Responsible Consumption of Water

Akshay Verma, Anupam Samanta, Deepak Gupta, Srikant Panda
CSE 545 Fall '17 Final Project Report
Computer Science Department, Stony Brook University

1. ABSTRACT

Less than 3 per cent of the worlds water is fresh water, of which 2.5 per cent is frozen in the polar caps. Humanity must therefore rely on 0.5 per cent for all of mans ecosystems and fresh water needs. Man is polluting water faster than nature can recycle and purify water in rivers and lakes. More than 1 billion people still do not have access to fresh water. Excessive use of water contributes to the global water stress. Water is freely available in nature but the infrastructure needed to deliver it is expensive.

With two thirds of the earth's surface covered by water and less than one percent available for drinking, it is evidently clear that water is one of the most important elements responsible for life on earth. Water circulates through the land just as it does through the human body, transporting, dissolving, replenishing nutrients and organic matter, while carrying away waste material. And recent trends show a uneven distribution of water in various levels of the geological surface.

2. INTRODUCTION

Responsible Consumption of water aims at utilizing the big data techniques to efficiently predict the water related problems in the near future. There are two approaches to build a model which can identify patterns from water level data and can predict the future values.
1. The first approach is physically based i.e. collecting all the data which is correlated to water depth level that is precipitation, wind, temperature, pressure etc. The second method is a data driven approach.
2. The goal of data-driven approach (also called machine learning model) is to find the relationship between the input attributes and the water level.
3. Data-driven model is quickly developed and easily implemented for building the forecasting model, it is useful for real-time and accurate water level forecasts. Our approach is based on data-driven model using machine learning method to forecast the groundwater level for each state.

3. SUSTAINABLE DEVELOPMENT GOAL

The Sustainable Development Goals (SDGs), otherwise known as the Global Goals, are a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. These are 17 SDG goals including Responsible Consumption of Water.

Our project aims at the following targets of the sustainable development goal
12.2: By 2030, achieve the sustainable management and efficient use of natural resources.
12.4: By 2020, achieve environmentally sound management of chemicals and all wastes and significantly reduce their release to water.

4. DATASET

4.1 Data set from United States Geological Survey (USGS)

The USGS provide access to water-resources data collected at approximately 1.9 million sites in all 50 States of USA. They investigates the occurrence, quantity, quality, distribution, and movement of surface and underground waters and disseminates the data to the public, State and local governments, public and private utilities, and other Federal agencies involved with managing our water resources. These includes various sub categories
Surface Water: Nationally, USGS surface-water data includes from more than 850,000 station, time-series data that describe stream levels, streamflow (discharge), reservoir and lake levels, surface-water quality, and rainfall.
Groundwater: The Groundwater database consists of more than 850,000 records of wells, springs, test holes, tunnels,drains, and excavations in the United States. Available site information includes well location information such as latitude and longitude, well depth, and aquifer.
The USGS annually monitors groundwater levels in thousands of wells in the United States. Groundwater level data are collected and stored as either discrete field-water-level measurements or as continuous time-series data from automated recorders. Data from some of the continuous record stations are relayed to USGS offices nationwide through telephone lines or by satellite transmissions providing access to current groundwater data.

4.2 Data set from Joint Research Centre Global Surface Water

The Water Occurrence dataset shows where surface water occurred between 1984 and 2015 and provides information concerning overall water dynamics. This product captures both the intra and inter-annual variability and changes. The occurrence is a measurement of the water presence frequency (expressed as a percentage of the available observations over time actually identified as water). The provided occurrence accommodates for variations in data acquisition over time (i.e. temporal deepness and frequency density of the satellite observations) in order to provide a consistent characterization of the water dynamic over time. An example of this dataset is shown below:

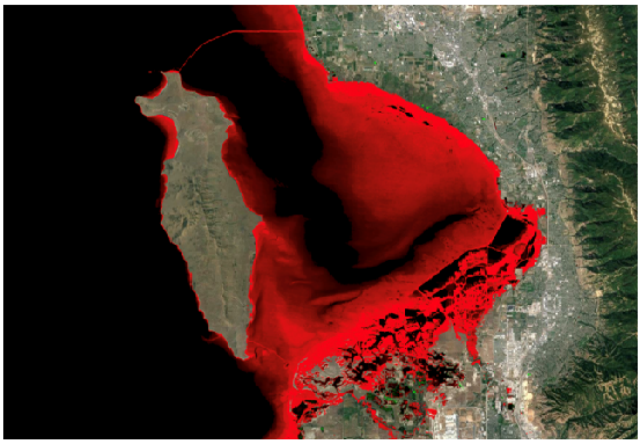
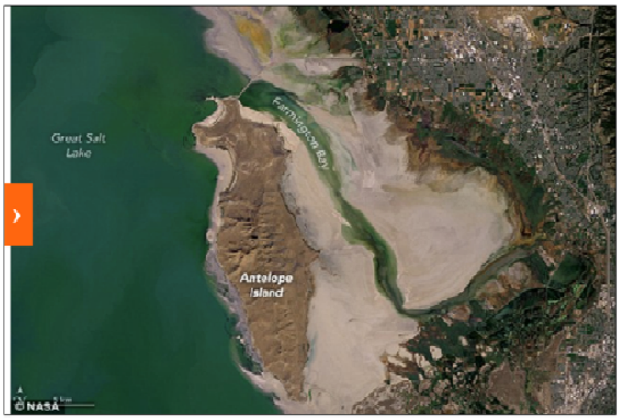


Figure 1: Water change in Great salt lake of Utah [10]



Before-and-after images taken in 2011 and 2016 show dramatic reductions in the Farmington Bay basin of Great Salt Lake. the images were released by NASA, which warned 'Both nature and man have a hand in the change'.

Figure 2: Actual satellite image from 2011

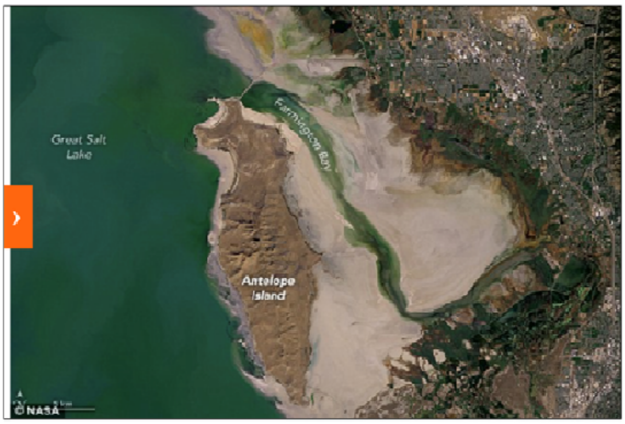


Figure 3: Actual satellite image from 2016

Lets consider the case of a particular water body, namely great salt lake of Utah. Figure 1 shows the image of Great salt late of Utah, 2011 and the same in Figure 2 in 2016. If we look into the satellite image data for receding water boundary of the salt lake, the red regions in figure 3 determine the regions where surface water has decreased. If we look at finer level, we find that the locals have diverted the water from these regions for extensive agricultural consumption. This has increased the salinity of the water and poses major health and economic concerns for the population in the region. Also the bear river project development to dam and divert further from the lakes main feeds with will cause aggravated decrease in surface water. Big data could help in these places to drive and gain insights about projects that require to maintain water surface levels below a certain level.

The following processing was done on the dataset:-

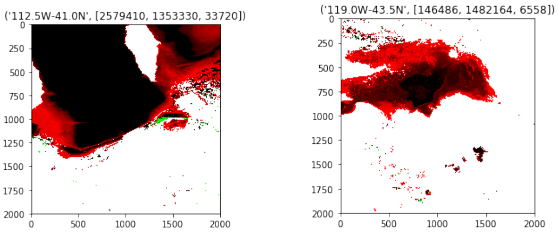
Data Preprocessing: Dataset for many hydrological regions was very sparse so we decided to analyse the ground water level for every state. We started with cleaning the data files for each hydrologic region, since for each region file consist of administration information and site information, we cleaned the dataset and converted it into csv format. Using pandas library we created dataframe for each state.

Missing value imputation: As various weather parameters affect the precipitation hence water level and the weather trends are more likely to recur each year. We found it most suitable to replace missing values with the value for same date in previous years. Since data was available for each of the hydrological region from 1980 to 2016, for each day, we merged the water level for all of the hydrological regions and took the average of it. The final dataframe we obtained was the average water level for all of the hydrological regions for each day from 1980 to 2016.

5. METHODOLOGY

5.1 Clustering

We developed a model to cluster all images based on



(a) Great Salt Lake

(b) Lake Malheur

Figure 4: Comparison of two similar images in the same cluster based on their surface water depletion

their similarity of the water loss/water inundated areas since 1984. We use K-means cluster to cluster images to find out similar regions.
We treat one cluster as one dataset. In this reduced search space, we use Nearest Neighbor to calculate similar images to other hydrologic region.

