ClimateWatcher

Environmental Studies Capstone Paper

Debargha Ganguly
debargha.ganguly_ug20@ashoka.edu.in

Abstract : *In this paper we explore computational prediction and visualization methods for environmental damage that can be independently verified to legitimize research and increase public environmental concern. We explore a previously mined dataset to derive analytics, and attempt to use an LSTM-RNN to predict trends in pollution and its factors from around India. Climate modelling techniques to best understand and predict extreme weather patterns are also discussed in this paper.*

Introduction :

Understanding the problem:

Most of the research used to quantify climate change, and make predictions is not replicable outside the lab without the usage of supercomputers. The methods used in them are also not available in the form of usable APIs or code libraries to complement other research.

An implication of this is that the since these results and practices aren't easily available to people who would be interested in this, the awareness for climate change is much lower.

The aim for this open source project would be to lower the barrier to entry for understanding how diverse the changes on the environment could be by making it experimentally replicable in some ways.

The lowering of the barrier to entry needs to be on three fronts:

1.) Development of libraries so that scientific advances in the field can be used by more general developers.
2.) Getting more data in the form of pollution data.
3.) A knowledge based research to be able to break down such research into higher level understandings and implementations that can be made on top of all this pollution data.

Our initial hypothesis stands that data science, specifically machine learning can be used for climate change modelling and environmental damage protection.

Overall, we will be using the general approach for such problems is through data acquisition, then modelling followed by observing trends and analyzing the data.

Methods and Improvements :

Data mining: The following methods were used for mining the dataset from Central Pollution Control Boards' web servers :

1.) Curl Script - PHP :
   - *Advantages for this method :* This allows quick mining of the data embedded in the webpage.
   - *Observed problems with this method:* The entire webpage is written in Javascript which requests the data from the database. The database is however not directly linked hence making it impossible to directly access the server contents without having access credentials.

2.) Selenium on Python :

   This method allows the browser to be controlled by a certain preprogrammed driver. This allows the web pages that need to be accessed to be controlled in certain ways by clicking on elements in a pre programmed setting. The data that is being displayed on the page can be dumped into a CSV file, thereby establishing a dataset from the displayed content. This process when repeated, allows repeatedly requesting the data just like humans would be when accessing the webpage. The program, couldn't however work on the webpage because the code the webpage was scripted in was highly unprofessional and unclean. The elements on the webpage didn't have unique ids therefore making it completely impossible for the program to differentiate between different sets of data. Hence the program kept running iteratively just mining the data from the first page.

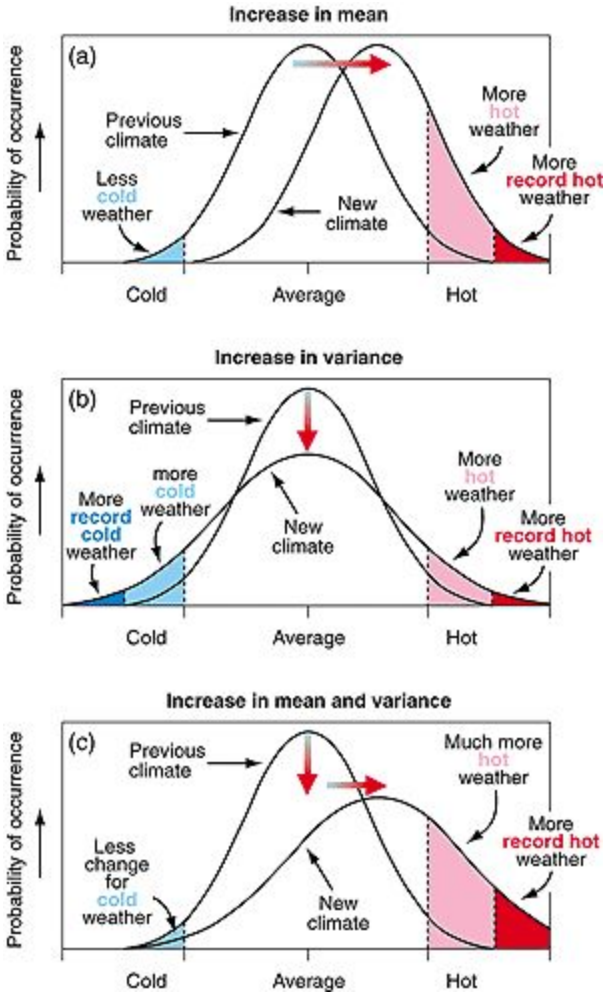
Finally however a pre-mined dataset from a third-party source that had been made open access in the past was found. That dataset has been used for the exploratory analysis.

