Establishing a pollution hotspot prediction model for Taiwan's river water quality monitoring stations using factor analysis and artificial neural networks

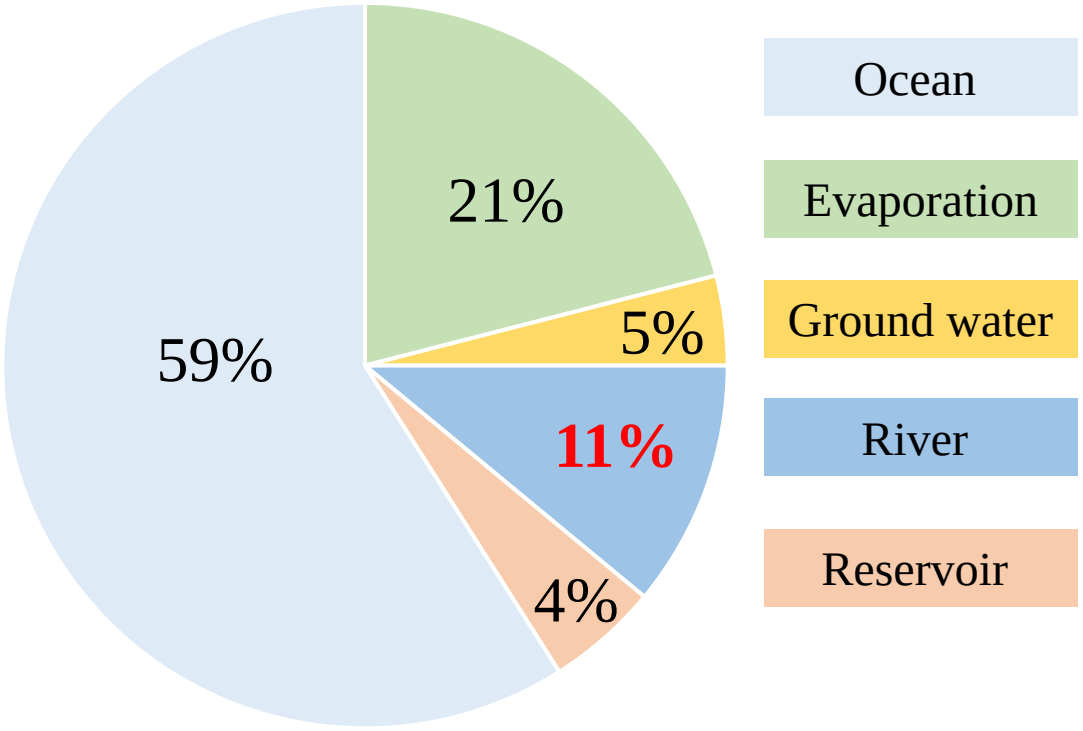
Name : Rui-Yu Zhang
Advisor : Jui-Sheng Chen, Prof. Ching-Ping Liang
Date :2025/03/21

Outline

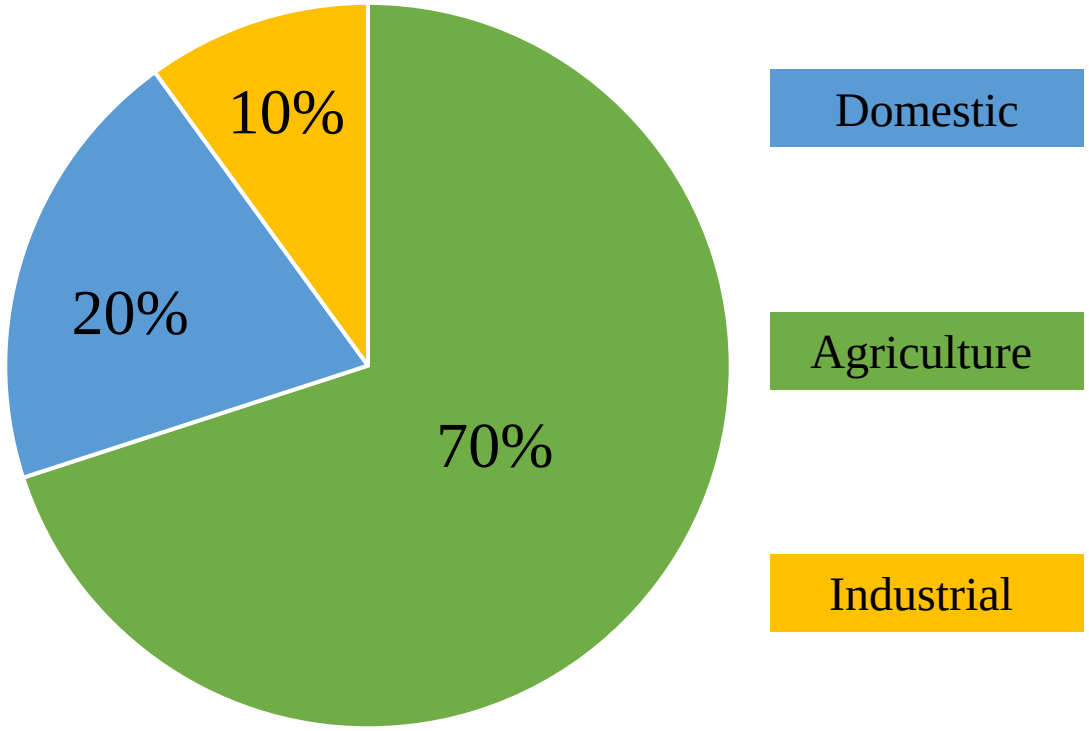
1. Introduction
2. Methodology
3. Preliminary result
4. Conclusions and Future work

