WATER POLLUTION ANALYSIS

Group-12
NaingSoe (13960333)
Tuan Dat Tran (13977666)
Aung Kaung Myat (13727540)
Khun Htun Myint Myat (13921403)

Abstract

Data mining is an area of computer science that is multidisciplinary. It is a computational technique that finds patterns in relatively big data sets by combining artificial intelligence, machine learning, statistics, and databases. And it's been used to sift through massive volumes of data in order to uncover business insight that can help people solve problems, mitigate risks, and capture new possibilities(Investopedia, n. d).

The purpose of this report is to provide our findings, recommendations and solutions to relevant water quality research institutions through data mining. Provide effective solutions and accurate statistics for severely polluted areas.

Our team will use Wake data mining software for data description, data cleaning, data transformation, data reduction, data modelling and other methods to conduct in-depth research on data. Through k-MEAN and DBSCAN studies, we find that water pollution is closely related to national population density. The higher the population density, the more serious the water pollution. To understand the impact and harm of water pollution on our daily life, and put forward our team's suggestions and methods for the areas with serious water pollution. Because we cherish the principle of protecting the ecology and maintaining the healthy life of the river.

The dataset: https://www.kaggle.com/ozgurdogan646/water-quality-dataset

Table of Contents

1 Introduction

Water pollution is when a water source is contaminated with substances that make it impossible to use for drinking, cooking, cleaning, swimming and other activities. Pollutants include chemicals, garbage, bacteria and parasites. All forms of pollution end up in the water (HSPH, n. d).

First, we choose an area with water pollution problem, and we have to solve and dig into this problem. Before deciding to study water quality analysis, we must find suitable data sets to select water quality data sets, because some data do not have relevant information, such as the area of water source and the time period of analysis. In the whole report, we will use data description, data cleansing, data transformation, data reduction, data modelling and other methods to conduct in-depth research on data step by step.

According to these methods, K-MEAN and DBSCAN can help us to know the areas and regions with serious water pollution more quickly and better. Results and charts for critical relational water quality analysis will also be obtained. During the modelling phase, we used the country and population density of the dataset to map areas with high water pollution. So we can categorize this data to change the local environment. The report will also be broken down by water pollution type and severity, so that local governments can provide information when improving the environment.

1.1 Business Scenario

Our business vision is to use the data mining services we provide to solve local water pollution problems. Because we know the harm and severity of water pollution, and water pollution is not easy to control, which requires a lot of manpower, material resources and time to deal with, our team hopes to provide useful data to relevant organizations through data mining and improve efficiency. In this case, we worked with the relevant surface water research team, which provided us with 20,000 water records and 29 data sets of water quality attributes. With these data, our team will use Weka data mining software to further understand the relationship between these water resources and give some suggestions and solutions based on our results.

2 Method

2.1 Min-Max Normalization

Min-max normalization is one of the most common methods of data normalization. For each function, the minimum value of the function is converted to 0, the maximum value to 1, and each other value is converted to a decimal between 0 and 1. Because it is an essential step in the feature processing process. Data standardization is to eliminate the dimensional influence of different indicators and facilitate the comparability between indicators. Dimensional differences will affect the results of distance calculation in some models （Codecademy, n. d).

2.2 k - means

K-means clustering is a basic unsupervised learning method. It works based on a simple algorithm that divides a given data set into clusters represented by the letter "K," which is predetermined. In this report, we use K-means to position clustering as points, all observation points or data points are associated with the most recent clustering, calculate and adjust, and then repeat the process with new adjustments until the desired results are obtained (Techopedia, n. d).

2.3 DBSCAN

DBSCAN is a density-based clustering technique that works on the premise that low-density regions in space separate dense regions in space. In "dense grouping," it groups data points into a single cluster. By examining the local density of data points, it can detect clusters in big spatial data sets. The DBSCAN cluster's most fascinating characteristic is that it excels at detecting outliers. Therefore, it helps us to find noise data and analyze water pollution (Analytics Vidhya, 2020).

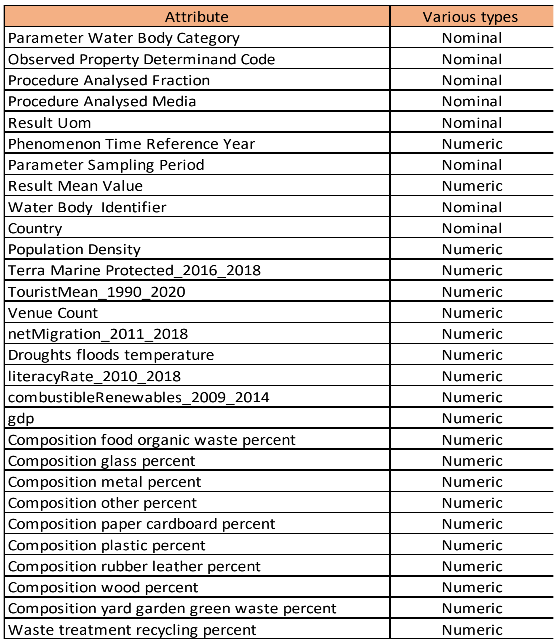
3 Data Pre-processing and Description

3.1 Dataset Used for the Report

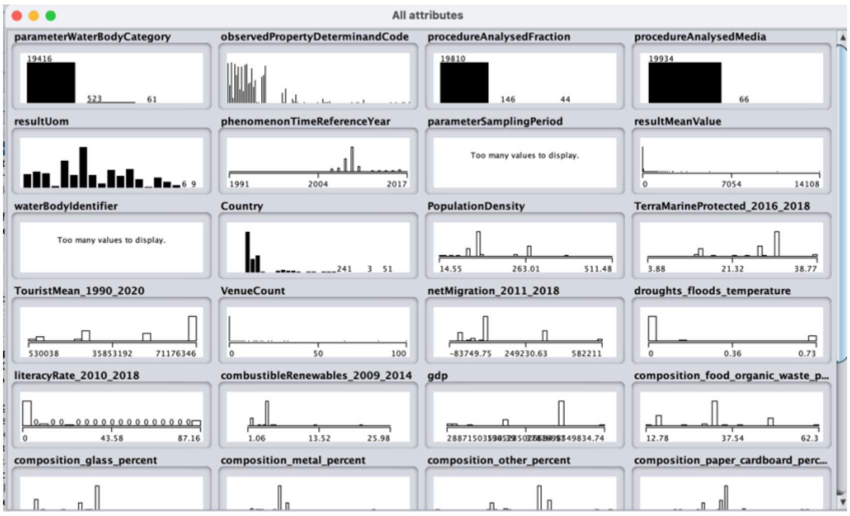
The data set is based on water-based water quality data shared on EAA and edited, deleted and optimized by Kaggle. Source： https://www.kaggle.com/ozgurdogan646/water-quality-dataset

3.2 Dataset Description

The new dataset contains a large amount of data on water-based water quality, including 20,000 records and 29 attributes. The data set mainly analyzes three categories of water quality, namely River water(RW),Ground water(GW) and Lake water(LW). The other detailed properties are shown in Figure 1 (Kaggle, n. d).



Attribute	Various types
Parameter Water Body Category	Nominal
Observed Property Determinand Code	Nominal
Procedure Analysed Fraction	Nominal
Procedure Analysed Media	Nominal
Result Uom	Nominal
Phenomenon Time Reference Year	Numeric
Parameter Sampling Period	Nominal
Result Mean Value	Numeric
Water Body Identifier	Nominal
Country	Nominal
Population Density	Numeric
Terra Marine Protected_2016_2018	Numeric
TouristMean_1990_2020	Numeric
Venue Count	Numeric
netMigration_2011_2018	Numeric
Droughts floods temperature	Numeric
literacyRate_2010_2018	Numeric
combustibleRenewables_2009_2014	Numeric
gdp	Numeric
Composition food organic waste percent	Numeric
Composition glass percent	Numeric
Composition metal percent	Numeric
Composition other percent	Numeric
Composition paper cardboard percent	Numeric
Composition plastic percent	Numeric
Composition rubber leather percent	Numeric
Composition wood percent	Numeric
Composition yard garden green waste percent	Numeric
Waste treatment recycling percent	Numeric



Too many values to display.

