Identification of urban water supply patterns across 627 cities in China based on supervised and unsupervised statistical learning

Andrew Gonzalez, Brandon Chou, Charles Li, Djavi De Clercq, Rinitha Reddy

Introduction

Urbanization, one of the prevailing trends of the 21st century, places great stress on the water resources of cities across the globe. Our project aims to investigate the most important variables underlying urban water supply patterns in China, a region which has seen rapid urban growth in the past few decades. We applied statistical learning methods to 12 years of urban water supply data for 627 cities across China. In addition, a PCA-informed urban water sustainability index was developed in order to benchmark cities. The implications of our research effort will be useful for decision makers in water-stressed urban areas who seek novel insights into urban water supply patterns using statistical learning techniques.

The first major contribution of our work was the identification of variables most responsible for variance in patterns of urban water distribution and management. The second innovation was the identification of statistical learning algorithms which provided high accuracies in prediction and classification problems related to China's urban water data. Thirdly, a general, systems-level perspective of major urban drinking water use trends in China was provided for the benefit of public-sector stakeholders. Lastly, an urban water use sustainability index was developed in order to benchmark cities against each other and identify areas where water sustainability was lacking.

Our findings can be summarized as follows: (1) PCA showed that approximately 46.8% of variability in the data could be explained by two principal components. (2) Random forest (88.3% test accuracy) and XGBoost (87.7% test accuracy) algorithms were effective in classifying provinces using numerical features of the dataset. (3) Chinese cities have consistently suffered water loss/leakage rates above 20% since 2001, and water prices are closely associated with leakage. (4) Based on the sustainability index, problem cities/regions were identified.

Data

The dataset is comprised of 87 variables grouped into six categories: (1) water supply and sale (e.g. daily supply capacity); (2) supply pipelines (e.g. pipe length and pipeline coverage area); (3) supply service (e.g. pressure standard compliance); (4) supply operations management (e.g. electricity consumption per unit of water supplied); (5) water supply finance (e.g. company revenues and costs); and (6) water supply prices. These yearbooks do not exhaustively cover all the cities in China. However, cities included in the yearbooks account for a total urban population of 332.4 million, which is around 43% of China's urban population.

Methodological Approaches and Results

First, the most recent data (2001 to 2013) on urban water supply was collected from the China Urban Water Association, synthesized, and pre-processed. This involved the merging of all years' data into a single dataframe, removing outliers from all continuous features (≥ 3 standard deviations from the mean), removing features and observations with excessive missing data, and imputing missing values using feature means. Our final data frame consisted of 5,667 observations and 50 variables.

Second, Python's visualization packages (including matplotlib, Seaborn, plotly, and cufflinks) were used to generate comprehensive swarm plots and a correlation heat map of all continuous features, which revealed important associations, e.g. between water price and year.

We then employed a combination of supervised and unsupervised statistical learning techniques to describe the data. The primary unsupervised method used was principal component analysis (PCA). The PCA algorithm results showed that five principal components (PCs) cumulatively accounted for 57.85% of variance in the dataset; component 1 (C1) and component 2 (C2) accounted for 37.26% and 9.51% of variance, respectively. Based on a

correlation heat map (with PCs on the *y*-axis and variables on the *x*-axis), C1 seems to be closely associated with variables related to water supply and sale, supply pipelines, and water supply finance; C2 is related to urban water prices and average per capita water use.

Our first application of supervised learning was to assess which classification algorithms had the best accuracy in classifying observations (i.e. provinces at a particular point in time) into their respective provinces. The motivation behind doing this was as follows. Firstly, annual data on urban water quality is particularly hard to come, and data from the China Urban Water Association often contains missing values. Our algorithm offers a solution for these missing values. Secondly, this is the first time any machine learning has been applied to this dataset, and it useful to understand which algorithms perform well on urban water data.

Of the algorithms we tested (logistic regression, SVC, perceptron, KNN, random forest, XGboost), XGboost and random forest performed best: their accuracies were 87.69% and 88.32, respectively. The top three variables that were most important in classifying data using XGBoost were total wages, turbidity rate meeting standards, and chemical oxygen demand meeting standards; for the random forest, top variables were total electricity use, sales revenue, and total wages.

Furthermore, given its high performance in classifying regions by province, we ran random forests to predict each feature of the data (constructing regression trees for continuous response variables and classification trees for categorical ones). We then selected a few features of particular importance to policy makers (e.g. leakage rate, water quality) as prediction targets and identified top variables that dominated their respective feature importance computations. For leakage rate, the top three variables were administrative water price, water quality, and population with water access; for water quality, top variables were total disinfectant consumption, industrial sector water price, and total electricity use.

