Group Tuvalu
01/06/2021

Report Modern Data Analytics: Water Security

1. Context

The water security and scarcity of regions around the world are currently challenged by climate change. The increased occurrence of extreme weather events, such as droughts, floodings and a generally warmer climate requires countries to use their water resources more carefully. According to UNICEF, at least two billion people experience extreme water scarcity for at least one month a year and are projected to worsen in the coming years. In addition to climate factors, also socio-economic factors play an important role. Indeed, with the increasing urbanisation around the world, water consumption will concentrate more in densely populated areas. Furthermore, the availability of proper infrastructure and distribution, which depends on the wealth of the country, are also key factors in determining whether water security can be achieved. An interview that was conducted with a water resources system modeller at Anglian Water Services pointed out the difficulties in ensuring water security around the world depends on very specific nuances. For example, a large water project in Angola was dealing with the following dilemma: Angola wanted to build a large dam to allow greater agricultural and human development. But the dam would prevent water from flowing in the Okovango Delta, one of the most biodiverse regions in southern Africa. The issue lies in deciding whether human development or biodiversity takes priority. In sum, drawing detailed conclusions on individual countries is a task that requires a thorough understanding of individual circumstances.

Therefore, the Tuvalu team has decided to create a tool that predicts future water stress on a country level by considering both climate factors and socio-economic factors. Following research questions were formulated to which this project seeks to find an answer:

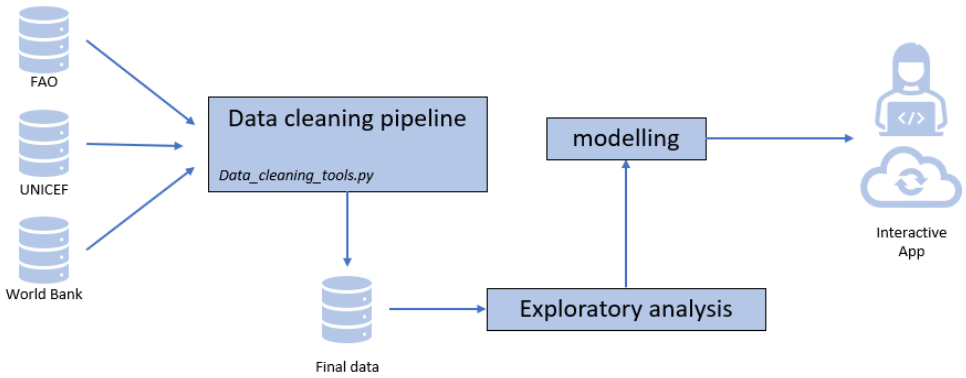
1. How important are climate factors versus socio-economic (or human-induced) factors in determining water security for countries around the world?
2. What effect does climate change have on water security in the coming decades?
3. Which countries will experience the greatest threat in terms of water security in the coming years?

2. Data assimilation

Three types of data on a country level are required to answer the research questions: (1) current and projected climate data, (2) socio-economic data and (3) water stress indicators as a proxy for water security and scarcity. Whilst many resources are available, a major issue with an extensive number of data sources arose: data regarding water security is typically already summarized data (i.e. qualitative) and not quantitative. This has several implications, especially with respect to the type of modelling. The climate data comprised of current and projected rainfall and temperature data and available natural water resources data and were extracted from the World Bank Climate Change knowledge portal [1] and the AQUASTAT database from the Food and Agricultural Organisation from the UN (FAO) [2]. For the socio-economic data, which included population data (urban and rural, population growth, mortality rate, human development index, …) and economic data (GDP) two sources were used: the AQUASTAT [2] and the UNICEF database [3]. Finally, since ‘water security’ is a broad definition that cannot be modelled directly, we decided to use two proxies, i.e. water scarcity and water use efficiency which were also extracted from the AQUASTAT database. The combination of these two targets encompasses the goal of achieving water security. The raw data files are available in the *raw data* map in the Jupyter Notebook.

3. Overall methodology

Given that 6 individuals worked on this project remotely, a critical factor was to keep everything well-organised. This means that all Python code adheres to the PEP8 coding style. Moreover, weekly update meetings and a Slack channel greatly facilitated communications and progress. An overview of the pipeline of the data is shown below and includes: (1) data cleaning (outliers, missing data), (2) combining the data in a final dataset, (3) exploratory data analysis using clustering and principal component analysis, (4) testing multiple predictive models and (5) making a simple and interactive app that allows users to use the final prediction model to visually investigate the impact of climate and socio-economic factors on a country level.



