DESENVOLVIMENTO DE UM FRAMEWORK PARA A ANÁLISE INICIAL DE DADOS PROVENIENTES DO SISTEMA NACIONAL DE DADOS AMBIENTAIS (SINDA)

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

Glenda Paola Barboza Lami (Fatec Cruzeiro - Prof. Waldomiro May, Bolsista PIBIC/CNPq)
E-mail:glenda.lami@inpe.br

Eugenio Sper de Almeida (CPTEC / SESSS)
E-mail: eugenio.almeida@inpe.br

Julho de 2017

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

GLENDA PAOLA BARBOZA LAMI

DESENVOLVIMENTO DE UM FRAMEWORK PARA A ANÁLISE INICIAL DE DADOS PROVENIENTES DO SISTEMA NACIONAL DE DADOS AMBIENTAIS (SINDA)

Relatório final de projeto de
iniciação científica
(PIBIC/INPE/CNPq)

Orientador: Prof. Dr. Eugênio Sper de Almeida

Cachoeira Paulista, 2019

Sumário

Lista de Figuras

Lista de Símbolos e Abreviaturas

INPE - Instituto Nacional de Pesquisa Espaciais

CPTEC - Centro de Previsão de Tempo e Estudos Climáticos

SINDA - Sistema Integrado de Dados Ambientais

PCD - Plataforma de Coleta de Dados

DCS - transportador de coleta de Dados

ANA - Agência Nacional de Água

INMET - Instituto Nacional de Meteorologia

CETESB - Companhia Ambiental do Estado de São Paulo

ID - Número de identificação

Resumo

Os dados ambientais ou observacionais possuem um grande importancia para as atividades de previsão do tempo, agricultura e de desastres naturais entre outras previsões, e esses dados são provenientes e coletas feitas por sensores de plataformas de coletas de dados. No Brasil existe um sistema integrado de coleta de dados observacionais chamada SINDA responsável por armazenar e transmitir os dados para as principais instituições de pesquisas. No entanto essa base de dados permite apenas a seleção de dados de um período de no máximo de um ano e a visualização dos dados é via arquivo EXCEL, para melhorar esse sistema Silva e Silva (2017) desenvolveram um framework que permite a visualização de todos os dados. Este trabalho propõe acrescentar ao framework desenvolvido por Silva e Silva (2017) estatística descritiva e aplicação de bibliotecas gráficas para visualização dos dados.

Introdução

Os dados ambientais possuem importância fundamental em atividades de previsão meteorológica, monitoramento de recursos hídricos e de qualidade das águas, agricultura de precisão, prevenção de desastres naturais e diversas outras atividades(AEB, 2018). Os resultados das medições feitas pelos sensores dos sistemas de observação revelam características sobre o ambiente (Gunther, 1997).

O SINDA é parte integrante do Sistema Brasileiro de Coleta de Dados Ambientais. possui a finalidade de processar, armazenar e disseminar os dados das PCD's ambientais. (Santos et al., 2013)

O acesso aos dados da PCD (http://sinda.crn.inpe.br) é realizado através do número da estação/estado/ nome da estação. Em seguida o usuário fornece o intervalo de datas que deseja selecionar, a partir de informações disponibilizadas pelo SINDA. O sistema permite apenas a seleção de dados de um período de no máximo de um ano, fornecendo os dados solicitados em formato EXCELL.

Silva e Silva (2017) desenvolveram um *framework* para acesso e visualização das informações coletadas pelas PCD's, sem limitação de período de tempo. Apesar de não analisarem os dados, os gráficos gerados pelo *framework* permitiram verificar a existência de informações faltantes na base de dados do SINDA.

A obtenção da estatística descritiva das variáveis ambientais armazenadas no SINDA, antes do início da análise dos dados, pode ser bastante útil e permite uma visão global do conjunto de dados das PCD's. Outra forma de verificar a disponibilidade dos dados seria através da utilização de ferramentas gráficas (*histogram, box plot, etc.*).

