waste_not_the_water:
A Data Science Tool for Urban Waste Water Treatment Plants

Yi-Yu Lin, Caitlin Parke, Yuening Wang, Sijia Xiao

Background

- Treatment of waste water is fundamental to ensure public health and environmental protection.
- In northern countries has more than 70 % waste water receiving treatment. In central countries about 75% are receiving treatment. Southern, south-eastern and eastern countries has only 70 % of waste water receiving treatment.
- Data reported sometimes provide an incomplete picture of inhabitants connected to waste water treatment.
- There are variations in definitions of different classes of treatment between countries that lead to differences in the level of purification.
- Approximately € 22 billion investment is needed to achieve the implementation of the urban waste water treatment. The total yearly investment in the renewal, improvement and extension of existing infrastructure are expected to reach € 25 billion per year.

Goals

- To predict capacity for new European waste water treatment plants.
- To visualize of database from the Urban Waste Water Treatment Directive with various visualization packages.

Use Cases

1. Machine Learning Model:
- Data Refining: Finalize proper data structure with no empty data and packages for model.
- Build and Train Model: Takes final dataframe, train and test the model for prediction of a new treatment plant.

2. Data Visualization:
- Data Mining: Data cleaning of the primary database and create final combined dataframe with necessary columns.
- Visualization: Color graphs, historgram, map plots with interactive map graphs.

3. Prediction of a new treatment plant by the machine learning model:
- User input to the model: Desired location, waste water entry load and treatment type.
- Output from the model: Capacity of the treatment plant. Comparison of the user input with the primary database, if similar case exists, suggest user to contact existing plant for more information.

User Interface

- A 3- page app from dash that runs on local browser.

1st page – Default home page with description of project and use cases involved and links.

2nd page – Data visualization with interactive graph and a dropdown menu for updates.

3rd page – Machine learning model with input of agglomeration size, location and output of what capacity is needed and the closest plants in terms of all inputs.

- The user interface was created with Dash by Plotly.
- The Dash platform allows users to specify components in Python, which are then converted into html code and rendered on the web.
- The user interface for the predictive machine learning model utilizes such components as input text boxes, check boxes, and submit buttons. When the user interface python script is called, the user hosts it locally in a web browser.

Simple Linear Regression

- We considered the load volume entering the waste water treatment plant and its location to build a machine learning model for capacity prediction. Therefore, the load volume is the most important parameter.
- We plotted the load volume versus capacity (Figure 1) and realize that there is an obvious linear relationship between the predictor and response. However, the statistics of multilinear regression shows that latitude and longitude also affect the capacity. To further analyze the linear relationship, Figure 2 was plotted to analyze high leverage points and outliers of our dataset.

