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

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

Background

Waste water treatment is especially vital in large cities where industrial pollution can affect the water supply. In Europe, the Urban Waste Water Treatment Directive mandates that areas where there is sufficient concentration of industrial waste there needs to be a waste water treatment plant. While some countries have over 90% of population receiving water treatment, southern European countries have only reached about 70% of implementation. Approximately, €22 billion investment is needed to achieve the implementation of the urban waste water treatment, and the total yearly investment in the renewal, improvement, and extension of existing infrastructure is expected to reach €25 billion per year. The database from the European Environment Agency on urban waste water treatment was used to create a model that predicts capacity for new plants and data visualization on the size and location of existing water treatment infrastructure.

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, Figures 2 and 3 were plotted to analyze high leverage points and outliers of our dataset.

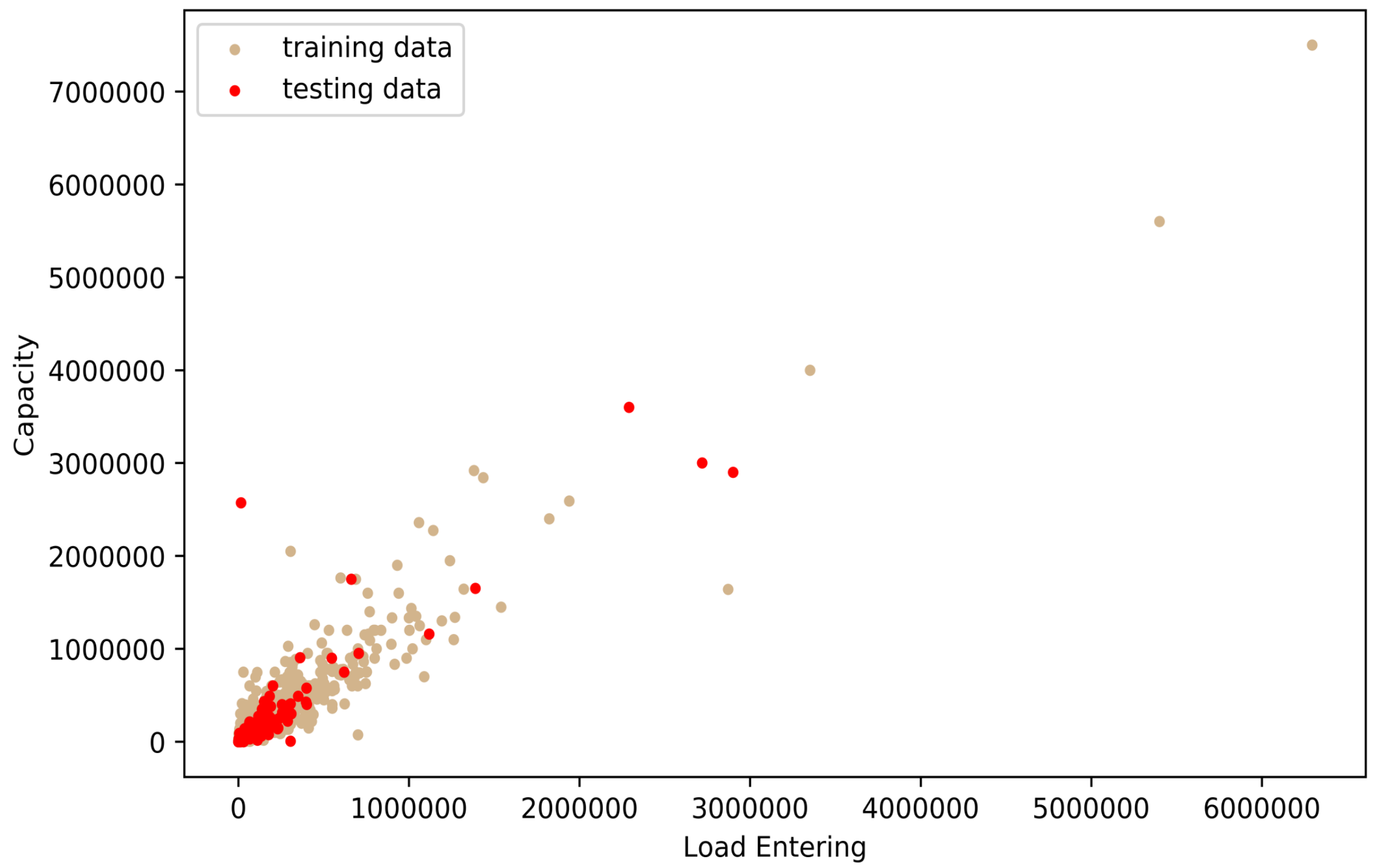


Figure 1. Load entering and capacity of treatment plants.

