Estimating the relationship between Human Development Index level and the tap water prevalence in countries

Tapio Sivonen
Oulu 2017-06-18

Portfolio work

Intent in this work was to model tap water prevalence in different countries compared to their Human Development Index level to compare if there would happen to exist a given HDI level that could be visioned to enable tap water access to a country. At the same time this allows the writer the opportunity to exercise his ability of statistical inference with multiple tools, in this work mainly LibreOffice, Java 8 and Apache Commons Math 3.6.1.

Results

The resulting models show that in year 2015 the HDI level associated with wider tap water prevalence is higher than in year 1990, which is somewhat surprising.

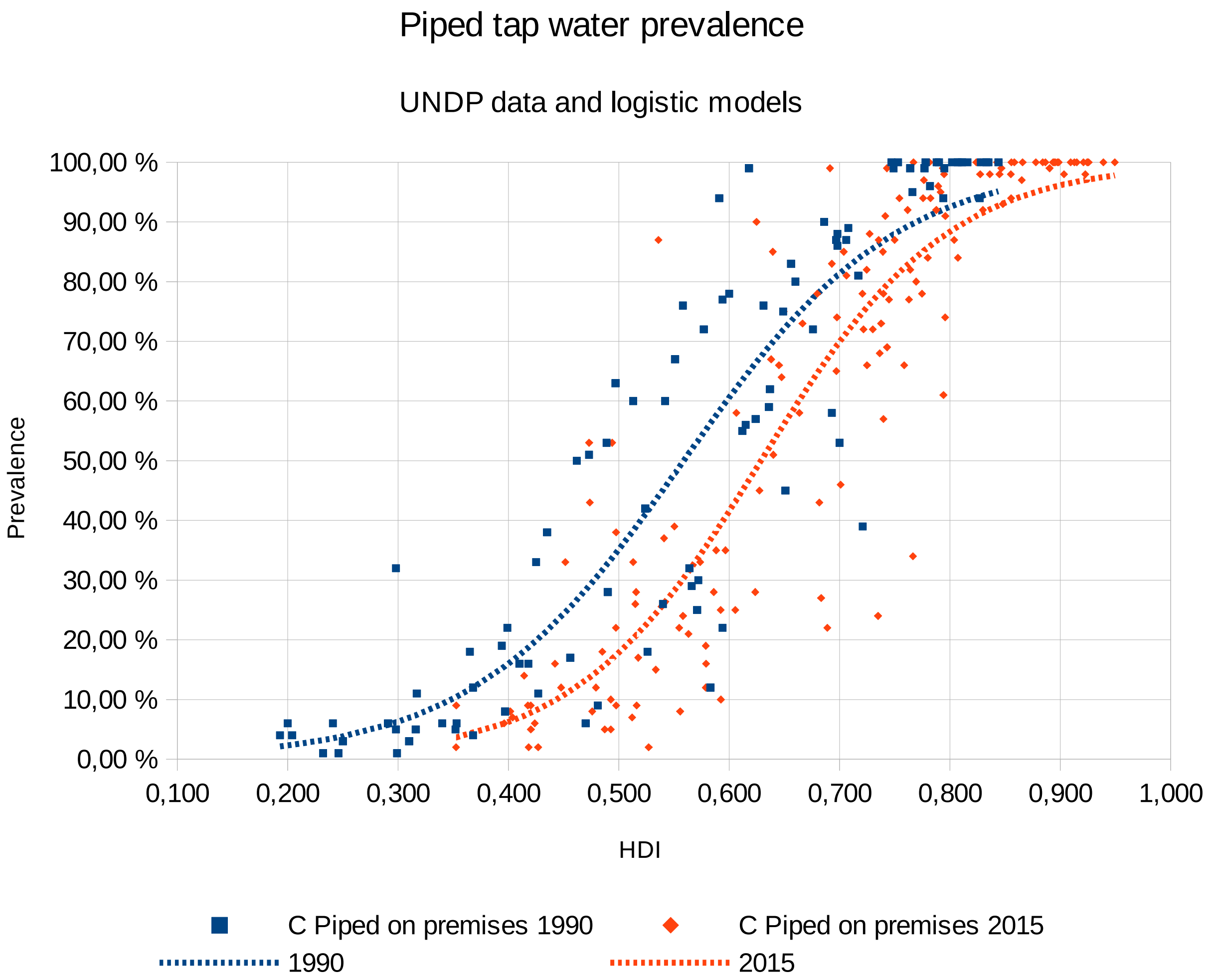

Illustration 1

The results suggest that while HDI levels of countries of the world have improved clearly from 1990 to 2015, the availability of piped tap water has not increased in the pace as one could have expected if tap water was clearly linked to the HDI level of the countries.

Work outline

The work was done in multiple steps utilizing UNDP data sources, LibreOffice, Java 8 and Apache Commons Math.

On previous work, the author had gathered UNDP data on HDI in different countries including the year 1990. This data was copied from previous spreadsheet into new spreadsheet. UNDP HDI data for year 2015 was gathered from UNDP web database in http://hdr.undp.org/en/composite/HDI . The data was imported to the LibreOffice Calc worksheet prepared for this purpose. Data sources were combined into one Calc worksheet.

