# DS 5500 Final Project Report
# Netherlands Energy Consumption Analysis

Cheng Peng Xuexian Li Yizhen Chen

## Motivation

The Netherlands is known by being one of the countries that want to ban petrol and diesel cars in the coming few years. Last year, Dutch government presented its detailed plan for the coming years and it includes making all new cars emission-free by 2030 – virtually banning petrol-fueled & diesel-fueled cars in favor of battery-powered vehicles. And at the end of 2016, the Dutch government presented its "Energy Agenda" which indicates the policies that should lead to an almost carbon-neutral economy in 2050. So such policies mean there will be an increasing demand of new energy across Netherlands and a growth of electric vehicles as a sustainable transport alternative.

Given this background information, we would like to analyze the energy consumption amount and overall pattern for all the cities in Netherland. We will aim at enlightening questions such as:

(1) How the percentage of the net consumption of electricity changed along the years?
(2) Did more people save energy by using renewable resources?
(3) Find the hidden pattern between gas usage and electric usage.
(4) Find the trend of smart meters spreading.

By investigating and visualizing these points, we hope viewers of our project, especially the general public, may get deeper recognition about the change of energy consumption pattern / structure in Netherlands.

## Data & Preprocessing

The dataset can be found & directly downloaded at Kaggle.com with the following link: [https://www.kaggle.com/lucabasa/dutch-energy](https://www.kaggle.com/lucabasa/dutch-energy).

Every data file is related to an energy company in Netherlands (Enexis / Liander / Stedin) and is saved depending on specific year (2009-2018) as well as specific energy type (gas / electricity). The columns in each file include:

- net_manager: code of the regional network manager

- purchase_area: code of the area where the energy is purchased
- street: Name of the street
- zipcode_from and zipcode_to: 2 columns for the range of zip codes covered, 4 numbers and 2 letters
- city: Name of the city
- num_connections: Number of connections in the range of zip codes
- delivery_perc: percentage of the net consumption of electricity or gas. The lower, the more energy was given back to the grid (for example if you have solar panels)
- perc_of_active_connections: Percentage of active connections in the zipcode range
- type_of_connection: principal type of connection in the zipcode range. For electricity is # fuses X # ampère. For gas is G4, G6, G10, G16, G25
- type_conn_perc: percentage of presence of the principal type of connection in the zipcode range
- annual_consume: Annual consumption. Kwh for electricity, m^3 for gas
- annual_consume_lowtarif_perc: Percentage of consume during the low-tarif hours. From 10 p.m. to 7 a.m. and during weekends.
- smartmeter_perc: percentage of smart meters in the zip code ranges

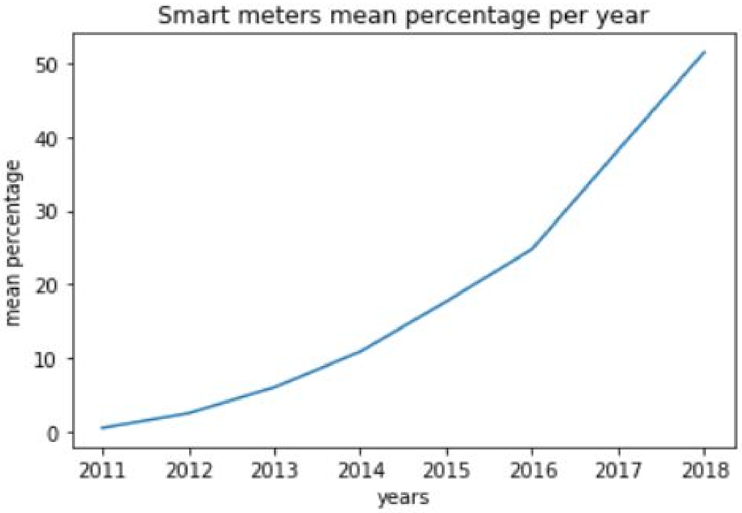
And we are working with both categorical (zip code, street, city, purchase_area, etc) and quantitative (annual_consume, num_connections, delivery_perc, etc) data.

All the data files are in the .csv format (dataframe). This dataset was already cleaned by creator but still has several problems: First, the size of dataset is large: there are multiple files in this dataset and each file represent energy consumption per year per company. Second, Enexis company's gas and electricity files for year 2009 are missing. Third, the "type_conn_perc" and "type_of_connection" values are missing in enexis_electricity_2010.csv and enexis_gas_2010.csv. Fourth, the "smartmeter_perc" column is filled with NaN for all gas files of Enexis company.

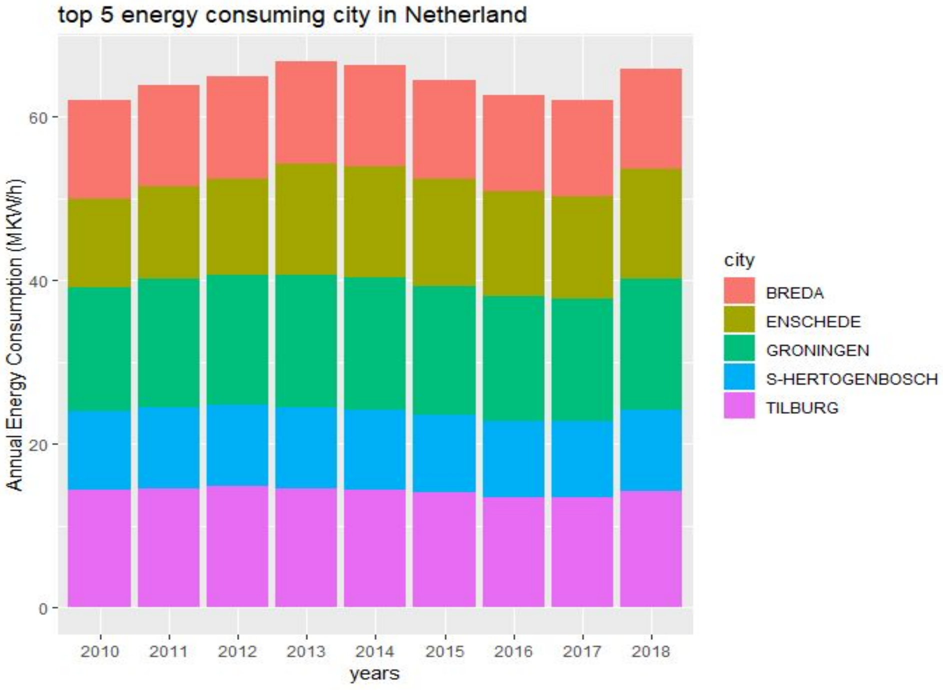
In order to clean the data, we firstly deleted all data files of year 2009 for better comparison. For the missing values of "type_of_connection" column, we looked up the defination in the description file and it told us this column represents the principal type of tube connection and it is hardly ever changed after installation. So we decided to copy the same column on next year's file (enexis_electricity_2011) dataset to fill it up. After filling the "type_of_connection" column by next year's data, there is only "type_conn_perc" column left blank. We decided to perform a simple regression task on that column to fill it up: The training set is enexis gas (or electricity) data from 2011 to 2019 and the testing set is enexis_2010_electricty/gas file. Lastly, for the NaN values under "smartmeter_perc" column for all gas files of Enexis company, we again implemented linear regression to fill them up. The training sets are all of the Liander and Stedin company's gas files from 2010 to 2019 and the testing sets are the corresponding Enexis company's gas files from 2010 to 2019. Besides that, we also applied min-max normalization on all numerical features to ensure singular values will not affect the results.