Segundo Silvestre (2007), estatística descritiva é o ramo da estatística que visa sumarizar e descrever qualquer conjunto de dados, como medidas de posição e medidas de variabilidade ou dispersão. Medidas de posição incluem média, mediana e moda. Medidas de variabilidades incluem desvio padrão, variância, o valor de máximo e mínimo, obliquidade e curtose (Bussab, Morettin, 2017)

Desta forma, o objetivo deste trabalho é estender o framework de Silva e Silva (20170 incluindo informações de estatística descritiva e gráficos estatísticos referentes às diferentes variáveis ambientais coletadas pelas PCD's e armazenadas no SINDA.

A partir da informação fornecida pelo framework um pesquisador pode determinar *a priori* quais dados ambientais são mais adequados à sua pesquisa e os responsáveis pelas PCD's poderiam monitorar o funcionamento dos sensores e componentes das PCD's.

1.Fundamentação Teórica

Nos dias atuais a dependência de informação tem aumentado cada vez mais, portanto a estatística se tornou uma importante ferramenta para a tomada de decisão. (SANTOS, 2007)

O objetivo principal da estatística descritiva é sintetizar uma série de valores de mesma natureza, permitindo uma visão global da variação desses valores, além de organizar e descrever os dados por meio de tabelas, de gráficos e de medidas descritivas. (GUEDES, 2015)

Na estatística existem muitos métodos e técnicas que possibilitam a análise de dados e servem para recolher, organizar, sintetizar e descrever os dados, além dos que são a base na Teoria das Probabilidades, permitindo a análise e a interpretação dos dados, assim como efectuar inferências sobre uma população com base no estudo de uma amostra. (SANTOS, 2007)

Com os métodos estatísticos não é possível adivinhar com precisão acontecimentos futuros, no entanto com eles é possível medir as variações ao longo do tempo e fazer previsões prováveis. (REIS, 1999)

Uma importante ferramenta da estatística é o gráfico, que é recursos visual utilizado nos meios de comunicação social, técnica e científica, devido à facilidade de interpretação e a eficiência com que resume informações dos mesmos. (GUEDES, 2015)

2. Materiais e Métodos Utilizados

Nesta etapa da pesquisa foram realizadas as seguintes etapas:

1. Revisão bibliográfica sobre o INPE e Dados ambientais;
2. Estudo de estatística descritiva;
3. Familiarização com o SINDA;
4. Análise e avaliação das bibliotecas estatísticas e métodos gráficos para análise de dados;
5. Atualização do código de Silva e Silva (2017) para python 3 para ser possível a utilização das bibliotecas PANDAS, MATPLOTLIB e NUMPY;
6. Aplicação das bibliotecas estatísticos e de métodos gráficos;
7. Aprendizado sobre ferramentas como Jupyter e Docker;

2.1. Ferramentas utilizadas

Esta pesquisa está sendo desenvolvida em linguagem Python, juntamente com as bibliotecas Pandas, Numpy E Matplotlib. O ambiente escolhido para a codificação foi a ferramenta Jupyter Notebook dentro de um container Docker.

2.1.1. Python

O Python é uma linguagem de programação que está sendo muito utilizada em aplicações científicas tradicionalmente dominadas por MATLAB, Stata, SAS, assim como em pesquisa comerciais ou de código aberto, devido a qualidade da documentação e a maturidade de bibliotecas como NumPy, SciPy e outras. (SANNER, 1999)

Embora a modelagem estatística utilizando python tem sido relativamente lenta se comparada com outras áreas da ciência computacional, a integração da biblioteca matplotlib com as ferramentas de codificação notebook fornece um ambiente de pesquisa interativo com visualização dos dados adequada a grande parte dos usuários. (SANNER, 1999)

Devido ao crescimento e qualidade das bibliotecas de análise de dados e da própria linguagem python que esta linguagem foi escolhida para o desenvolvimento deste trabalho.

2.1.2. Pandas

Pandas é uma biblioteca de código aberto e licença BSD que fornece ferramentas de análise de dados para a linguagem de programação Python, muito utilizada em trabalhos acadêmicos, comerciais incluindo finanças, economia, estatística, análise, etc. Possui funções que fornecem estruturas de dados de alto desempenho e relativamente fáceis de usar devido a documentação bem desenvolvida.

