2024年臺灣國際科學展覽會
優勝作品專輯

作品編號　160035

參展科別　物理與天文學

作品名稱　以數據驅動方式探究類星體於可見光與無線電光譜之性質關聯

得獎獎項　四等獎

就讀學校　臺北市立建國高級中學
指導教師　賴奕帆、藍鼎文
作者姓名　謝濟遠

關鍵詞　類星體、無線電、機器學習

作者簡介

大家好，我是建國中學科學班的二年級學生，謝濟遠。一場有關暗物質的演講在我一年級的時候點燃了我對天文學的熱情，推動我開始用資料科學探索類星體奧秘的旅程。

感謝教授、老師、朋友以及家人在這段旅程中的支持和陪伴。這份支持使我能夠擁有現在的研究成果，也讓我有機會參與台灣國際科展。期待與其他優秀的參展者們交流。

2024 年臺灣國際科學展覽會
研究報告內容

摘要

通過 LOFAR 望遠鏡與斯隆數位巡天計畫之資料釋出，我們得以在大量、清晰的資料庫中發掘許多天體的特殊性質。本研究結合兩個大型研究計劃的無線電波源目錄與類星體光學目錄，分析 LOFAR-detected 類星體在可見光觀測上與其他類星體的光譜性質差異，利用決策樹與相關的集成學習模型建立對於缺乏無線電紀錄之類星體的光學資料分類模型，並期望在未來驗證學界目前對於類星體產生無線電原理的各種假說。

Through the release of data from the LOFAR telescope and the Sloan Digital Sky Survey (SDSS), we have been able to discover unique properties of many celestial objects within extensive and high-quality databases. This study combines two large-scale research initiatives: the radio source catalog from LOFAR and the optical catalog of quasars from SDSS. We analyze the spectral differences between LOFAR-detected quasars and other quasars in visible light observations. We utilize decision trees and related ensemble learning models to establish a classification model for quasars lacking radio records based on optical data. We aim to validate various hypotheses in the field of astrophysics regarding the radio emission mechanisms of quasars in future .

壹、前言(含研究動機、目的、文獻回顧）

一、研究動機

類星體是一種極度明亮的活躍星系核，其在許多無線電波段均顯示出高光度的特性。現在，隨著資料科學的蓬勃發展，從大量資料中探勘科學事實已成為可能。在類星體蘊含的眾多謎團中，我很想知道為甚麼一類星體能夠產生無線電訊號，但並非所有類星體都能產生訊號。成功解決這個問題將不僅有助於解決單一問題，更能成為天文學許多開放性問題的跳板，並且為電波天文台的建置、天空地圖、類星體目錄的建立帶來重要影響。

二、研究目的

(一) 整合 LOFAR 望遠鏡與斯隆數位巡天計畫的類星體與無線電波源目錄。

(二) 分析 Radio Loud 類星體與 Radio Quiet 類星體在可見光譜的性質差異。

(三) 以機器學習方法建立從類星體光學資料預測無線電波強度的模型。

三、文獻探討

(一) AGN Unified 模型與類星體(Quasar)[1]

活躍星系核(Active Galactic Nucleus, AGN)定義為包含在一星系中，質量龐大(>$10^5 M_\odot$)的黑洞，並且其愛丁頓比率$L_{AGN}/L_{Edd} > 10^{-5}$，其中$L_{AGN}$為星系核的全波段光度(bolometric luminosity)，$L_{Edd} = 1.5 \times 10^{38} M_{BH}/M_\odot \; erg\; s^{-1}$為恆星組成氣體成分的愛丁頓光度，為向外作用的輻射力與向內作用的引力之間達到平衡時，可以達到的最大光度，M_{BH}代表黑洞質量。

AGNs 通常具有以下特性：

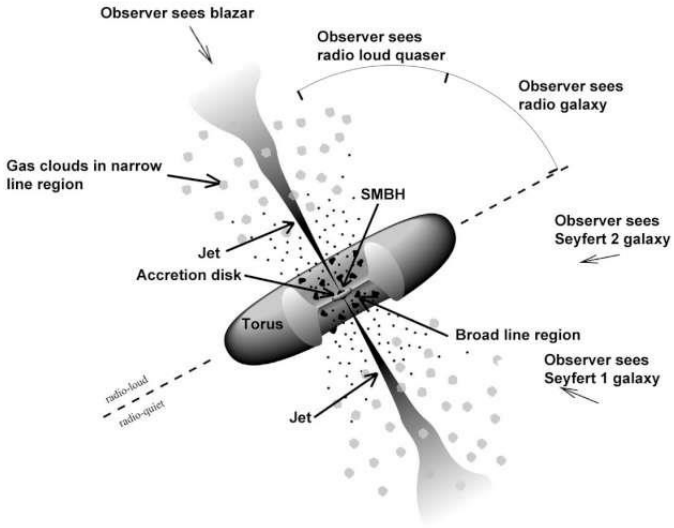
- 具有由物質圍繞 AGN 運動所形成的盤狀結構-吸積盤。當物質旋轉並被向黑洞運動時，在重力與摩擦力的作用下，盤中物質壓縮且溫度升高，進而向外發射電磁輻射。觀測到的吸積盤通常為 sub-pc 尺度(小於 1 秒差距)。本研究將吸積盤之結構視為厚度相較於寬度可忽略的薄吸積盤。

- 高密度、無塵的氣體雲以與亮度相關的距離從 0.01 到 1 秒差距的位置以開普勒速度($v \propto r^{-0.5}$)移動，此氣體雲稱為寬線區域(Broad Line Legion，BLR)

- 低密度、低速度的電離化氣體 (稱為窄線區域，Broad Line Legion，BLR) 從環狀結構外部延伸到數百或數千秒差距範圍內，沿著環狀結構的開口方向（"電離錐"）延伸。

目前根據觀測型將類星體分為兩類，Type-I 與 Type-II。

Type-I 類星體可以觀測到由寬線區域所發出的線寬較大光譜線，意即觀測到的光譜主要為吸積盤等 AGN 核心構造所發射的光譜，此類 AGN 被稱為類星體，即本研究聚焦的天體類型。

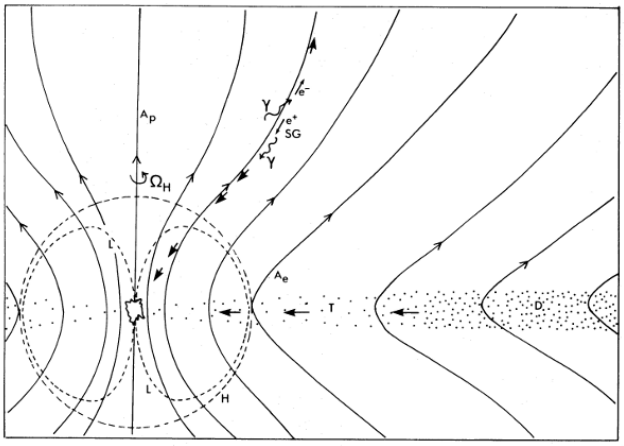


圖：AGN 構造簡圖[2]

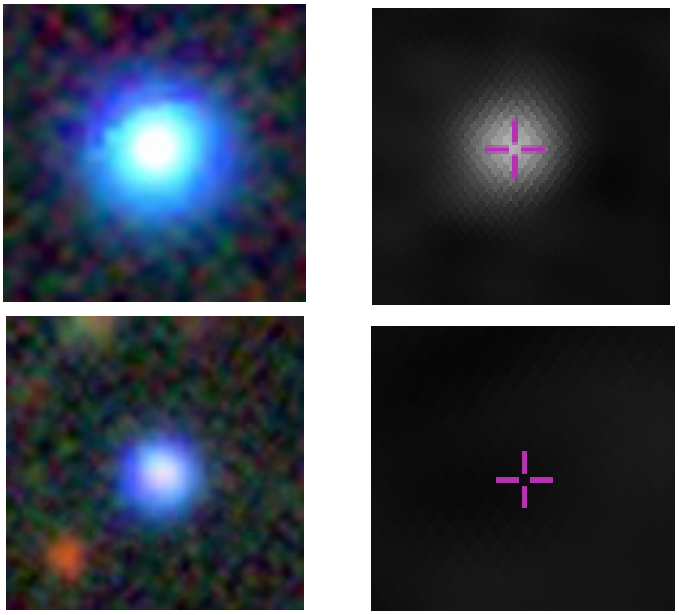
(二) Radio Loud 類星體與 Radio Jets

當吸積盤旋轉時，其中帶電粒子運動帶來外加電流，製造穿過旋轉中黑洞的磁場線。黑洞的重力扭曲了空間與磁場線，電流隨著磁場線將能量傳遞到黑洞兩極附近的電漿態物質，使其以噴流的形式自黑洞兩極流出，此過程為 Blandford–Znajek process[3]，為目前 AGN Jets 形成假說之一。

在大約 1/10 的類星體中，觀測光譜中包含了強烈的無線電波段訊號，其餘則是相當微弱[4]，由此分類為 Radio Loud 類星體與 Radio Quiet 類星體[5]。觀測上以無線電響度參數$R = L_{6cm}/L_{2500\text{Å}} \geq 10$作為一類星體是 Radio Loud 類星體的標準[6]，L_{6cm}、$L_{2500\text{Å}}$ 代表波長為 6 cm 與2500Å 的單色光波長(以無線電光度與短波紫外光做比較)。目前對於 Radio Loud 類星體如何形成強烈無線電訊號的所有原因尚未確定，但普遍認為 Radio Jets，也就是發出無線電訊號的 AGN Jets 為主要因素[4]。使用 RL 與 RQ 簡稱代表 Radio Loud 與 Radio Quiet。