# Data Analysis

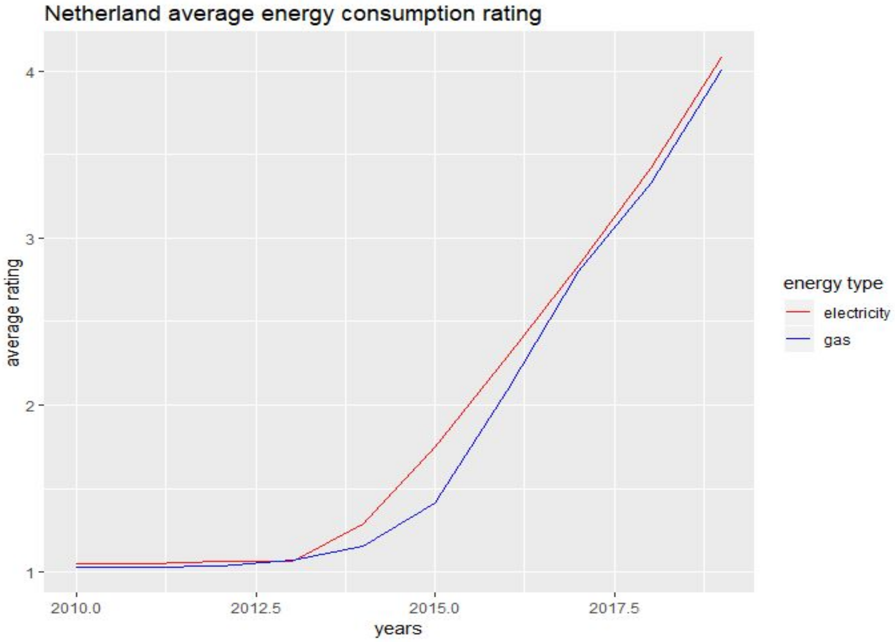
Firstly, we find that mean percentage of smart meters shows clear trend of increasing as the year progressed, indicating that Netherlands indeed follows the new energy policy in some level.



Then we observe the annual consumption amount in top 5 energy consuming cities in Netherland. It is quite exciting that all of these cities are not big cities. They are all industrial cities.



Also, we provide a glimpse of energy rating across Netherlands. The rating is generated by our machine learning algorithm. The blue line represents gas rating and red line represents electricity rating. In this plot we find the average rating grows slowly from 2010 to 2015, then grows rapidly from 2016 to 2019, which means there might be an energy structure reform or new technique implementation in this period.



# Task Analysis

Here is the updated task table:

| Task ID | Domain Task Summary | High-Level Task | Mid-Level Task | Low-Level Task |
|---|---|---|---|---|
| 1 | Examine change of annual energy consumption in each city | Discover | Locate | Compare |
| 2 | Provide a general understanding of Netherland energy consumption | Present | Browse | Summarise |
| 3 | Find the change of smart meter percentage in different cities and years. | Discover | Explore | Compare |
| 4 | Find each city's energy consumption rating | Discover | Explore | Summarise |
| 5 | Find the change of low-tarif energy consumption percentage/consumption per connection in different cities and years. | Discover | Explore | Summarise |

Based on our tasks, our visualizations are primarily developed for "discover" consumption (exploratory visualization). In the above table, most of the Analyze (High-Level) Tasks are categorized as "Discover" after our discussion. Because these tasks all contain keywords such as "change", "trend", "spread", "correlation", which satisfies the definition and expectation of "discover" consumption. And the primary consumer of our visualizations is the general public. All our visualizations are designed in straightforward but impressive styles, which should be easily memorized and comprehended by the public.

# Model Description

The machine learning part in our project is implemented to solve domain task 4. The original dataset does not provide us direct information about energy consumption ratings. So we need to apply some machine learning algorithms to generate it. We tried many different approaches to generate labels but only few of them worked. I will record each machine learning algorithm we tried and discuss them separately.

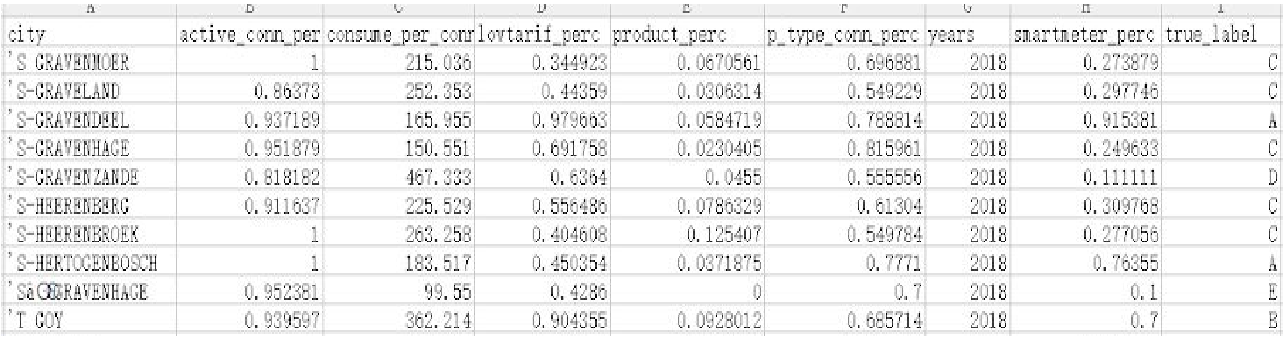
## Label generating

We generated a set of energy rating labels for our machine learning algorithm. These labels are created manually according to the smart meter percentage.

