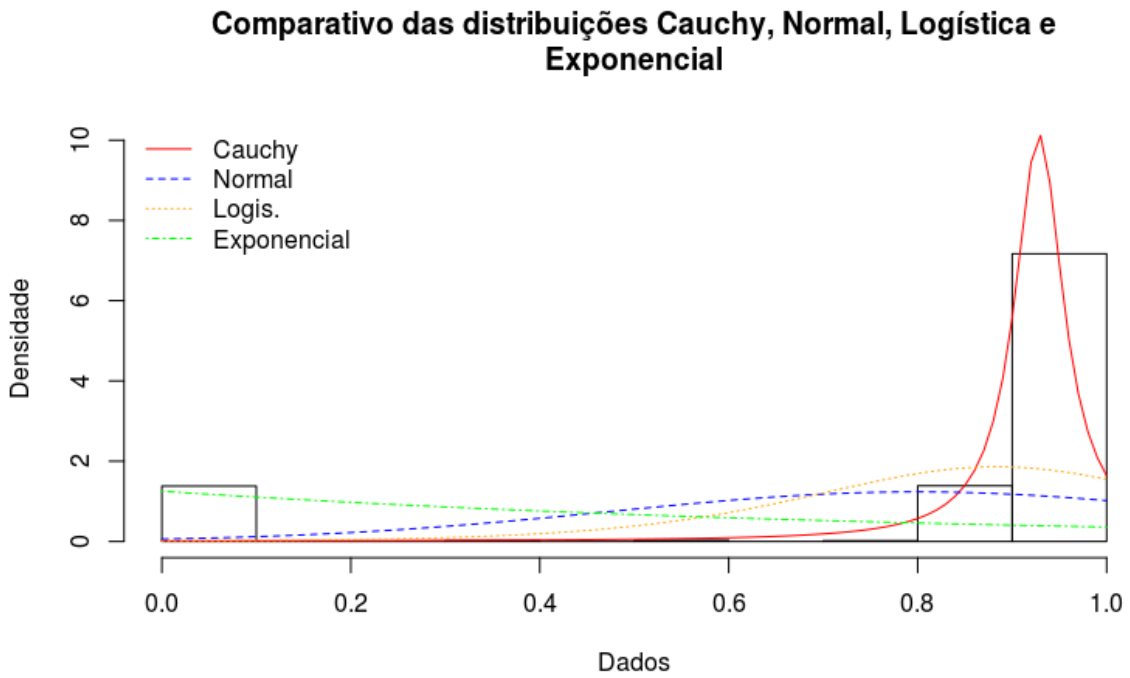

Figura 3. Comparação gráfica dos ajustes das distribuições de Cauchy, Normal, Logística e Exponencial

Através da comparação feita na Figura 3, pode-se observar que a distribuição de Cauchy, representada pela linha em vermelho, obteve um melhor ajuste em relação às outras distribuições, que são a Normal, apresentada por uma linha na cor azul, a Logística, apresentada por uma linha na cor laranja, e pela Exponencial, representada por uma linha na cor verde. Logo no início do eixo "x" e "y" do gráfico da Figura 3, destaca-se um volume visível de densidade de valor zero, isso se deve ao fato da presença de possíveis erros nos dados. Para verificar se a análise gráfica está correta, foi necessário utilizar testes estatísticos, que comprovam numericamente se a distribuição realmente se ajustou aos dados.

Na Tabela 2 será apresentada os testes elaborados, onde será exibido os valores dos testes de Kolmogorov-Smirnov, Cramer-von Mises e Anderson-Darling. Esses testes foram feitos através de um programa de computador na linguagem de programação R, onde os valores obtidos podem ser interpretados da maneira em que o menor valor encontrado dos testes com relação às distribuições em análise é considerado o melhor.

Tabela 2. Testes estatísticos elaborados nos dados de observação com relação aos ajustes de distribuição de probabilidade

Teste	Cauchy	Normal	Logística	Exponencial
Kolmogorov-Smirnov	0.13417	0.40035	0.29796	0.50179
Cramér-von Mises	5.56150	46.38206	31.13285	64.72519
Anderson-Darling	59.17689	244.77563	215.83641	519.13214

Analisando os valores, pode-se observar que a distribuição de Cauchy obteve os menores valores com relação às outras distribuições, ou seja, os valores analisados estão bem

próximos, podendo assim, ser definida como a melhor distribuição que se ajusta aos dados de observação.