2.1.3. Matplotlib

O Matplotlib é uma biblioteca de plotagem 2D para a linguagem python em vários formatos de impressos e pode ser usado em scripts Python, shells Python e IPython, notebooks Jupyter, entre outros, permitindo ao usuário gerar gráficos histogramas, espectros de potência, gráficos de barras, gráficos de erros, diagramas de dispersão, etc., com apenas algumas linhas de código. A relativa rapidez de leitura e plotagem dos dados fornecida por essa biblioteca possibilitou a visualização e facilitou a análise dos dados do SINDA.

2.1.4. Numpy

O NumPy é uma importante biblioteca de computação científica para a linguagem Python, fornece um objeto de matriz multidimensional, vários objetos derivados e uma variedade de rotinas para operações rápidas em matrizes, incluindo manipulação matemática, lógica, de formas, classificação, seleção, I / O , transformada discreta de Fourier, álgebra linear básica, operações estatísticas básicas, simulação aleatória e muito mais.

2.1.4. Jupyter

O Jupyter notebook une todo o processo de computação: desenvolvimento, documentação e execução de código, bem como a comunicação dos resultados, combinando um aplicativo da Web baseado em um navegador para criação interativa de documentos unindo texto explicativo, matemática, cálculos e sua saída e a documentos do Notebook que é

uma representação de todo o conteúdo visível no aplicativo da Web, incluindo entradas e saídas das representações de computação, texto explicativo, matemática, imagens e rich media de objetos.

2.1.5. Docker

Docker é uma plataforma Open Source escrito em Go, que é uma linguagem de programação de alto desempenho desenvolvida pela Google, que facilita a criação e administração de ambientes isolados e é um sistema de virtualização com recursos isolados que utilizando bibliotecas de kernel em comum (entre host e container), possível pois o Docker utiliza como backend o LXC.

A escolha de utilizar o Docker no desenvolvimento deste projeto é devido a característica do docker de possibilitar o empacotamento de uma aplicação ou ambiente inteiro dentro de um container, tornado-o portável para qualquer outro Host que contenha o Docker instalado, reduzindo o tempo de deploy de alguma infraestrutura ou até mesmo da aplicação, por não haver a necessidade de ajustes de ambiente para o correto funcionamento da aplicação.

2.2. Base de dados

As bases de dados meteorológicos que foram utilizadas no projeto são as do INPE, onde o CPTEC é responsável por receber as informações de todas as PCDs do INPE e estão armazenadas no SINDA.

Os dados armazenados no SINDA são provenientes da coleta pelos sensores de mil e cinco PCD's e a Tabela 1 exibe quais são as variáveis de estudo do SINDA.

Tabela 1 - Variáveis de estudo do SINDA

Parâmetro	Sigla	Unidade	Descrição
Temp. do Ar	TempAr	°C	valor instantâneo a cada 3 H
Temp. Máx. do Ar últ. 24 H	TempMax	°C	valor a cada 3 H com a Máx. das últ. 24 H, amostragem a cada 1 minuto.
Temp. Mín. do Ar últ. 24 H	TempMin	°C	valor a cada 3 H com a Mín. das últ. 24 H, amostragem a cada 1 minuto.
Umidade Relativa do Ar	UmidRel	%	valor instantâneo a cada 3 H
Pressão Barométrica	PressaoAtm	mB	valor instantâneo a cada 3 H
Veloc. do Vento	VelVento	m/s	valor a cada 3 H, calculado da média de 200 amostras com 3 seg de intervalo, 10 min antes de cada 3 H
Dir. do Vento	DirVento	° NV	valor a cada 3 H, calculado da média de 200 amostras com 3 seg de intervalo, 10 min antes de cada 3 H
Veloc. Máx. do Vento (Rajada)	VelVentoMax	m/s	valor máximo(rajada) cada 3 H, amostras cada 3 seg
Dir. do Vento na Veloc. Máx.	DirVelVentoMax	° NV	valor(Dir. da rajada) cada 3 H, amostras cada 3 seg
Radiação Solar Global	RadSolAcum	MJ/m²	valor acumulado a cada 3 H, integração de 1080 amostras de 10s de intervalo
Radiação Solar Líquida	RadSolLiq	W/m²	valor instantâneo a cada 3 H
Precipitação Acumulada	Pluvio	mm	valor acumulado mensal a cada 3 H (zera o acum. automat. todo dia 01 de cada mês)
Temp. do Solo 100mm, 200mm, 400mm (*)	TempSolo100 TempSolo200 TempSolo400	°C	valor instantâneo a cada 12 H amostrado às 06 e 18 Hs GMT
Conteúdo Água no Solo 100mm, 200mm, 400mm (*)	ContAguaSolo100 ContAguaSolo200 ContAguaSolo400	(m³/m³)	valor instantâneo a cada 12 H amostrado às 06 e 18 Hs GMT
Fluxo de Calor no Solo	FluxCalSolo	W/m²	valor instantâneo a cada 3 H
Bateria	Bateria	Volts	valor instantâneo a cada 3 H
Corrente Painel Solar	CorrPSol	Lógico	valor instantâneo a cada 3 H
Umidade Interna	UmidInt	%	valor instantâneo a cada 3 H