| Smart meter percentage | Rating (actual meaning) |
|---|---|
| < 10% | E (Terrible) |
| 10% to 20% | D (Bad) |
| 20% to 40% | C (Normal) |
| 40% to 70% | B (Good) |
| >70% | A (Excellent) |

Here we didn't separate the smart meter percentage by equal length. Because after we go through all data points we find a pretty interesting fact: Before 2015 almost every city in Netherlands has poor smart meter cover rate. If we label energy rating by using equal length of smart meter percentage, then the result will be more than 95% of cities are labeled as level E. This kind of imbalance will be challenging for our ML algorithm to predict and classify these labels. After we applied the increasing interval, we find the boundaries between different levels become more explicit than previous ones and the results are better.

# Model



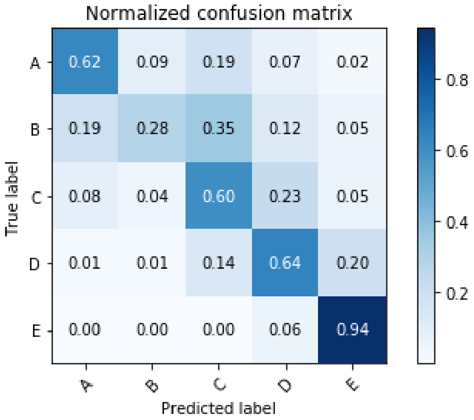
| A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|
| city | active_conn_per | consume_per_conr | lowtarif_perc | product_perc | p_type_conn_perc | years | smartmeter_perc | true_label |
| 'S GRAVENMOER | 1 | 215.036 | 0.344923 | 0.0670561 | 0.696881 | 2018 | 0.273879 | C |
| 'S-GRAVELAND | 0.86373 | 252.353 | 0.44359 | 0.0306314 | 0.549229 | 2018 | 0.297746 | C |
| 'S-GRAVENDEEL | 0.937189 | 165.955 | 0.979663 | 0.0584719 | 0.788814 | 2018 | 0.915381 | A |
| 'S-GRAVENHAGE | 0.951879 | 150.551 | 0.691758 | 0.0230405 | 0.815961 | 2018 | 0.249633 | C |
| 'S-GRAVENZANDE | 0.818182 | 467.333 | 0.6364 | 0.0455 | 0.555556 | 2018 | 0.111111 | D |
| 'S-HEERENBERG | 0.911637 | 225.529 | 0.556486 | 0.0786329 | 0.61304 | 2018 | 0.309768 | C |
| 'S-HEERENBROEK | 1 | 263.258 | 0.404608 | 0.125407 | 0.549784 | 2018 | 0.277056 | C |
| 'S-HERTOGENBOSCH | 1 | 183.517 | 0.450354 | 0.0371875 | 0.7771 | 2018 | 0.76355 | A |
| 'SaGRAVENHAGE | 0.952381 | 99.55 | 0.4286 | 0 | 0.7 | 2018 | 0.1 | E |
| 'T GOY | 0.939597 | 362.214 | 0.904355 | 0.0928012 | 0.685714 | 2018 | 0.7 | B |

We used random forest algorithm as machine learning approach to predict labels. The target variable is the true_label generated by us. And rest columns in the picture are features used to predict the true label. For better results, we treated the column "ture_label", "city" and "year" as dummies rather than numerical values. And we also normalized all numerical columns to avoid small changes dominated by big changes.

Then we adopted these into our model with randomly picking 60% of data as training set and 40% as testing set. Here we tried different set of hyperparameters. This include number of tree estimators from 10 trees to 1000 trees, "Gini" and "Entropy" as criteria to measure the quality of spilt and whether using out-of bag score as a support measure to estimate the generalization accuracy or not. Then we find the best result is generated by 100 trees, using Gini index to measure the quality of spilt and using out-of bag score as a support measure to estimate the generalization accuracy.

# Evaluation Method

After finding the best set of hyperparameters and getting result from random forest classifier, we used the normalized confusion matrix to evaluate the model performance.



## Previous attempts

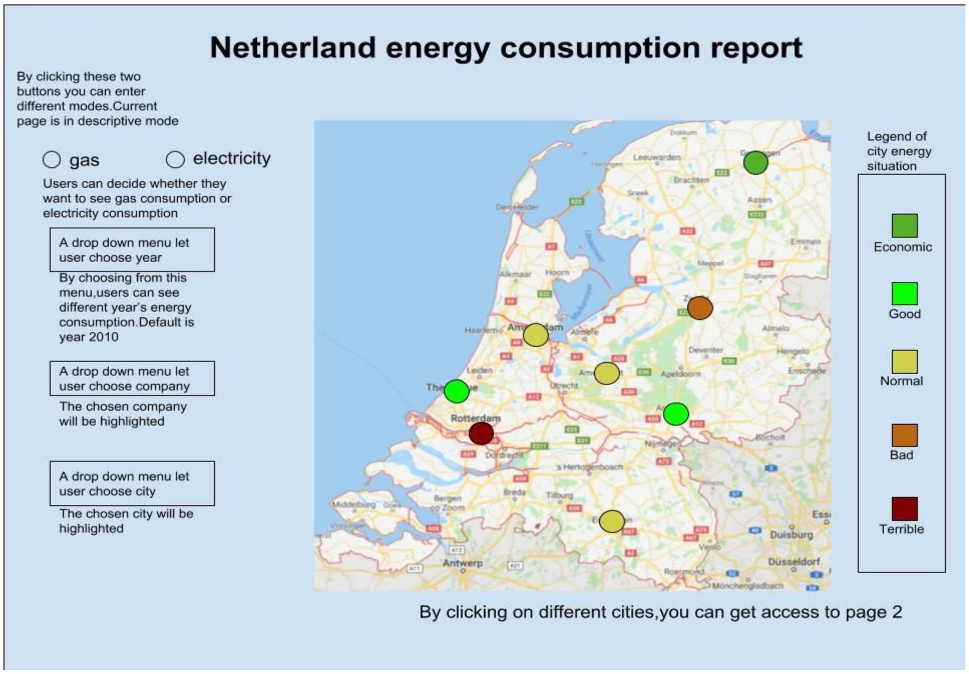
We had tried many different machine learning approaches before we finally decided to use random forest algorithm. At the beginning, the ML task was clustering and we planned to use DBSCAN to find cluster structures from the data. But the result was undesirable. There was only a huge cluster and some single points on the PCA plane. So it is obvious that we can not find any interesting structure from it. After we visualized the whole process we found the DBSCAN kept absorbing data points and never got stopped.