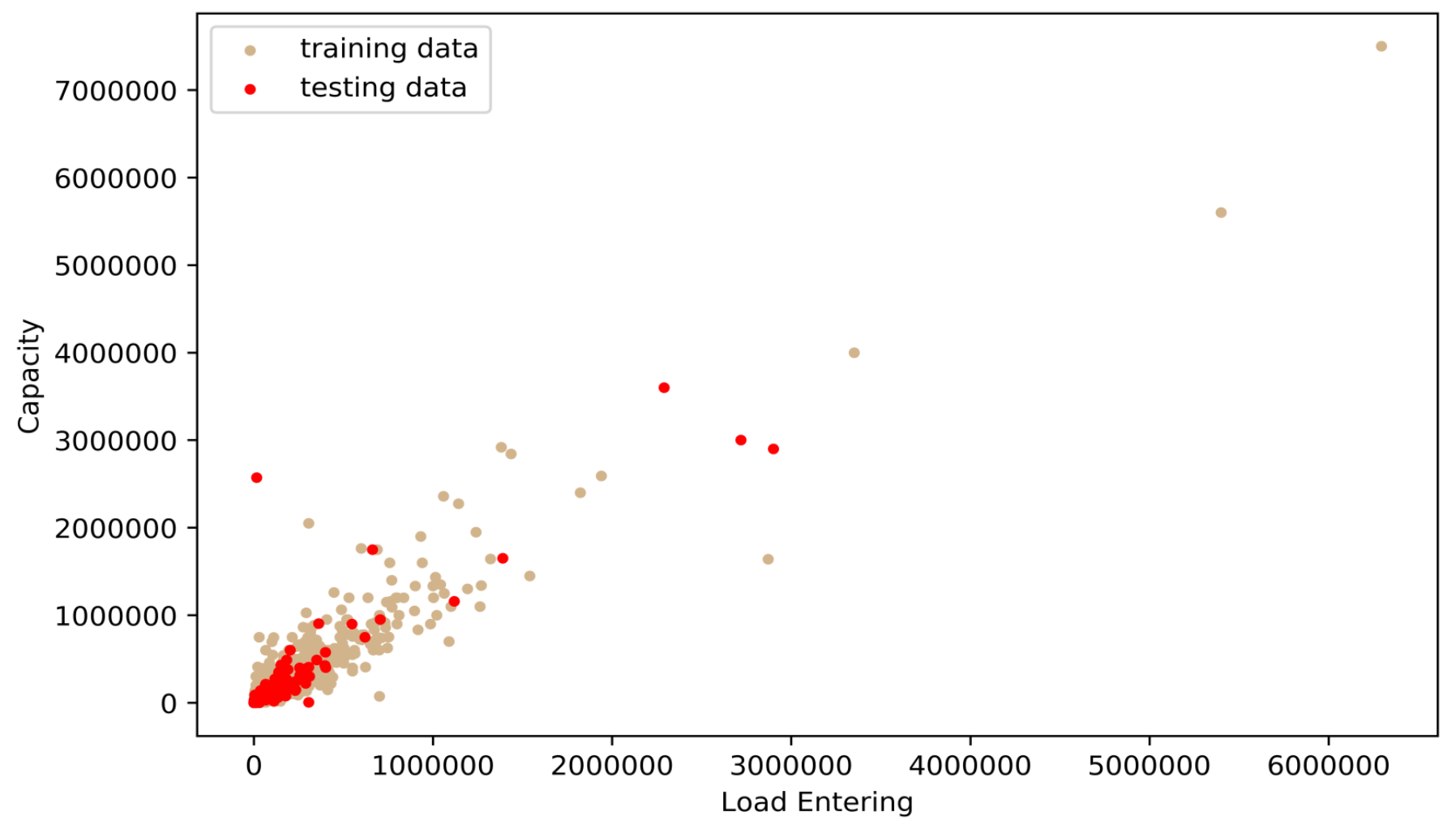


Figure 1. Load entering volume and capacity of treatment plants

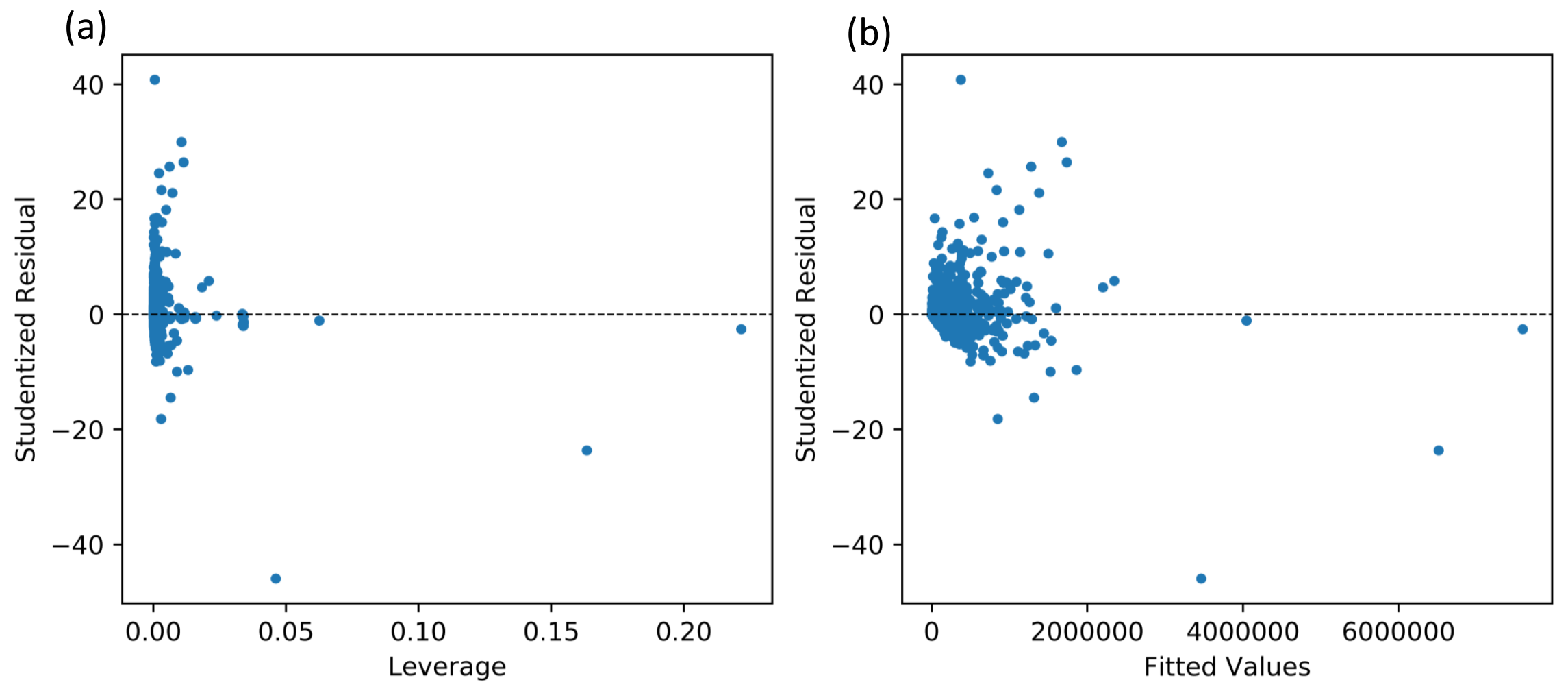
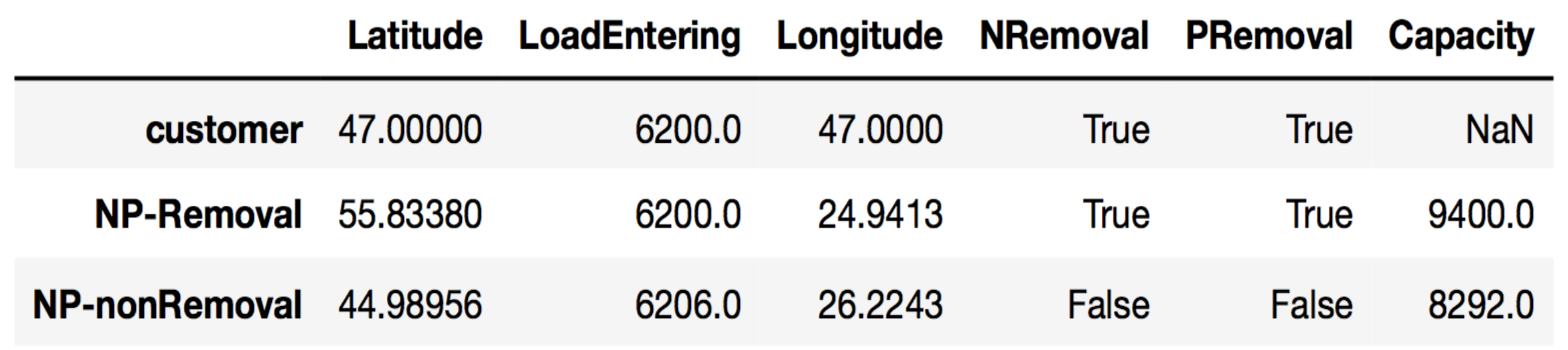


Figure 2. Analysis of multilinear regression a) Studentized Residual with Leverage. b) Studentized Residual with the fitted values.

Ridge Regression

- Although SLR is easy to interpret, the high variance results in the introduction of ridge regression.
- Figure 3 is the result of the prediction. The MSE of RR model is 0.09 and R^2 is 0.91 which is better than SLR model. However, the capacity predicted by RR is slightly lower than observed value while SLR prediction is much higher. We return to the user interface with a range.