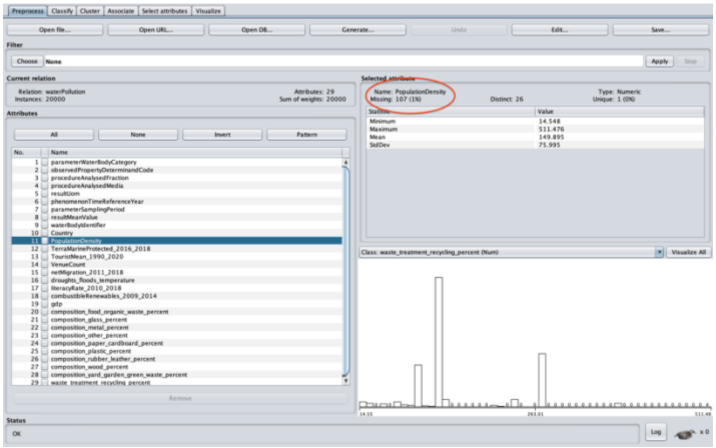
Too many values to display.

Figure-1

3.3 Data Cleaning

Once we get the initial data set, we need to edit the data set with 20,000 records and 29 attributes. Because there are missing values and duplicate values in the data set that need to be deleted and optimized. We use Weka for this series of processes.

First, we load the 'CSV' file into Weka and store it as' arff 'for our future identification, checking each attribute and eliminating missing values. We can observe that many attributes have 107 (1%) missing values, such as attribute 11, as shown in Figure 2. To maximize the accuracy of the study results, we used unsupervised filters named " Remove with Value " and "Remove Duplicates" in Weka, which delete missing values and duplicate values from the data set. After deletion, 19,814 records remain. See Figure 3 and 4.



Filter: None

Current relation
Relation: waterPollution
Instances: 20000
Attributes: 29
Sum of weights: 20000

Selected attribute
Name: PopulationDensity
Missing: 107 (1%)
Distinct: 26
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	14.548
Maximum	511.476
Mean	149.895
StdDev	75.995

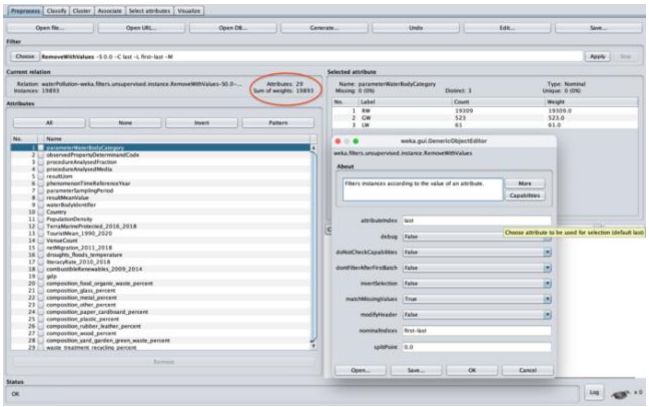
Class: waste_treatment_recycling_percent (Num)

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	phenomenonTimeReferenceYear
7	parameterSamplingPeriod
8	resultMeanValue
9	waterBodyIdentifier
10	Country
11	PopulationDensity
12	TerraMarineProtected_2016_2018
13	TouristMean_1990_2020
14	VenueCount
15	netMigration_2011_2018
16	droughts_floods_temperature
17	literacyRate_2010_2018
18	combustibleRenewables_2009_2014
19	gdp
20	composition_food_organic_waste_percent
21	composition_glass_percent
22	composition_metal_percent
23	composition_other_percent
24	composition_paper_cardboard_percent
25	composition_plastic_percent
26	composition_rubber_leather_percent
27	composition_wood_percent
28	composition_yard_garden_green_waste_percent
29	waste_treatment_recycling_percent

Status
OK

Figure-2



Filter: RemoveWithValues -S 0.0 -C last -L first-last -M

Current relation
Relation: waterPollution-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-...
Instances: 19893
Attributes: 29
Sum of weights: 19893

Selected attribute
Name: parameterWaterBodyCategory
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	RW	19309	19309.0
2	GW	523	523.0
3	LW	61	61.0

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	phenomenonTimeReferenceYear
7	parameterSamplingPeriod
8	resultMeanValue
9	waterBodyIdentifier
10	Country
11	PopulationDensity
12	TerraMarineProtected_2016_2018
13	TouristMean_1990_2020
14	VenueCount
15	netMigration_2011_2018
16	droughts_floods_temperature
17	literacyRate_2010_2018
18	combustibleRenewables_2009_2014
19	gdp
20	composition_food_organic_waste_percent
21	composition_glass_percent
22	composition_metal_percent
23	composition_other_percent
24	composition_paper_cardboard_percent
25	composition_plastic_percent
26	composition_rubber_leather_percent
27	composition_wood_percent
28	composition_yard_garden_green_waste_percent
29	waste_treatment_recycling_percent

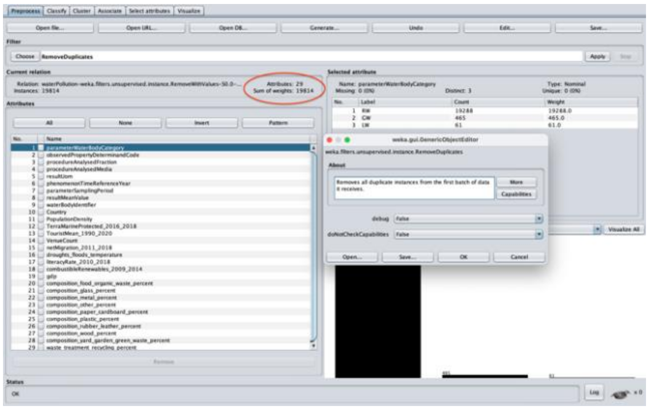
weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveWithValues

About
Filters instances according to the value of an attribute.

attributeIndex: last
debug: False
doNotCheckCapabilities: False
dontFilterAfterFirstBatch: False
invertSelection: False
matchMissingValues: True
modifyHeader: False
nominalIndices: first-last
splitPoint: 0.0

Choose attribute to be used for selection (default last)

Status
OK



Filter: RemoveDuplicates

Current relation
Relation: waterPollution-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-...
Instances: 19814
Attributes: 29
Sum of weights: 19814

Selected attribute
Name: parameterWaterBodyCategory
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	RW	19288	19288.0
2	GW	465	465.0
3	LW	61	61.0

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	phenomenonTimeReferenceYear
7	parameterSamplingPeriod
8	resultMeanValue
9	waterBodyIdentifier
10	Country
11	PopulationDensity
12	TerraMarineProtected_2016_2018
13	TouristMean_1990_2020
14	VenueCount
15	netMigration_2011_2018
16	droughts_floods_temperature
17	literacyRate_2010_2018
18	combustibleRenewables_2009_2014
19	gdp
20	composition_food_organic_waste_percent
21	composition_glass_percent
22	composition_metal_percent
23	composition_other_percent
24	composition_paper_cardboard_percent
25	composition_plastic_percent
26	composition_rubber_leather_percent
27	composition_wood_percent
28	composition_yard_garden_green_waste_percent
29	waste_treatment_recycling_percent

weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveDuplicates

About
Removes all duplicate instances from the first batch of data it receives.

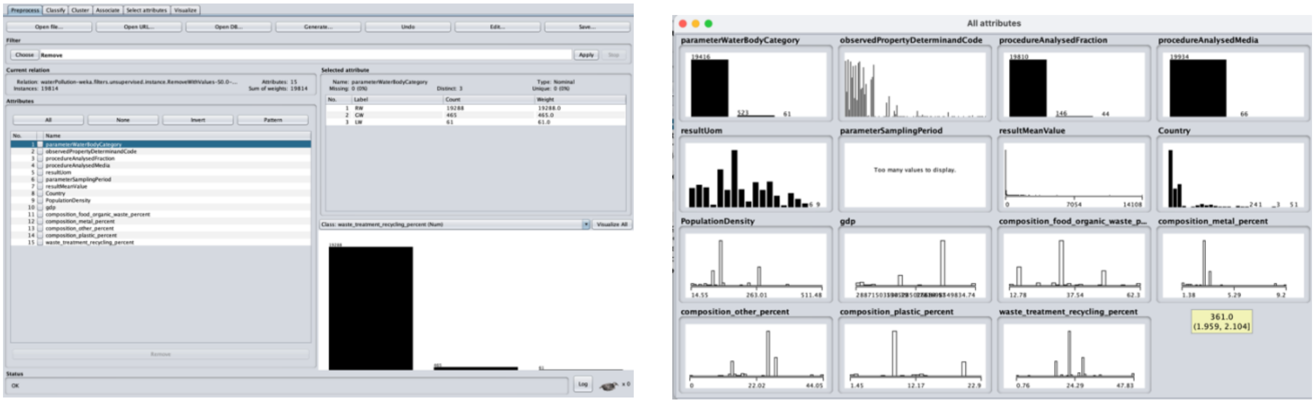
debug: False
doNotCheckCapabilities: False

Status
OK

Figure-3

Figure-4

After editing, we deleted some unnecessary attributes and selected 15 attributes for in-depth research because many attributes were not very relevant to what we wanted to study, and there were also many values missing in it. See Figure 5. To confirm that the current data set is complete, unsupervised filters of "Remove Duplicates" for 15 attributes were implemented again, resulting in a reduction of 2,359 incomplete records as shown in Figure 6.



