UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS

DEPARTAMENTO DE FÍSICA DE LA TIERRA Y ASTROFÍSICA

TRABAJO DE FIN DE GRADO

Código de TFG: [Código TFG]

[Relaciones estructurales de galaxias remotas a partir de los catálogos CANDELS]

[Structural relations of remote galaxies from the CANDELS catalogues]

Supervisor/es: [Nombre del/os supervisores]

Jesús Gallego Maestro

Grado en Física

Curso académico 20[24-25]

Convocatoria XXXX

Resumen:

La busqueda de patrones en los datos de experimentos cientificos es una de las principales fuentes de informacion en el campo de la astrofísica. El universo a grandes escalas está compuesto por procesos muy complejos, por lo que la observación de estos patrones nos permite entender mejor el funcionamiento de estos procesos.

Durante muchos años, se ha observado el firmamento con nuevas y mejores tecnologías con el fin de tener una mayor cantidad de datos de los que obtener información. El catálogo CANDELS contiene una gran cantidad de datos sobre galaxias, a varios redshifts diferentes, que contiene una gran cantidad de datos sobre propiedades físicas acercas de estas.

El objetivo es usar este catálogo para analizar los datos con un algoritmo de regresión simbólica para obtener relaciones ya conocidas, y ver que otras relaciones se pueden obtener a partir de estos datos.

[Rsumen de lo encontrado] [Conclusiones] [Prespectiva]

Abstract:

The search of patterns in scientific experiments data is one of the main sources of information in astrophysics. The vast universe is characterized of complex, that makes the observation of these patterns can help us to understand better the working of these processes.

For many years, the cosmic dust has been observed with new and better technologies to obtain a higher amount of data. The CANDELS catalogue contains a large amount of data about galaxies, at different redshifts, that contains a large amount of data about physical properties of these objects.

The goal is to use this catalogue to analyze the data with a symbolic regression algorithm to obtain known relations, and see that other relations can be obtained from these data.

[Abstract of what has been found] [Conclusions] [Outlook]

Índice

1. Introducción

El objetivo principal de la física es tener la capcaidad de entender, y por consecuente, poder explicar los procesos que ocurren en nuestro universo. Desde hace ya varios siglos la observación ha sido una de las herramientas más útiles para poder tener indicios del comportamiento de estos procesos, como la relación $r \propto T^3$ descubierta por Kepler, o la predicción de nuevas partículas basandose en las propiedades y simetrias de un determinado conjutno de partículas.

Es por ello que las relaciones que podemos obtener a partir de los datos experimentales propician el descubrimiento de nueva física, especialmente en la astrofísica. Las relaciones entre los diferentes parámetros de las galaxias, como el radio, el tamaño, la masa, etc, nos permiten clasificarlas según como estas se relacionan entre si. Este tipo de clasificaciones, meramente observacionales, son muy habituales en la astrofísica extragaláctica entre ellos destacan algunos puramente observacionales como puede ser el diagrama de Hubble, y otros más analíticos como puede ser el índice de Sersix. Ambos tienen un único objetivo, clasificar las galaxias con el fin de poder estudiar y comparar los sistemas con otros que sufren procesos similares debido a que pertenecen al mismo grupo.

Este trabajo se centra en estudiar un grupo conrecto de galaxias, aquellas galaxias que pertenecen al *Cosmic Noon* o amanezer cósmico (Ver figura **??**). En este periodo del universo (con valores de redshift entre 1,5 y 2,5) se formaron la mayoría de galaxias que observamos en el universo actual, es por eso que su estudio puede revelar una gran cantidad de información no solo acerca de los orígenes de estas estructuras, si no también de su evolución hasta llegar a las galaxias del universo cercano.

Dada la morfología de este grupo de galaxias, la cual es muy diferente de las actuales, es de esperar que sus relaciones estructurales no tengan porque ser las mismas, o que en caso de que si sean la misma no se relacionen de la misma forma.