Table 1. Nearest neighbor of the given input by the user



	Latitude	LoadEntering	Longitude	NRemoval	PRemoval	Capacity
customer	47.00000	6200.0	47.0000	True	True	NaN
NP-Removal	55.83380	6200.0	24.9413	True	True	9400.0
NP-nonRemoval	44.98956	6206.0	26.2243	False	False	8292.0

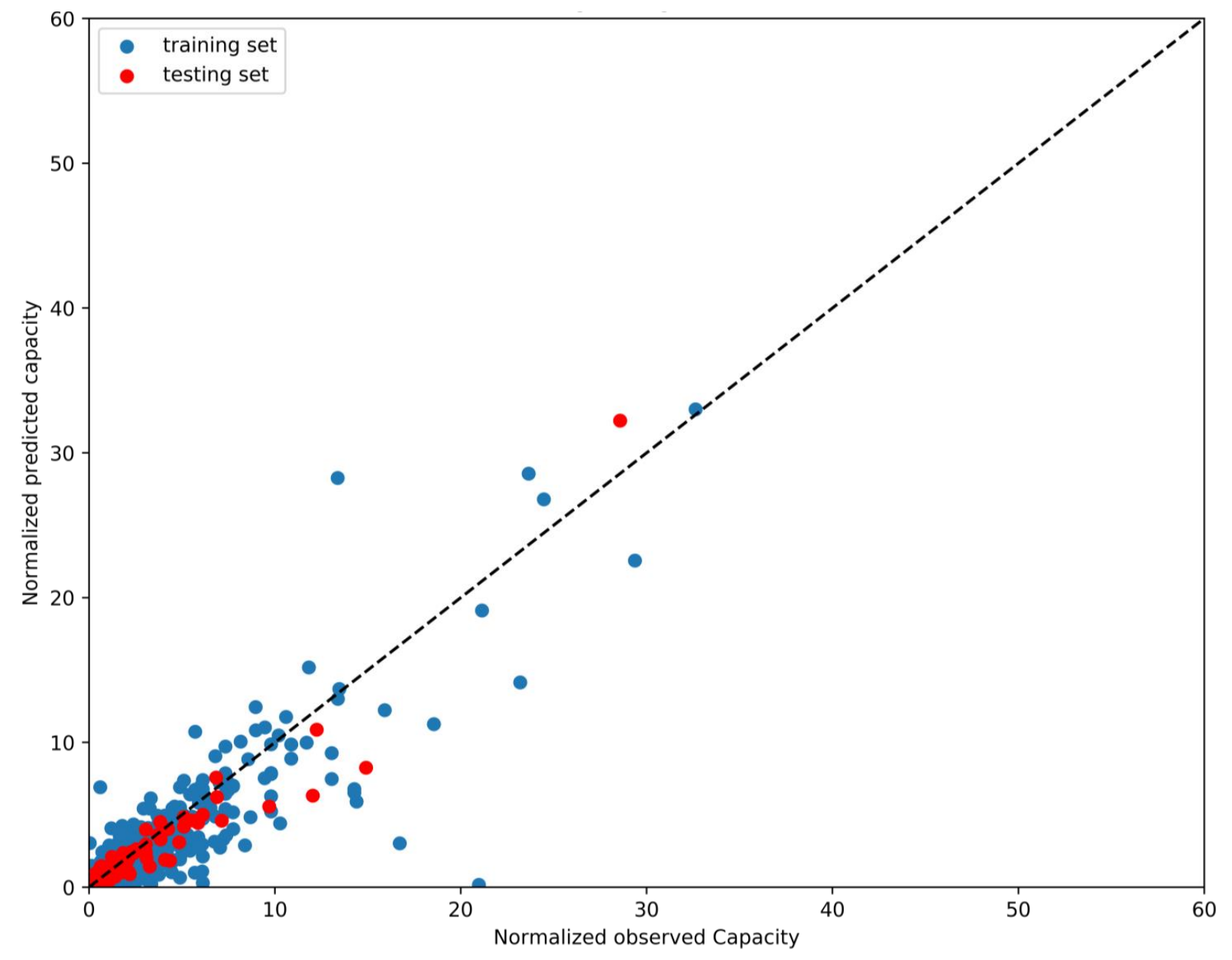


Figure 3. Ridge regression of the normalized observed capacity and normalized predicted capacity.

Nearest Neighbor

- We locate the treatment plants in the database which the users may be interested based on their desire. We use phosphate and nitrogen removal as conditions to filler and spatial function from scipy package to calculate the distance in between.