Rainfall distribution and Water usage proportion in Taiwan

Rainfall Distribution in Taiwan



Water Usage Proportion in Taiwan



How does land use affect water quality?

Domestic

- **Source :** Waste from households, schools, and commercial areas.
- **Affect :** Environmental sanitation and odor issues around water bodies.

Agricultural

- **Source :** Fertilizers used for crops.
- **Affect :** Water eutrophication.

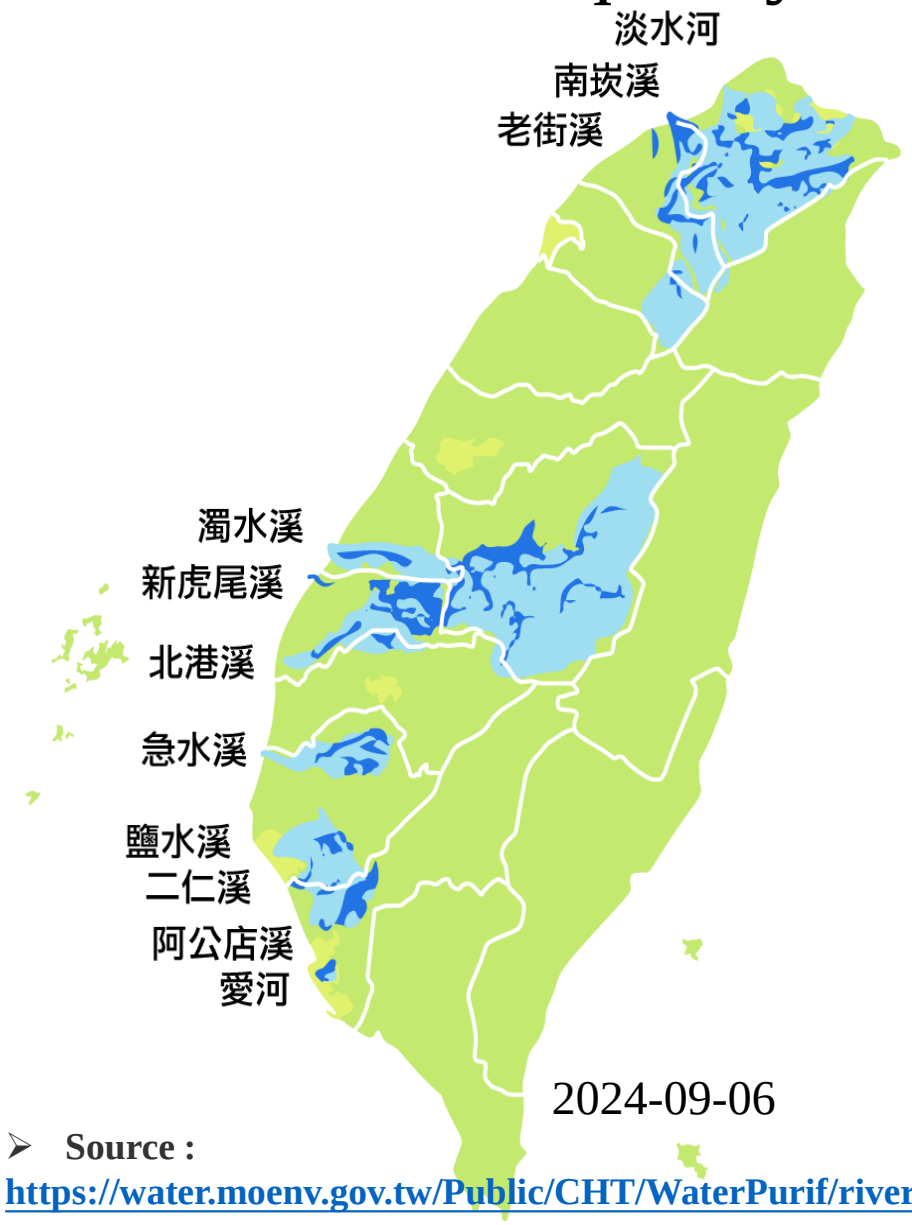
Industrial

- **Source :** Wastewater, heavy metals, and chemicals discharged from factories.