Data was filtered out to extract country and HDI observations when HDI was present and year was either 1990 or 2015. Tap water prevalence data was imported to worksheet from WHO from https://data.unicef.org/topic/water-and-sanitation/drinking-water/ . Tap water prevalence data was combined to HDI data by using country name and year as a key.

Data and proposed model was quickly sanity checked by transforming the tap water prevalence data into logit form and then plotting the logit over HDIs (Illustration 2). As it seems visually meaningful to make linear regression over the logit results, the premise was deemed potentially viable and process was continued.

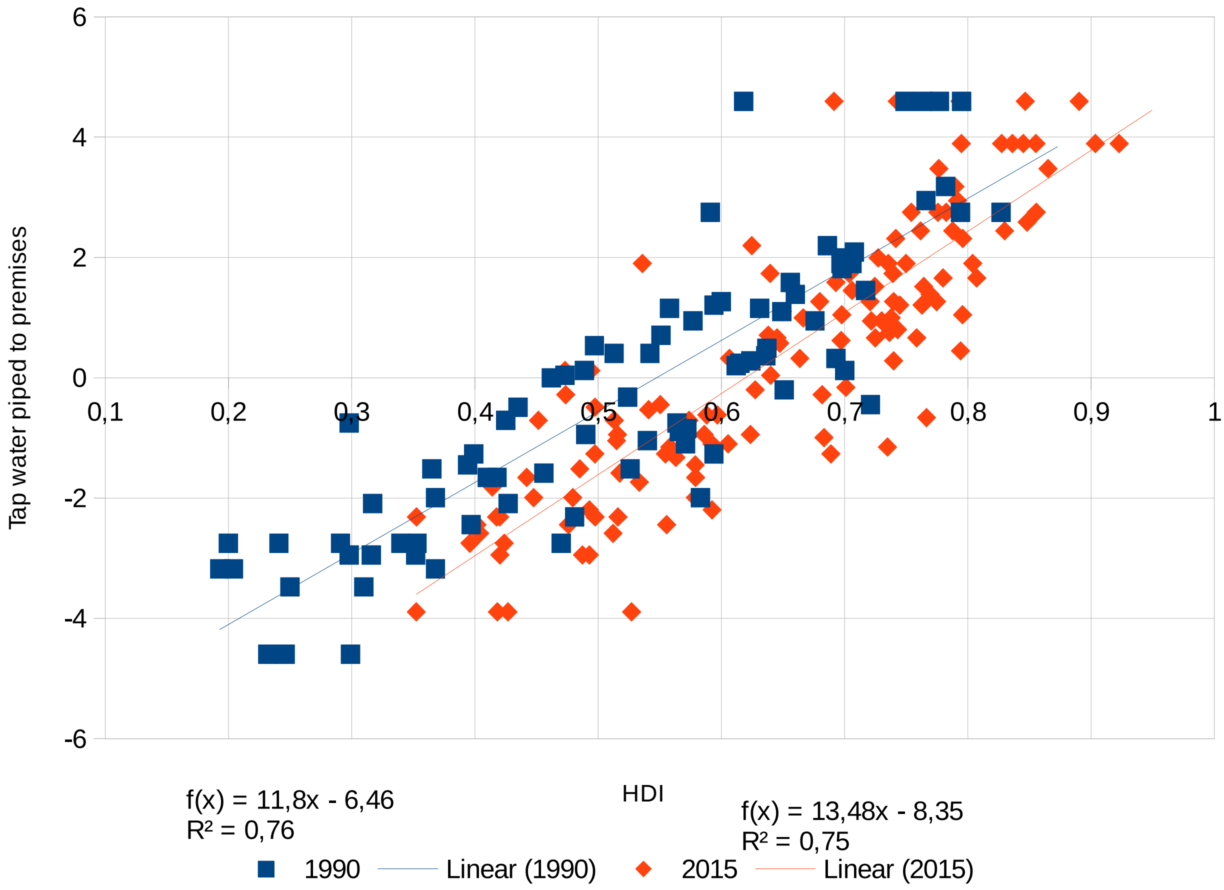

Illustration 2

Combined data was formatted to CSV format using semicolon as separator in LibreOffice Calc using concatenation. Project folder structure was created and git version control was initialized.

Java program was created to do logistic regression on the CSV file. The program consists of 5 classes, HDITapCSVReader to read the CSV data in, HDITapEntry to store single observation values in, HDITapObservationMapper for transforming the data into Apache Math Commons usable format, HDITapModelFunction to create a two parameter logistic function instead of baseline 6 and HDITapFitter as the main class to manage reading and mapping the data and fitting the model to the data with least square loss function.

Analysis program was then run with the CSV data, which resulted in models for year 1990 (m=0,558790332564035; b=10,4324676235407) and year 2015 (m=0,628836874480114; b=11,8503859589139). In plain words, these models mean that in 1990 at about HDI of 0,559 a country had about 50% of population within piped water and in 2015 the same HDI requirement was about 0,629.

The observation data was then supplemented with modeled data and the result was used to draw Illustration 1.

The results, charts and work method descriptions were then written into work report on the subject. The production took about 6 hours for data preparation and visualizing, about 8 hours of programming and about 4 hours for reporting.