5.2 Machine learning

The dataframe for each state was divided into train and test set. We made 70-30 train test split of the dataset. **Lasso Regression**: After tuning the alpha parameter, we fixed it at 0.1. We used Sklearn machine learning library for applying lasso regression on the training data and predicting the water level on the test data using that model.
Random Forest Regression: The second model we used was random forest, the hyper parameter max_depth=3 and random_state=0 was used.
Support Vector Regression: Another model was SVM regressor. The penalty parameter of the error term we taken was 1 and epsilon-tube within which no penalty is associated is 0.2

5.3 Satellite Image analysis using Spark

We analyzed satellite image analysis from coordinate 130W, 60N - 80W, 30N. It consists of 18 images, with each image spanning a latitude of 10 degrees latitude and 10 degrees longitude. We only used the Surface water change image dataset, that indicated the change in surface water levels with a value of ff0000(red) or 00ff00(green). Red indicates a permanent loss of water body from the year to 1984, whereas green has indicated addition of water body in the area. The resolution of each image is 40000x40000. To increase computation efficiency, we split the images into smaller size, i.e, size of 400 images of 2000x2000.

Now each image represents a land area of 100 sq. km.

We computed the size of regions that lost water, size of regions that have gained new water bodies or area where there is no change in surface water. Now each tile of map of 100 sq km can now be represented by a feature vector of (unchanged water area, area of lost water body, area of newly inundated regions).

We filtered out regions that did not contain any changes, that is images whose 2nd column and 3rd column is zero. We now cluster the images based on these features using K-means clustering. We did an elbow curve analysis and found that k = 10 is the best possible number of clusters. Now we have designed a system, that when queried upon could return all areas that have similar water trends since 1984. Using a top 10 neighbor search, we find the regions that are most similar to a particular query coordinates.

6. RESULTS

Our analysis shows several insights into our problem statement. From analysis of USGS data we calculated the error metric (Root Mean Square Error) of each model which we applied on all states. Below table is a sample for 5 states.

We used elbow analysis to chose clusters of 10 size. This is based on the analysis of plot of Within Set Sum of Squared Error(WSSSE) with cluster size, which showed a sharp decline at around 10.

Top 10 neighbors on satellite image after clustering shows similar regions that even though are separated by huge distance show alike in their water depletion/inundation patterns.

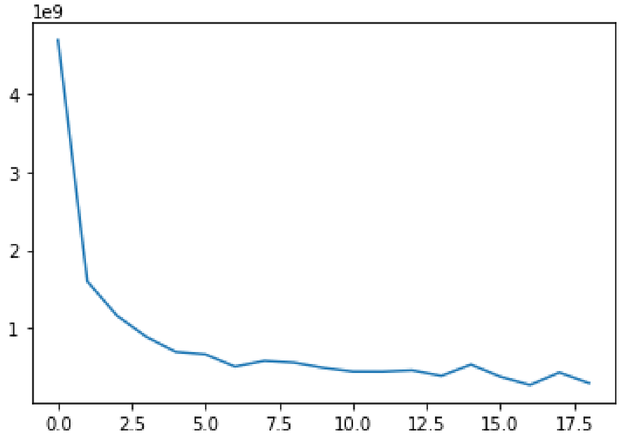


Figure 5: Elbow Analysis, using 10 size cluster

7. DISCUSSION

We considered 2 machine learning methods to forecast the water level: Using atmospheric features like temperature, pressure, precipitation, humidity, wind speed to predict water level which is the dependent variable.
Using water levels of previous ten days corresponding to day d as feature variables to predict the water level of the day d+5.

Table 1: Model Accuracy

States	Lasso	SVM	Random Forest
Minnesota	0.654	0.822	1.139
Massachusetts	0.887	1.013	1.934
Kentucky	0.434	1.306	0.668
Louisiana	4.876	18.790	11.657
Wisconsin	2.979	9.7915	3.873