7. Conclusões

Com os arquivos que continham os dados meteorológicos, foi elaborado a distribuição espacial utilizando a ferramenta Matplotlib da linguagem de programação Python, em que foi criado um mapa de cobertura com a localização das estações meteorológicas.

Nas análises dos dados do tipo SYNOP, foram identificados os erros, que podem ter sido originados por diversos fatores. Através de scripts na linguagem de programação Python, foi possível filtrar alguns desses erros dentro dos arquivos, criando uma nova tabela composta pela contagem das observações. Com a contagem das observações, foi possível produzir a distribuições de frequências com scripts na linguagem de programação R, e, com isso, foi elaborado uma série de testes com o intuito de identificar a distribuição de probabilidade que representa da melhor forma os dados, ou seja, o melhor ajuste.

Em trabalhos futuros iremos avaliar os outros tipos de dados, onde será analisada a melhor distribuição de probabilidade que representa cada tipo de dado. Após a conclusão dessas etapas futuras, será elaborado um ambiente computacional que irá integrar a visualização de mapa de cobertura (de superfície terrestre, superfície do mar e aeronaves), com as análises estatísticas, onde será possível observar o estado em que os dados se encontram ao usuário em tempo real, informando se há algum erro ou atraso na chegada.

8. Referências Bibliográficas

MATTES, E. **Python crash course: a hands-on, project-based introduction to programming.** 560 p.

MILLER, P. A.; BARTH, M. F. Ingest, integration, quality control and distribution of observations from state transportation departments using MADIS. In: 19th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology (IIPS), 2003, Long Beach, California.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. 8a. Ed. Editora Saraiva, 568 p., 2014.

GRIMMETT, G. R.; STIRZAKER, D. R. **Probability and Random Processes**. 2nd ed. Clarendon Press, Oxford, 583 p., 1992.

WORLD METEOROLOGICAL ORGANIZATION. **Guide to meteorological instruments and methods of observations**. WMO-No. 8, Geneva, Switzerland, 2008.

TOSI, S. **Matplotlib for developers: build a remarkable publication quality plots the easy way.** 305 p.

VANDERPLAS, J. **Python data science handbook: essential tools for working with data.** 1st. Ed. Editora O'Reilly Media, 2017. 530 p.

STEVENSON, M.; KAMPEL, M. **Utilização de bóias oceanográficas rastreadas por satélite no Brasil**. Rev. Bras. Geof. [online]. 1997, vol.15, n.1, pág. 59-64. ISSN 0102-261X. Disponível em:http://dx.doi.org/10.1590/S0102-261X1997000100006BLUESTEIN, H. B. **Synoptic-dynamic Meteorology in midlatitudes: observations and theory of weather systems.** Oxford University Press, 1993, vol. 2. 593 p.

WORLD METEOROLOGICAL ORGANIZATION. **Guide to marine meteorological services.** WMO-No. 471, Geneva, Switzerland, 2017.

WORLD METEOROLOGICAL ORGANIZATION. **Aerodrome reports and forecasts**. WMO-No. 782, Geneva, Switzerland, 2014.

WORLD METEOROLOGICAL ORGANIZATION. **Guide to aircraft-based observations**. WMO-No. 1200, Geneva, Switzerland, 2017.

DUBOIS, P. F. **Python: Batteries Included**. Rev. IEEE Comput. Sci. Eng. 2007, vol. 9, pág. 7-9.

CINTRA, R. S. C. **Preparação de dados de observações para o sistema de assimilação de dados PSAS do CPTEC**. INPE, São José dos Campos, 2005. INPE-14108-RPQ/255.

ROSENBLATT, B. **Learning the Korn Shell**. 1st. Ed. Editora O'Reilly Media, 1993. 336 p.

MATLOFF, N. **The Art of R Programming: A Tour of Statistical Software Design**. Ed. No Starch Press, 2011. 400 p.