Then we find it is necessary to generate some initial labels before we apply the cluster method. So we generated a very small bootstrap, around 50 data points, and labeled these data randomly. We planned to use K-means algorithm and expected to find some interesting patterns from it. It worked but the result was still poor. The adapted method can generate different labels but it is heavily relied on the initial values. The random labels fail to bring us enough information about energy consumption pattern.

Based on these previous attempts, we found the unsupervised machine learning methods might not be suitable for our project. So we finally decided to change the ML task from clustering to classification problem.

# Design Process

Our original Page 1 sketch was like the below image. This page is designed to solve Task 2 and Task 4. The domain task 2 is to provide a general understanding of Netherland Energy consumption. So we thought the best way is to present visualization on a real Netherland map. In the dataset we have energy consumption data and electricity consumption data. Each kind of data is divided by years, cities and companies. So we decided to get the geocode of each city and draw a circle as mark, then use vertical and horizontal position as channels to control marks (here are the latitude and longitude). And we planned to use radio boxes, drop down menus and buttons to give our user more freedom to decide which' year's data he/she want to see on the visualization. On the other hand, the domain task 4 is about using ML algorithm to classify/label data. So we designed to add color on each circle as marks and channels of visualize energy rating.



Netherland energy consumption report

By clicking these two buttons you can enter different modes.Current page is in descriptive mode

○ gas ○ electricity

Users can decide whether they want to see gas consumption or electricity consumption

A drop down menu let user choose year

By choosing from this menu,users can see different year's energy consumption.Default is year 2010

A drop down menu let user choose company

The chosen company will be highlighted

A drop down menu let user choose city

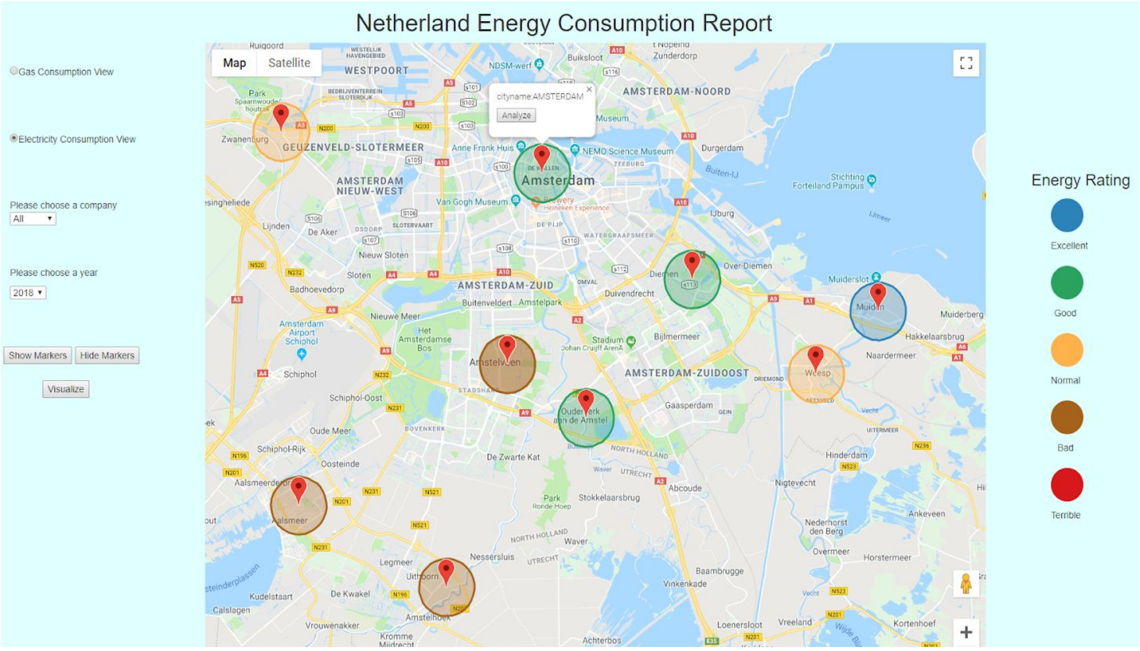
The chosen city will be highlighted

By clicking on different cities,you can get access to page 2

Based on the feedback from Milestone 2 and usability testing, we summarises the suggestions for better visualization on this page:

1. The color gradient of legend for energy rating was said to be lack of logic.
2. We used circles to show the energy rating but the legends are rectangles.This would lead to misunderstanding.
3. The filling color blocks city name and it is hard for users to figure out each city.
4. It would be better to translate all Chinese characters into English.
5. The size of circle is too big and might cause overlapping if two cities are close enough.