2. La exploración CANDELS

La exploración (*Cosmic Assembly Near-IR Deep Extragalactic Legacy Survey (CANDELS)*)(Grogin et al. 2011; Koekemoer et al. 2011)

Los objetivos de la exploración pueden verse en la tabla 2 en (Grogin et al. 2011)

2.1. El catálogo EGS

La exploración CANDELS brinda una gran fuente de información de multitud de fuentes con las que trabajar, sin embargo, este trabajo se va a centrar el estudio en el catálogo *Extended Growth Strip (EGS)* el cual se centra en

(Stefanon et al. 2017; Kodra et al. 2023)

3. Regresión simbólica

La regresión simbólica (SR) es un tipo de ajuste a los datos en el que no se parte de un modelo concreto, como puede ser la regresión lineal o ajustar a una exponencial, sino que se parte de una

serie de operadores o funciones, como puede ser el uso de funciones trigonométricas o usar funciones con exponentes, que busca el mejor modelo como combinación de esos operadores.

Los algoritmos de SR despuntaron en la decada de los 80 como respuesta a la difícil tarea de encontrar ecuaciones que ajustaran a unos datos experimentales que crecian tanto en volumen como en dimensión. Es por eso que muchos métodos diferentes se han presentado desde entonces con diferentes puntos de vista. La idea detrás de estos algoritmos es facilitar la búsqueda de expresiones analíticas con las que luego poder hacer otros cálculos (ecuaciones de continuidad, condiciones de contorno, etc) y derivar conclusiones de los posibles mecanismo causantes del comportamiento de los datos.

En este caso concreto se ha usado la librería PySR la cual ha sido elegida por su impletenciación en Python, por los "benchmarks" o resultados que se describen en (Cranmer 2023). Además, al tratarse de una libreria reciente y de codigo abierto, PySR se encuentra en constante crecimiento haciendo que la comunidad científia en su conjunto pueda apoyar su desarrollo y mejora.

A parte, aunque no se entrará en gran detalle esta libreria tiene una gran cantidad de funcionalidades que la hacen muy polivalente para una gran cantidad de usos, además de una integración con varias librerias que permiten un futuro tratamiento de las expresiones obtenidas. A continuación se explica el mecanismo de funcionamiento de este algortimo en concreto.

3.1. Fundamento

PySR se basa en un algoritmo de SR general al que sus autores le han realizado algunas modificaciones con el fin de facilitar su versatilidad en cuanto a la aplicación a diversos campos.

A continuación se describe el algoritmo de convergencia, en lineas generales, para la obtención de expresiones matemáticas; junto con las modificaciones que presenta PySR. El algoritmo consiste en el siguiente conjunto de pasos:

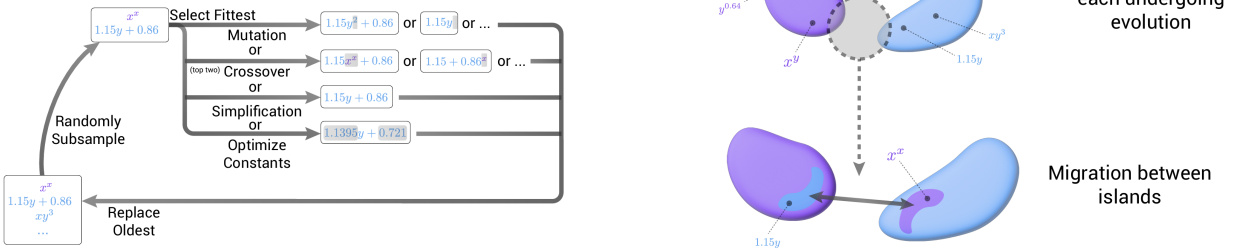
- Bucle interno

  El bucle interno consiste en los siguientes pasos:

  1. Supongamos una expresión matemática con un conjunto N de combinaciones de operadores que queremos evaluar y vamos a comparar cada conjunto de operadores entre si, esto se denomina **torneo**. Cada torneo consiste en enfrentar n_s subconjuntos de operadores (tipicamente $n_s = 2$) y enfrentamos a cada operador O_i entre si.
  2. Asignamos una probabilidad p al operador que mejor ajuste tras compararlos todos. Si p es muy bajo, quitamos ese operador del subconjunto y repetimos este paso.
  3. Nos quedamos con aquel operador con mayor probabilidad y lo metemos en otro de los subconjuntos de operadores O para repetir el proceso hasta que converga del cual obtendremos el conjunto de operadores que define el mejor ajuste.