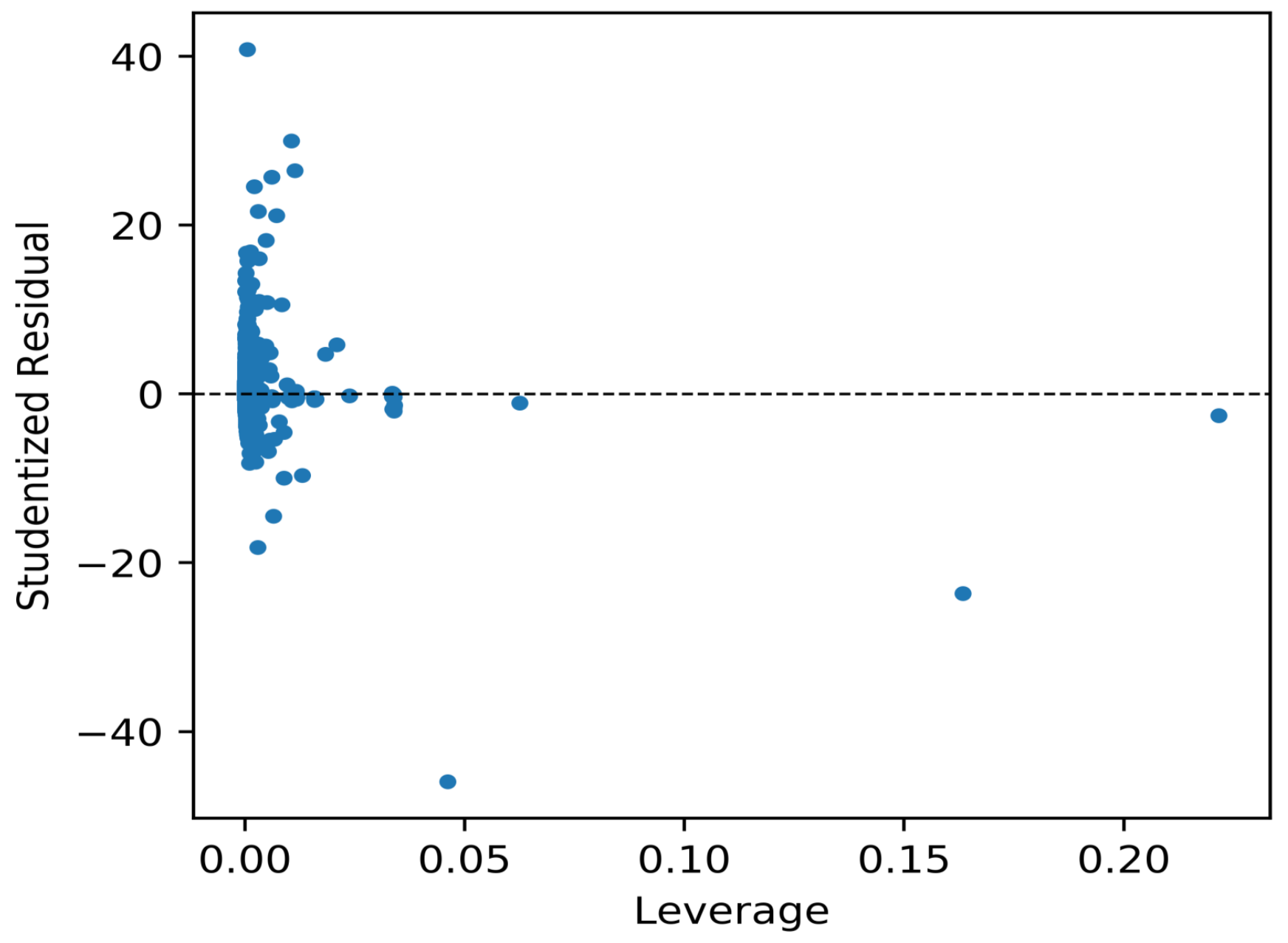


Figure 2. Analysis of multilinear regression of Studentized Residual with Leverage.