Remove

Current relation
Relation: waterPollution-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-... Attributes: 15
Instances: 19814 Sum of weights: 19814

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	parameterSamplingPeriod
7	resultMeanValue
8	Country
9	PopulationDensity
10	gdp
11	composition_food_organic_waste_percent
12	composition_metal_percent
13	composition_other_percent
14	composition_plastic_percent
15	waste_treatment_recycling_percent

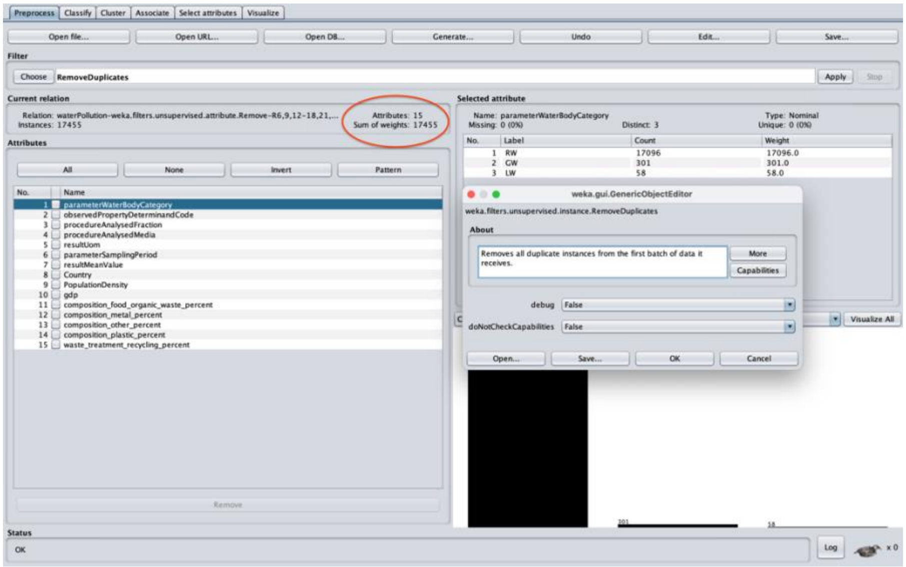
Selected attribute
Name: parameterWaterBodyCategory Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	RW	19288	19288.0
2	CW	465	465.0
3	LW	61	61.0

Class: waste_treatment_recycling_percent (Num)

Status
OK

Figure-5



RemoveDuplicates

Current relation
Relation: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,... Attributes: 15
Instances: 17455 Sum of weights: 17455

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	parameterSamplingPeriod
7	resultMeanValue
8	Country
9	PopulationDensity
10	gdp
11	composition_food_organic_waste_percent
12	composition_metal_percent
13	composition_other_percent
14	composition_plastic_percent
15	waste_treatment_recycling_percent

Selected attribute
Name: parameterWaterBodyCategory Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	RW	17096	17096.0
2	GW	301	301.0
3	LW	58	58.0

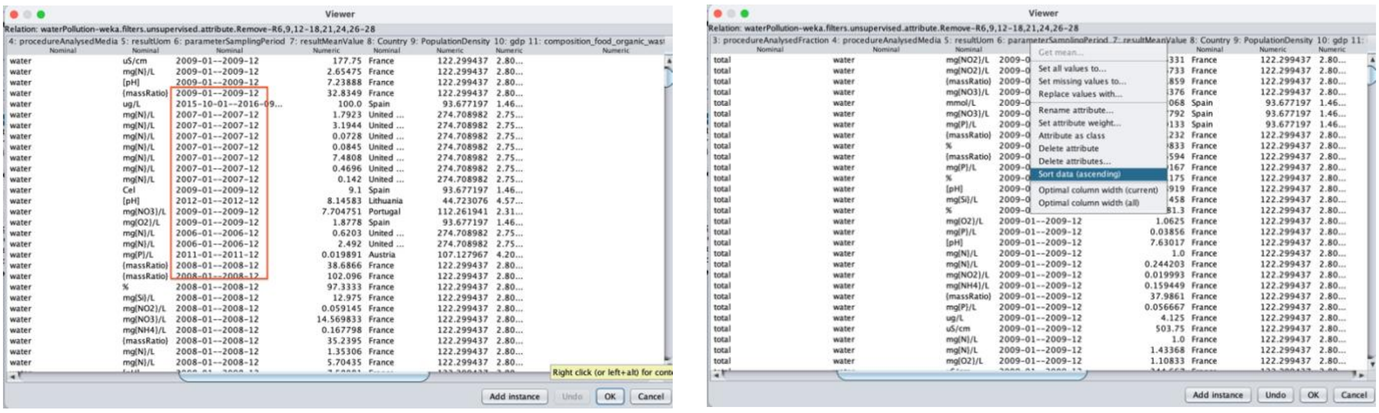
weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveDuplicates
About
Removes all duplicate instances from the first batch of data it receives.
debug False
doNotCheckCapabilities False

Status
OK

Figure-6

3.4 Data Transformation

After the data set is cleaned, we can clearly see that the date of the parameter sampling period is irregular. So we sort the data in 'Edit' by ascending the parameter sampling period. The min-Max Normalization approach is then used to narrow down the data to the specified extent. See Figure 7