- Bucle externo

  El bucle externo consiste en realizar el proceso para diferentes conjuntos N (**islas**) y así tener la posibilidad de migrar de una isla a otra y aumentar las combinaciones posibles.



(a) Bucle interno del algoritmo

(b) Bucle externo del algoritmo

Figura 1: Diagrama de los bucles utilizados en PYSR. (Cranmer 2023)

A continuación vamos a ver como se ha implementado el modelo de SR para obtener resultados.

Listing 1: Implementación del modelo de regresión simbólica

```
def symbolic_regression(df, input_cols, target_col, threshold=0.2):
    """
    Esta funcin procesa los datos, ajusta el modelo de regresin simblica (SR) y cuantifica su ajuste.

        ----------ENTRADAS----------
    df: DataFrame con los datos originales.
    input_cols: Lista con 1 o 2 columnas del DataFrame usadas como variables independientes.
                - Si es 1, se asume una relacin 2D.
                - Si son 2, se asume una relacin 3D.
    target_col: Nombre de la columna objetivo (variable dependiente).
    threshold: Umbral para la varianza explicada por la tercera componente (en caso 3D).
               Por defecto es 0.2, es decir, el modelo es aceptable si el 80 % de los datos
                estn contenidos en el plano de entrada.

        ----------SALIDAS----------
    - best_expr: Mejor expresin simblica encontrada (en formato SymPy) si cumple el umbral.
    - mtrica: Varianza explicada o RMSE segn el caso.
    - datos: Datos combinados usados para visualizacin o evaluacin.
    """

    # Extraer variables independientes y dependientes
    X = df[input_cols].values
    y = df[target_col].values

    # Dividir los datos en entrenamiento y prueba
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.3, random_state=12
    )

    # Ajustar el modelo de regresin simblica con los datos de entrenamiento
    model = PySRRegressor(
        binary_operators=["+", "-", "*", "/", "^"],
        unary_operators=["exp", "log", "sqrt"],
        model_selection="best",
        verbosity=0,
        constraints={'^': (-2, 2)}
    )
    model.fit(X_train, y_train)

    # Predecir con el modelo ajustado
    y_pred = model.predict(X_test)

    # Normalizar los datos para evaluacin (evita sesgos por escalas)
    scaler_X = StandardScaler()
    X_test_scaled = scaler_X.fit_transform(X_test)

    scaler_y = StandardScaler()
    y_test_scaled = scaler_y.fit_transform(y_test.reshape(-1, 1)).ravel()
    y_pred_scaled = scaler_y.transform(y_pred.reshape(-1, 1)).ravel()

    # Juntar las predicciones con los datos de entrada
    PCA_data = np.hstack((X_test, y_pred.reshape(-1, 1)))

    if len(input_cols) == 2:
        # Caso tridimensional: evaluar el plano de entrada y su capacidad explicativa

        # Combinar entradas normalizadas y prediccin para PCA
        data_combined = np.hstack((X_test_scaled, y_pred_scaled.reshape(-1, 1)))
```

```
        # Aplicar PCA a 3 componentes
        pca =PCA(n_components=3)
        pca. fit (data_combined)
        explained_variance =pca. explained_variance_ratio_

        # Varianza explicada por la tercera componente (z)
        third_component_variance =explained_variance[2]

        print(f"Varianza explicada por la tercera componente: {third_component_variance:.4f}")

        # Si la varianza es menor al umbral, aceptamos el modelo
        if third_component_variance ¡=threshold:
            best_expr =model.sympy()
            return best_expr, third_component_variance, PCA_data
        else:
            return None, third_component_variance, PCA_data

    else:
        # Caso bidimensional: usar error cuadrtico medio como mtrica
        rmse =root_mean_squared_error(y_test_scaled, y_pred_scaled)
        print(f"RMSE: {rmse:.4f}")

        # Retornar la mejor expresin encontrada junto con la mtrica
        best_expr =model.sympy()
        return best_expr, rmse, y_pred, y_test
```

De esta forma podemos obtener las expresiones, evaluarlas y cuantificar su calidad.