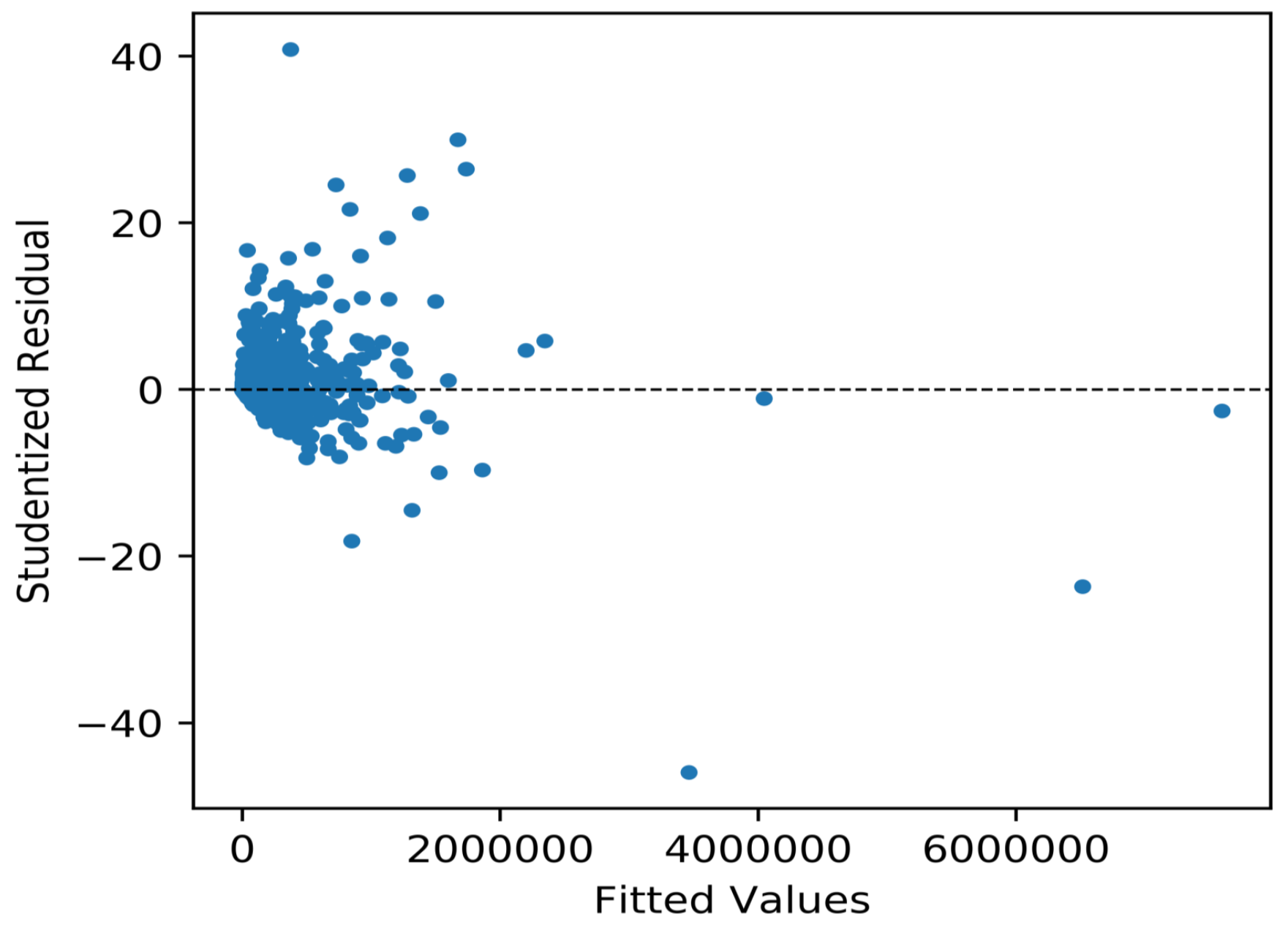


Figure 3. Analysis of multilinear regression of Studentized Residual with the fitted values.