- **Affect :** Toxic to aquatic organisms.

Animal husbandry

- **Source :** Animal excreta (feces and urine)
- **Affect :** Environmental sanitation and odor issues around water bodies.

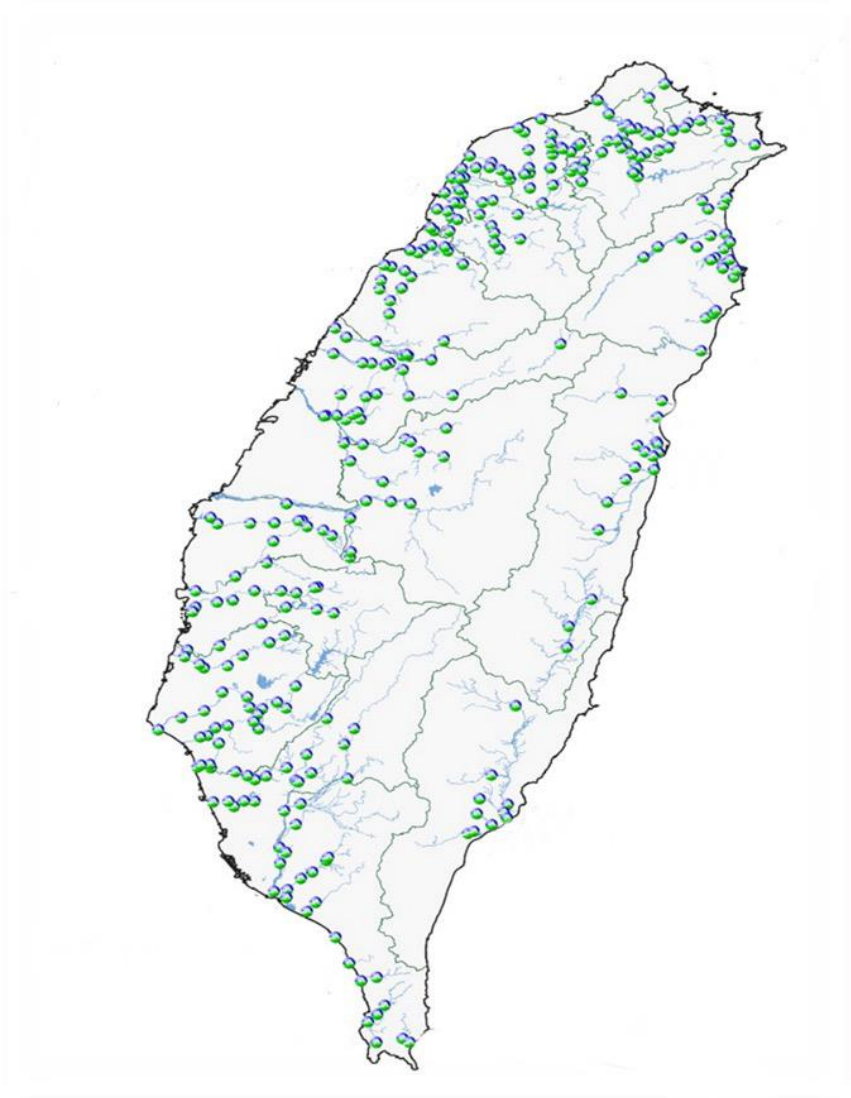
The river water quality in Taiwan



1. Total quantity control.
2. Promote on-site treatment facilities.
3. Strengthen inspection and regulation.
4. Domestic waste resource utilization.
5. Accelerate the implementation of sewerage projects.
6. In-Depth enterprise audits and functional evaluations.
7. Promote river surface garbage removal.