Data Visualization

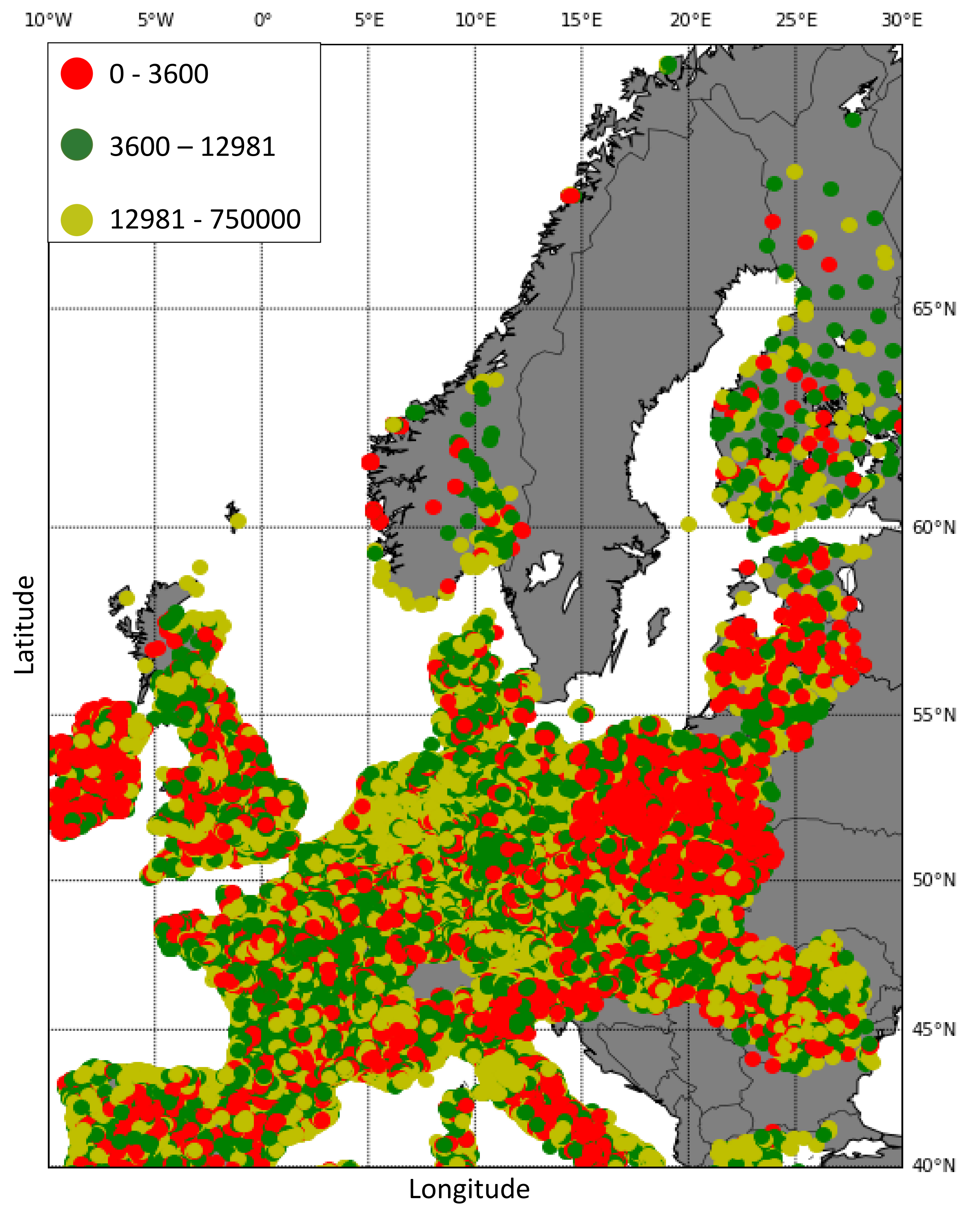


Figure 3. Map of Europe with capacity of waste water treatment plant in groups of color