Ridge Regression

Although SLR is easy to interpret, the high variance results in the introduction of ridge regression. Figure 4 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.

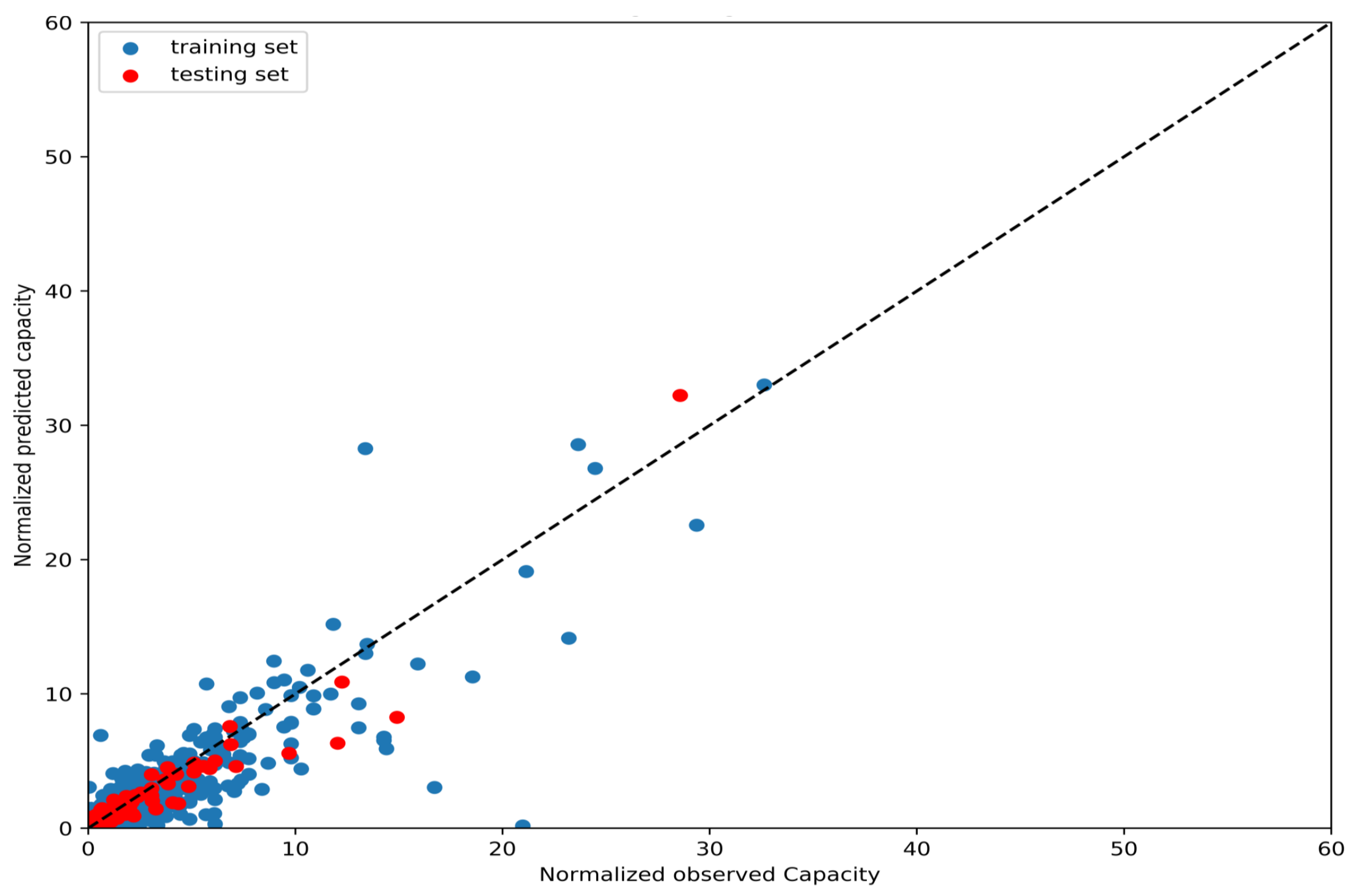
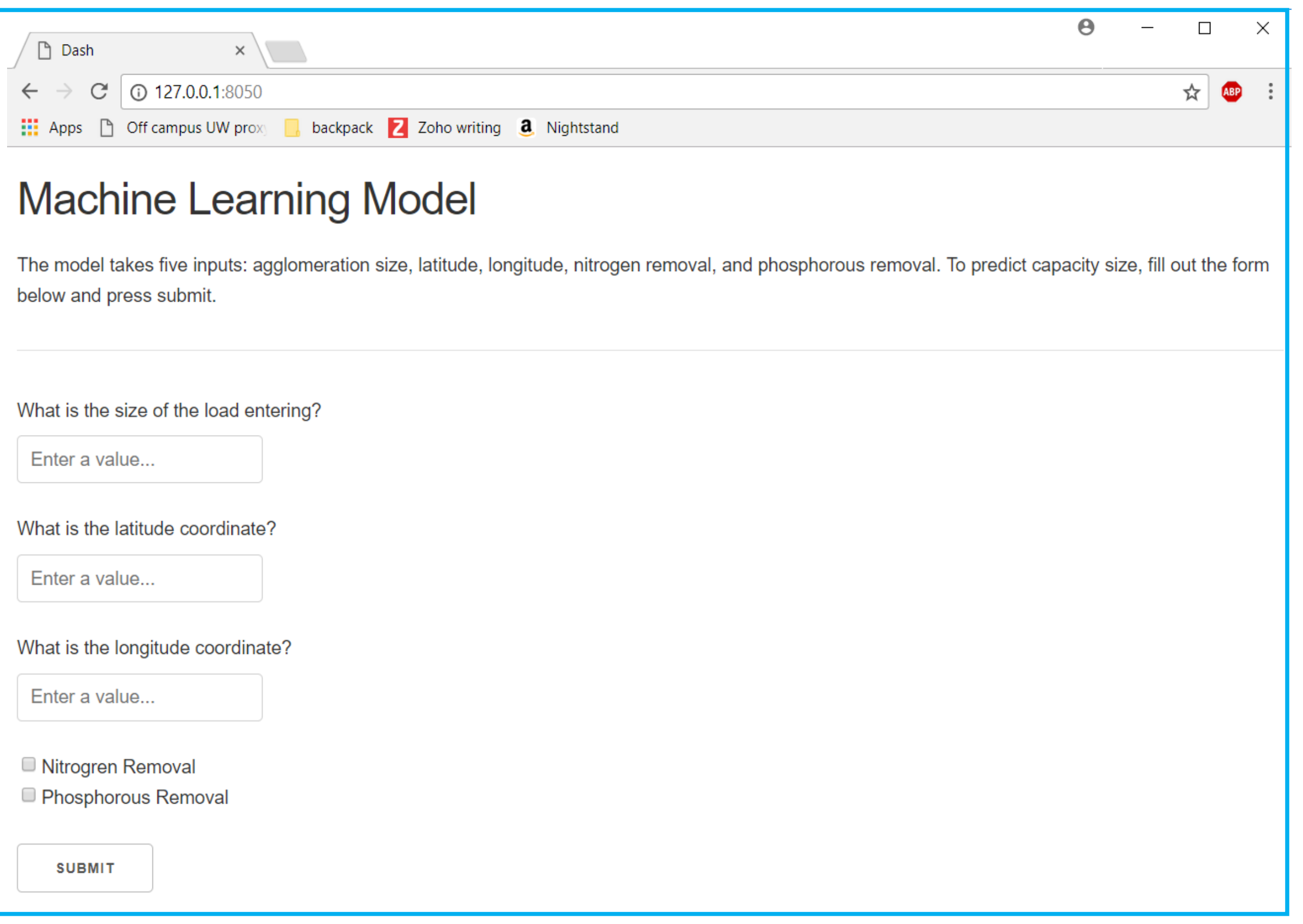