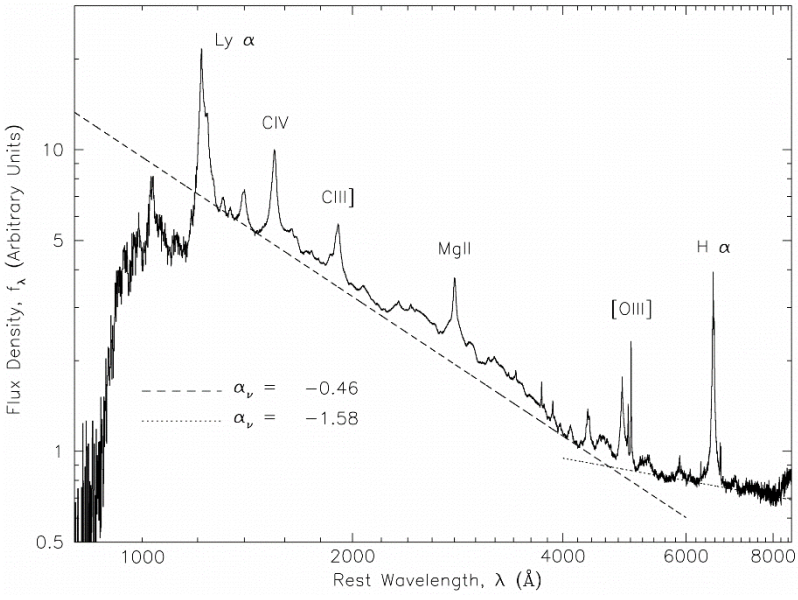
圖：磁場線受黑洞扭曲示意圖[3]



圖(左上)：Radio Loud 類星體光學影像[7]
圖(左下)：Radio Quiet 類星體光學影像[7]
圖(右上)：Radio Loud 類星體無線電影像[8]
圖(右下)：Radio Quiet 類星體無線電影像[8]

類星體上具有 Radio 訊號所帶來的分別十分明顯，因此本研究認為在可見光光譜中應存在 Radio Loud 與 Radio Quiet 類星體的性質差異，能驗證 Radio Quasar 成因或是噴流對於 AGN 各種構造發光機制影響的關聯假說。

(三) SDSS Quasar catalog 與 LoTSS 所提供的資料



圖：類星體觀測光譜觀測圖[10]

1.光譜資料與 SDSS Quasar Catalog[9]

在大部分類星體觀測中，光譜資料通常是唯一、極為重要的資料。其中分為連續光譜與發射光譜。連續光譜(continuum)如圖中虛線所展示，分布型態由光源本身的熱輻射所決定。在圖中偏離虛線的通量密度突起則是發射光譜(Emission spectrum)，由單一的分子、原

子或離子上的電子離開激發態到達低能量軌道所發出的特定頻率光。

本研究將著重發射光譜為主，連續光譜為輔的資料進行研究。本研究分析時主要使用目錄中提供的變數種類如下：

(1) 全波段熱光度(Bolometric luminosity)：由可見光譜特定波段光度擬合的涵蓋所有波段、該類星體電磁輻射的總光度。一個類星體具有一筆。

(2) 紅移值(Redshift, z)：紅移是電磁波由於相對運動、相對論性引力場、空間的度規膨脹等原因導致波長增加、頻率降低的現象。紅移值便是用來衡量紅移程度的數值，定義為$z = (\lambda - \lambda_0)/\lambda_0$，其中 z 是紅移值，$\lambda$是觀測到的波長，$\lambda_0$是光波原波長。一個類星體具有一筆。

(3) 譜線光度(Line luminosity)：發射光譜中單一譜線範圍的光度。目錄提供H_α、H_β、MgII、CIV、NiI、SII、OIII 譜線的光度數據。

(4) 等效寬度(Equivalent width, EW)：對於每一條光譜線，一個高度為連續光譜強度，面積為發射光譜線超過連續光譜部分的面積的矩形，其寬度為等效寬度。可用來衡量其發射光譜相對於連續光譜的能量差異。目錄提供H_α、H_β、MgII、CIV、NiI、SII、OIII 譜線的 EW 數據。

(5) 半峰全寬(Full width at half maximum, FWHM)：對於每一條光譜線，由於類星體旋轉運動使各處對於觀察者的相對運動速度不同，使得譜線因紅移值不一而寬度增加，目錄提供的半峰全寬為光度降至峰值一半的波長差所推算得到的相對運動速度差值。目錄提供H_α、H_β、MgII、CIV 譜線的 FWHM 數據。

(6) 連續光譜冪律斜率(Power-law slope, α_λ)：對於類星體的連續光譜分布，可以局部地以冪律來描述：$f_\lambda \propto \lambda^{\alpha_\lambda}$，$\lambda$為波長、$f_\lambda$為在該波長的電磁波通量密度。目錄提供$H_\alpha$、$H_\beta$、MgII、CIV 譜線附近的$\alpha_\lambda$。

(7) 連續光譜光度(continuum luminosity)：利用沒有譜線的區域，得到連續光譜所帶來的光度。目錄提供位於 5100 Å, 3000 Å 與 1350 Å 的連續光譜光度。

由於光度變化巨大，因此本研究都將使用其對數值進行分析。

2.LoTSS 提供無線電資料[8]

本研究使用的無線電波源目錄來自 LoTSS(LOw-frequency array Two-metre Sky Survey)的第二次資料釋出。該目錄涵蓋了大概 27%北方天空，包含 120-168 MHz 無線電訊號波源。該目錄所帶來的資料如下：

(1)Total_flux：該類星體於 120-168 MHz 的總輻射通量

(2)Peak_flux：該類星體所發出最大單色輻射通量(120-168 MHz)

本階段研究以 LoTSS 資料與 SDSS Quasae Catalog 資料交集標記 Radio Loud 類星體為主。

(四) 紅移與 SDSS 觀測限制[11]

本研究使用可見光資料為 SDSS 第七次資料釋出提供的類星體目錄，此時期該計畫所能捕捉的光波長範圍為 3800-9200 Å，當一顆類星體的紅移值過大時，我們所關注的某些可見光譜線將會移動到可觀測範圍外，產生資料缺失。

為確保我們之後所做的明暗比例分析不受無效資料的影響，我們針對各譜線須設定一個極限紅移值$z_{limit}(\lambda)$，紅移值低於$z_{limit}(\lambda)$才為有效資料。

$$\lambda(z+1) < \lambda_{range} = 9100(\text{Å})$$

$$z_{limit}(\lambda) = 9100/\lambda - 1$$

使用 9100 作為波長最大值，是為了確保光譜線不被觀測限制部分截斷。以下是各譜線波長λ與其$z_{limit}(\lambda)$值：

	波長(Å)	z_{limit}
H alpha	6563	0.3866
H beta	4861	0.8719
NiI6585	6585	0.3667
SII6718	6718	0.3397
SII6732	6732	0.3368
OIII4959	4959	0.8149
OIII5007	5007	0.8175

MgII	2803	2.2465
CIV	5812	3.3605

貳、研究方法或過程

一、使用的 python package

本研究以程式操作資料集為主，以下為使用到的 python package：

	Numpy[12]
	Matplotlib[13]
	Astropy[14]
	Sklearn[15]
	pandas[16]

二、合併 SDSS 與 LoTSS 資料

(一) match table 之使用

本研究的資料來自兩個不同的研究計畫，所使用之資料格式 FITS(Flexible Image Transport System)，研究中將會時常需要 Topcat 軟體和 Astropy 與資料互動。其中 match table 功能能夠使用座標資料合併資料。其中重要的參數為 Max Error。選定一個 Max Error 值，軟體會將 LoTSS 與 SDSS Quasar Catalog 中座標相差小於 Max Error 的各一筆資料合併為新表格的同一筆資料。意即將兩個不同研究計畫的天體資料視為同一個類星體之數據。

本研究使用的目錄定位類星體使用的是赤道坐標系統。RA 代表赤經，DEC 代表赤緯，單位為 arcsec(角秒)。所使用的曆元(座標系統之參考時間點)為 J2000.0。

(二)Max error 選擇優化

合併兩個天文台資料時，由於測量誤差，同個星體在天空中的座標不會完全相同，因此 Max Error 的選擇十分重要，不適當的 Max Error 可能會帶來以下兩種影響：

(1)Max Error 過高：將兩個不同類星體誤列為同一個，汙染資料。

(2)Max Error 過低：剔除過多正確資料，不利後續分析。

上述兩個情況中，成功配對的資料數量與 Max error 的關係應呈現不同的增長趨勢，因此我們將使用以下方法找出最佳 Max error 值：

Step 1：使用不同的 Max error(arcsec)值，並紀錄配對資料數量。

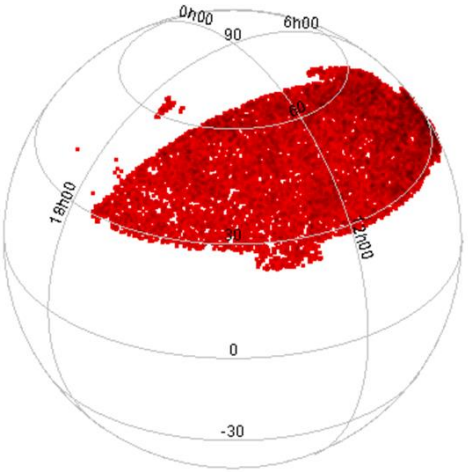
Step 2：將 Max error 與配對資料數量均取對數。

Step 3：遍歷 0.05 至 100 間的所有 Max error 值，將大於該值的資料與小於該值的資料進行對數線性回歸，計算加權R^2平均值。其中最大值出現的點便為趨勢轉折點，也就是我們認為的最佳 Max error 值。