4. Data pipeline

4.1 Pre-processing

The data pre-processing for all data is performed in the *Preprocessing* notebook. In order to promote code reusability and readability the cleaning functions are coded into a separate script *data_cleaning_tools.py*. This notebook also combines all data in one final dataset (*clean data/final_data.csv)* and includes multiple steps. Since the goal is to present the current (and future) state of water security for countries, the data were filtered to only include a snapshot of 2017. And because the data sources had different ways of encoding countries, a standardised format (ISO3) was implemented. Variables were also renamed for improved data manipulation. Overall, the data quality was more important than quantity, because the data are already summarised on a country level. Therefore, a significant amount of time was spent exploring the nature of missing values and manipulating variables to a suitable format for further analysis. Missing values originated mostly from smaller island countries, such as the Falkland Islands and were removed, because imputations based on continental countries for geographically remote regions can be inaccurate. The *final_data.csv* file contains data for 123 countries and 19 predictor variables and 3 target variables.

4.2 Exploratory data analysis

In the exploratory analysis the goal is to analyse the data on a variable level and observation level. In terms of variables, by using boxplots the distribution of the different predictors illustrates that there are multiple outliers for most variables, most notably in the climate related variables. The data is therefore log-scaled in the predictive analysis. Secondly, basic correlations demonstrate that most socio-economic variables correlate highly - pointing to a potential issue of multicollinearity. The second step in the exploratory analysis consists of identifying clusters of countries. K-means and spectral clustering are explored to discover underlying patterns. As could be expected, a number of intuitive clusters form,

though some surprising clusters form as well. For example, Peru, Colombia and Indonesia cluster together both in terms of socio-economic factors and climate factors. The enhanced flexibility of spectral clustering compared to k-means also yields better error statistics.

4.3 Predictive modelling

The three log-scaled water stress indicators were predicted using (1) only the climate factors, (2) only the socio-economic factors and (3) both the climate and socio-economic factors together. First the data was split in a training and a test set. The training set was put into a pipeline (sklearn) which exists of first a scaler function, then a dimensionality reduction and finally a regression model. The best performing model was obtained by testing a total of four scaler functions, a range of selected principal components for the dimensionality reduction and three regression models, including their hyper-parameter optimisations (see *Pipeline* notebook). This last step was performed using cross validation with the sklearn gridsearchCV function. The best predictive model was selected based on the performance of the test set. The final model uses a log scaler, only on selected variables. Followed by a PCA analysis which only slightly reduced the dimension of the predictor variables from 15 to 12-14 (depending on the target variable), indicating that this step was mostly redundant. And at the end, a ridge regression gave the best model performance, which had the advantage of dealing with the previously observed multicollinearity, especially for the socio-economic variables.

5. App

The app is built using dash and deployed with Heroku. The functionality of the app is contained within two python scripts in the *dash* folder. The *app.py* script contains the frontend of the app, the html component and the interaction between the user, frontend and backend. The *tools.py* script contains the backend components of the app. The backend functionality is contained within one class. This allows for the data to be stored within the class as a self-object, requiring fewer arguments for the functions as the data does not have to be provided every time. The remaining files are the datasets, and files required for the deployment (*Procfile* and *requirements.txt*), and the favicon in the assets folder.

6. Results and conclusion

The results of the final model, i.e. ridge regression demonstrate rather low predictive power. The disappointing results are mainly to blame on the size of the data set and the type of data that was used. Moreover, thanks to the app deployment it is also possible to assess the impact of each variable intuitively. Variables such as Urbanisation have a large weight in a model. Indeed, in the coming decades megacities in countries with high population growth and high rates urbanization water security is a pressing issue [4]. Furthermore, playing around with the model demonstrates that using very high values for the different parameters results in very volatile results, suggesting the model is unstable. The results with respect to the two research questions are mostly inconclusive. Though both climate and socio-economic factors have an undeniable effect on the water security of a country, the current, broad approach may not be the best way to answer the research questions. Indeed, circling back to the Okavango-issue from the introduction suggests that dealing with water security cannot solely be approached based on broad indicators. One should consider the variety of nuances of specific cases in order to conclude and make predictions. Finally, a major issue of modelling water security is the available data. Summary statistics are not particularly useful to make predictions. As such, targeting specific regions or issues related to a specific part of water security with more extensive data may be a less ambitious, yet more practical approach.

Bibliography

[1] World Bank database, https://climateknowledgeportal.worldbank.org/ [as of 31/05/2021]

[2] FAO AQUASTAT database, http://www.fao.org/aquastat/en/ [as of 31/05/2021]

[3] UNICEF database, https://data.unicef.org/ [as of 31/05/2021]

[4] Kookana, Rai S, et al. "Urbanisation and Emerging Economies: Issues and Potential Solutions for Water and Food Security." The Science of the Total Environment, vol. 732, 2020, p. 139057.