Figure 4. Ridge regression of the normalized observed capacity and normalized predicted capacity.

Nearest Neighbor

The nearest neighbor model locates the treatment plants in the database which the users may be interested based on their input features. The phosphate and nitrogen removal are used as filter conditions, and the spatial function from scipy package is called to calculate the distance in between.



Machine Learning Model

The model takes five inputs: agglomeration size, latitude, longitude, nitrogen removal, and phosphorous removal. To predict capacity size, fill out the form below and press submit.

What is the size of the load entering?

Enter a value...

What is the latitude coordinate?

Enter a value...

What is the longitude coordinate?

Enter a value...

□ Nitrogren Removal
□ Phosphorous Removal

SUBMIT

User Interface

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.

Data Visualization

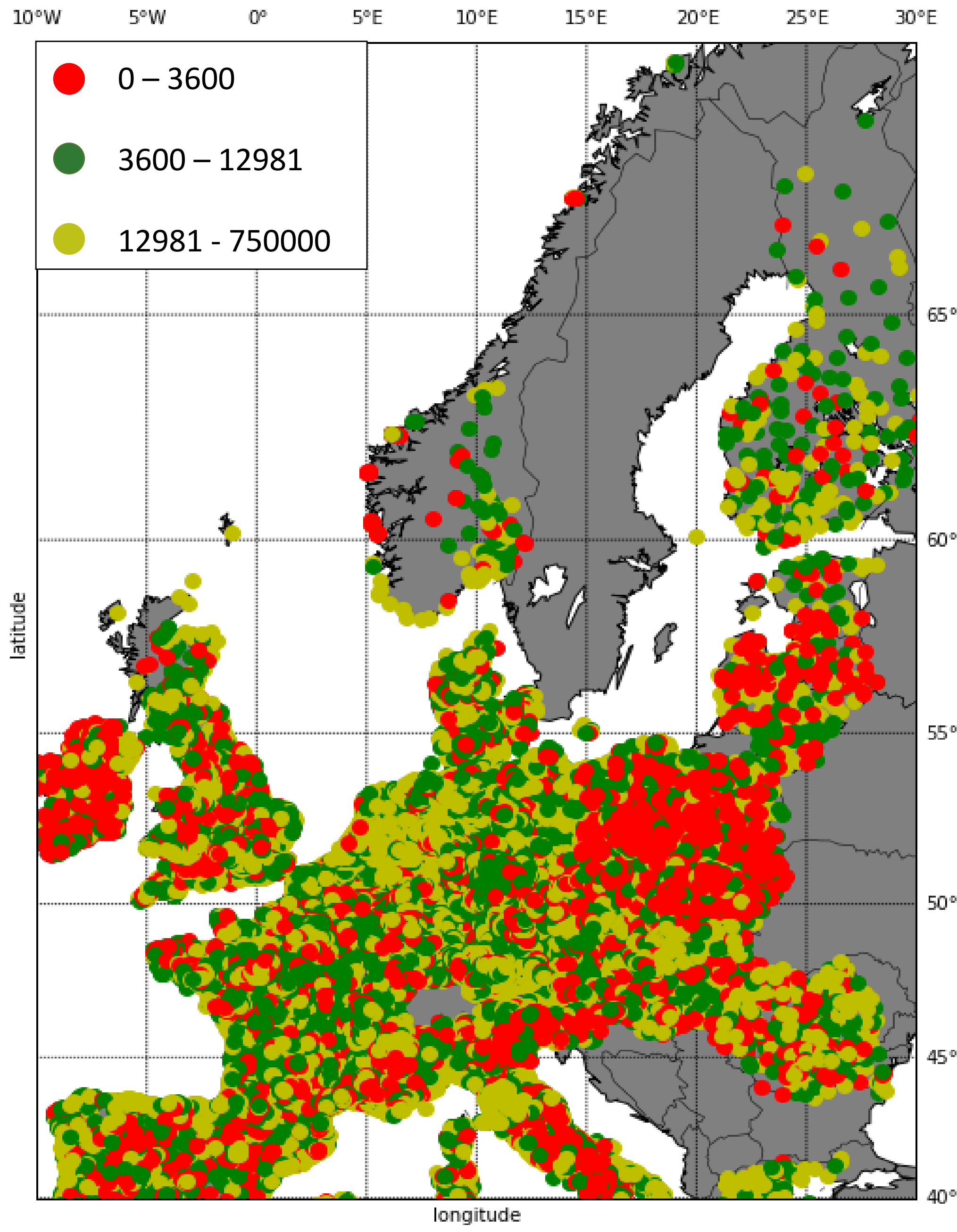
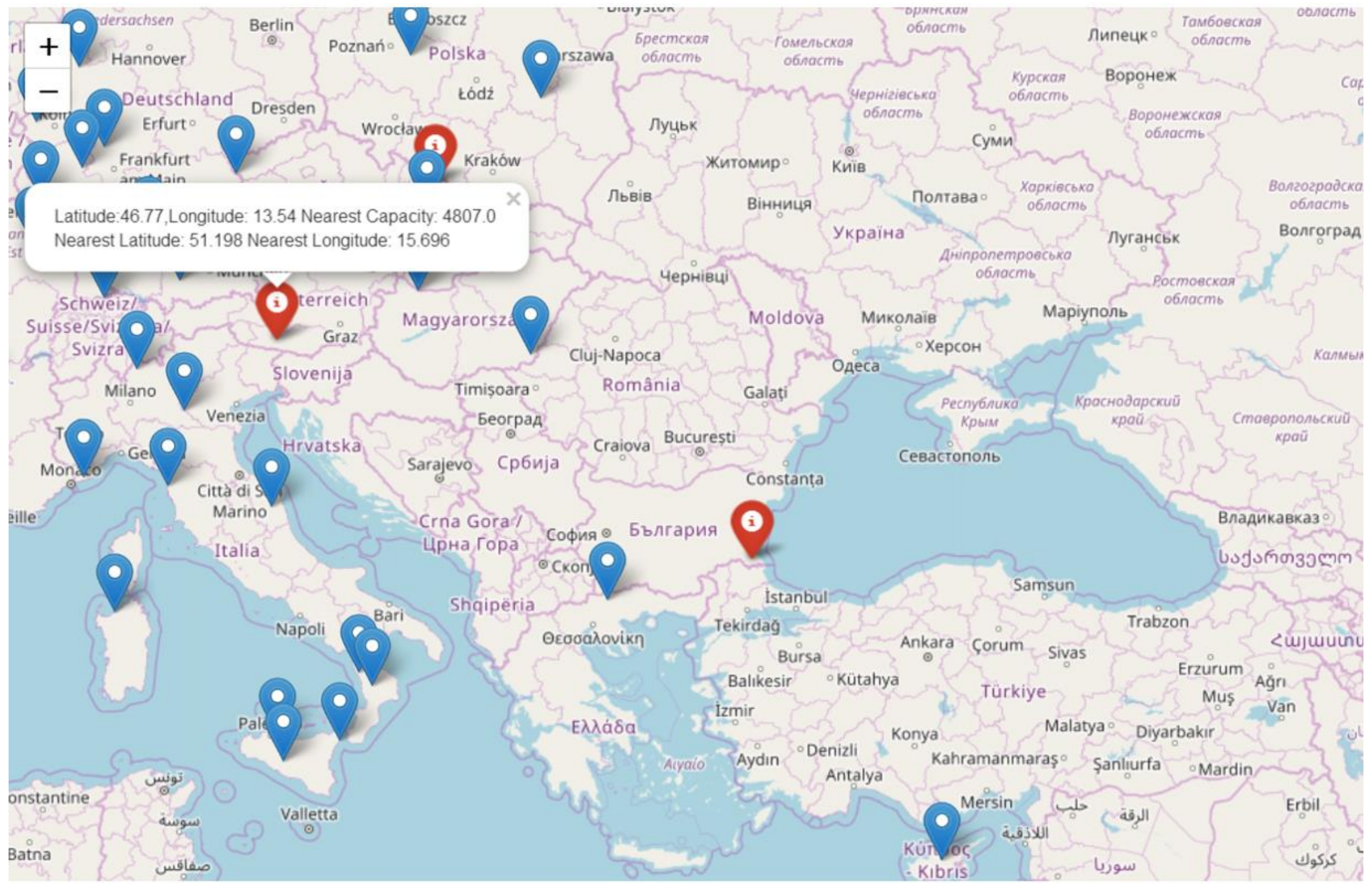


Figure 5. Map of Europe with capacity of waste water treatment plant in groups of color.

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.



Latitude:46.77,Longitude: 13.54 Nearest Capacity: 4807.0 Nearest Latitude: 51.198 Nearest Longitude: 15.696

Figure 6. Interactive map with original data and customer data.

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

	Latitude	LoadEntering	Longitude	NRemoval	PRemoval	Capacity
Customer	51.198	7000.0	15.696	True	True	NaN
NP-Removal	51.198	7000.0	15.696	True	True	4807.0
Non NP-Removal	53.707	7000.0	18.175	False	False	12000.0

References

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