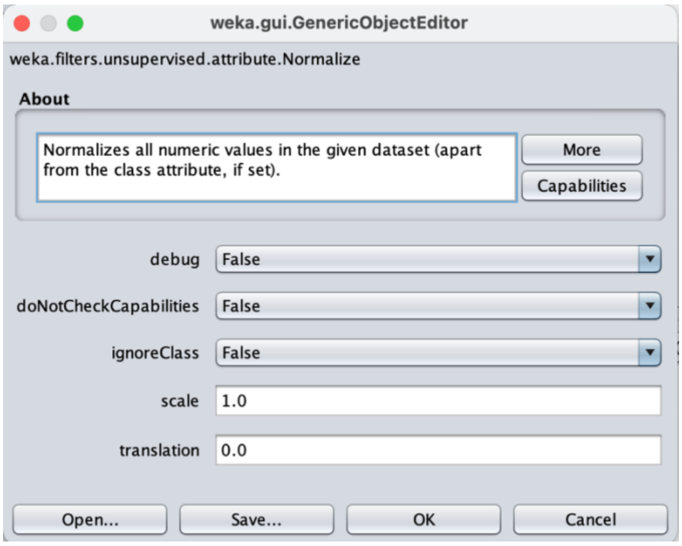


Relation: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28

4: procedureAnalysedMedia Nominal	5: resultUom Nominal	6: parameterSamplingPeriod Nominal	7: resultMeanValue Numeric	8: Country Nominal	9: PopulationDensity Numeric	10: gdp Numeric	11: composition_food_organic_wast Numeric
water	uS/cm	2009-01--2009-12	177.75	France	122.299437	2.80...	
water	mg{N}/L	2009-01--2009-12	2.65475	France	122.299437	2.80...	
water	[pH]	2009-01--2009-12	7.23888	France	122.299437	2.80...	
water	{massRatio}	2009-01--2009-12	32.8349	France	122.299437	2.80...	
water	ug/L	2015-10-01--2016-09...	100.0	Spain	93.677197	1.46...	
water	mg{N}/L	2007-01--2007-12	1.7923	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	3.1944	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	0.0728	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	0.0845	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	7.4808	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	0.4696	United ...	274.708982	2.75...	
water	mg{N}/L	2007-01--2007-12	0.142	United ...	274.708982	2.75...	
water	Cel	2009-01--2009-12	9.1	Spain	93.677197	1.46...	
water	[pH]	2012-01--2012-12	8.14583	Lithuania	44.723076	4.57...	
water	mg{NO3}/L	2009-01--2009-12	7.704751	Portugal	112.261941	2.31...	
water	mg{O2}/L	2009-01--2009-12	1.8778	Spain	93.677197	1.46...	
water	mg{N}/L	2006-01--2006-12	0.6203	United ...	274.708982	2.75...	
water	mg{N}/L	2006-01--2006-12	2.492	United ...	274.708982	2.75...	
water	mg{P}/L	2011-01--2011-12	0.019891	Austria	107.127967	4.20...	
water	{massRatio}	2008-01--2008-12	38.6866	France	122.299437	2.80...	
water	{massRatio}	2008-01--2008-12	102.096	France	122.299437	2.80...	
water	%	2008-01--2008-12	97.3333	France	122.299437	2.80...	
water	mg{Si}/L	2008-01--2008-12	12.975	France	122.299437	2.80...	
water	mg{NO2}/L	2008-01--2008-12	0.059145	France	122.299437	2.80...	
water	mg{NO3}/L	2008-01--2008-12	14.569833	France	122.299437	2.80...	
water	mg{NH4}/L	2008-01--2008-12	0.167798	France	122.299437	2.80...	
water	{massRatio}	2008-01--2008-12	35.2395	France	122.299437	2.80...	
water	mg{N}/L	2008-01--2008-12	1.35306	France	122.299437	2.80...	
water	mg{N}/L	2008-01--2008-12	5.70435	France	122.299437	2.80...	

Relation: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28

3: procedureAnalysedFraction Nominal	4: procedureAnalysedMedia Nominal	5: resultUom Nominal	6: parameterSamplingPeriod	7: resultMeanValue	8: Country Nominal	9: PopulationDensity Numeric	10: gdp Numeric
total	water	mg{NO2}/L	2009-0	331	France	122.299437	2.80...
total	water	mg{NO2}/L	2009-0	733	France	122.299437	2.80...
total	water	{massRatio}	2009-0	859	France	122.299437	2.80...
total	water	mg{NO3}/L	2009-0	376	France	122.299437	2.80...
total	water	mmol/L	2009-0	068	Spain	93.677197	1.46...
total	water	mg{NO3}/L	2009-0	792	Spain	93.677197	1.46...
total	water	mg{P}/L	2009-0	133	Spain	93.677197	1.46...
total	water	{massRatio}	2009-0	232	France	122.299437	2.80...
total	water	%	2009-0	833	France	122.299437	2.80...
total	water	{massRatio}	2009-0	594	France	122.299437	2.80...
total	water	mg{P}/L	2009-0	167	France	122.299437	2.80...
total	water	%	2009-0	175	France	122.299437	2.80...
total	water	[pH]	2009-0	919	France	122.299437	2.80...
total	water	mg{Si}/L	2009-0	458	France	122.299437	2.80...
total	water	%	2009-0	81.3	France	122.299437	2.80...
total	water	mg{O2}/L	2009-01--2009-12	1.0625	France	122.299437	2.80...
total	water	mg{P}/L	2009-01--2009-12	0.03856	France	122.299437	2.80...
total	water	[pH]	2009-01--2009-12	7.63017	France	122.299437	2.80...
total	water	mg{N}/L	2009-01--2009-12	1.0	France	122.299437	2.80...
total	water	mg{N}/L	2009-01--2009-12	0.244203	France	122.299437	2.80...
total	water	mg{NO2}/L	2009-01--2009-12	0.019993	France	122.299437	2.80...
total	water	mg{NH4}/L	2009-01--2009-12	0.159449	France	122.299437	2.80...
total	water	{massRatio}	2009-01--2009-12	37.9861	France	122.299437	2.80...
total	water	mg{P}/L	2009-01--2009-12	0.056667	France	122.299437	2.80...
total	water	ug/L	2009-01--2009-12	4.125	France	122.299437	2.80...
total	water	uS/cm	2009-01--2009-12	503.75	France	122.299437	2.80...
total	water	mg{N}/L	2009-01--2009-12	1.0	France	122.299437	2.80...
total	water	mg{N}/L	2009-01--2009-12	1.43368	France	122.299437	2.80...
total	water	mg{O2}/L	2009-01--2009-12	1.10833	France	122.299437	2.80...



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Normalize

About

Normalizes all numeric values in the given dataset (apart from the class attribute, if set).

debug False

doNotCheckCapabilities False

ignoreClass False

scale 1.0

translation 0.0

Figure-7

3.5 *Data Reduction*

There is a large amount of data in the dataset, so in order to get an accurate result, we decided to focus on the 2010 data and delete other dates from the new dataset.

First, we open the 'Edit' screen, select the data for 2010 from the 'Parameter Sampling period' property, and delete the data for other dates. And save the file again as another 'arff' file. Then delete the 'parameter sampling period' to make the data clean. See Figure 8.

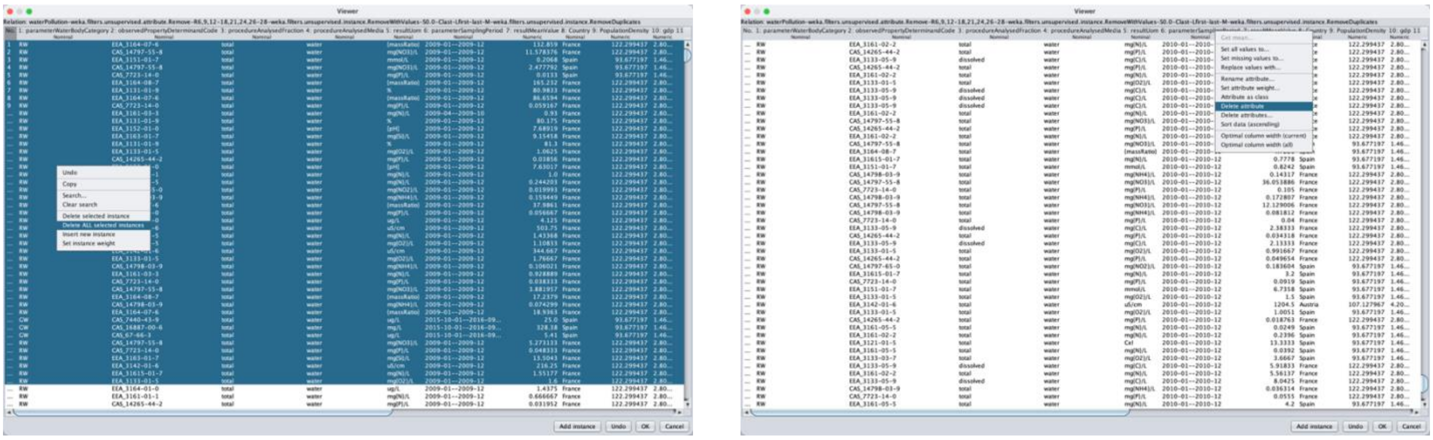
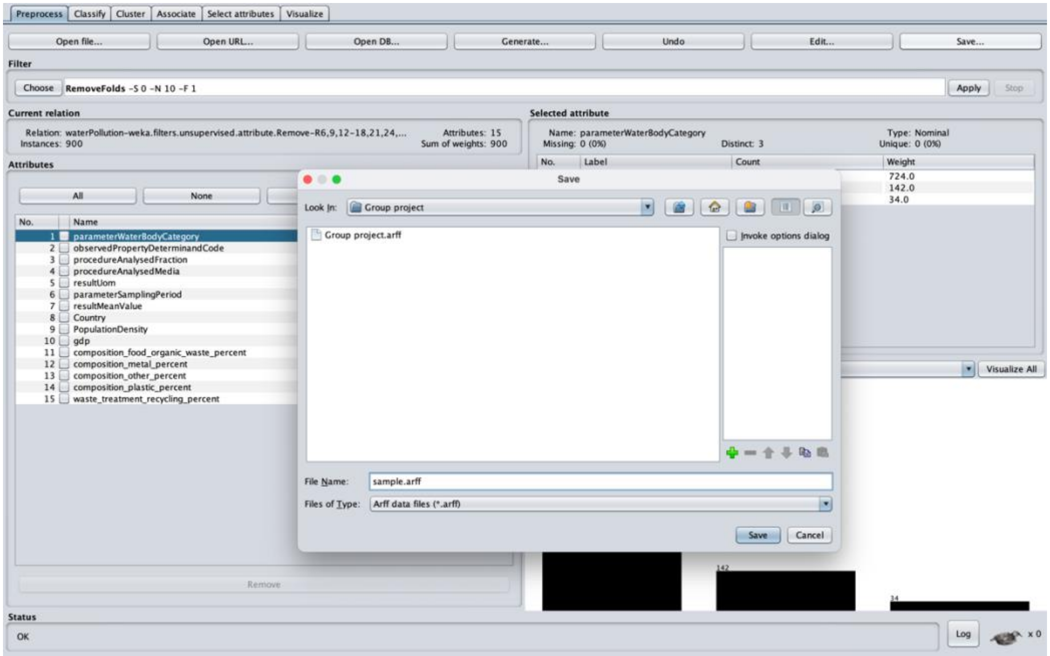


Figure-8

After selecting the 2010 water quality sampling survey data, there are more than 1500 records. The data set is still difficult and complex to study because the survey subjects are so large, so we randomly selected 900 data from it. Name it 'sample.arff' and store it. See figure-9.