(三)範圍控制

本研究合併資料集目的為找出同時是無線電波源與類星體的 Radio Loud 類星體資料。經過前述步驟選出最佳 Max Error 後，仍需要確保選擇研究的類星體都位於兩研究資料皆大量涵蓋區域，以確保：

1. 兩資料同時包含的類星體為 radio loud 類星體
2. 僅在 SDSS Quasar Catalog 中出現的類星體為 Radio Quiet 類星體



圖：研究座標範圍天球圖

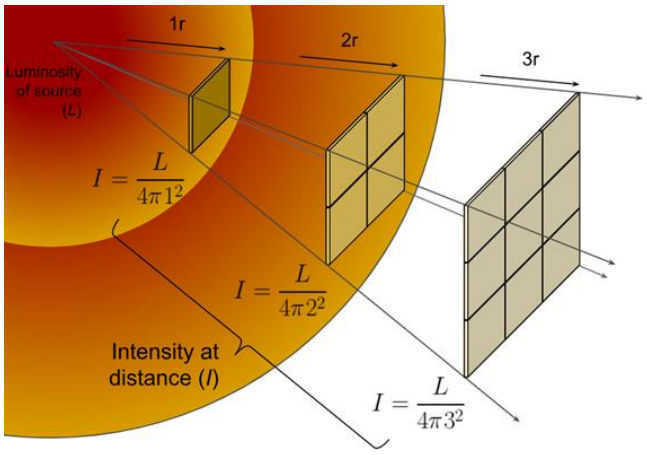
因此將研究範圍限定於兩研究目錄皆主要涵蓋的區域：

RA:9-15(hr)、DEC:30-60(度)

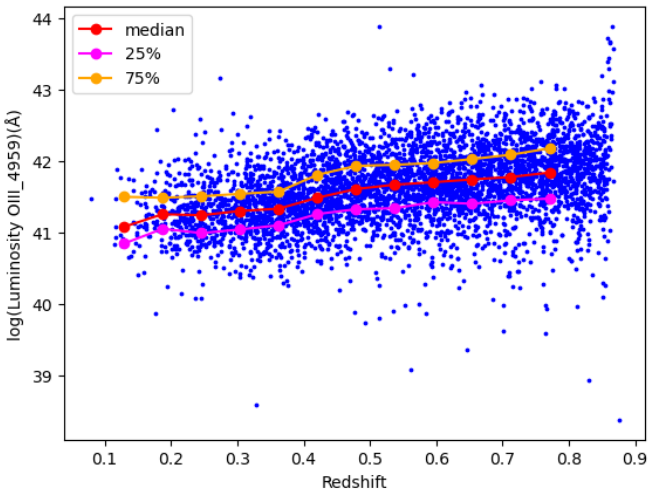
三、 單變數分析：紅移分布控制方法與初步觀察

(一)紅移分布控制

在研究過程中，由於天文台觀測星體時，其接收到的亮度需大於一個固定極限。因此當一類星體距離地球越遠時，我們觀測到的類星體光度便會提升，而與距離有關的紅移值也因此變成資料集中與光度相關的變數。



圖：光度相同下，距離越遠亮度越小[17]



圖：OIII4959 譜線光度隨紅移變化

為了確保紅移作為控制變因，我們使用以下流程自 Radio Quiet 的資料中選擇做為觀察對照組的類星體：

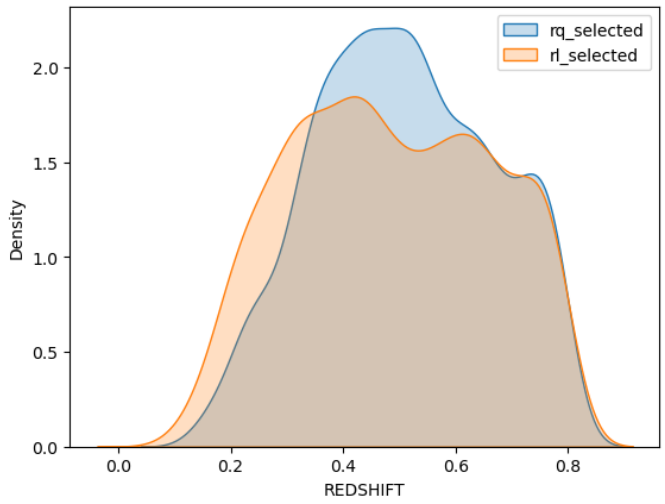
Step 1：篩除 Radio Loud 資料中所有 z>0.8 的類星體資料。

Step 2：對於 Radio Loud 篩選後資料中的每一筆類星體資料，選擇 Radio Quiet 中紅移值最接近的一筆類星體資料。

Step 3：篩除 Radio Quiet 資料中未在第二步被選擇者。

```
# 選擇具有紅移小於0.8的radio_loud數據，並將其存儲在rl_select中
rl_select = radio_loud[radio_loud['REDSHIFT'] < 0.8]

# 複製radio_quiet數據到rq_select
rq_select = radio_quiet.copy()

# 創建一個新列'sel'，並將其初始化為0
rq_select['sel'] = 0

# 遍歷rl_select中的每一筆數據X
for _, X in rl_select.iterrows():
    # 創建一個新的DataFrame "datacopy"，複製rq_select中所有'sel'等於0的數據
    datacopy = rq_select[rq_select['sel'] == 0].copy()

    # 按照abs(X['REDSHIFT']-Y['REDSHIFT'])的值從小到大排序datacopy
    datacopy['diff'] = abs(X['REDSHIFT'] - datacopy['REDSHIFT'])
    datacopy.sort_values('diff', inplace=True)

    # 選擇其中第一項（差值最小的數據），將rq_select中對應的數據的'sel'設置為1
    min_diff_index = datacopy.index[0]
    rq_select.loc[min_diff_index, 'sel'] = 1

# 最後刪除rq_select中所有'sel'等於0的數據
rq_select = rq_select[rq_select['sel'] == 1]
```

圖：紅移分布控制的程式碼



圖：控制後的紅移分布

控制後的分布資訊如下：

	類星體數量	紅移中位數	紅移標準差
Radio Loud	2000	0.4773	0.1794
Radio Quiet	2000	0.5060	0.1580

在 z<0.8 的條件限制下，可以用於分析的譜線為：

H beta, OIII4959, OIII5007, MGII, 。

(二)初步觀察

為了檢驗 RL 類星體在那些觀測變數上與 RQ 有較顯著差異，以利後續建立模型可用來輔助判斷之依據。我們使用以下流程進行觀察分析：

Step 1：選定紅移最大值，只採取紅移值小於最大值的類星體進行分析

Step 2：繪製 RL 與 RQ 於一變數的直方分布圖，直接觀察兩分布的差異

Step 3：計算兩分布平均值與標準差，確認是否有顯著、可用於判斷之差異

(三)LOFAR-detected 類星體比例變化：單變數與多變數分析

1.單變數分析

從上部分的分析結果得知，當一類星體能在這些譜線上具有較高光度時，其為 Radio Loud 類星體的機率應更高。本研究使用以下流程進行分析：

Step 1：計算該光譜線變數的平均值μ與標準差σ(RL 與 RQ 合併計算)

Step 2：從$\mu - 3\sigma$到$\mu + 3\sigma$將變數分為 12 個區段。

Step 3：計數 Radio Loud 類星體與 Radio Quiet 類星體於各區段的數量。

Step 4：計算於該區段抽樣一類星體屬於 Radio Loud 機率$p = N_{RL}/(N_{RL} + N_{RQ})$

Step 5：

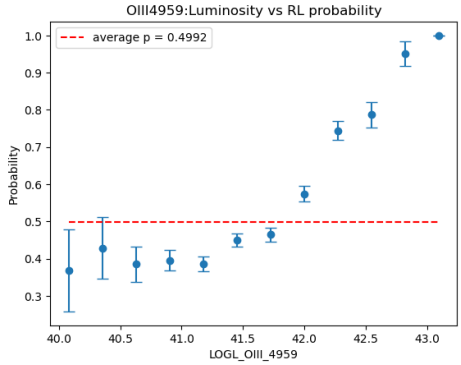
計算於該變數無缺失值所有類星體抽樣一類星體屬於Radio Loud的機率$p_{average}$

Step 6：比較各區段p與$p_{average}$的差異，即可檢驗該譜線變數大小與無線電訊號的關聯性。

為了量化該變數所為判斷帶來的有利程度，本研究採用 Gini Impurity 變化ΔG來衡量，Gini Impurity 之定義為：

$$Gini\ Impurity(S) = 1 - p_{loud}^2 - p_{quiet}^2$$

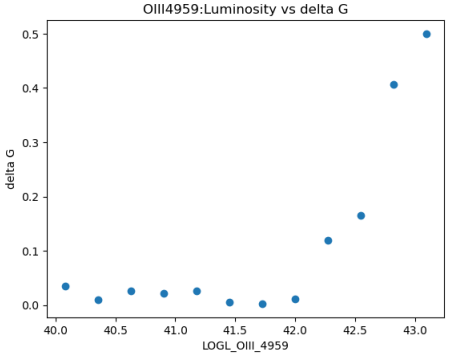
S 為任意一個集合。p_{loud}、p_{quiet}代表在該集合中 radio-loud、radio-quiet 比例。$\Delta G = G(average) - G(A)$。$G(average)$、G(A)分別為平均情況下、區段 A 中的 Gini Impurity。(Gini Impurity 高代表資料欠缺分類，因此ΔG越大代表對分類情形帶來的改善越多)為了避免一個區段中類星體數量過少而導致比例變化過大，本研究另外計算該比例的二項式標準誤：$\Delta p = \sqrt{p \times (1-p)/N_A}$，$N_A$為區段中資料量，$p$為其中 Radio Loud 類星體比例。