Usage of Machine learning in Climate Change Analysis:

Although there is well established consensus in the scientific world about the existence of climate change, fundamental uncertainties about the scale at which the extreme weather patterns would be affected remain. The following are the possibilities about how extreme weather patterns will be affected by climate change :

Considering a gaussian normal distribution to represent the probability of weather events of all kinds happening from extreme cold weather to extremely hot weather.

1.) A shifted mean - This would mean more extreme hot weather, more hot weather, less extreme cold weather and less cold weather in general.
2.) Increased variability - This would mean that there is more extreme hot weather, more hot weather, more extreme cold weather and more cold weather.
3.) A changed probability shifting the symmetry of the curve towards hot weather means that more extreme hot weather, more hot weather and near constant cold and extreme cold weather.



Here the weather predictions on the basis of the curves mean that there is a lot of uncertainty in a given prediction when considering extreme weather patterns in regional spaces.

The World Climate Research Programme's grand challenges program in understanding extreme weather practices quotes the following, *" Weather and climate extremes are an inherent part of climate. There is overwhelming evidence that the climate and its extremes are changing. As extremes affect every aspect of our society, ….."*.

Such extreme climate events are very rare according to the probability distribution defined above, however since we established that climate change will be changing the curve, predicting such events are extremely challenging.

Understanding Climate Models :

Climate prediction models usually consider the land, oceans and the atmosphere to be different entities that interact amongst each other through forcings. Forcings in a climate change model is a certain factor that affects all of these parameters in a certain way. Natural climate change forcings include volcanic eruptions, and on the other hand, other climate change forcings include greenhouse gas addition to the atmosphere.

Climate Modelling in Research:

A climate model is built on a complex system of the interacting smaller mathematical models . This is not a data - driven statistical process, but instead based on the scientific principles grounded in Meteorology, Oceanography and Geophysics.

Using Non-linear dynamical systems for climate prediction:

The generic view towards approaching climate prediction through timesteps until the recent switch to using statistical distributions were usually governed by this generic model :

$$x_{t+1} = \boldsymbol{G}(x_t, F_t)$$

Where,

x_t - the current state of the system and can be a vector representation of the current weather.

F_t - are the external forcings on the system

The function **G** is usually grounded in physics to emulate complex behaviour however is made sure to be deterministic.

The goal towards establishing an open source project in this specific field:

Climate change adaptation and mitigation are problems that are extremely important to society, potentially very adversely affecting the entire population of the earth. It also happens to be a data rich field to work in, especially with a lot of public datasets. With machine learning, there are a lot of low hanging fruit that can be exploited. On the other hand, climate scientists work on the forefront of the High performance Computing (HPC) and usually write huge code-bases. This allows for fruitful partnerships and collaborations in the space. Using laws grounded in physics also provide predictability and repeatability to the results.

Predicting pollution levels through tIme series forecasting:

Method reasoning:

Long Short Term Memory (LSTM) recurrent neural networks can be used to model problems with a lot of input variables. Classical methods like using regression perform very poorly when it is a multivariate or multi-input problem.

The proposed network architecture is composed of two kinds of layers: LSTM layer and fully connected dense layer. LSTM layer is utilized to model the time series relationship.
Fully-connected layer is utilized to map the output of LSTM layer to a final prediction.

Results:

Trends in data observed (general):

Past statistics are not sufficient to predict future climate trends, because the data is not nearly high dimensional enough or dense enough to extract understandable trends out of it. These statistics can be amalgamated with other statistics from climate models, to create high dimensional datasets with high density and volume.

Datasets that are recorded in the past show that they're highly heterogeneous, and in limited quantities. With the big data explosion right now, we are accumulating huge amounts of data. However some of this data can be sparse, unlabeled or at a higher resolution than their other parameters (therefore this means that the data can't be utilised with the best of it's potential).

Climate model simulations from the past, present and future can be used to make huge, high-dimensional datasets, The drawbacks to this are that while models are made on the basis of discrete processes, some of the data is lost. Also, on the other hand the future values which are being used for the machine learning models haven't been validated yet.

