元智大學 工業工程與管理學系所
Yuan Ze University, Department of Industrial Engineering and Management

結合決策樹與迴歸模型之廢水處理

指導教授：黃晧
組員：1081201游家齊
1081212吳柏彥

研究背景與目的

水資源使用量逐年增加，相對的與水資源相關的一些設備也會需要調整更換的時間，預測未來的該更換的時機點，以使過程中發生錯誤的決定的機會最小。

研究方法

研究變項：產水電導度(決策樹、樞紐)、壓差(樞紐)。

決策樹：把相關性高的數據聚在一起，則比較不會產生極值。

Python迴歸模型：
將我們所使用的自變數及應變數放Python中建構迴歸模型。

散布圖：快速的看出資料的分配性。

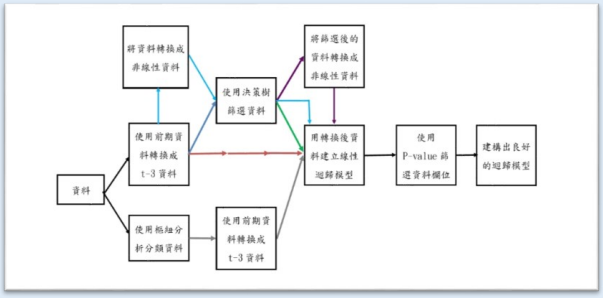
資料轉換：前期資料增加feature。非線性轉換資料是指把資料相乘得到新的一個欄位。

樞紐分析：資料分成最大值、最小值及平均值。

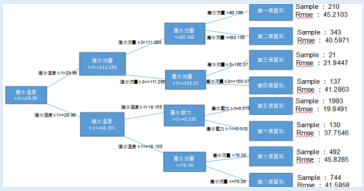
P值篩選：
利用P值去篩選對於迴歸無用的欄位，可以使迴歸模型的調整判定係數提高。

研究過程

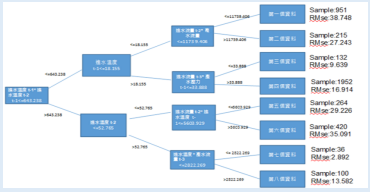
決策樹決定之變數做迴歸分析



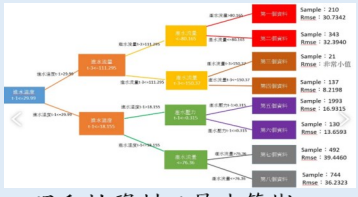
分析流程



線性資料三層決策樹



非線性資料三層決策樹

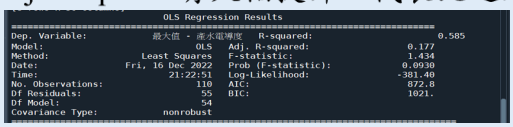


混和性資料三層決策樹

線性三層決策樹	sample	R-Square	Adj-R-square	Rmse
第一個資料	210	0.063	-0.014	45.21
第二個資料	343	0.056	0.01	40.6
第三個資料	21	0.973	0.891	21.94
第四個資料	137	0.150	0.037	41.29
第五個資料	1993	0.066	0.059	19.85
第六個資料	130	0.365	0.275	37.75
第七個資料	492	0.049	0.016	45.83
第八個資料	744	0.072	0.051	41.59
非線性三層決策樹	**sample**	**R-Square**	**Adj-R-square**	**Rmse**
第一個資料	951	0.224	0.114	38.748
第二個資料	215	0.629	0.213	27.243
第三個資料	132	0.958	0.695	9.639
第四個資料	1952	0.219	0.169	16.914
第五個資料	264	0.623	0.316	29.226
第六個資料	420	0.312	0.042	35.091
第七個資料	36	1	趨近 1	2.892
第八個資料	100	0.916	0.078	13.582
混合三層決策樹	**sample**	**R-Square**	**Adj-R-square**	**Rmse**
第一個資料	210	0.567	0.027	30.73
第二個資料	343	0.399	0.083	32.39
第三個資料	21	1	趨近 1	極小
第四個資料	137	0.966	0.817	8.22
第五個資料	1993	0.32	0.278	16.93
第六個資料	130	0.917	0.513	13.66
第七個資料	492	0.295	0.072	39.45
第八個資料	744	0.295	0.162	36.23
前期轉換資料	**sample**	**R-Square**	**Adj-R-square**	**Rmse**
t-3	4070	0.116	0.112	39.18

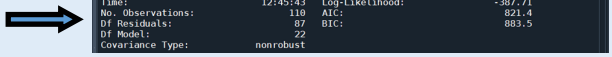
樞紐分析

資料分成最大值、最小值及平均值。並透過P值去篩選資料，提升迴歸模型的預測效果。篩選後的Adj.R-squared有大幅提升，代表此迴歸模型可以解釋的變異上升，迴歸模型優化提升了。



OLS Regression Results

Dep. Variable:	最大值 - 產水電導度	R-squared:	0.585
Model:	OLS	Adj. R-squared:	0.177
Method:	Least Squares	F-statistic:	1.434
Date:	Fri, 16 Dec 2022	Prob (F-statistic):	0.0930
Time:	21:22:51	Log-Likelihood:	-381.40
No. Observations:	110	AIC:	872.8
Df Residuals:	55	BIC:	1021.
Df Model:	54		
Covariance Type:	nonrobust		



OLS Regression Results

Dep. Variable:	最大值 - 產水電導度	R-squared:	0.534
Model:	OLS	Adj. R-squared:	0.417
Method:	Least Squares	F-statistic:	4.537
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	1.85e-07
Time:	12:45:43	Log-Likelihood:	-387.71
No. Observations:	110	AIC:	821.4
Df Residuals:	87	BIC:	883.5
Df Model:	22		
Covariance Type:	nonrobust		

結論

透過前期資料增加feature、決策樹篩選資料、非線性轉換資料、樞紐分析去轉換資料，我們以分鐘為單位之預測模型以及樞紐分析以天為單位之模型分開去估計，而這樣的目的是為了長時間的預警。而透過這些預測結果，我們可以預測一個合適的時間點去更換產水設備，以便產水的效率一直都會維持在同樣的水準。