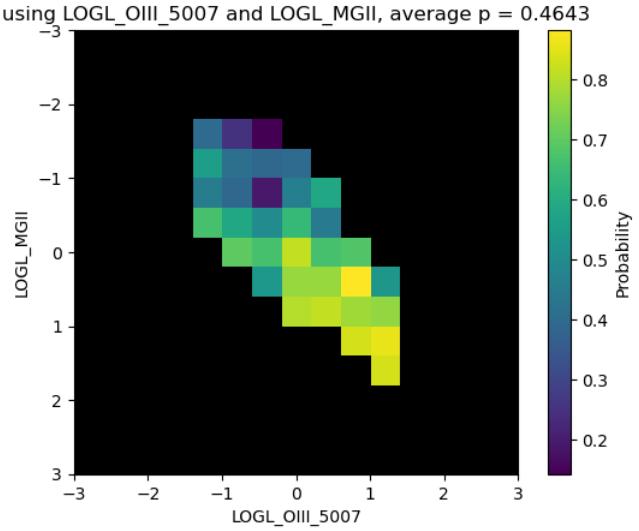
圖(左)：
RL 機率(比例)對光度趨勢

圖(右)：
轉換為 Gini 不純度
下降值對光度趨勢

(以 OIII4959 為例)



2.多變數分析



圖：考慮雙光度影響的 RL 比例熱力圖

延續上部分結果，同時考慮兩種不同譜線光度對 Radio Loud 類星體比例的影響，並將其繪製為熱力圖以便觀察趨勢，x,y 軸標示為變數 z-score。採用$14 \times 14 = 196$個區段進行觀察，並刪除所有資料量< 5的區段。並將考慮雙變數的區段ΔG與單變數得到的ΔG進行比較。依照改善情況與兩譜線光度的相關係數，推論是否有超過一種因素影響 Radio Jets 的形成。

(四)以去除分布影響之中位數比較法驗證前述比例分析方法

1. 中位數比較法

在前述分析中，我使用的都是經過預先篩選控制紅移，RL 與 RQ 類星體數量相同，事實上目前觀測到的 RL 與 RQ 類星體數量大致為 10:1。為了避免因為控制資料數量，而帶來統計分析的誤差。因此我採取以下方法控制紅移進行分析，並以此驗證前部分之結果，此部分將使用完整、未經紅移分布控制的 Radio Loud 與 Radio Quiet 類星體資料：

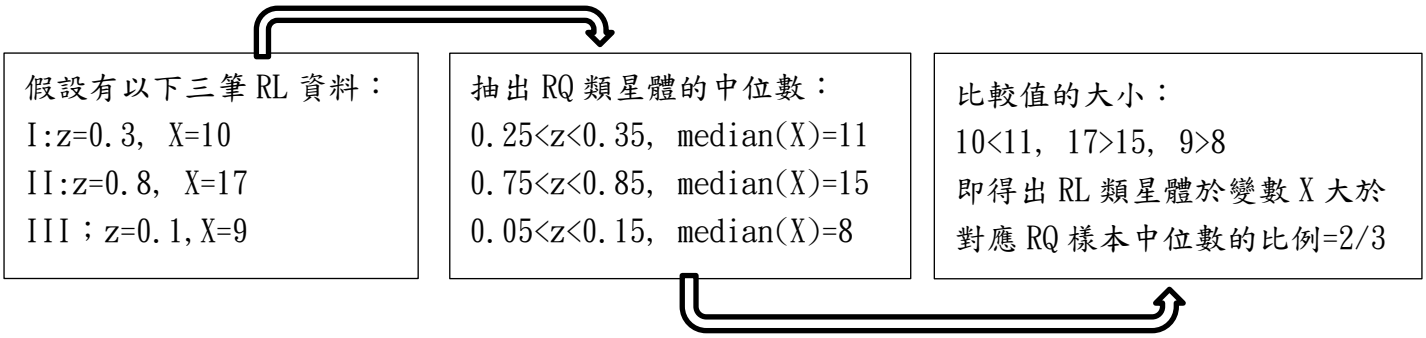
Step 1：選定要分析的觀測變數，選擇一個 RL 類星體 X。

Step 2：計算 RQ 類星體中紅移與 X 相差不超過 0.05 之樣本的變數中位數，並比較其與 X 該變數值的大小。

Step 3：重複前述步驟直到遍歷所有 RL 類星體，計算 RL 類星體於該變數值大於對應 RQ 樣本中位數的比例。

對於其中控制範圍採用 0.1 的原因，我們檢視(一)紅移分布控制中將 MgII 對數光度與紅移值做迴歸直線，斜率約為 0.713。在 $0.7<z<0.8$ 的區段中，MgII 對數光度標準差為 $0.327>0.0713$，因此我們認為此時紅移對光度的影響非主要因素。

另外，為探討 Radio Jets 與窄線區域或寬線區域何者較有關連，本研究另外控制 Broad H beta 與 Narrow H beta 進行中位數分析。



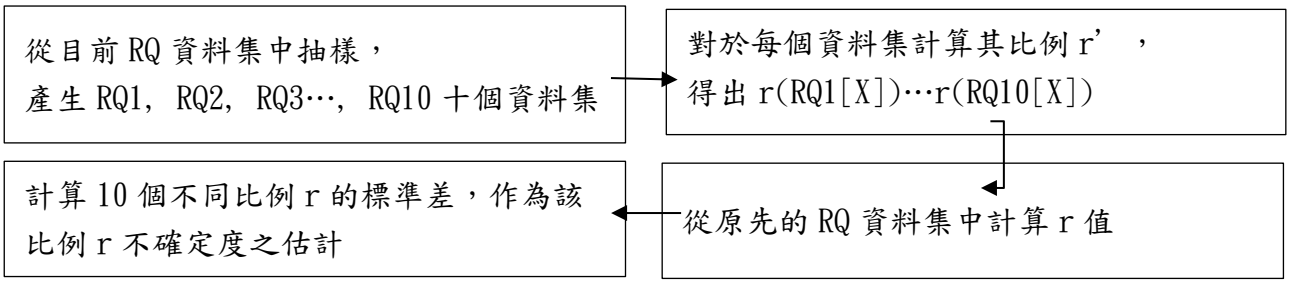
2. 使用 bootstrap method 估計不確定度

為了估計前述比例值條件控制穩定性，本研究採用以下流程估計不確定度：

Step 1：從 RQ 資料中隨機抽樣，產生十個新的 RQ 資料測試集。

Step 2：利用前述流程計算十個 RQ 資料集對應的比例值。

Step 3：計算十個比例值的標準差，即為我們採用的比例不確定度。



(五) 決策樹預測：訓練過程與參數優化

1.決策樹訓練方法

決策樹分類器是一種利用樹狀資料結構來進行一系列數據條件判斷操作的分類器。在每一步的訓練中，找到一個數據分類點，使資料盡可能的分類一致，並在枝葉長度到達閥值或分類一致性到達標準停止訓練。本研究採用 Gini Impurity 用來衡量分類一致程度：

$$Total\ Gini\ Impurity = Gini\ Impurity(A) + Gini\ Impurity(B)$$

其中 A、B 代表經過該節點分類後的兩個資料集合。訓練流程如下：

(1)資料標籤與預處理：將 RL 資料與 RQ 資料標註為兩個類別，並將缺失值設

定為全體平均。

(2)隨機抽取資料作為測試集：在每一批次訓練中，隨機抽取 20%資料作為評估模型表現的測試資料集。剩餘資料則做為訓練模型的資料集。

(3)批量訓練決策樹：為保持決策樹多樣化，每棵決策樹訓練時將隨機抽樣訓練集的 80%資料作為訓練集，由於不同紅移值可用的參數維度不同，因此此研究將會以提到的紅移極限作為不同批次分別訓練。

(4)選擇測試集表現佳的決策樹：藉由測試集資料判斷上一步訓練出的決策樹，選擇出測試集正確率最高的決策樹。

(5)可視化與決策邊界圖：上一步篩選出的若干決策樹，我們利用內建可視化工具了解決策樹判斷的條件。

2.優化參數

為了提升性能以及避免 overfitting，本研究對決策樹的以下兩個訓練參數進行隨機搜索優化。

(1) Max depth：該決策樹最大層數。

(2) min samples split：若一個節點的資料量小於該值則不再分枝。

3. 利用 sklearn 的 feature importance 模組進行變數重要性評估。

由於紅移影響，本研究對不同紅移範圍的類星體資料以不同的決策樹模型進行分析，將分為：

I：0<z<0.3(無紅移分布控制),

II：0<z<0.3(有紅移分布控制),

III：0.3<z<0.8(無紅移分布控制),

IV：0.3<z<0.8(有紅移分布控制)。

加入 SDSS Quasar Catalog 提供的估計黑洞質量作為變數之一。

(七) 使用多決策樹之集成學習

單一的弱分類器可能會產生分類成效不佳、過擬合之問題。因此我們可以集合大量的弱分類器，運用集成學習的力量打造更好的分類器，本研究考慮採用的方法有以下兩種：

1.隨機森林：此演算法是一種 Bootstrap 聚合（Bagging）集成學習。透過隨機的選擇部分訓練資料、部分特徵，我們能夠訓練出許多棵不同的決策樹，透過讓所有決策樹公平的投票決定分類結果，我們能夠避免過擬合，提升預測能力。