And the below image is the final version of Page 1. As mentioned in previous part, the main task of this page is to show the energy consumption rating (represented by five labels/levels) of Netherlands cities. And it is suitable for our Domain Tasks 2 & 4. Here we zoom in the map for a better view:



Netherland Energy Consumption Report

○ Gas Consumption View

● Electricity Consumption View

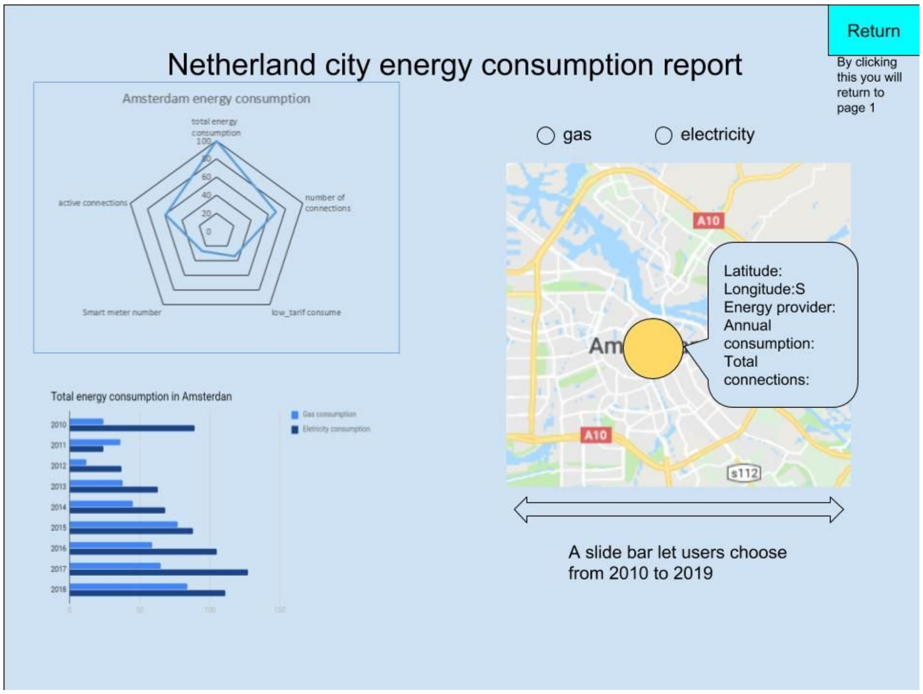
Please choose a company: All

Please choose a year: 2018

After carefully considering the previous advices, we made five main updates for this page:

1. The color of legend (for energy consumption rating) should be diverging and differences between any two colors should be obvious enough for users. So we redesigned the colors: Blue means excellent, green means good, light orange means normal, brown means bad and red means terrible.
2. The color is only used to show the energy rating and it should not block city information. We set the filling color with high transparency so that our users would have better experience when using this page.
3. We changed shape of legend into circle and language to English for consistency.
4. To avoid overlapping, we applied small radius for the circles on each map.

Our original Page 2 sketch was like the below image:



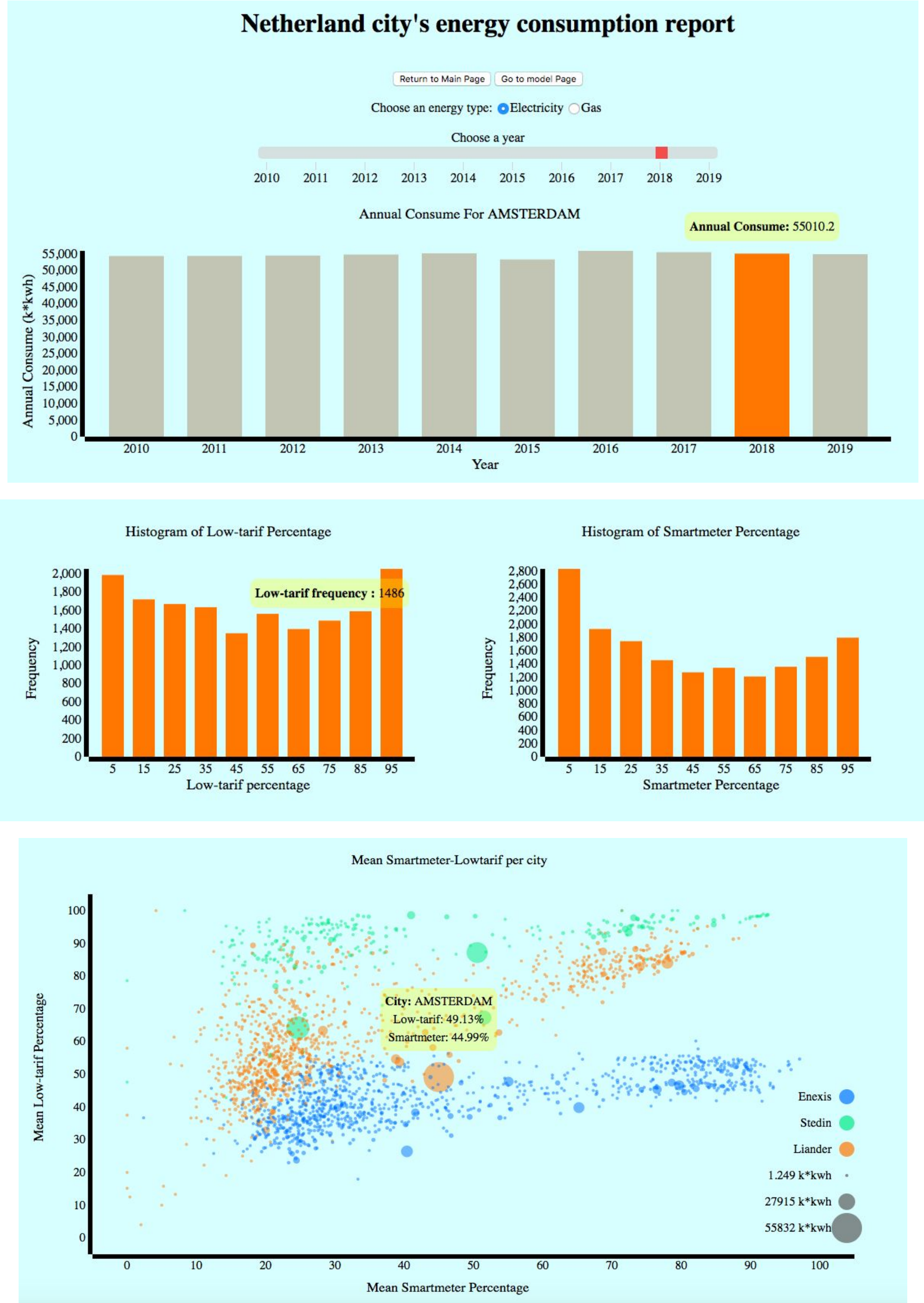
Based on the feedback from Milestone 2 and usability testing, we actually reconstructed this page and constantly made several changes for better visualization:

1. We removed the radar chart since the selected features are all in different units and difficult to be unified then shown on the radar chart.
2. We added a tick under the slide bar (used to select a specific year) for operation convenience.
3. We included a new bar chart to show annual gas/electricity (based on user input) consumption in selected city. We used bar chart since it can show clear comparison of energy consumption among different years. The bars are colored by company names for consistency. When user selects a specific year on the slider, the corresponding bar will be highlighted and other bars will be faded out for “pop-out” effect (some bars might be missing since the dataset doesn’t include corresponding data for that specific years). This visualization is suitable for Domain Task 1.
4. We included a new histogram to either show annual low-tarif consumption percentage in selected city (user input is electricity) or show annual consumption per connection in selected city (user input is gas). We adopted histogram since it is widely used for presenting distribution of continuous variables, which is what we expect to view. Also, when user selects a specific year on the slider, the histogram will be updated accordingly based on data from that year. This visualization is suitable for Domain Task 5.
5. We included a new histogram to show annual smart meter percentage in selected city (works for both user inputs: gas and electricity). We adopted histogram since it is widely used for presenting distribution of continuous variables, which is what we expect to view.

And when user selects a specific year on the slider, the histogram will be updated accordingly based on data from that year.This visualization is suitable for Domain Task 3.

6. We included a new bubble chart to either show each city's mean smart meter percentage vs. mean low-tarif percentage in each year (user input is electricity) or show each city's mean smart meter percentage vs. mean annual consumption per connection in each year (user input is gas). We used bubble chart since it can help us discover correlation between two continuous variables as well as encode one more desired variable (annual consumption amount). The bubbles are colored by company names (with a legend shown) for consistency and the size of the bubbles represents the annual consumption amount (with a legend shown). Also, when user selects a specific year on the slider, the chart will be updated accordingly based on data from that year. This visualization is suitable for Domain Task 2.
7. We enable the mouse hover interactivity for all of the visualizations on this page. Users can examine specific data/information when hovering mouse on the bars or the bubbles in corresponding charts.
8. If the users want to select another city to analyze, they can either go back to Page 1 or take the shortcut provided on this Page. We enable the users to switch city by clicking the bubbles in bubble chart since each bubble represents a city (will show city name when mouse hovering on it). And after clicking any of the bubbles, all visualizations on this page will be updated accordingly.

And here is the final version of Page 2. The main task of this page is to show detailed energy consumption information for specific city in Netherland. So according to the above description, it is suitable for our Domain Tasks 1, 2, 3 and 5:



Netherland city's energy consumption report

Choose an energy type: ● Electricity ○ Gas

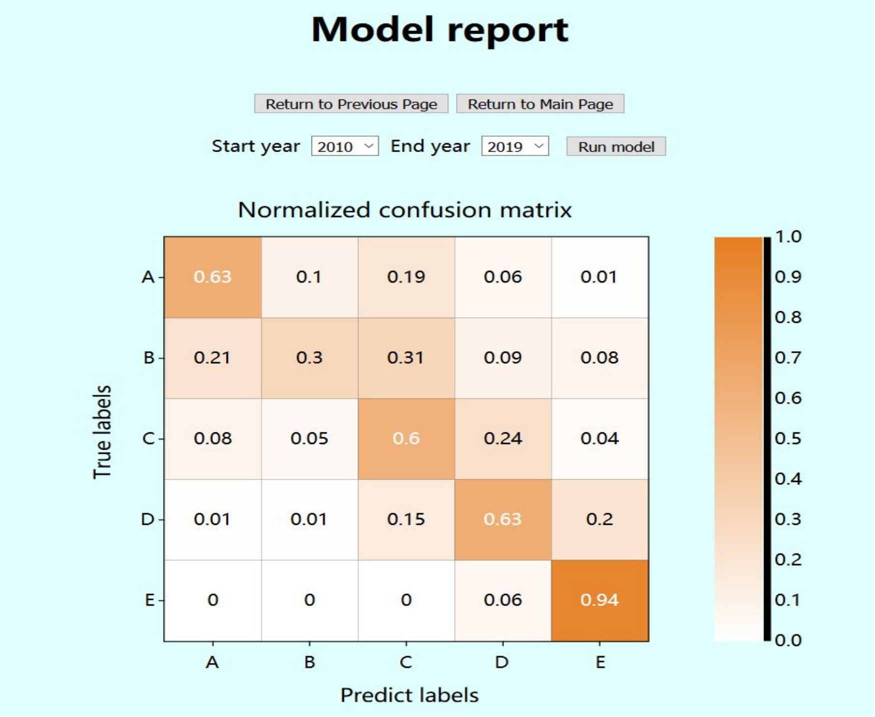
Choose a year: 2018

Annual Consume: 55010.2

Low-tarif frequency : 1486

City: AMSTERDAM
Low-tarif: 49.13%
Smartmeter: 44.99%

And we also created a whole new Page 3. The main task of this page is to show the performance of our machine learning models. We enable users to select start year and end year so that only data from that time period is kept and applied to ML model (we automatically change the year order if the start year is later than end year). Then we adopted the confusion matrix to evaluate model performance since it is widely accepted as the evaluation tool for classification problem. It contributes to solve Domain Task 2 in some level. Here is the final version of Page 3:



Model report

Start year 2010 End year 2019

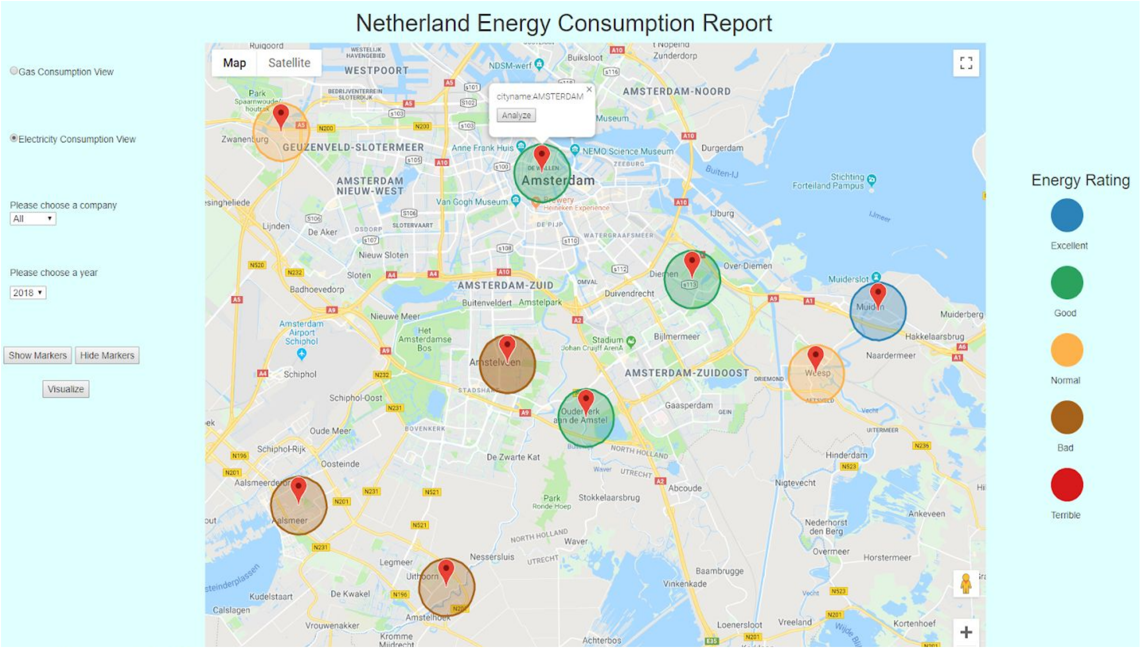
Normalized confusion matrix

| True labels \ Predict labels | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0.63 | 0.1 | 0.19 | 0.06 | 0.01 |
| B | 0.21 | 0.3 | 0.31 | 0.09 | 0.08 |
| C | 0.08 | 0.05 | 0.6 | 0.24 | 0.04 |
| D | 0.01 | 0.01 | 0.15 | 0.63 | 0.2 |
| E | 0 | 0 | 0 | 0.06 | 0.94 |

# Final Visualization

We used D3 and Google Cloud Platform package in our project.

## Page 1 Interactive Designs (From top left to bottom right):



Netherland Energy Consumption Report

○ Gas Consumption View

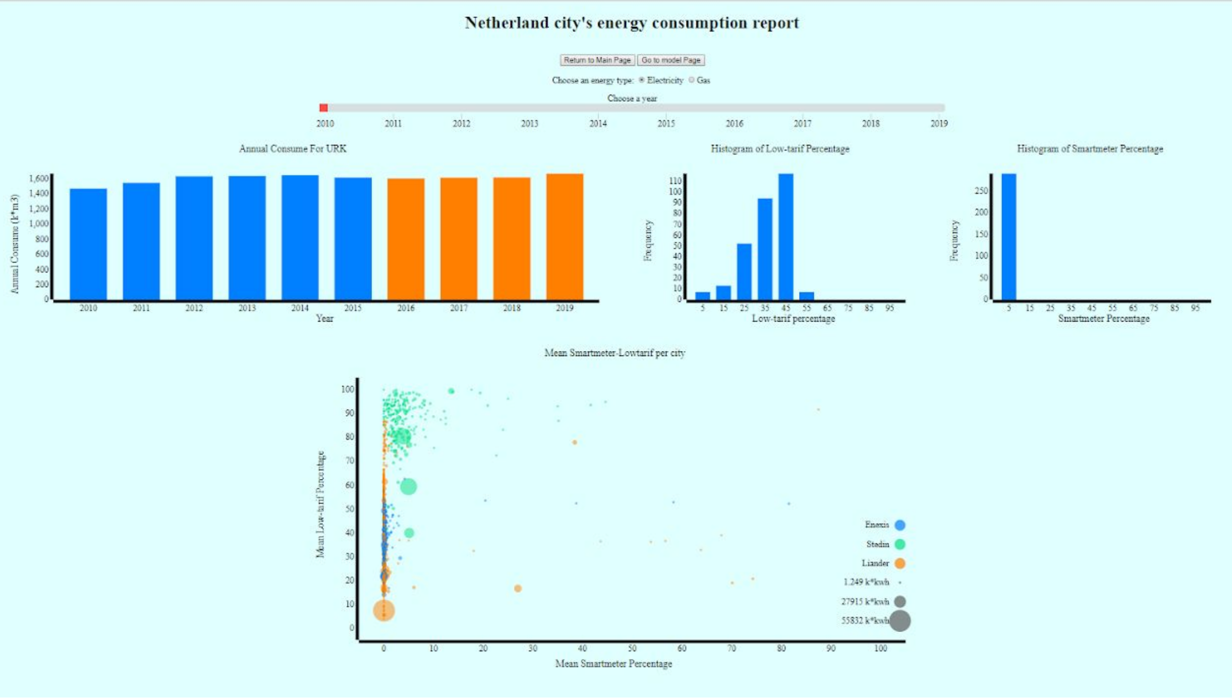
● Electricity Consumption View

Please choose a company: All

Please choose a year: 2018

1. Radio buttons Gas/Electricity consumption view: This is a pair of radio buttons control which energy our users want to see.
2. Drop down menu “please select company”: User can select from three energy provider in Netherlands by choosing company name or see all data by choosing “all”
3. Drop down menu “please select year”: User can decide which year’s data they want to see by selecting from 2010 to 2019.
4. Button show markers: By clicking this user can see hidden markers on the map.
5. Button hide markers: By clicking this user can hide all existed markers on the map.
6. Button Submit: By clicking this our script will collect all parameters from parts 1, 2 and 3 and start to plot on the map. User can submit various combination of parameters to see different types of energy in different years.
7. Markers on the map: By clicking this on the map, users can see a info-window containing city name and a button called “analyze”.
8. Analyze button in the info-window: By clicking this button users will be directed to page 2. Page 2 is the detailed analysis of selected city’s energy consumption.