Filter: RemoveFolds -S 0 -N 10 -F 1

Current relation

Relation: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,... Attributes: 15
Instances: 900 Sum of weights: 900

Attributes

No.	Name
1	parameterWaterBodyCategory
2	observedPropertyDeterminandCode
3	procedureAnalysedFraction
4	procedureAnalysedMedia
5	resultUom
6	parameterSamplingPeriod
7	resultMeanValue
8	Country
9	PopulationDensity
10	gdp
11	composition_food_organic_waste_percent
12	composition_metal_percent
13	composition_other_percent
14	composition_plastic_percent
15	waste_treatment_recycling_percent

Selected attribute

Name: parameterWaterBodyCategory Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
			724.0
			142.0
			34.0

Look In: Group project

Group project.arff

□ Invoke options dialog

File Name: sample.arff

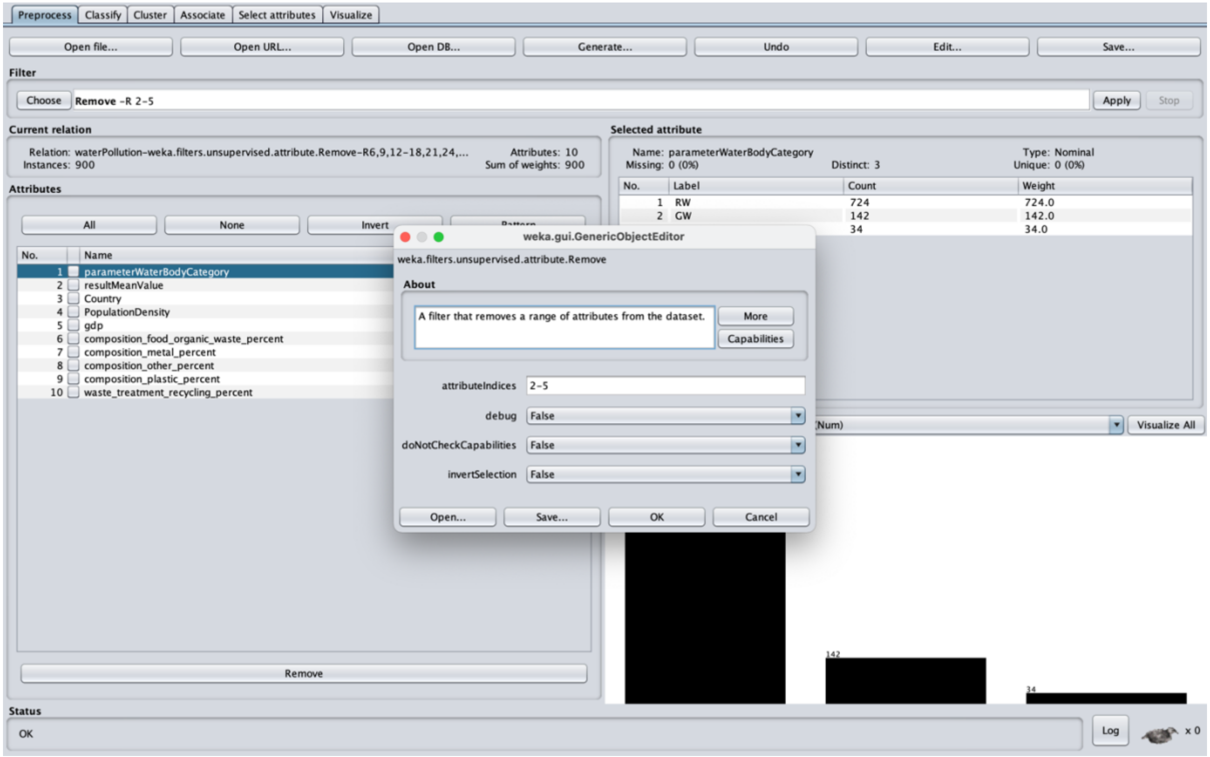
Files of Type: Arff data files (*.arff)

Status

OK

Figure-9

As the data for these columns named 'Procedure Analysed Fraction', 'Procedure Analysed Media' and 'Result Uom' are not relevant to our target, they are deleted. See Figure-10.



Filter: Remove -R 2-5

Current relation

Relation: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,... Attributes: 10
Instances: 900 Sum of weights: 900

Attributes

No.	Name
1	parameterWaterBodyCategory
2	resultMeanValue
3	Country
4	PopulationDensity
5	gdp
6	composition_food_organic_waste_percent
7	composition_metal_percent
8	composition_other_percent
9	composition_plastic_percent
10	waste_treatment_recycling_percent

Selected attribute

Name: parameterWaterBodyCategory Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	RW	724	724.0
2	GW	142	142.0
		34	34.0

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Remove

About

A filter that removes a range of attributes from the dataset.

attributeIndices: 2-5
debug: False
doNotCheckCapabilities: False
invertSelection: False

(Num)

Status

OK

Figure-10

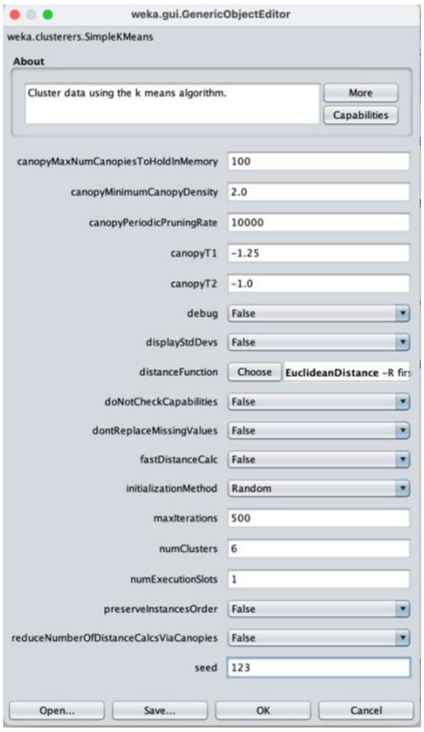
4 Data Modelling

4.1 Partitioning Clustering Approaches

In order to find out the serious water pollution areas in Europe in a better and faster way, we decided to use k-mean and DBSCAN to do the research. However, in order to make it more specific, we need to use k-mean algorithm to divide the clustering into 6 groups with 123 seeds. As show in figure 11.