4. Metodología

Para proceder al análisis de los datos lo primero es centrarse en posbles relaciones ya conocidas de forma que podamos reducir el número de posibles relaciones, como ya se ha comentado en la sección (EGS), no todas las relaciones son interesante, ya sea por el tipo de dato que tenemos o porque no haya una motivación física detrás con la que se pueda argumentar la existencia de dicha relación [1].

Una forma sencilla, visual y efectiva de iniciar la busqueda de relaciones es mediante el uso de *pairplots*. Este tipo de gráficos nos permiten tener de un primer vistazo una imagen del comportamiento de los datos.

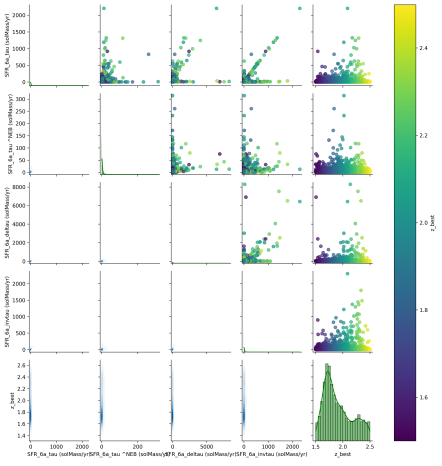


Figura 2: Ejemplo de Pairplot. En este caso se muestra un Pairplot para los datos de Fontana

Como se puede ver en la figura **??** con este tipo de gráficos podemos obtener una gran cantidad de información para entenderla naturaleza de los datos:

- **Triángulo inferior**: En el tríangulo inferior podemos observar un diagrama KDE *Kernel*

[1] Aunque lo idela fuera no tener ningún sesgo, dado la dimensionalidad del conjunto de datos lo óptimo es filtrar por relaciones ya conocidas y hacer variaciones de las mismas para explorar otras nuevas

Density Estimator el cual nos permite ver la distribución de los datos, es especialmente útil para poder observar distintos grupos de datos y su distribución.

- **Diagonal**: En la diagonal es posible observar la distribución de los datos. Como es de esperar los datos no tienen porque seguir una distribución normal o gaussiana, esto se debe al *Sesgo del superviviente*, dado que estamos estudiando una región muy concreta del espacio, no solo por el campo de la exploración si no también por centrarnos en el *Cosmic Noon*, es muy probable que no tengamos galaxias muy diferentes entre si
- **Triángulo superior**: El triangulo superior consiste un simple gráfico de puntos donde podemos ver mejor cual es el comportamiento de los datos.

Haciendo uso de estos gráficos podemos obtener de forma visual información sobre el comportamiento de los datos, no solo en cuanto a la dependencia de una variable con respecto a otra; si no también como se distribuyen o si froman agrupaciones.

Dadas las dimensiones de estos gráficos y el número límitado de páginas de este trabajo a continuación se muestran las conclusiones que se ha obtenido con ellos[2]:

- Los diferentes modelos de ajuste a los datos (denotados con tau, invtau, lin, cons, etc) suelen obtener datos similares, en los pairplots se ve que (en muchos casos) forman una recta cuando se gráfican entre sí. Esto nos permite poder comparar diferentes modelos entre si y cambiar el modelo sencillamente usando la recta de ajuste.
- El punto anterior solo es aplicable cuando nos comparamos modelos de un mismo autor, ya que aunque siguen formando una recta la dispersión es mucho mayor y no es un procedimiento preciso en este caso.
- Al igual ocurre si la medida tiene en cuenta la emisión nebular o no, en este caso la dispersión es mucho mayor y no se puede comparar entre si.
- A continuación una lista de posibles relaciones que quedan excluidas una vez analizados los gráficos:

  1.