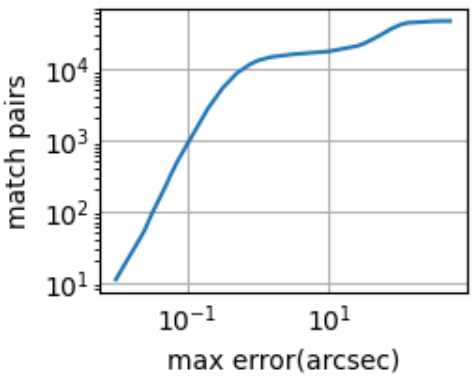
2.Adaboost：Adaboost 代表”Adaptive Boosting”，其中最重要的自適應(Adaptive)部分實現方法是將前一個弱分類器分錯的樣本加強權重，加權後的全體樣本再次用來訓練下一個弱分類器。並在每一輪訓練加入一個新的弱分類器，訓練直到訓練集錯誤率小於預先設定的值或是到達最大訓練次數為止。本研究中使用的弱分類器為決策樹。

此部分同樣使用利用 sklearn 的 feature importance 模組進行變數重要性評估，並分為同樣的 I, II, III, IV

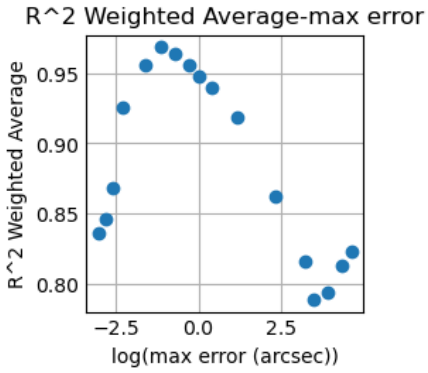
伍、研究結果與討論

(一)Max error 選擇

通過前述的分析，我們得到以下 Max Error 與配對數量之關係：



圖：配對數量與 Max Error 關係



圖：加權 R^2 與對數 Max Error 關係

在前述過程中我們發現 Max Error = 0.316 arcsec 為加權R^2最大值發生時，但實

際上位於 Max Error = 1 arcsec 時加權R^2依舊大於 0.95。

通過圓孔繞射的 Rayleigh criterion：

$$\theta_{min} = \frac{1.22\lambda}{a}$$

λ代表觀測光波長，a代表圓孔孔徑。

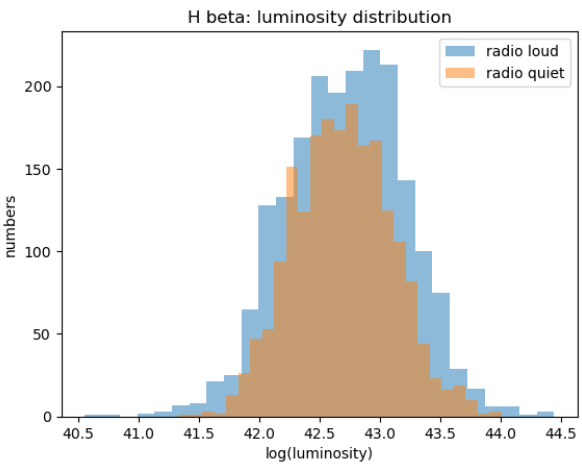
帶入解析度較小的 LOFAR 相關數據計算後，解析度範圍為 0.3~7.54 (arcsec)，因此我們有足夠信心採用 Max Error = 1 arcsec，作為可接受並且使資料量足夠的選擇。

(二)初步觀察-各譜線光度

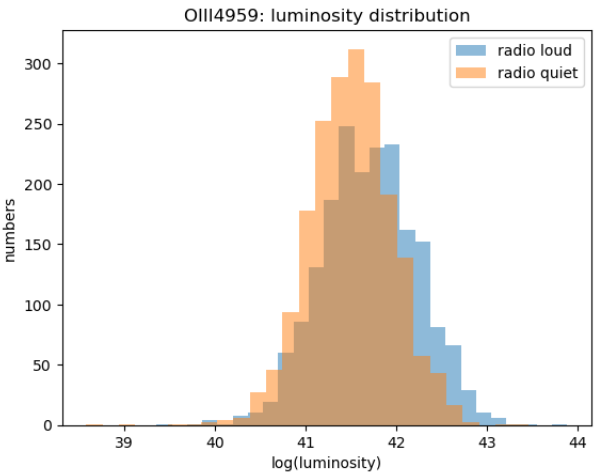
在經過紅移控制後，我們觀察 H beta, OIII4959, OIII5007, MGII 譜線的光度分布在 Radio Loud 與 Radio Quiet 中的差異。

類星體的平均值與標準差(僅計算有訊號的類星體)。

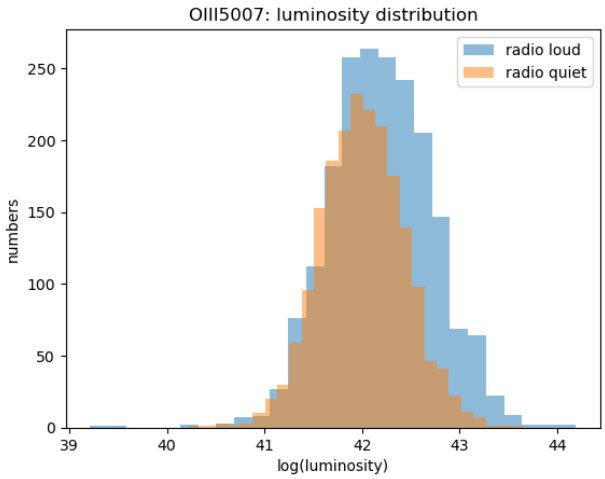
	Radio Loud average	Radio Quiet average	Radio Loud std	Radio Quiet std
H beta	42.7157	42.6924	0.4962	0.4041
OIII4959	41.7076	41.5042	0.5286	0.4626
OIII5007	42.2022	42.0044	0.5269	0.4288
MgII	43.0377	42.9462	0.4308	0.3790



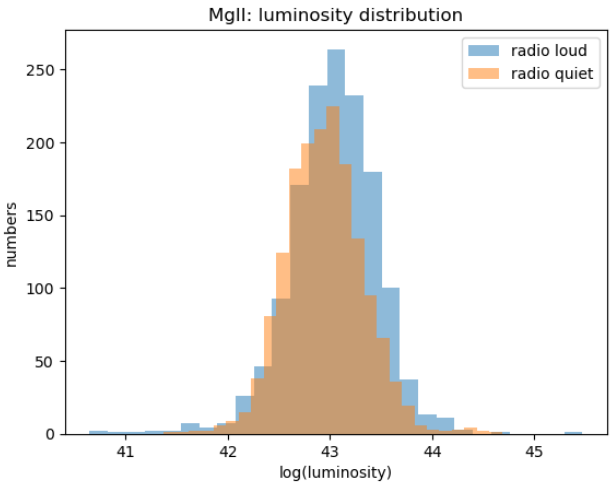
圖：H beta 單色光度分布



圖：OIII4959 單色光度分布



圖：OIII5007 單色光度分布



圖：MgII 單色光度分布

從分布圖可觀察出在 OIII4959, OIII5007, MgII 光譜線上，Radio Loud 的光度有更高的趨勢，但由於分布高度重疊，我們並無法直接利用單頻率光度區分一類星體是否為 Radio Loud。

(四) 初步觀察-z<0.8 各譜線等效寬度(EW)

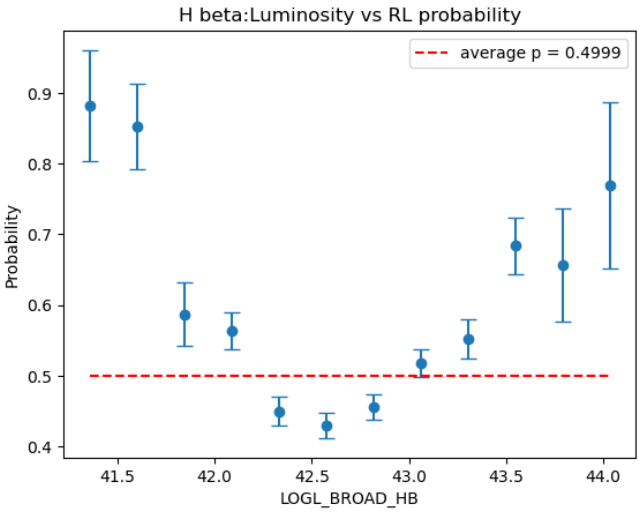
在經過紅移控制後，我們觀察 H beta, OIII4959, OIII5007, MGII 譜線的 EW 分布在 Radio Loud 與 Radio Quiet 中的差異。

	Radio Loud average	Radio Quiet average	Radio Loud std	Radio Quiet std
H beta	70.6173	75.9656	56.9650	40.9064
OIII4959	10.8138	6.1028	19.6911	5.9062
OIII5007	32.4033	18.7016	43.9822	17.1241
MgII	71.0821	65.3636	671.0297	288.7063