4.1.1 K-means clustering

As can be seen from figure 12, among the six clusters, cluster 2 is the highest, accounting for 36%, followed by cluster 4, accounting for 31%, followed by cluster 0,3,1 and 5, accounting for 33%. This suggests that water pollution in the United Kingdom and Spain is more serious than in other countries and recommends special action to address water pollution in these waters.



weka.gui.GenericObjectEditor

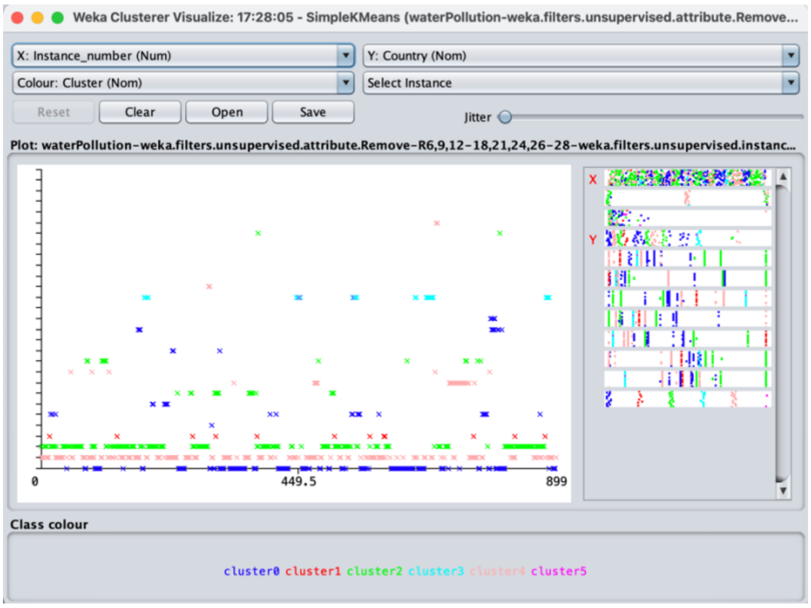
weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

canopyMaxNumCanopiesToHoldInMemory	100
canopyMinimumCanopyDensity	2.0
canopyPeriodicPruningRate	10000
canopyT1	-1.25
canopyT2	-1.0
debug	False
displayStdDevs	False
distanceFunction	Choose EuclideanDistance -R firs
doNotCheckCapabilities	False
dontReplaceMissingValues	False
fastDistanceCalc	False
initializationMethod	Random
maxIterations	500
numClusters	6
numExecutionSlots	1
preserveInstancesOrder	False
reduceNumberOfDistanceCalcsViaCanopies	False
seed	123

Figure-11



Weka Clusterer Visualize: 17:28:05 - SimpleKMeans (waterPollution-weka.filters.unsupervised.attribute.Remove...

X: Instance_number (Num)

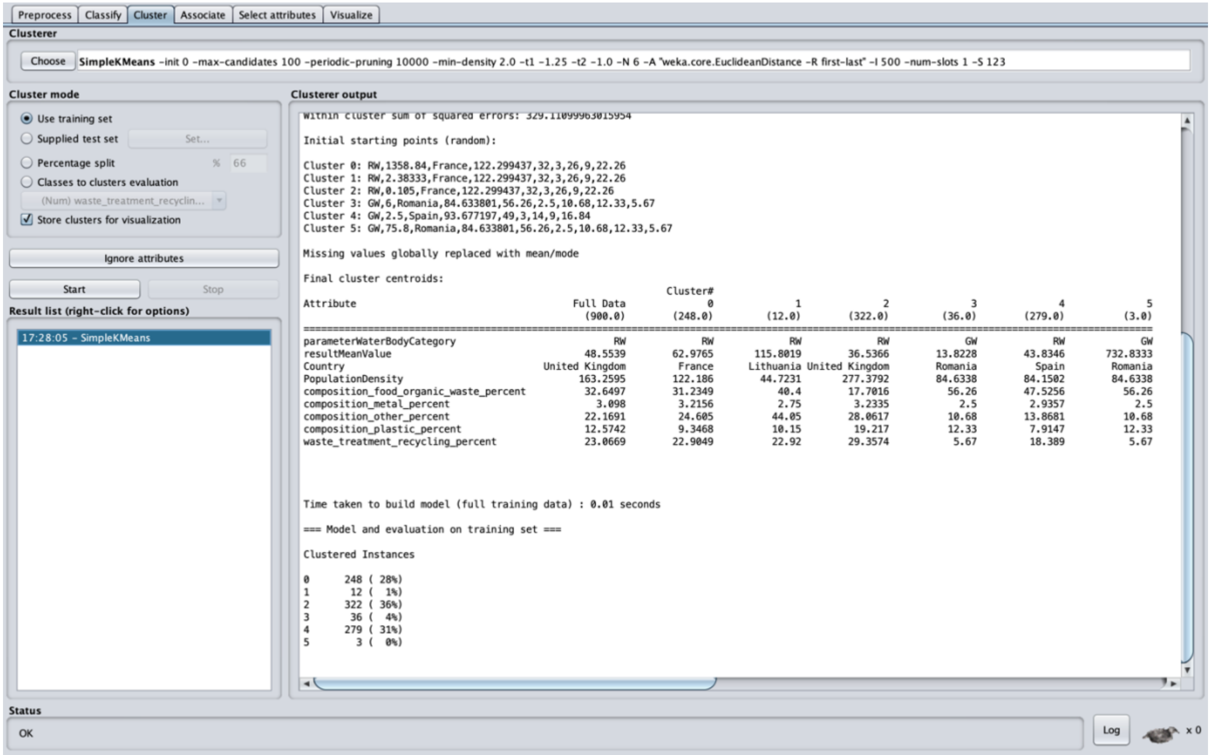
Y: Country (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28-weka.filters.unsupervised.instanc...

Figure-12



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 6 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 123

Cluster mode

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) waste_treatment_recyclin...
☑ Store clusters for visualization

Result list (right-click for options)

17:28:05 - SimpleKMeans

Clusterer output

```
Within cluster sum of squared errors: 329.11099963015954

Initial starting points (random):

Cluster 0: RW,1358.84,France,122.299437,32,3,26,9,22.26
Cluster 1: RW,2.38333,France,122.299437,32,3,26,9,22.26
Cluster 2: RW,0.105,France,122.299437,32,3,26,9,22.26
Cluster 3: GW,6,Romania,84.633801,56.26,2.5,10.68,12.33,5.67
Cluster 4: GW,2.5,Spain,93.677197,49,3,14,9,16.84
Cluster 5: GW,75.8,Romania,84.633801,56.26,2.5,10.68,12.33,5.67

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                      Cluster#
Attribute                                 Full Data          0          1          2          3          4          5
                                            (900.0)    (248.0)     (12.0)    (322.0)     (36.0)    (279.0)      (3.0)
=====================================================================================================================
parameterWaterBodyCategory                       RW         RW         RW         RW         GW         RW         GW
resultMeanValue                             48.5539    62.9765   115.8019    36.5366    13.8228    43.8346   732.8333
Country                              United Kingdom     France  Lithuania United Kingdom  Romania      Spain    Romania
PopulationDensity                          163.2595    122.186    44.7231   277.3792    84.6338    84.1502    84.6338
composition_food_organic_waste_percent      32.6497    31.2349       40.4    17.7016      56.26    47.5256      56.26
composition_metal_percent                     3.098     3.2156       2.75     3.2335        2.5     2.9357        2.5
composition_other_percent                   22.1691     24.605      44.05    28.0617      10.68    13.8681      10.68
composition_plastic_percent                 12.5742     9.3468      10.15     19.217      12.33     7.9147      12.33
waste_treatment_recycling_percent           23.0669    22.9049      22.92    29.3574       5.67     18.389       5.67




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       248 ( 28%)
1        12 (  1%)
2       322 ( 36%)
3        36 (  4%)
4       279 ( 31%)
5         3 (  0%)
```

Status

OK

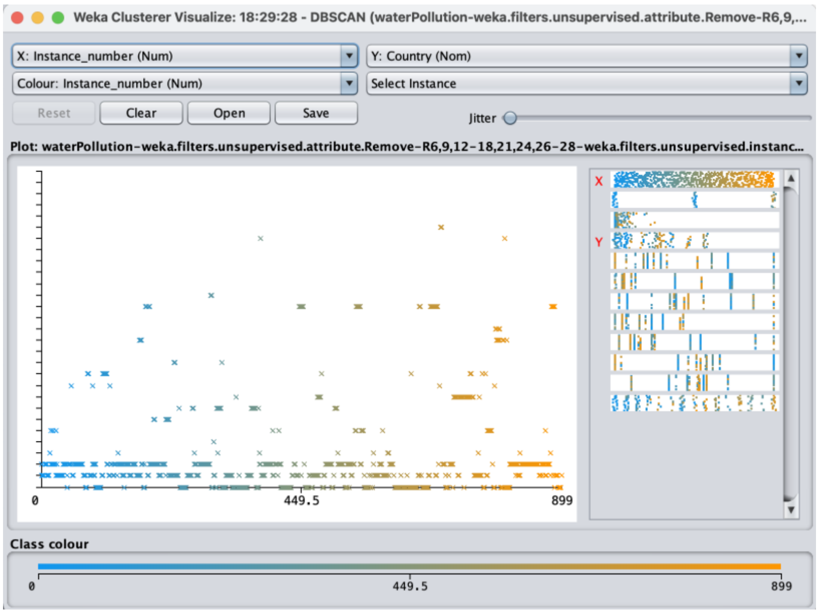
Figure-13

4.1.2 DBSCAN

DBSCAN is a clustering method used for machine learning to separate high-density clusters from low-density clusters. Two important parameters of DBSCAN are epsilon and midpoints. (Elutins, 2017).