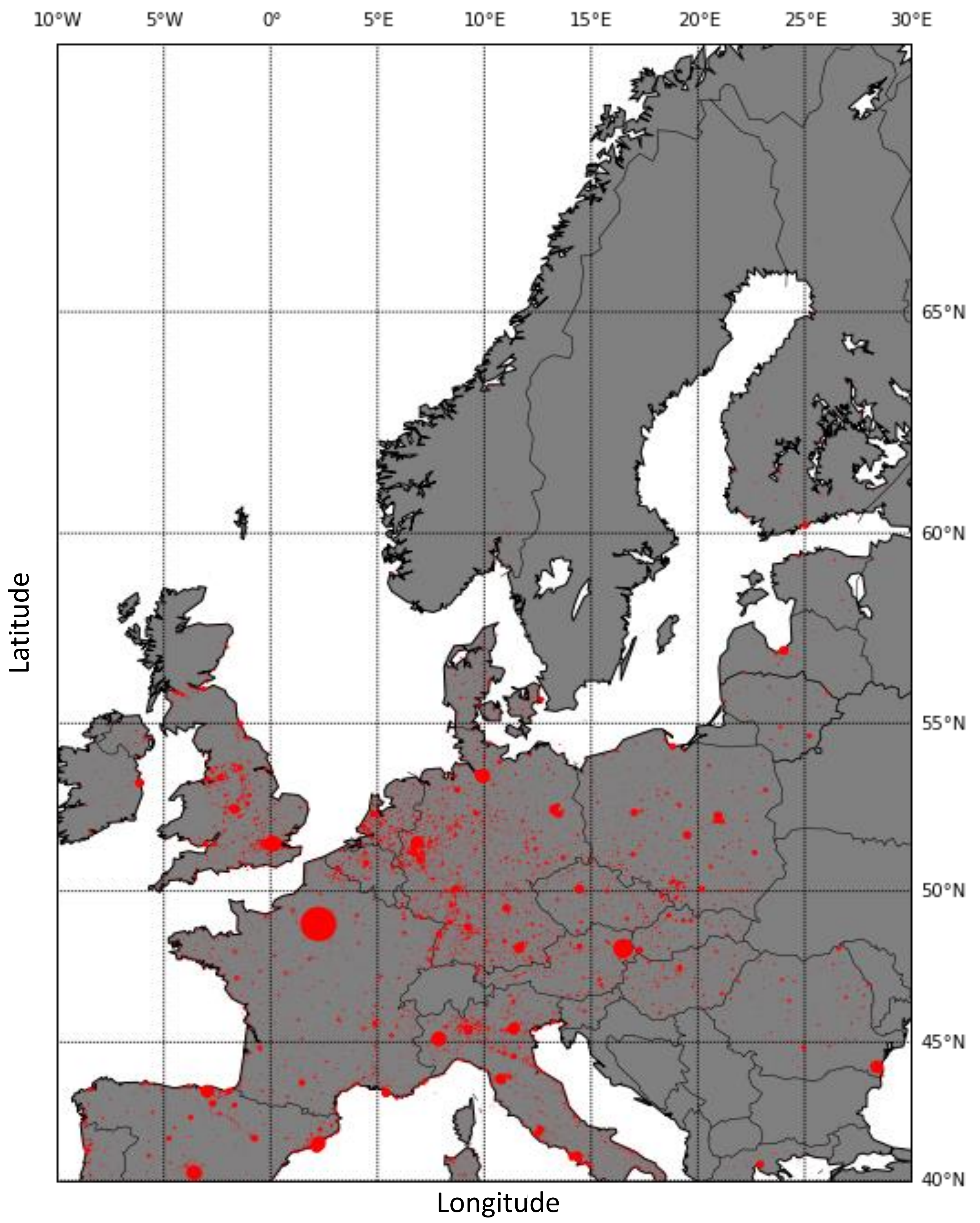


Figure 4. Map of Europe with capacity of waste water treatment plant in size of dots

- The maps was created by folium and mpl_toolkits.basemap.
- The figures allow us to clearly identify the area where most waste water treatment plants are located in Europe and where in Europe still lack the development of urban waste water treatment.

Reference

1. https://www.eea.europa.eu/data-and-maps/data/waterbase-uwwtd-urban-waste-water-treatment-directive-5
2. http://ec.europa.eu/environment/water/water-urbanwaste/implementation/factsfigures_en.htm
3. https://www.eea.europa.eu/data-and-maps/indicators/urban-waste-water-treatment/#tab-data-references-used

DIRECT
Data Intensive Research
Enabling Clean Technologies

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

European Environment Agency

CHEMICAL ENGINEERING
UNIVERSITY of WASHINGTON