➢ **Source :**
https://water.moenv.gov.tw/Public/CHT/WaterPurif/river_remed.aspx

Study area



- Selecting 303 river water quality monitoring stations from the Ministry of Environment.
- Calculating the annual average concentration of pollutants.
- Using a GIS system to calculate land use around the water quality monitoring stations.

➢ **Source :**
https://wq.moenv.gov.tw/EWQP/zh/ConService/DownLoad/AnnReport.aspx

Literature review

- Li et al. (2009) used factor analysis to examine the impact of land use on river water quality in the Han River basin.

	Rainy season						Dry season					
	FOR	SHR	AGR	URB	BAR	VEG	FOR	SHR	AGR	URB	BAR	VEG
T	−0.45	0.12	−0.14	0.33	**0.68**[a]	−0.42	−0.50	0.32	0.26	−0.47	0.20	−0.26
pH	0.07	0.44	−0.33	−0.28	−0.38	0.56	−0.51	0.49	−0.05	−0.55	0.32	−0.09
EC	**−0.73**[a]	0.32	0.27	−0.28	0.56	−0.55	−0.53	−0.14	0.44	0.04	**0.67**[a]	**−0.81**[b]
Turbidity	−0.17	0.46	0.09	−0.07	−0.50	0.29	−0.02	0.46	−0.31	0.31	−0.42	0.48
SPM	−0.11	0.32	0.23	0.00	−0.55	0.21	−0.32	0.23	0.48	0.16	−0.36	−0.14
DO	0.13	−0.33	0.40	0.42	−0.26	−0.19	−0.23	−0.13	0.02	0.30	0.51	−0.42
COD_{Mn}	−0.33	0.44	0.10	0.13	−0.26	0.08	−0.43	0.14	0.35	0.10	0.12	−0.38
NH_4^+-N	−0.57	0.33	0.03	−0.47	0.56	−0.34	−0.43	0.17	0.23	−0.13	0.27	−0.35
NO_3^--N	**−0.73**[a]	0.36	0.12	−0.32	0.65	−0.50	−0.53	−0.07	0.14	0.02	**0.89**[b]	**−0.73**[a]
DP	0.31	−0.58	0.36	0.54	−0.16	−0.25	0.33	−0.63	0.20	0.31	0.11	−0.28
Cl^-	−0.54	0.24	0.07	−0.30	0.58	−0.41	−0.66	−0.08	0.56	−0.11	**0.69**[a]	**−0.89**[b]
SO_4^{2-}	**−0.73**[a]	0.32	0.18	−0.33	0.66	−0.55	−0.65	0.03	0.28	−0.13	**0.82**[b]	**−0.77**[a]
HCO_3^-	−0.43	0.21	0.28	−0.15	0.15	−0.30	−0.39	−0.17	0.45	0.07	0.45	**−0.67**[a]
K	−0.47	0.22	−0.19	−0.40	**0.79**[a]	−0.33	−0.40	−0.08	0.07	−0.28	**0.82**[b]	−0.58
Ca	−0.59	0.23	0.28	−0.19	0.42	−0.47	−0.29	−0.35	0.52	0.22	0.44	**−0.74**[a]
Na	**−0.70**[a]	0.16	0.15	−0.21	**0.88**[b]	**−0.69**[a]	−0.55	−0.13	0.26	−0.02	**0.89**[b]	**−0.81**[b]
Mg	−0.66	0.29	0.24	−0.32	0.51	−0.49	−0.52	−0.06	0.34	−0.08	0.66	**−0.71**[a]