Epsilon: the radius used to measure the distance between data points.

MinPts: the minPts in the EPS neighbourhood at that point.



Weka Clusterer Visualize: 18:29:28 - DBSCAN (waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,...

X: Instance_number (Num)

Y: Country (Nom)

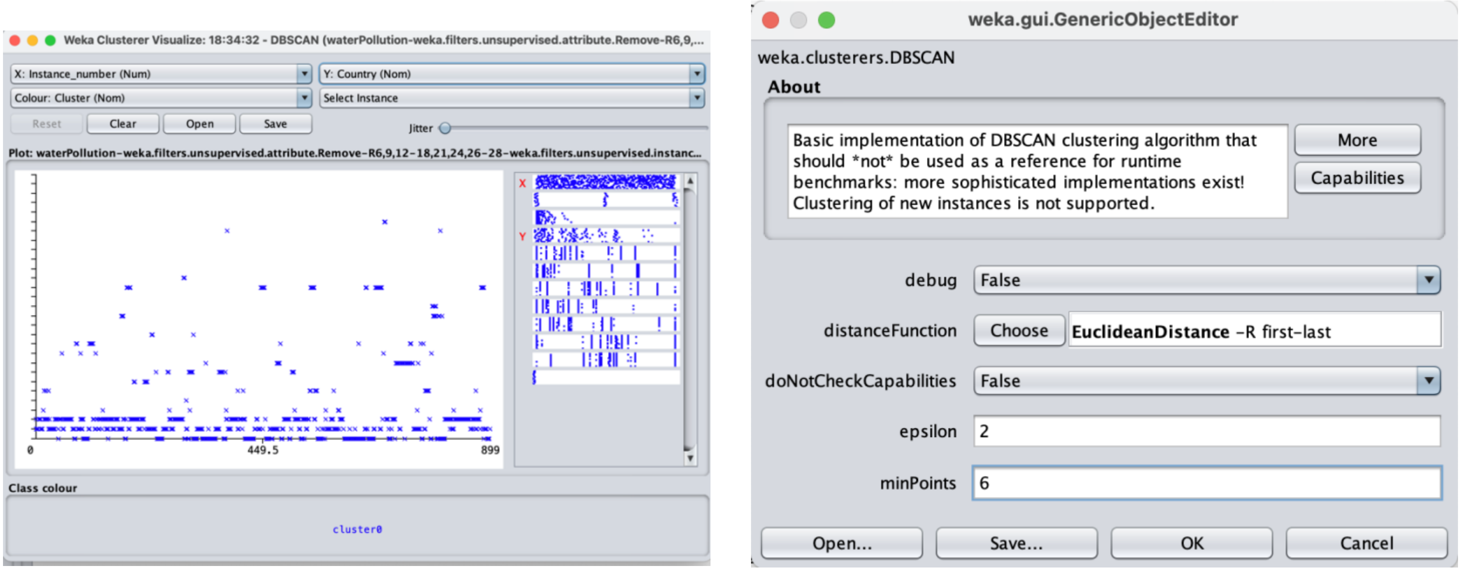
Colour: Instance_number (Num)

Select Instance

Plot: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28-weka.filters.unsupervised.instanc...

Class colour

First, we set Epsilon to 2 and Minpts to 6, as shown in Figure 14. Then we will set Epsilon to 1 and MinPts to 6, and the results are shown in Figure 15. By comparison, we found that the smaller the Epsilon, the more 'M' values appeared on the image, and conversely, the smaller the Minpts, the less 'M' values appeared on the image. As can be seen from the figure, the areas with heavy colour represent serious water pollution, and the areas with 'M' can be ignored. Countries concerned should treat specific water pollution according to the high frequency areas of K-mean and DBSCAN to protect the environment of the local area.


Weka Clusterer Visualize: 18:34:32 - DBSCAN (waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,...

X: Instance_number (Num)

Y: Country (Nom)

Colour: Cluster (Nom)

Plot: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28-weka.filters.unsupervised.instanc...

weka.gui.GenericObjectEditor

weka.clusterers.DBSCAN

About

Basic implementation of DBSCAN clustering algorithm that should *not* be used as a reference for runtime benchmarks: more sophisticated implementations exist! Clustering of new instances is not supported.

debug: False

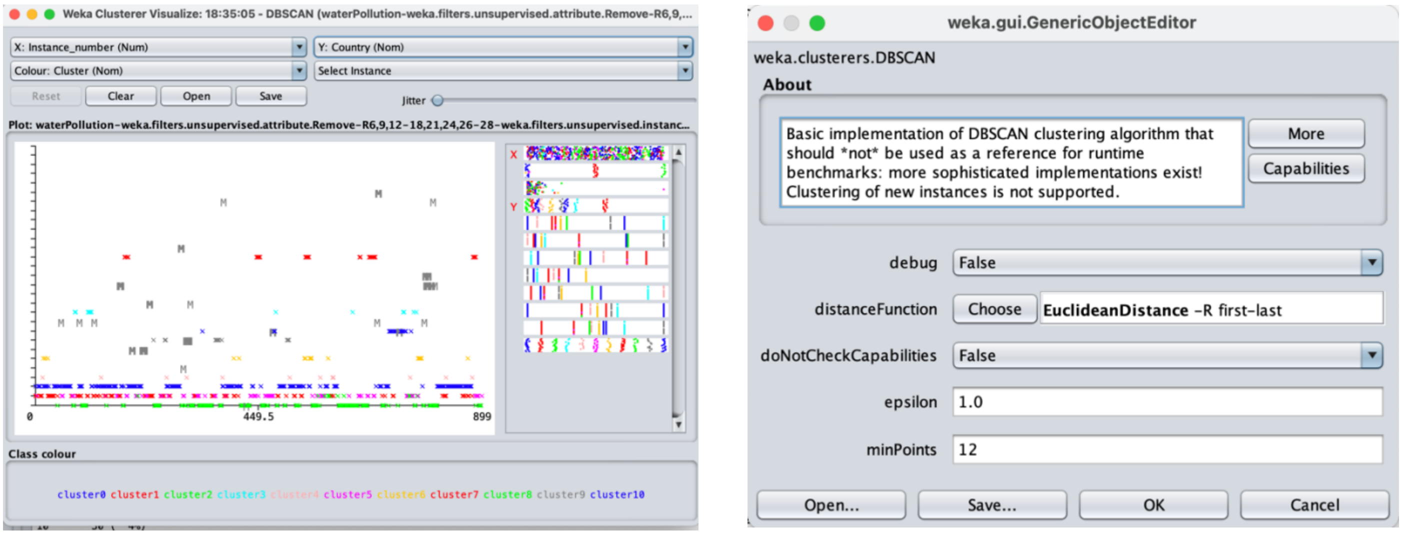
distanceFunction: EuclideanDistance -R first-last

doNotCheckCapabilities: False

epsilon: 2

minPoints: 6

Figure-14


Weka Clusterer Visualize: 18:35:05 - DBSCAN (waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,...

X: Instance_number (Num)

Y: Country (Nom)

Colour: Cluster (Nom)

Plot: waterPollution-weka.filters.unsupervised.attribute.Remove-R6,9,12-18,21,24,26-28-weka.filters.unsupervised.instanc...

Select Instance
Reset
Clear
Open
Save
Jitter
X
Y
M
0
449.5
899
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4 cluster5 cluster6 cluster7 cluster8 cluster9 cluster10

weka.gui.GenericObjectEditor

weka.clusterers.DBSCAN

About

Basic implementation of DBSCAN clustering algorithm that should *not* be used as a reference for runtime benchmarks: more sophisticated implementations exist! Clustering of new instances is not supported.

debug: False

distanceFunction: EuclideanDistance -R first-last

doNotCheckCapabilities: False

epsilon: 1.0

minPoints: 12

Figure-15

4.2 *Data Visualization*

After data mining, data visualization is needed. In order to convey information clearly and effectively, we use statistical graphs and other tools to express it. See Figure 0. Because effective visualization helps us analyze and reason about data and evidence. It makes complex data easier to understand and use (Tableau, N. D.).