Exploring the Dataset :

The Indian pollution data from the 1990s when presented after some data-cleaning, looks like the following :

Out[5]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	1990-02-01
1	151	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01
5	152	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.4	25.7	NaN	NaN	NaN	NaN	1990-03-01

In [6]:
```
df.rename(columns={'stn_code': 'station_code', 'location': 'city', 'location_monitoring_station': 'monitoring_statio
df.head()
```

Here were the parameters and the number of data points :

Residential, Rural and other Areas 179014
Industrial Area 96091
Residential and others 86791
Industrial Areas 51747
Sensitive Area 8980
Sensitive Areas 5536
RIRUO 1304
Sensitive 495
Industrial 233
Residential 158
Name: type, dtype: int64

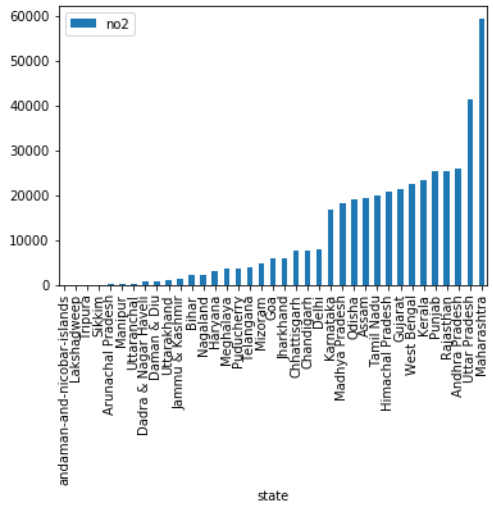
After pre-processing data and clubbing different classes through assumptions :

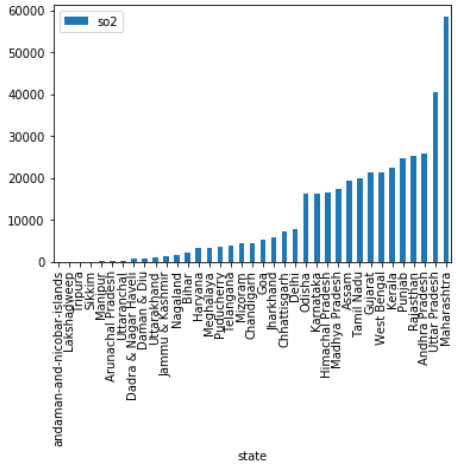
Residential, Rural and other Areas	*179014*
Industrial	*148071*
Residential and others	*86791*
Sensitive	*15011*
RIRUO	*1304*
Residential	*158*

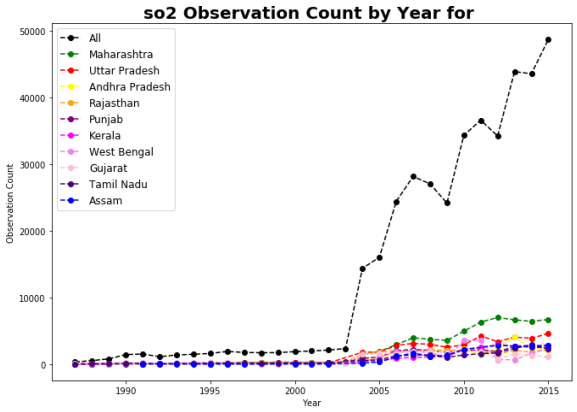
Name: type, dtype: int64

The graphs obtained when the dataset was analysed for every state :

Sulphur Dioxide and Nitrogen Dioxide data was found to be much less sparse than other parameters and hence are used as a quantifier of the broader term of pollution here.