Nota: O horário de Coleta das PCDs é sincronizado com a Hora Universal **GMT** (**G**reenwich **M**ean **T**ime) = Hora de Brasília + 3 horas (horário normal) ou Hora de Brasília + 2 horas (horário de verão).

Fonte - Centro de Pesquisa Tecnológica e Estudos Climáticos (CPTEC)

2.3. O projeto

Este projeto de pesquisa é uma continuação do projeto desenvolvido por (Silva e Silva, 2017) acrescenta métodos estatísticos para uma visão macro dos dados e métodos gráficos para visualização e análise de dados provenientes do SINDA.

Nesta etapa do projeto foram incorporados ao framework desenvolvido por (Silva e Silva, 2017) os métodos de medidas de posição como média, mediana e moda e de variabilidades como desvio padrão, variância, o valor de máximo e mínimo.

Para a visualização dos dados foram utilizados os gráficos de barra e linha para todas as PCD's e suas variáveis de estudo.

3. Análises e Resultados

Neste projeto foram analisadas e gerados gráficos das mil e cinco PCD's existentes na base de dados do SINDA, no entanto para representar as análises realizadas foi escolhido uma das PCD's, a de Cachoeira Paulista cuja número de identificação é 31000. A Tabela 2 traz os valores das medidas de posição e de variabilidade.

Tabela 2 - Medidas de posição e variabilidade dos dados do SINDA

	Valor Mín.	Valor Máx.	Média	Mediana	Quartis 50%	Moda	Amplitude	Variância	Desvio Padrão	Desvio Absoluto
Bateria	11.50	13.20	12.50	12.30	12.30	12.20	1.70	0.18	0.42	0.39
CorrPSol	0.00	1.00	0.44	0.00	0.00	0.00	1.00	0.25	0.50	0.49
DirVelVento Max	0.00	430.00	133.32	100.00	100.00	20.00	430.0	10501.62	102.48	88.29
DirVento	0.00	360.00	139.08	110.00	110.0	20.0	360.00	9820.41	99.10	84.91
Pluvio	0.00	92.50	29.44	23.00	23.00	11.75	92.50	742.49	27.25	23.51
PressaoAtm	845.00	976.00	958.90	959.00	959.00	958.00	131.00	77.17	8.78	4.65
RadSolAcum	0.00	8.80	2.23	0.60	0.60	0.00	8.80	8.82	2.97	2.52
TempAr	0.00	31.50	20.24	20.50	20.50	21.00	31.50	27.97	5.29	4.22
TempMax	1.00	35.00	27.67	28.00	28.00	28.00	34.00	12.37	3.52	2.67
TempMin	4.50	27.00	15.41	16.00	16.00	18.00	22.50	21.64	4.65	3.89
UmidInt	0.00	35.00	11.07	10.00	10.00	10.00	35.00	62.68	7.92	5.94
UmidRel	0.00	100.00	75.84	80.00	80.00	94.00	100.00	306.82	17.52	14.75
VelVento10m	0.00	12.70	3.36	2.40	2.40	12.70	12.70	8.36	2.89	2.00
VelVentoMax	1.60	51.10	11.28	8.00	8.00	51.10	49.50	121.81	11.04	6.90