# Page 2 Interactive Designs (From top to bottom):



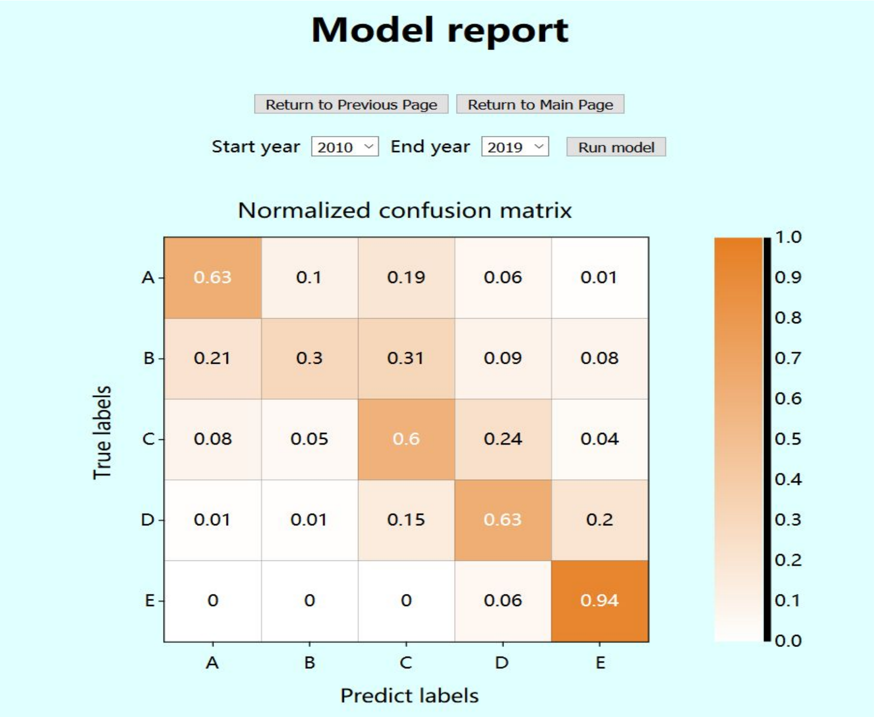
Netherland city's energy consumption report

Choose an energy type: ◉ Electricity ○ Gas

Choose a year

1. Return to main page & go to model page button: Enable users to either go back to Page 1 or direct to Page 3.
2. Radio buttons for choosing energy type: Enable users to choose an energy type, either gas or electricity.
3. Slide bar for choosing year: Enable users to choose a specific year from 2010 to 2019.
4. Mouse hover interactivity: Enable users to examine specific data/information when hovering mouse on the bars or the bubbles in corresponding charts.
5. Click interactivity: Enable users to switch city by clicking the bubbles in bubble chart, as mentioned in previous part.

## Page 3 Interactive Designs (From top to bottom):



Start year 2010 End year 2019

Normalized confusion matrix

| True labels \ Predict labels | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0.63 | 0.1 | 0.19 | 0.06 | 0.01 |
| B | 0.21 | 0.3 | 0.31 | 0.09 | 0.08 |
| C | 0.08 | 0.05 | 0.6 | 0.24 | 0.04 |
| D | 0.01 | 0.01 | 0.15 | 0.63 | 0.2 |
| E | 0 | 0 | 0 | 0.06 | 0.94 |

1. Return to previous page & Return to main page button: Enable users to either go back to Page 2 (previous page) or go back to page 1 (main page).
2. Drop down menu for start year and end year: Enable users to select start year and end year so that only data from that time period is kept and applied to ML model (we automatically change the year order if the start year is later than end year).

## Link to the running demo:

Users can find our demo video either on Github:
https://github.com/cyz5469872/DS5500-Final-Project/blob/master/Demo_video.mp4

Or on Google Drive:
https://drive.google.com/drive/u/1/folders/1XgiMwXDh60YC8hEo4tsDOJLwI_5IR53A

# Conclusion & Future Work

Based on our project, we find Netherlands government and energy providers have been making great effort on reconstructing and reforming the structure of Netherlands energy consumption. On one hand, the average energy consumption amount is becoming higher and higher. On the other hand, the overall energy consumption rating is becoming better and better. The average rating from slightly better than E in 2010 to around B in 2019. There are three reasons on why Netherlands can make such a huge leap. The first reason is Netherlands energy providers emphasize on energy management. The use of smart meters can provide more detailed information of each user's energy consumption history, which makes energy providers have more efficiency on energy control and budget. The average percentage of smart meter raises from 0.1% in 2010 to about 60% in 2019. The second reason is Netherlands has been encouraging people to use clean energy rather than traditional energy since 2016. Based on the average energy rating plot we provided in data analysis part, we can find there is a rapid growth after 2016. And when we search for supplement material, we find Netherlands government established a large amount of wind power plants in 2016 so that more people are willing to set up solar panels in their houses. Besides, the Netherlands government also signed an act about stop using petrol and diesel in the coming years. The last reason is that energy consumption in low tarif period (from 10 P.M. to 6 A.M.) has been increased significantly since 2010, which means a more balanced / reasonable distribution of energy consumption .

For the improvement / future work, we came up with several directions: First, we expect to adopt an alternative criterion to generate labels for consumption rating. In our present project, we generate the rating labels only based on smart meter values since the original dataset doesn't have other useful label-related information. So we hope to find out another criterion or supplement datasets that contain related data, which should be helpful for providing more persuasive labels. Second, we expect to implement dynamic filtering on the Page 1 like we saw in the "zip code search" case, keeping narrowing down locations as users type in zip code. We think this function can contribute more for user interaction and avoid potential visual clutter on the map. Third, we expect to fix some geocoding values in the Google Map for more accurate geographical coordinates. On the current Page 1, several circles are not plotted on the center of corresponding cities because of the inaccurate geocoding values. So we hope this problem can be fixed in future for higher accuracy. Fourth, we hope to include zoom in/out effect for the bubble chart on Page 2. Because some bubbles in the current chart are small and hard to be clicked or examined by users. We expect such zoom in/out function will contribute for operation convenience as well as diminish the disadvantage of overlap.

# Related Works

1. Dutch electricity: EDA, FS, clustering, maps Bojk
2. Using rule mining to understand appliance energy consumption patterns Sami Rollins and Nilanjan Banerjee.
3. Recent Techniques of Clustering of Time Series Data: A Survey T. Warren Liao, Becky Bolt
4. Hierarchical Clustering of Time-Series Data Streams Pedro Pereira Rodrigues
5. DMCA A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series Michail Vlachos
6. A density based method for multivariate time series clustering in kernel feature space S. Chandrakala