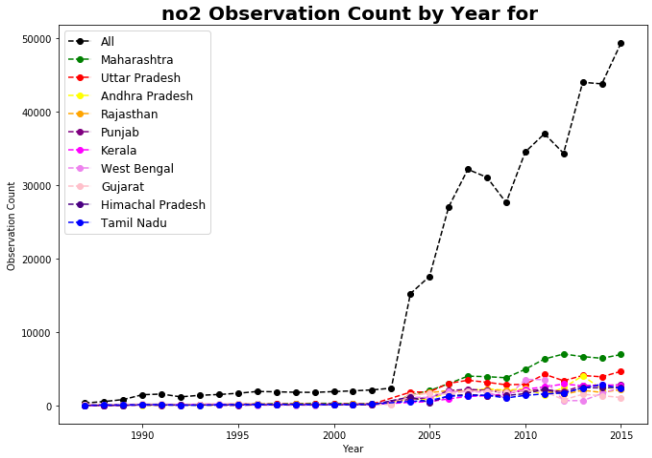






The broad trend of the increase in pollution levels with time is very clearly visible, with a high amount of variance thereby making predicting future trends through different forms of regression increasingly difficult. RMSE based methods for evaluating the accuracy of such predictions are also questionable for such problems as this is a multi-dimensional time series prediction problem. (This would mean that the pollution level depend on a lot more factors than just the time)

The surprising reality pointed by these graphs are how small increments at the state level count up to make much bigger impacts at the national level. Understanding how these trends behave at the global level would require other datasets to be fused into the same, hence wasn't tried.



LSTM based failure to predict :
Since the size of the data was so huge (about a total of 5.6 mn points of data) the problem of predicting a general trend grew to be an extremely high dimensional problem which needed all the parameters of all the cities being needed to analysed together.
The data being considered was from a total of 304 cities/ towns from 36 states/ union territories cumulated to 435741 rows being processed all at once for prediction.
The end result for this was the algorithm being used became too memory intensive for machines with even a 12 gigabyte RAM on board, force stopping the program from running midway, although it starts executing with a memory error. A way to be able to solve this would be to computationally optimise the code to use less system resources, and break the problem into smaller sections thereby computing parts, such as state by state, or city by city individually, and not all together.

Discussion and Other Innovation:

On Comparison with the state of the art techniques:

The Information Lab at Univ of Cal, Berkeley produced state of the art results in the field of using LSTM to model time period based changes caused by pollution. Their results have somehow been much better because they had access to better data. Their dataset was much more dimensionally rich in the form that it allowed the model to learn complex behavioral patterns followed by the variables about air quality in Beijing.

The dataset used from India is not as highly dimensional as the Beijing dataset and in turn is much more sparse also than the Beijing dataset used by the team at Berkeley.

Prediction of ocean surface temperatures using LSTMs has also been tried, and have produced particularly amazing results in predicting short term trends in ocean surface temperature. The architecture for the model used in this paper is inspired in many ways by these researchers.

At a future point in time once the dataset is received, this program will be recreated. A request has already been written to the concerned governmental authorities asking for access to specific datasets, however I've not received a response yet because turnaround time is high.

References:

- IPCC (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change AdaptaHon. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plakner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.
- IPCC (2013). Climate Change 2013: The Physical Science Basis. ContribuHon of Working Group I to the Fith Assessment Report of the Intergovernmental Panel on Climate Change. [Stocker, T.F., D. Qin, G.-K. Plakner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.
- E. Lorenz (1996). Predictability – A problem partly solved. Seminar on Predictability, Vol. I, ECMWF.
- C. Monteleoni, G.A. Schmidt, F. Alexander, A. Niculescu-Mizil, K. Steinhaeuser, M. Tippek, A. Banerjee, M.B. Blumenthal, A.R. Ganguly, J.E. Smerdon, and M. Tedesco (2013). Climate InformaHcs. ComputaHonal Intelligent Data Analysis for Sustainable Development; Data Mining and Knowledge Discovery Series. Yu, T., Chawla, N., and Simoff, S. (Eds.), CRC Press, Taylor & Francis Group. Chapter 4, pp. 81–126, 2013.
- IPCC Fith Assessment Report: www.ipcc.ch/report/ar5/
- World Climate Research Program Grand Challenges: www.wcrp-climate.org/grand-challenges
- Claire Monteleoni. Climate Informatics: Recent Advances and Challenge Problems for Machine Learning in Climate Science.
- Qin Zhang, Hui Wang, Junyu Dong, Member, IEEE Guoqiang Zhong, Member, IEEE and Xin Sun Member, IEEE. Prediction of Sea Surface Temperature using Long Short-Term Memory https://ieeexplore.ieee.org/document/8008749/
- Air Pollution in China: Mapping of Concentrations and Sources http://berkeleyearth.org/wp-content/uploads/2015/08/China-Air-Quality-Paper-July-2015.pdf
- S. Hochreiter and J. Schmidhuber, “Long short-term memory.” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas,A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay,“Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.