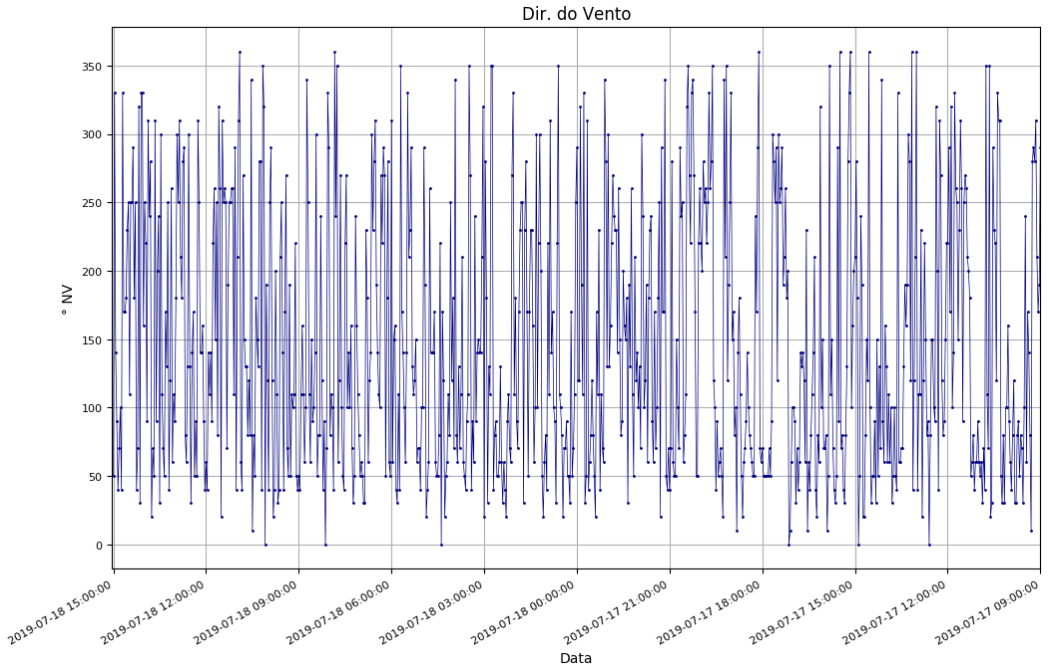
Fonte - O autor

Com os valores das medidas de posição obtidos e expressos na Tabela 2 pode-se verificar que os dados se encontram-se bastante dispersos, pois os valores de amplitude estão muito altos na maioria das variáveis, e também é possível dizer que os dados possuem alta variabilidade, o que que pode ser comprovado pelos valores obtidos na variância e desvio padrão.

O motivo da discrepância dos dados só poderia ser encontrado mediante análises mais profundas o que não é o foco deste trabalho e sim o apontamento de tais ocorrências.

A figura 1 mostra o gráfico da direção do vento e pode-se notar que a direção varia de 0° a aproximadamente 350°.