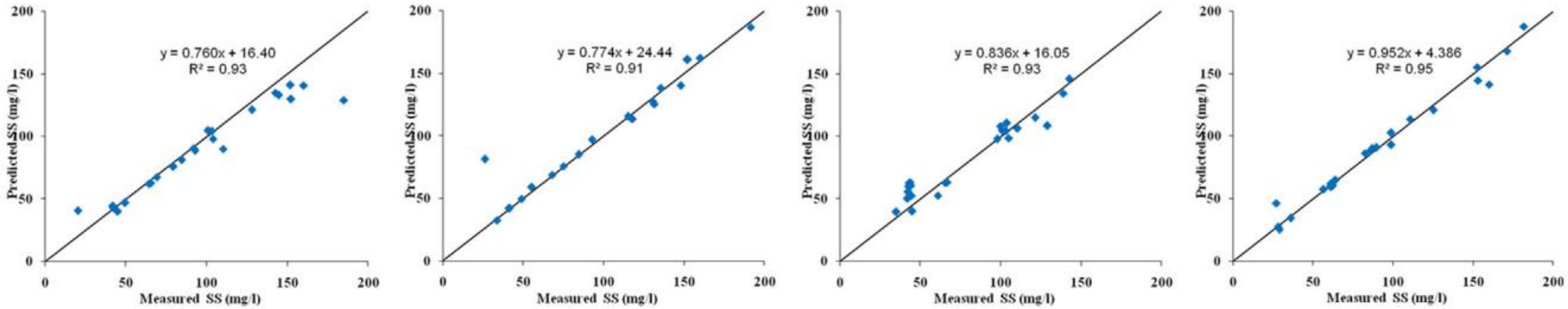
Literature review

- Kazi et al. (2009) used principal component analysis (PCA) to understand the relationship between pollution sources and water quality changes.

Parameters	PC1	PC2	PC3
EC	0.188	−0.065	−0.191
Salinity	0.186	−0.112	−0.071
TS	0.188	−0.096	−0.047
TDS	0.190	−0.076	−0.051
TSS	0.172	−0.180	0.089
T-Hard	0.189	−0.087	−0.066
Ca-Hard	0.185	−0.124	−0.046
DO	0.024	−0.312	−0.248
BOD	0.120	−0.283	−0.205
COD	0.176	−0.137	0.262
F	0.186	−0.114	−0.036
Cl	0.190	−0.081	−0.033
T-Alk	0.185	−0.051	−0.264
PO_4	0.169	0.088	0.199

Literature review

- Ahmed et al. (2019) applied a multilayer backpropagation neural network to predict the concentrations in the Johor River basin.



➢Factor analysis can identify the relationships between different parameters, enhancing the accuracy of machine learning models

Motivation

- Due to Taiwan's limited water resources, preventing river pollution is crucial. Thus, developing effective and accurate methods for monitoring and predicting water quality is an urgent necessity.

Objective

- This study applies factor analysis to extract key water quality factors and establish a machine learning model to enhance prediction accuracy and pollution hotspot identification.

River water quality monitoring data

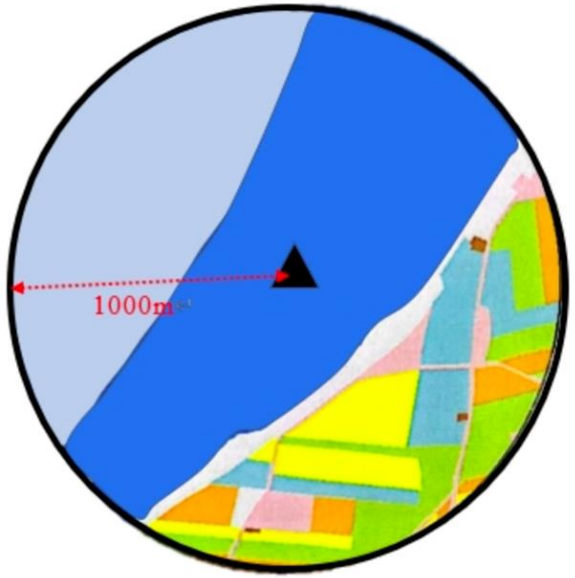
Metals
Lead, Silver, Copper, Zinc, Manganese, Nickel, Chromium VI
Non-Metals
pH value, Electrical Conductivity, Suspended Solids, Dissolved Oxygen, Chemical Oxygen Demand, Biochemical Oxygen Demand, Chlorine, Escherichia coli, Total Organic Carbon, River Pollution Index, Arsenic
Nutrients
Ammonia Nitrogen, Nitrate Nitrogen, Nitrite Nitrogen, Total Phosphorus

環境部
Ministry of Environment

River water quality monitoring data

- Data that cannot be measured is recorded as 0.
- Values below the detection limit are assigned the minimum value.
- The annual average of each data point is calculated to reduce the impact of short-term fluctuations.