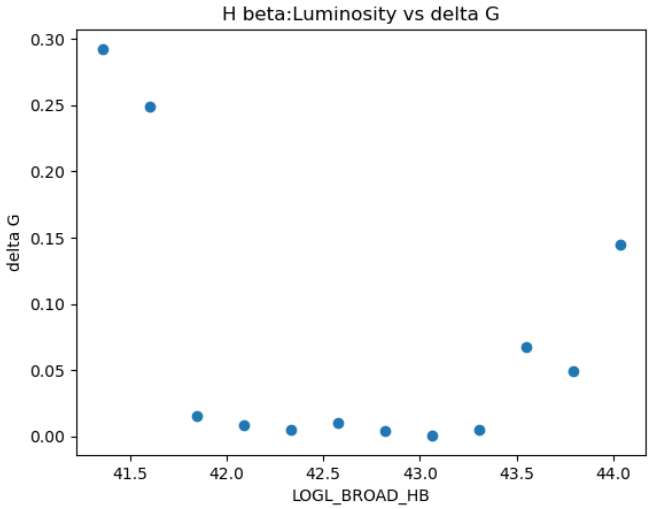
於 OIII4959、OIII5007 譜線上發現可能的 Radio Loud 大於 Radio Quiet 趨勢，H beta 可能具有相反的趨勢，但由於資料離散程度過大，只能於決策樹部分分析其重要性。

(六) LOFAR-detected 類星體-光度影響 RL 比例分析

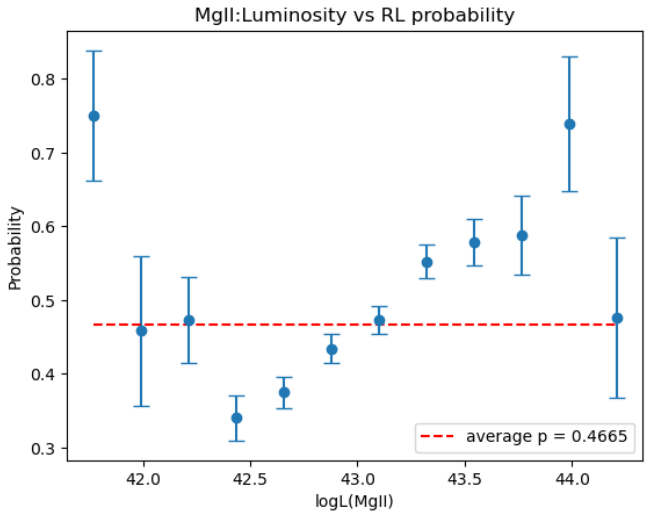
使用前述的分析方法後，我們得到以下 RL 比例(機率)與其轉換成ΔG的關係圖。



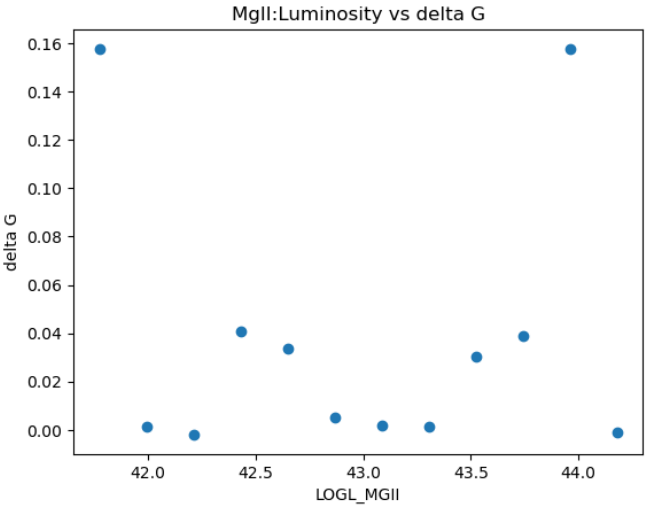
圖：H beta 單色光度 p-L 圖



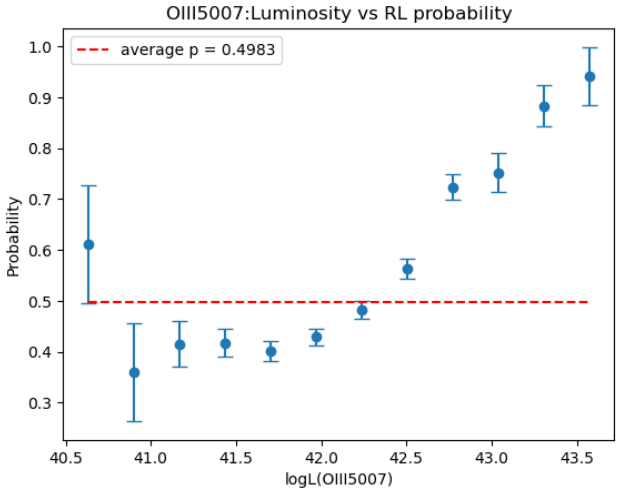
圖：H beta 單色光度ΔG-L 圖



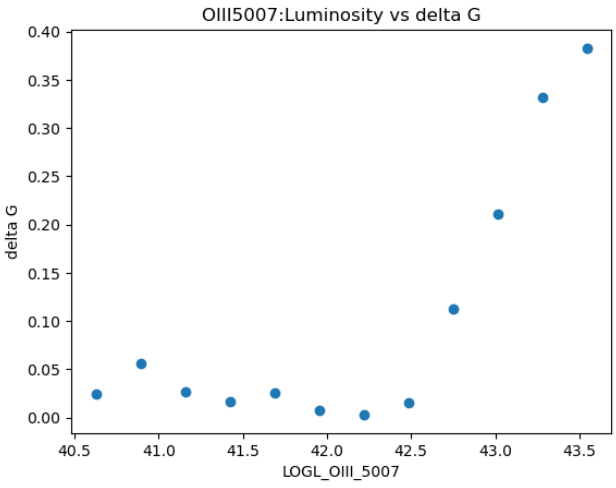
圖：MgII 單色光度 p-L 圖



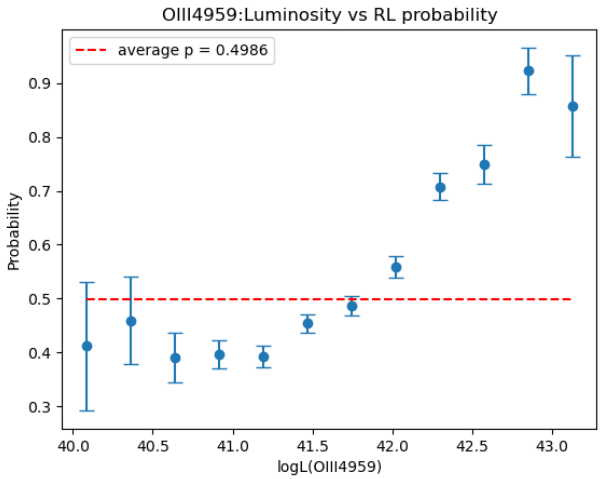
圖：MgII 單色光度ΔG-L 圖



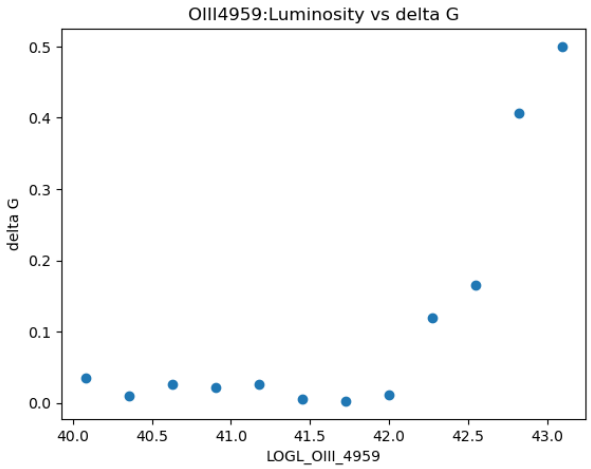
圖：OIII5007 單色光度 p-L 圖



圖：OIII5007 單色光度ΔG-L 圖



圖：OIII4959 單色光度 p-L 圖

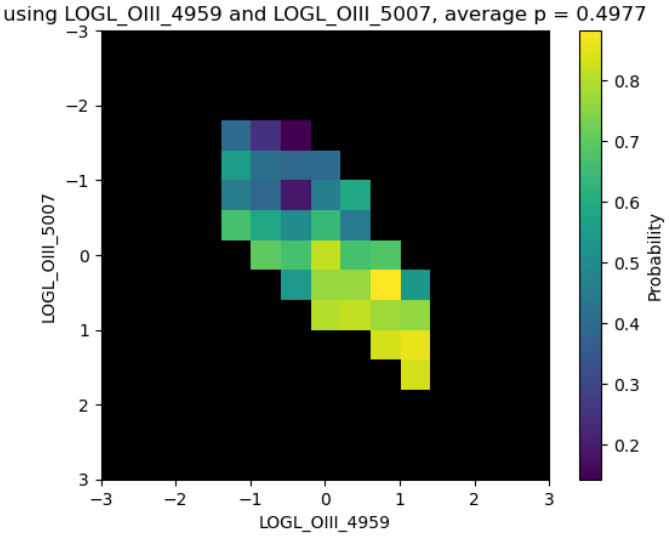


圖：OIII4959 單色光度ΔG-L 圖

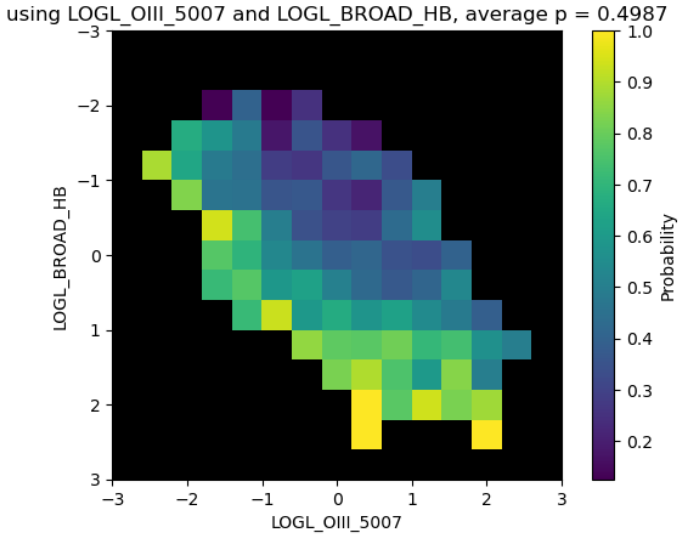
光譜線名稱	平均比例	平均比例的 Gini Impurity	最大ΔG	前三高的ΔG平均
H beta	0.4999	0.5000	0.2924	0.2288
MgII	0.4665	0.4978	0.1576	0.1186
OIII4959	0.4983	0.5000	0.5000	0.3574
OIII5007	0.4986	0.5000	0.3828	0.3088