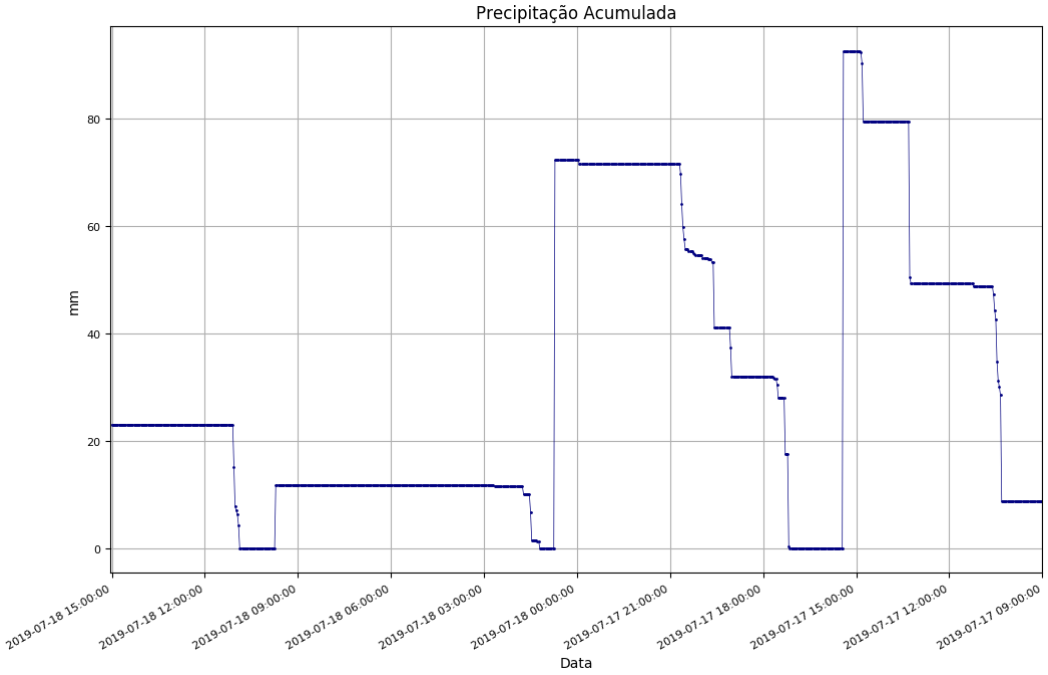
Figura 1- Gráfico de Direção do vento



Fonte - O autor

A Figura 2 mostra o gráfico de precipitação acumulada e ao observarmos o gráfico podemos ver que a quantidade de chuvas diminuiu drasticamente de 2017 para 2018 no mês de julho.

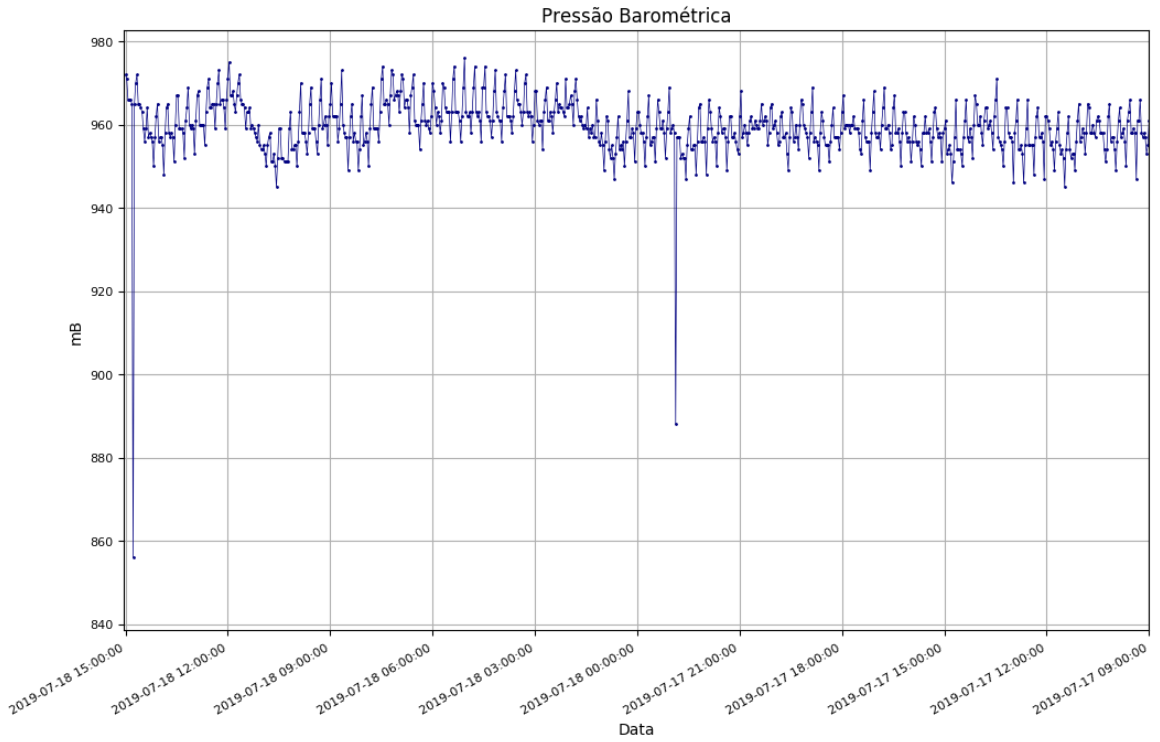
Figura 2 - Gráfico de Precipitação Acumulada