4.1. Previo y limpieza de los datos

Una vez acotadas las posibles relaciones que esperamos encontrar es necesario profundizar en la naturaleza de los datos con el fin de detectar posibles errores en la medida o valores atípicos que, aunque válidos, se alejen del comportamiento de la mayoría de las medidas. Este descarte de los valores atípicos se raliza con el fin de asegurar que los datos puedan ajustar bien a un modelo, en este caso una SR, y poder así analizar correctamente las predicciones del modelo.

La mayor fuente de valores atípicos son medidas de mala calidad, que generalmente se indicaban con el número $-99{,}0$. Estos valores han de ser eliminados por la simple razón de carecer de sentido físico, en ninguno de los parámetros estudiados tiene sentido que puedan tomar ese valor (ej, masas, edades, magnitudes, ...).

Otra fuente de valores atípicos viene de la mano de los núcleos activos. Aunque en los datos se señaliza que galaxias podrían contener un AGN (Active Galactic Nuclei) el número de estos es mucho menor que el tamaño de la muestra, por lo que estudiarlos de la misma forma siendo tan poco representativos y teniendo un comportamiento tan diferente al de una galaxía que no lo tienen (o cuya imagen no tiene la resolución para identificarlo) carece de sentido si el objetivo es ajustar a un modelo.

[2]Los gráficos pueden verse en este enlace

4.2. Análisis de relaciones estructurales 2D

El análisis de la relaciones entre 2 parámetros es un caso, que aunque simple, muy poderoso. Estas relaciones han sido muy estudiadas para diferentes parámetros como pueden ser la relación Masa-Luminosidad o los diagramas H-R que relacionan propiedades observacionales entre si, todo ello con el fin de encontrar relaciones, que aplicadas a muestras no estudiadas, nos permitan obtener información.

En el caso más simple de todos, una relación lineal, podemos ver que tan buena es esa relación mediante el *Coeficiente de Pearson* R^2; pero estas relaciones no siempre tienen que ser lineales, es posible que sean cuadráticas, logarítmicas, exponenciales, etc. En esos caso, cuando la relación no es puramente lineal, el R^2 puede ser bajo pero la relación sigue estando ahí, es por eso que para estudiar estas relaciones sin sesgar por aquellas relaciones que sean lineales se ha hecho uso de lo que se denomina *Distancia de Correlación* (Ramos-Carreño y Torrecilla 2023).

La distancia por correlación es una medida de la dependencia entre dos variables aleatorias, en la que se mide la distancia entre los valores de las variables (distancia entre puntos en un diagrama de dsipersión), por lo que no se ve tan afectado pot la linealidad de los datos, es decir, captura la dependencia entre las variables sin importar si esta es lineal o no.

Para estudiar esta relación se han usado los gráficos de dispersión de la figura **??** en donde se han representado los datos en diferentes formas funcionales típicas en astrofísica.

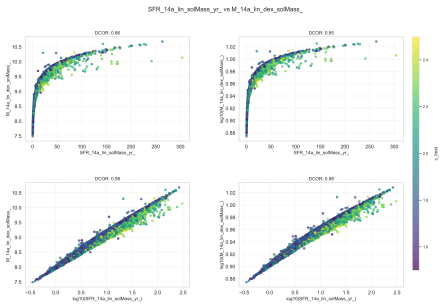


Figura 3: Relación entre la masa y la tasa de formación estelar. Pueden verse más gráficos aquí

Para obtener las siguientes relaciones finales se ha filtrado el gráfico que presenta el menor *Root Mean Square Error (RMSE)* (Ver **??**). El RMSE es una medida de la diferencia entre los valores predichos por un modelo y los valores observados.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

donde y_i son los valores observados, $\hat{y}_i$ son los valores predichos por el modelo y n es el número total de observaciones. Un RMSE bajo indica que el modelo se ajusta bien a los datos, mientras que un RMSE alto sugiere que el modelo no se ajusta bien, por lo que nos quedamos con el modelo que tiene el menor RMSE.