可以發現當一類星體在 OIII4959、OIII5007 亮度小於對數平均值時，RL 類星體佔整體比例並無顯著差異。但當光度大於對數平均值時，一類星體為 RL 的機率大致遞增且大於平均機率。 在 H beta 則是不論增加或減少光度偏離平均值一標準差以上時 RL 比例增加。MgII 上則無明顯趨勢。

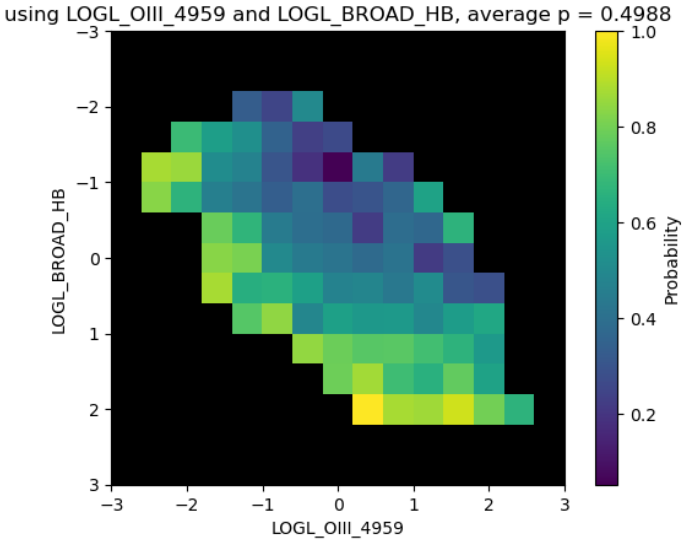
(四) LOFAR-detected 類星體比例變化(多變數)



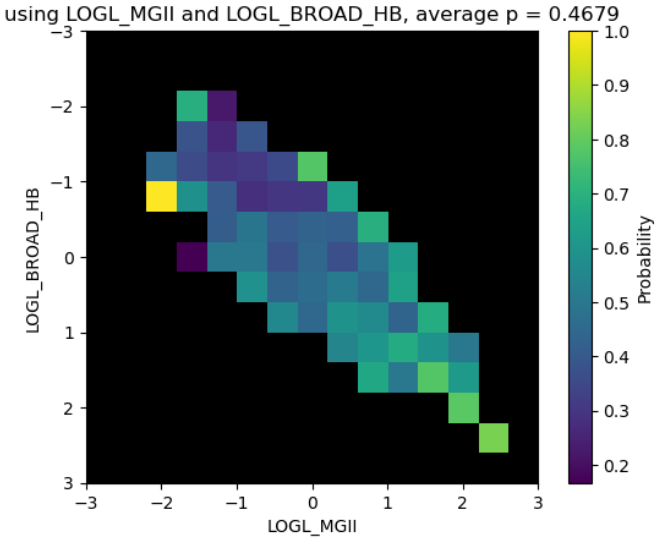
圖：OIII4959、OIII5007 單色光度 p-L 圖



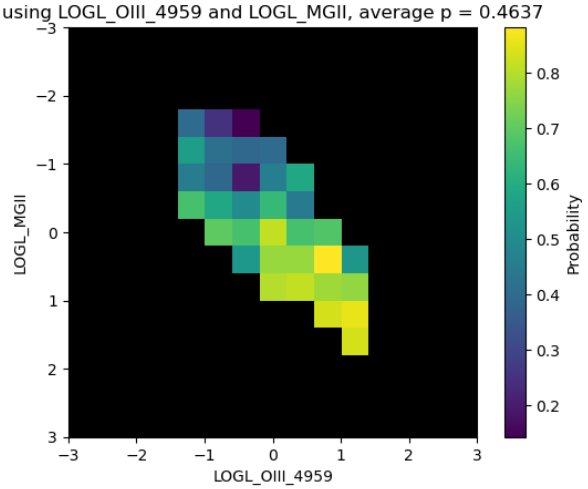
圖：H beta、OIII5007 單色光度 p-L 圖



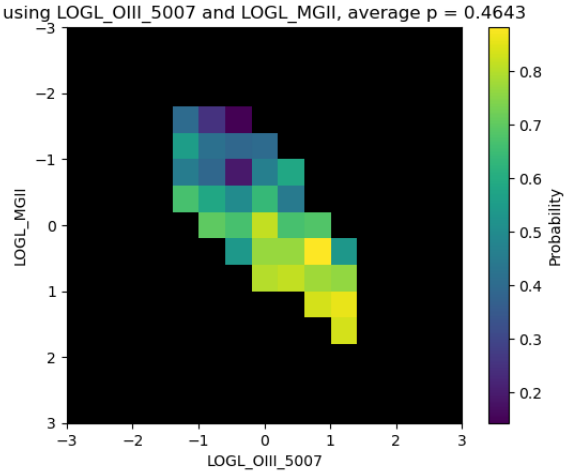
圖：OIII4959、H beta 單色光度 p-L 圖



圖：MgII、H beta 單色光度 p-L 圖



圖：OIII4959、MgII 單色光度 p-L 圖



圖：OIII5007、MgII 單色光度 p-L 圖

可以觀察到，當一類星體能同時在兩條光譜線上獲得大於對數平均值的光度，屬於 Radio Loud 的機率相較於只考慮一條光譜線有提升。 檢驗其前三高的ΔG平均(相同代表僅考慮單條譜線)：

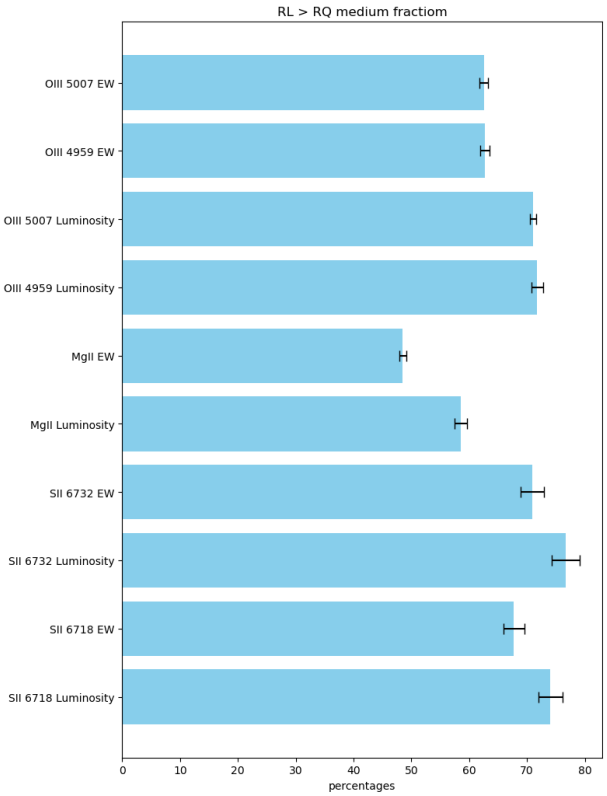
光譜線名稱	H beta	OIII4959	OIII5007	MgII
H beta	0.2288			
OIII4959	0.3856	0.3574		
OIII5007	0.4599	0.4466	0.3088	
MgII	0.2931	0.4007	0.3220	0.1186

其中 OIII 系列的譜線顯示出相較另外兩條譜線更強的趨勢性。

(五) 單變數分析：控制紅移與 H beta 譜線光度

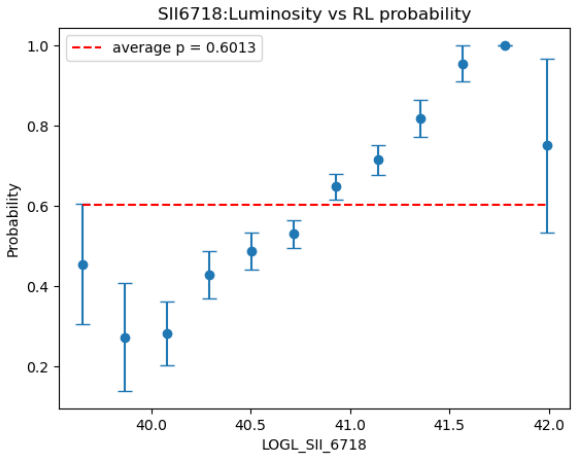
在經過前述中位數分析法後，我們得到以下的結果(控制住紅移後的結果)：

變數	RL 大於 RQ 中位數的比例
SII 6718 光度	74.06±2.08%
SII 6718 等效寬度	67.74±1.80%
SII 6732 光度	76.71±2.36%
SII 6732 等效寬度	70.89±2.05%
MgII 光度	58.58±1.11%
MgII 等效寬度	48.50±0.62%
OIII 4959 光度	71.77±1.04%
OIII 5007 光度	71.10±0.56%
OIII 4959 等效寬度	62.70±0.78%
OIII 5007 等效寬度	62.55±0.73%