Fonte - O autor

A figura 3 que traz o gráfico da pressão barométrica demonstra que a pressão se manteve próxima de 960 mB durante todos os anos de coleta dos dados, com exceção de uma medição que se encontra com valor bem abaixo, por volta de 890 mB, o que pode indicar uma leitura incorreta ou uma anomalia.

Figura 3 - Gráfico de Pressão Barométrica



Fonte - O autor

4. Conclusões

Com o desenvolvimento desta pesquisa, incorporando bibliotecas estatísticas e de métodos gráficos ao framework de Silva e Silva (2017) acreditasse que tenha favorecido a análise de dados ambientais somando ao banco de dados do Sinda uma ferramenta de análise e visualização dos dados.

O acréscimo de métodos estatísticos possibilitou o apontamento de possíveis falhas de medição e pode auxiliar os pesquisadores na escolha de quais variáveis seriam mais interessantes às suas pesquisas.

Os métodos gráficos somados ao feito de Silva e Silva (2017) de possibilitar a visualização de todos os dados armazenados pode auxiliar o processo de previsão.

Referências Bibliográficas

AEB Dados ambientais. Disponível em http://www.aeb.gov.br/servicos/dados-ambientais/, acessado em junho/2018.

Bussab, W.O.; Morettin, P.A. Estatística Básica, v., 2017.

CPTEC. Informações. Disponível em: <http://www.cptec.inpe.br/sobreocptec/pt> Acesso em: 12 de Fevereiro de 2019.

Docker. Informações. Disponível em: https://www.mundodocker.com.br/o-que-e-docker/> Acesso em 3 de Abril de 2019.
GUEDES, Terezinha Aparecida et al. Estatística descritiva. Projeto de ensino aprender fazendo estatística, p. 1-49, 2005.

Gunther, O. Environmental information system. ACM SIGMOD Record 26(1) 3-4, 1997.

Jupyter. Informações. Disponível em: https://jupyter-notebook.readthedocs.io/en/stable/notebook.html> Acesso em 20 de junho de 2019.

MATPLOTLIB. Informações. Disponível em: https://matplotlib.org/> Acesso em: 5 de Março de 2019.

NUMPY. Informações. Disponível em: https://docs.scipy.org/doc/numpy/user/whatisnumpy.html> Acesso em 5 de Março de 2019.

PANDAS. Informações. Disponível em: https://www.tutorialspoint.com/python_pandas/> Acesso em 5 de Março de 2019.

PYTHON. Informações. Disponível em: https://www.python.org/about/> Acesso em: 8 de Fevereiro de 2019.

REIS, Elizabeth et al. Estatística aplicada. Lisboa: Edições Sílabo, 1999.

Sanner, Michel F. et al. Python: a programming language for software integration and development. J Mol Graph Model, v. 17, n. 1, p. 57-61, 1999.

SANTOS, Carla. Estatística descritiva. Manual de auto-aprendizagem, v. 2, 2007.

Santos, M. A. F.; Francisco, M. F.M.; Yamaguti, W. O Sistema Nacional de Dados Ambientais e a Coleta de Dados por Satélite. Anais XVI Simpósio Brasileiro de Sensoriamento Remoto, 2013.

Silva, J.V.O.; Silva, W.C.D. Proposta de Sistema de Apoio a Análise de Dados Meteorológicos - Fatec Cruzeiro, 2010

Silvestre, A.L. Análise de dados e estatística descritiva. Escolar editora, 2007.

SINDA. Informações. Disponível em: http://sinda.crn.inpe.br/PCD/SITE/novo/site/sobre.php> Acesso em: 15 Fevereiro de 2019.

TOSI, Sandro. Matplotlib for Python Developers. Reino Unido: Packt Publishing, 2009.