4.3. Análisis de relaciones estructurales 3D

En el caso de las relaciones entre 3 parámetros, el análisis es más complicado, ya que no solo se tiene que tener en cuenta la relación entre los tres parámetros, si no también la forma en la que estos se distribuyen; especialmente para el caso en el que buscamos un *Plano fundamental*. Es por eso que es necesario cuantificar cuando aparece un plano al representar los 3 parámetros, para ello

se ha empleado el *Explained Variance Ratio (EVR)*. El EVR se define como:

$$EVR_i = \frac{\lambda_i}{\sum_j^n \lambda_j} \tag{2}$$

En donde λ_i es el autovalor de la componente $i \in \{x, y, z\}$ y n es el número de componentes principales, en este caso $n = 3$. El EVR es una medida de la varianza explicada por cada componente principal en un análisis de componentes principales (PCA). En este caso, el EVR se utiliza para evaluar la cantidad de varianza que se explica por el plano en comparación con la varianza total de los datos. Un EVR alto indica que el plano captura una gran parte de la variabilidad de los datos, lo que sugiere que los datos están bien representados por el plano.

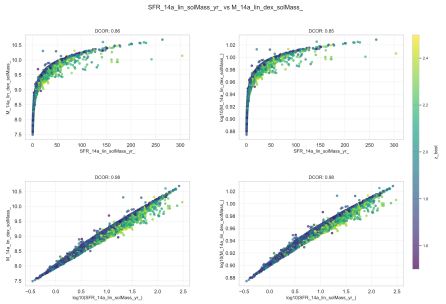


Figura 4: Relación entre la masa y la tasa de formación estelar. Pueden verse más gráficos aquí

Para obtener la expresión final se filtrado por el EVR de cada una de las gráficas del grid y seleccionado la menor de todas (Ver **??**).

5. Resultados

5.1. Relaciones estructurales en 2D

5.1.1. Relación Masa-Tasa de formación estelar

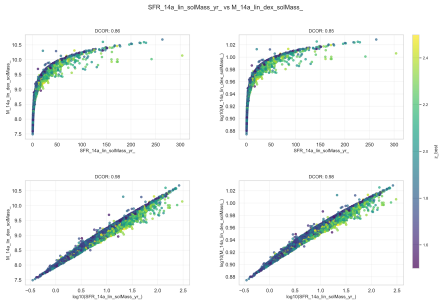


Figura 5: Relación entre la masa y la tasa de formación estelar. Pueden verse más gráficos aquí

5.1.2. Relación Masa-Luminosidad

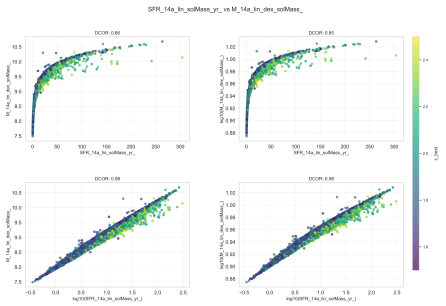


Figura 6: Relación entre la masa y la luminosidad. Pueden verse más gráficos aquí

5.1.3. Relación Masa-Magnitud

5.1.4. Relación Tasa de formación estelar-Magnitud

5.1.5. Relación

5.2. Análisis de relaciones estructurales 3D

5.2.1. Relación Masa-Redshift-Tasa de formación estelar

6. Conclusiones

Referencias

Cranmer, M. (mayo de 2023). *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*. arXiv:2305.01582 [astro-ph, physics:physics].

Grogin, N. A. et al. (dic. de 2011). «CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey». En: 197.2, 35, pág. 35.

Kodra, D. et al. (ene. de 2023). «Optimized Photometric Redshifts for the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS)». En: *The Astrophysical Journal* 942.1, pág. 36.

Koekemoer, A. M. et al. (dic. de 2011). «CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey—The Hubble Space Telescope Observations, Imaging Data Products, and Mosaics». En: 197.2, 36, pág. 36.

Ramos-Carreño, C. y J. L. Torrecilla (feb. de 2023). «dcor: Distance correlation and energy statistics in Python». En: *SoftwareX* 22.

Stefanon, M. et al. (abr. de 2017). «CANDELS Multi-wavelength Catalogs: Source Identification and Photometry in the CANDELS Extended Groth Strip». En: 229.2, 32, pág. 32.