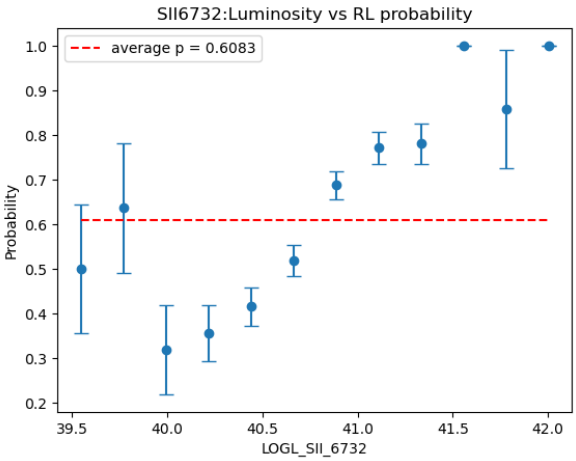


圖(左)：中位數分析法結果

從此部分結果，我們驗證了在 OIII 譜線上較強的趨勢性。SII 譜線也同樣展示出趨勢性，我們使用 z<0.3 的資料檢視 SII6718、SII6732 譜線的光度比例性質，也同樣具有光度越大，RL 比例越高趨勢。



圖：SII6718 單色光度 p-L 圖



圖：SII6732 單色光度 p-L 圖

(六)決策樹與隨機森林分析

precision（精確率）：指的是模型在預測為正例的樣本中實際為正例的比例。

recall（召回率）：召回率是模型在所有實際正例中正確辨識的正例的比例。

對於 I：0<z<0.3(無紅移分布控制)資料，我們得到以下結果：

使用隨機搜索，得到參數最佳化 best_max_depth=6,best_min_samples_split=20。得到準確率為 69%、67%、65%、73%、70%的五個決策樹模型，對於準確率最高(73%)的決策樹，我們利用分枝後的 Gini Impurity 評估各變數的重要性，其中重要性前二高的變數為 NII_6585 光度與 redshift。

	precision	recall
Radio loud	80%	80%
Radio quiet	56%	56%

使用由 100 個決策樹集合而成的隨機森林模型，正確率為 75%

	precision	recall
Radio loud	78%	89%
Radio quiet	64%	44%

使用由 50 個決策樹集合而成的 Adaboost 模型，正確率為 80%

	precision	recall
Radio loud	80%	94%
Radio quiet	80%	50%

對於 II：0<z<0.3(有紅移分布控制)資料，我們得到以下結果：

使用隨機搜索，得到參數最佳化 best_max_depth=6,best_min_samples_split=20。得到準確率為 65%、66%、69%、67%、75%的五個決策樹模型，對於準確率最高(75%)的決策樹，我們利用分枝後的 Gini Impurity 評估各變數的重要性，其中重要性前二高的變數為 SII_6718 光度與 redshift。

	precision	recall
Radio loud	80%	80%
Radio quiet	56%	56%

使用由 100 個決策樹集合而成的隨機森林模型，正確率為 75%

	precision	recall
Radio loud	82%	80%
Radio quiet	59%	62%

使用由 50 個決策樹集合而成的 Adaboost 模型，正確率為 76%

	precision	recall
Radio loud	85%	80%
Radio quiet	61%	69%

對於 III：0.3<z<0.8(無紅移分布控制)資料，我們得到以下結果：

使用隨機搜索，得到參數最佳 best_max_depth=22,best_min_samples_split=2。得到準確率為 65%、62%、61%、60%、63%的五個決策樹模型，對於準確率最高(65%)的決策樹，我們利用分枝後的 Gini Impurity 評估各變數的重要性，其中重要性前二高的變數為 H beta 等效寬度與 redshift。

	precision	recall
Radio loud	80%	80%
Radio quiet	56%	56%

使用由 100 個決策樹集合而成的隨機森林模型，正確率為 69%

	precision	recall
Radio loud	61%	53%
Radio quiet	72%	79%

使用由 50 個決策樹集合而成的 Adaboost 模型，正確率為 69%

	precision	recall
Radio loud	63%	53%
Radio quiet	73%	80%

對於 IV：0.3<z<0.8(有紅移分布控制)資料，我們得到以下結果：
使用隨機搜索，得到參數最佳 best_max_depth=22,best_min_samples_split=2。
得到準確率為 63%、64%、60%、59%、61%的五個決策樹模型，對於準確率最高(63%)的決策樹，我們利用分枝後的 Gini Impurity 評估各變數的重要性，其中重要性前二高的變數為 H beta 等效寬度與 redshift，與 III 相同。

	precision	recall
Radio loud	60%	57%
Radio quiet	61%	65%

使用由 100 個決策樹集合而成的隨機森林模型，正確率為 68%

	precision	recall
Radio loud	70%	61%
Radio quiet	67%	75%

使用由 50 個決策樹集合而成的 Adaboost 模型，正確率為 67%

	precision	recall
Radio loud	66%	70%
Radio quiet	66%	73%

綜合討論，在紅移值較小時，Radio Loud 類星體在變數上有較有利於分辨的特徵。但在紅移值較大時，由於紅移帶來的影響，使其光度變化受到干擾，因此模型預測的能力受到限制。在大紅移時的 H beta 等效寬度則具有寬度越大越可能為 RL 類星體的趨勢。對於目前研究顯示出的 EW 增強與光度增強現象，我認為與[18]中提到的噴流與 AGN 周圍氣體互動，進而促使其離子化。因此觀測到的 RL 類星體也出現了譜線等效寬度加大(氣體粒子能量受與噴流互動後升高)與光度增強(能量與離子化氣體粒子數量接受噴流影響增加)。

陸、結論與應用

(一)成功結合 LOFAR 資料以及 SDSS 類星體資料，並發現與說明座標距離誤差小於 0.316(arcsec)時為最佳，而 1(arcsec)的誤差選擇是合理且增加資料量的。

(二)在 SII 6718, SII 6732, OIII4959, OIII5007 等光譜發射線上，RL 類星體相較於 RQ 類星體有光度更大的趨勢，並經過兩種方法驗證。

(三) 在 SII 6718, SII 6732, OIII4959, OIII5007, H beta 等光譜發射線上，RL 類星體相較於 RQ 類星體有等校寬度更大的趨勢，其中 H beta 為利用決策樹分析得出結果。

(四)使用單一決策樹模型輔助判別重要變數，並於 z<0.3 的範圍建立準確率 75% 之模型，其餘範圍準確率均不低於 65%。

(五)使用決策樹集成於 z<0.3 的範圍建立準確率 75%的隨機森林模型、準確率 80%的 Adaboost 模型。其餘範圍隨機森林準確率不低於 68%，Adaboost 不低於 67%

柒、參考文獻

[1] Clive Tadhunter, An introduction to active galactic nuclei: Classification and unification, New Astronomy Reviews,Volume 52, Issue 6, 2008, Pages 227-239

[2] National Aeronautics and Space Administration Fermi Learning Center (2016, Feb 17),'Exploring Active Galactic Nuclei', https://fermi.gsfc.nasa.gov/science/eteu/agn/

[3] R. D. Blandford, R. L. Znajek, Electromagnetic extraction of energy from Kerr black holes, Monthly Notices of the Royal Astronomical Society, Volume 179, Issue 3, July 1977, Pages 433–456

[4] Blandford, R., et al. (2019). Relativistic Jets in Active Galactic Nuclei. Annual Review of Astronomy and Astrophysics, 57(1), 467-509.

[5] Tadhunter, C. (2008). An Introduction to Active Galactic Nuclei: Classification and Unification. New Astronomy Reviews, 52(6), 227-239.

[6] Gaur, H., et al. (2019). Properties of radio-loud quasars in the Sloan Digital Sky Survey. Astronomy & Astrophysics, 631, A46.

[7] Abazajian, Kevork N., et al. (2009). THE SEVENTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY. The Astrophysical Journal Supplement Series, 182(2), 543-558.

[8] Shimwell, T. W., et al. (2022). The LOFAR Two-metre Sky Survey: V. Second data release. Astronomy & Astrophysics, 659, A1.

[9] Shen, Yue, et al. (2011). A CATALOG OF QUASAR PROPERTIES FROM SLOAN DIGITAL SKY SURVEY DATA RELEASE 7. The Astrophysical Journal Supplement Series, 194(2), 45.

[10] Vanden Berk, Daniel E., et al. (2001). Composite Quasar Spectra from the Sloan Digital Sky Survey. The Astronomical Journal, 122(2), 549-564.

[11] Abdurro'uf, et al. (2022). The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. The Astrophysical Journal Supplement Series, 259, 35.

[12] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020)

[13] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[14] The Astropy Collaboration, et al. (2013). Astropy: A community Python package for astronomy. Astronomy & Astrophysics, 558, A33.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825-2830.

[16] McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

[17] Oliver Toogood, 'The magnitude scale', Mr Toogood's Physics, https://www.alevelphysicsnotes.com/astrophysics/classification.php

[18] Krause, Martin G. H. (2023). Jet Feedback in Star-Forming Galaxies. Galaxies, 11(1), 29.

【評語】160035

本研究使用 LOFAR 及 SDSS 的大量數據探討 radio loud 與 radio quiet 類星體在可見光輻射的可能差異。這樣的研究有可能可以幫助理解這些類星體之所以是 radio loud 或 radio quiet 的原因。作者有不錯的天文物理背景知識，本作品的數據處理也有條不紊。本研究並未發現該兩種類星體的可見光輻射有統計上的差異。建議未來可考慮使用中心黑洞質量正規化後的可見光輻射性質來做相關的探討。另外，單純以是否出現在 LOFAR 目錄裡來判定 radio loudness 可能是過於粗糙了點。