Reclassify land use type



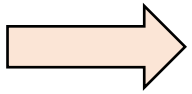
中華民國
內政部國土測繪中心
National Land Surveying and Mapping Center,
Ministry of the Interior, R.O.C (Taiwan)

- Forest
- Agriculture
- Building
- Public
- Bare land
- Waterbody
- Transportation
- Water conservancy
- Others
- Water quality station

	LCODE_C3	siteid	sitename	area
1	050204	1002	重陽大橋	13111
2	010402	1002	重陽大橋	302
3	060600	1002	重陽大橋	5438
4	060505	1002	重陽大橋	3244
5	060502	1002	重陽大橋	705
6	090802	1002	重陽大橋	9067
7	060503	1002	重陽大橋	9279
8	060100	1002	重陽大橋	434
9	050404	1002	重陽大橋	2033
10	050403	1002	重陽大橋	38580

Reclassify land use type

	LCODE_C3	siteid	sitename	percent
1	010101	1066	深坑橋	46
2	010101	1066	深坑橋	46
3	010101	1670	中山橋(新豐溪)	44
4	010101	1230	利澤簡橋	44
5	010101	1062	社子橋	43
6	010101	1678	十三份橋	39
7	010101	1052	茄苳溪橋	38
8	010101	1254	太平溪橋	37
9	010101	1096	南港溪橋	35
10	010101	1032	磺溪橋	34



Ministry of the Interior Land Use Classification Table

稻作	010101	係指從事稻米栽培之土地。包括水稻、陸稻
旱作	010102	係指從事雜糧作物、特用作物及園藝作物栽培之土地。雜糧作物包括小麥、黑麥、蕎麥、紅豆、大豆、玉米、粟（小米）、大麥、甘藷、花豆、綠豆、薏仁、落花生、蜀黍（高粱）；特用作物包括係指從事纖維料、油料、糖料（甘蔗）、嗜好料、香料、藥料及工業原料等特用作物栽培之土地。包括棕櫚、苧麻、亞麻、大甲蘭、芏苡（三角蘭）、向日葵、油菜籽、葛鬱金（粉薯）、甜菜、茶葉、菸草、胡椒、花椒、香茅草、芥末籽、杭菊、除蟲菊、枸杞、黃蓍、麥門冬、桑樹、棉花、瓊麻、黃麻、洋麻（鐘麻）、芝麻、蓖麻籽、樹薯、甜菊、咖啡、可可豆、蛇麻、茴香、仙草、洛神葵、薄荷、魚藤、當歸、山藥、柴胡、牧草、綠肥作物；園藝作物包括蔬菜、食用菌菇類（包括木耳、香菇、草菇、食用菌菇類菌種、靈芝、洋菇、金針菇）及花卉（包括盆花植物、觀葉植物、切花植物）
果樹	010103	係指從事水果及乾果種植、栽培而以收穫其果實為目的之土地。包括李、杏、柿、栗、枇杷、橄欖、木瓜、楊桃、鳳梨、檳榔、葡萄、椰子、柑桔類、番石榴、梅、桃、棗、梨、芒果、胡桃、蘋果、龍眼、香蕉、蓮霧、荔枝、番荔枝、百香果

Land use classification used in this study

Domestic

Agricultural

Industrial

Animal husbandry

Forest

Governing equations in factor analysis

- Fundamental Representation of the factor analysis

$$X_1 = l_{11} \cdot f_1 + l_{12} \cdot f_2 + \cdots + l_{1q} \cdot f_q + \varepsilon_1$$
$$X_2 = l_{21} \cdot f_1 + l_{22} \cdot f_2 + \cdots + l_{2q} \cdot f_q + \varepsilon_2$$
$$\vdots$$
$$X_i = l_{i1} \cdot f_1 + l_{i2} \cdot f_2 + \cdots + l_{iq} \cdot f_q + \varepsilon_i$$

Factor analysis model

X_i is the i-th observed variable
f_i is the j-th common factor
l_i is the factor loading
ε_i is the specific factor that cannot be explained by the common factors

Principal components analysis

- Pearson correlation matrix

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}_{p \times p} \qquad r_{ij} = \frac{\left[\sum (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) \right]}{\sqrt{\left[\sum (X_{ik} - \bar{X}_i)^2 \cdot \sum (X_{jk} - \bar{X}_j)^2 \right]}}$$

R is the correlation matrix
r_{ij} is the Pearson correlation coefficient between variable i and variable j
X_{ik} *and* X_{jk} are the observed values of the k-th sample.
$\bar{X}_i$ *and* $\bar{X}_k$ are the means of the variables.