Finally, the China Urban Water Sustainability Index (CUWSI) was developed to benchmark cities with regards to the sustainability of their urban water supply patterns. We did this by integrating the weighted average of different variables associated with sustainability. Based on secondary research of water sustainability methods, we first selected variables from the original dataset closely associated with sustainability and grouped them into the following subcategories: (1) water supply reliability, (2) water network efficiency, (3) water quality levels, and (4) socio-economic impact. Each individual variable was assigned a nonzero integer weight (ranging from -1 to 5), and each subcategory was assigned a percentage weight as follows:

Sub-category and weight (%)	Sub-category variables and weights	Weights
Water supply reliability (20%)	▪ 1.1. Total per capita production capacity (10,000 cubic metres/day/person)	5
	▪ 1.3. Number of plants per capita (units/person)	3
	▪ 2.2. Average supply per day per capita (10,000 cubic metres/day)	4
	▪ 4. Water sold / water produced (%)	1
	▪ 5.1. Coverage rate of urban water supply (fraction)	2
Water quality (40%)	▪ 14.1. Overall rate of water that meets standard (%)	2
	▪ 14.2. turbidity rate meeting standard (%)	1
	▪ 14.6. Bacterial contamination rate meeting standards (%)	3
	▪ 14.7. Total coliform contamination rate meeting standard (%)	4
	▪ 14.8. Oxygen demand (CODMn) meeting standard rate (%)	5
Water network efficiency (10%)	▪ 8.1. Total pipe length (km)	1
	▪ 13.1. Total number of water meters (units)	2
	▪ 10. Loss and leakage rate (%)	-1
Resource intensity (30%)	▪ 17.1. Total electricity use (10,000 kWh)	1
	▪ 17.2. Electricity use for water production (1,000 kWh/1,000 cubic metre)	-1
	▪ 23. Unit water sales cost (yuan / thousand m3)	-1
	▪ 25.1. Total employees (persons)	2

In order to arrive at these weights, the results of the PCA were considered. The first principal component contributed to 37.26% of the variation and hence, the relative importance of each feature was derived from this result. For each city, SI sub-scores for the four subcategories were first computed. For example,

$$SI_{water\ supply\ reliability} = \sum_{i=1}^{5} W_i I_i$$

$$where\ SI \rightarrow Sustainability\ Index\ and\ I \rightarrow Indicator$$

These sub-scores were all brought to a 0-1 scale using min-max normalization, weighted and summed up to obtain the cumulative CUWSI score:

$$CUWSI = 0.2 \times SI_{water\ supply\ reliability} + 0.4 \times SI_{water\ quality} + 0.1 \times SI_{water\ network\ efficiency} + 0.3 \times SI_{socio\ economic\ impact}$$

The process was iterated several times in order to fine-tune the weights and obtain an Index that was able to benchmark the cities as accurately as possible. Ultimately, the assignment of a high CUWSI would indicate that a city has a desirable and sustainable water culture. We found that across the years, the average CUWSI of various regions is China have increased slightly, but not at a desirable level. This confirms results found in previous results as well as those found during our project, which indicated that the water loss/leakage percentage has remained at a constant 20% all throughout the years. The lack of development in water supply facilities coupled with an exponential increase in water usage contributes to this trend of our index. Further, this index can be used as a tool for policyholders, as the impact of variations of indicators such as unit water price, number of plants etc., can be acquired and necessary steps can be taken.

Future Directions

The question of how to deal with water scarcity is an enormous one that is grounded in not just environmental but also social, economic, and geopolitical frameworks. As a critical first step, this project conducted a descriptive analysis of one of the most comprehensive datasets on China's urban water supply. Nonetheless, the scale and richness of this dataset enables a host of more targeted research questions that explicitly propose predictive models informed by potential causal relationships. These relationships could take into account or further examine issues that are specific to China, such as the uneven distribution of water resources in the different parts of the country, massive rural-to-urban migration patterns, effects of rapid economic development and rising wages, and the consequences of inter-province trade involving goods that require a substantial amount of water to produce. To probe these relationships, future research endeavors can also explore the possibility of integrating city-level data on topics such as urban, rural, and regional development; living conditions; agriculture; and industry readily available from the China Statistical Yearbook.

Team Contributions

Andrew Gonzalez conducted the exploratory data visualizations and feature importance analyses; Brandon Chou spearheaded the implementation of predictive statistical learning techniques and also conducted feature importance analyses; Charles Li was responsible for the overall brainstorming and internal review of methodological approaches; Djavi De Clercq led the data collection, processing/cleaning, and dimensionality reduction (PCA) efforts; and Rinitha Reddy led the development of the China Urban Water Sustainability Index. Although the work was partitioned according to our strengths and interests, the entire team invested an equal amount of effort into this project. We met often and brainstormed/coded together, and everyone had a good understanding of each other's workload. We mutually contributed code, ideas, and everyone contributed their fair share to the final paper. Our teamwork and communication was excellent, and working as a team was a pleasure. The research culminated into a comprehensive research paper; this will be submitted at some point in December 2017. If accepted, we will notify the course instructors, so they hopefully have a good example of a dataX "success story" in teamwork for the next iterations of this course.