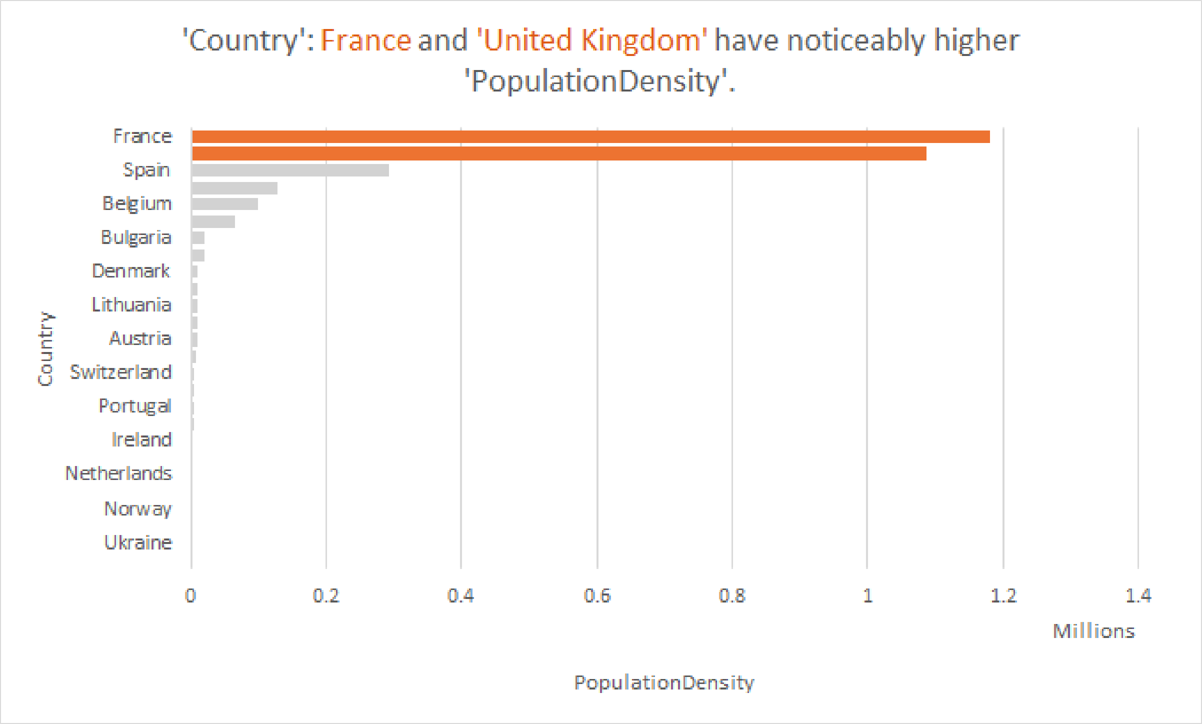
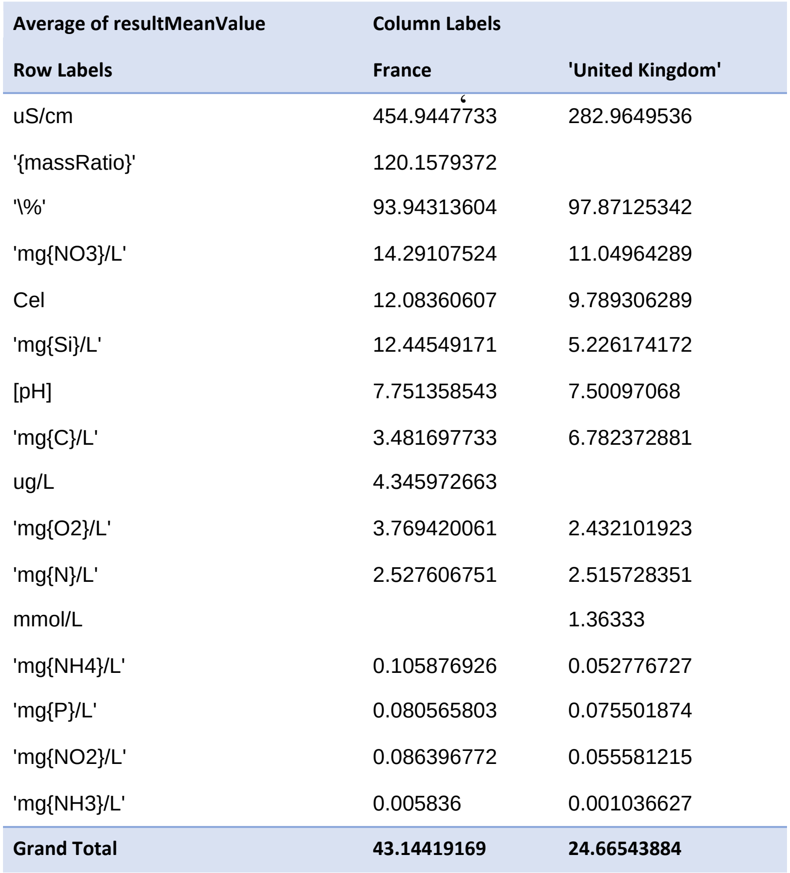


Figure-16

As shown in figure 16, we can see that France has the highest population density compared to other countries (see chart 0), ahead of The UK. Furthermore, the gap in population ratios between France, the United Kingdom and the third country, Spain, was almost double. Therefore, combined with the above data mining (as shown in Figure 0), we can conclude that there is a certain relationship between population density and water pollution. A densely populated country will have more water pollution than a less densely populated country. The reasons are also varied, such as factory sewage, agricultural sewage, and litter and other phenomena caused. Therefore, it is suggested that countries with high population density should reduce water consumption, establish urban sewage treatment system, adjust industrial structure, control agricultural non-point source pollution, rationally develop water resources and develop new water sources.



Average of resultMeanValue	Column Labels	
Row Labels	France	'United Kingdom'
uS/cm	454.9447733	282.9649536
'{massRatio}'	120.1579372	
'\%'	93.94313604	97.87125342
'mg{NO3}/L'	14.29107524	11.04964289
Cel	12.08360607	9.789306289
'mg{Si}/L'	12.44549171	5.226174172
[pH]	7.751358543	7.50097068
'mg{C}/L'	3.481697733	6.782372881
ug/L	4.345972663	
'mg{O2}/L'	3.769420061	2.432101923
'mg{N}/L'	2.527606751	2.515728351
mmol/L		1.36333
'mg{NH4}/L'	0.105876926	0.052776727
'mg{P}/L'	0.080565803	0.075501874
'mg{NO2}/L'	0.086396772	0.055581215
'mg{NH3}/L'	0.005836	0.001036627
Grand Total	**43.14419169**	**24.66543884**

Figure-17

In addition, water companies can strengthen operations in France and The UK, which are huge water needs. By identifying the two largest countries where companies can focus, we can delve deeper into the water resources of these regions. According to Figure 17, the records for each indicator unit in France and the United Kingdom are very similar. Moreover, with regard to all the numerical indicators of water quality, the water in these two countries can be considered to be of high quality (Naira Hassan Omar, 2019) and can definitely meet the needs of consumers.

5 Conclusion

Using public data sets obtained from Kaggle, this report examines three main water quality types in Europe: groundwater (GW), lake water (LW) and river water (RW). We know from data sets that water pollution is closely related to a country's population density, and the more people there are, the more emissions there are. As mentioned above, water pollution in Britain and France is the worst. we suggest that the government should establish a good sense of water pollution prevention and control in people's minds, if there is no good sense of prevention and control, it will lead to more and more serious water pollution. Therefore, everyone needs to improve the awareness of water pollution prevention, in daily life to minimize the consumption and reuse or recycling of plastic, the correct treatment of chemical cleaners, oil and non-biodegradable items. Everyone should give priority to economical and pollution control. Protect ecology, maintain river health life.

Reference

Analytics Vidhya.(2020, September 8). *How to Master the Popular DBSCAN Clustering Algorithm for Machine Learning.*

https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

Codecademy. (n. d). *Normalization.*

https://www.codecademy.com/articles/normalization

HSPH. (n. d). *Water Pollution*.

https://www.hsph.harvard.edu/ehep/82-2/

Investopedia, (n. d). Data mining.

https://www.investopedia.com/terms/d/datamining.asp

Kaggle. (n. d). Water Quality Dataset.

https://www.kaggle.com/ozgurdogan646/water-quality-dataset

Studio. (n. d). *Take control of your R code.*

https://www.rstudio.com/products/rstudio/

Techopedia. (n. d). *K-Means Clustering*.

https://www.techopedia.com/definition/32057/k-means-clustering