Principal components analysis

- Extract data eigenvalues and eigenvector

$$|R - \lambda I| = 0$$

$$RY = \lambda Y$$

R is correlation matrix
λ is data eigenvalues
I is unit matrix
Y is eigenvector

- Eigenvector matrix of principal components

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

a_{pp} is the coefficient of the original variable in the p-th principal component
x_p is observed variable

Principal components analysis

- Factor loading

$$L = \Lambda^{-\frac{1}{2}} Y \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\lambda_p}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

L is factor loading
Y is eigenvector matrix of the principal components
Λ is the corresponding eigenvalue diagonal matrix

Kaiser-Meyer-Olkin and Bartlett‘s Test (Water quality)

Unsuitable Parameters
SS, Cd, Ni, Hg, Cr^{6+}, Cl^{-}, pH, EC, RPI, Ag

Data Test Results	
Kaiser-Meyer-Olkin Test	0.82
Bartlett's Test	0.00

KMO value	Suitability levels
>0.90	Marvelous
0.80~0.89	Meritorious
0.70~0.79	Middling
0.60~0.69	Mediocre
0.50~0.59	Miserable
<0.50	Unacceptable

Preliminary result (Water quality)

Variable	Factor 1	Factor 2	Factor 3
NH_3-N	0.861	0.258	0.037
BOD	0.726	0.441	-0.081
E. Colie	0.545	-0.216	0.102
TP	0.521	0.409	0.065
NO_2-N	0.486	0.667	-0.113
NO_3-N	0.21	0.887	-0.119
DO	-0.745	-0.017	-0.026
COD	0.78	0.241	0.24
TOC	0.678	0.365	0.025
Mn	0.197	0.125	0.883
As	0.56	-0.028	0.654
Cu	0.089	0.738	0.417
Pb	-0.137	0.095	0.892
Zn	-0.027	0.621	0.311

- Positive and negative values indicate correlation direction.
- The variable with the highest absolute value is assigned to the corresponding factor.

BOD：Biological Oxygen Demand **DO**：Dissolved Oxygen **E. Colie**：Escherichia coli
COD：Chemical Oxygen Demand **TP**：Total Phosphorus **TOC**：Total Organic carbon

Kaiser-Meyer-Olkin and Bartlett‘s Test (Land use and Water quality)

Unsuitable Parameters
SS, Cd, Ni, Hg, Cr^{6+}, Cl^{-}, pH, EC, RPI, Ag

Data Test Results	
Kaiser-Meyer-Olkin Test	0.70
Bartlett's Test	0.00

KMO value	Suitability levels
>0.90	Marvelous
0.80~0.89	Meritorious
0.70~0.79	Middling
0.60~0.69	Mediocre
0.50~0.59	Miserable
<0.50	Unacceptable

Preliminary result (Land use and Water quality)

Variable	Factor 1	Factor 2	Factor 3
Domestic	0.211	0.12	-0.478
Indutural	0.21	0.546	-0.287
Agriculture	0.181	-0.133	0.453
Animal husbandry	0.267	-0.1	0.21
Forest	-0.437	-0.144	0.117
NH_3-N	0.815	0.287	0.061
BOD	0.702	0.448	-0.076
E. Colie	0.524	-0.153	0.09
TP	0.47	0.457	0.013
NO_2-N	0.492	0.622	-0.11
NO_3-N	0.207	0.851	-0.163
DO	-0.77	-0.052	0.049
COD	0.752	0.291	0.238
TOC	0.667	0.337	0.112
Mn	0.185	0.245	0.798
As	0.512	0.074	0.644
Cu	0.05	0.788	0.317
Pb	-0.155	0.221	0.78
Zn	-0.037	0.61	0.279

- Positive and negative values indicate correlation direction.
- The variable with the highest absolute value is assigned to the corresponding factor.

BOD：Biological Oxygen Demand **DO**：Dissolved Oxygen **E. Colie**：Escherichia coli
COD：Chemical Oxygen Demand **TP**：Total Phosphorus **TOC**：Total Organic carbon

Conclusion

- Preliminary analysis can know the relationship between water quality and land use, but improving the KMO value is needed for greater credibility.

Future work

- Attempt to improve the KMO value by using different land use radius.
- Develop a river water quality hotspot prediction model using machine learning.

Thanks for your listening