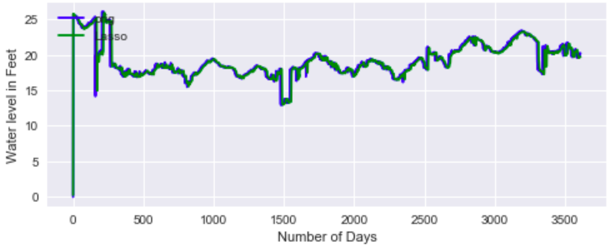


Figure 6: State of Massachusetts

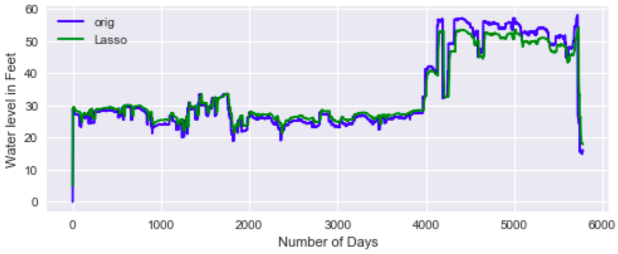


Figure 7: State of Wisconsin

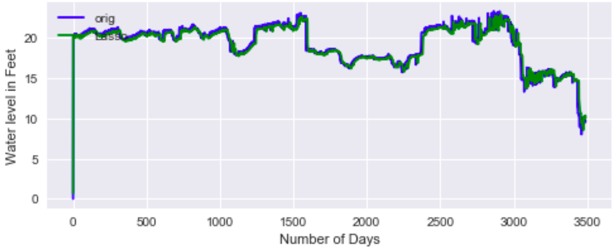


Figure 8: State of Minnesota

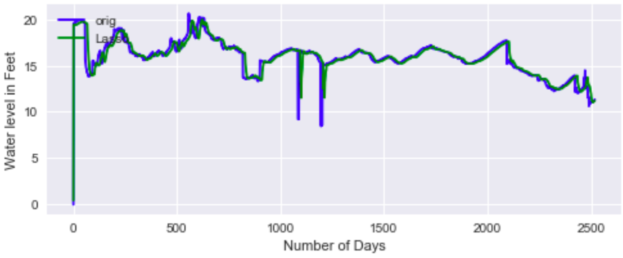


Figure 9: State of New Hampshire

For the first approach, the data for the atmospheric features was not available for many hydrologic regions. We couldnt impute the missing values as the entire feature column was absent. Thus we adopted the second approach in which we use water level of the previous days as the feature variables. We experimented with three machine learning models and finalized Lasso Regression as the best model. We plotted the predicted values and the original water level values on the test data and it is clear from the graphs that the predicted values are close to the actual values. By leveraging this pattern in the time series water level graph, we can predict the water level for the future days as the predicted water level values would be used as input for the subsequent days. In other words, it forms a recursive formula:

$$f(d+5) = f(d-10, d-9, ..., d-3, d-2, d-1)$$

Also from our satellite image analysis tells that we could extrapolate our findings based on high quality data from developed countries to countries with less infrastructure to maintain such records. We have clustered sites based on smaller features as observed in an area of 100 sq km and broken it down to three features. This reduces the search space to find similar regions in a more efficient methods.

8. CONCLUSION

Understanding and employing the techniques of big data is a challenging task. We have presented several machine learning models to forecast groundwater level using USGS dataset. Our contribution is to investigate the application of LASSO, Random Forests and Support Vector Regression for groundwater level analysis of each state in USA. Model performance was assessed in terms of the inputs used to the machine learning models and the output by our models.

On the other hand, we also explored satellite imagery dataset and identified similar regions based on surface water level depletion using K means clustering. Using this approach we can further identify similar regions in other part of the world for which we dont have surface water level data and which are prone to surface water level depletion.

9. TEAM MEMBER CONTRIBUTIONS

- **Analysis by clustering similar regions:** By Akshay and Anupam
- **Dataset merge and cleaning:**By Anupam and Srikant

- **Feature extraction and dimensionality reduction:**By Akshay, Srikant and Deepak

- **Prediction and insights using machine learning:**By Akshay and Deepak

10. REFERENCES

1. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber =7371798&tag=1
2. https://waterdata.usgs.gov/nwis/dv?referred_module =qw&search_criteria=state_cd&search_criteria=huc2_cd &search_criteria=lat_long_bounding_box&submitted_form=introduction
3. http://scikit-learn.org/stable/modules/linear_model.html
4. https://www.hindawi.com/journals/cin/2017/8734214/
5. https://sustainabledevelopment.un.org/sdg12
6. https://waterdata.usgs.gov/nwis/
7. http://i.dailymail.co.uk/i/pix/2016/09/23/18/38B9AD2200000578-3801519-image-a-39_1474652775701.jpg
8. http://www.dailymail.co.uk/sciencetech/article-3902686
9. https://global-surface-water.appspot.com/Jean-Francois Pekel, Andrew Cottam, Noel Gorelick, Alan S. Belward, High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422 (2016). (doi:10.1038/nature20584)
10. https://global-surface-water.appspot.com/