Estudio de los efectos sistemáticos de SOPHIE+ con algoritmos de aprendizaje automático

J. R. Serrano[1], R. F. Díaz[2]

[1] Facultad de Ciencias Astronómicas y Geofísicas, UNLP

[2] International Center for Advanced Studies (ICAS) and ICIFI (CONICET), ECyT-UNSAM

Grupo de exoplanetas ROCKY

contacto: juann.serrano@gmail.com

Facultad de Ciencias Astronómicas y Geofísicas
UNIVERSIDAD NACIONAL DE LA PLATA

Introducción

SOPHIE+ es un espectrógrafo echelle ubicado en el observatorio de Haute-Provence, Francia (figura 1). Está diseñado para obtener una gran precisión en velocidades radiales mediante IP control[1] y calibración simultánea de la longitud de onda.

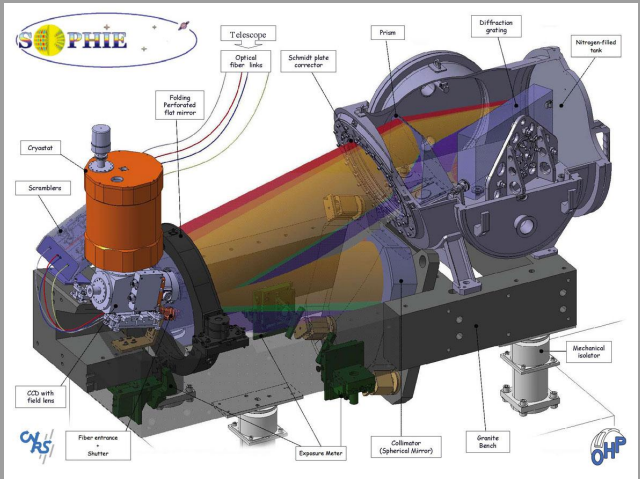

Figura 1: Esquema del espectrógrafo SOPHIE+.

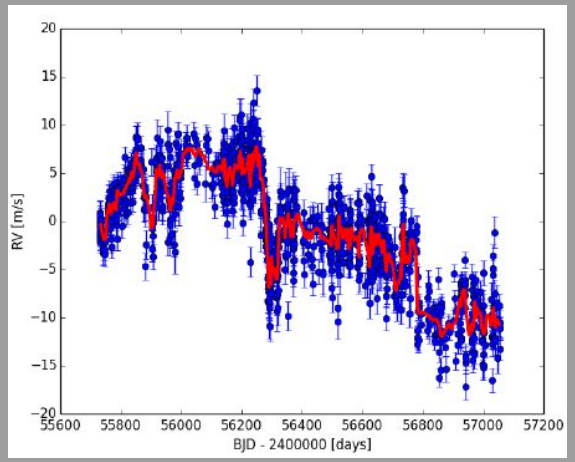

Figura 2: Estimación de la deriva con estrellas constantes.

Si bien alcanza precisiones de 1~2 m/s, se ha observado una deriva del punto cero a largo plazo, que actualmente se corrige mediante observación regular de estrellas estándar de velocidad radial constante[3] (figura 2). Con el objetivo de proporcionar un nuevo método de corrección y comprender las causas de estos efectos sistemáticos nos proponemos entrenar un algoritmo de aprendizaje automático sobre un dataset de estrellas constantes para predecir las velocidades radiales medidas por el instrumento.

Métodos

- **PREPROCESADO:** Recolectamos 645 observaciones de las estrellas HD 185144 (K0V), HD 89269 (G4V) y HD 9407 (G6.5V). Es importante incluir estrellas con distintos tipos espectrales para tener en cuenta el efecto de color[2]. Para cada observación extrajimos 83 variables (*features*) del header y 30 más fueron colectadas de archivos externos correspondientes a sensores de temperatura y presión en distintas partes del instrumento. Un 20% de los datos se separan para testeo del modelo (figura 4 y 5).
- **EXPLORACIÓN:** Visualizamos las características del dataset, detectamos los features más correlacionados con la VR (figura 3). Definimos nuevos features de interés como combinación de los primeros.

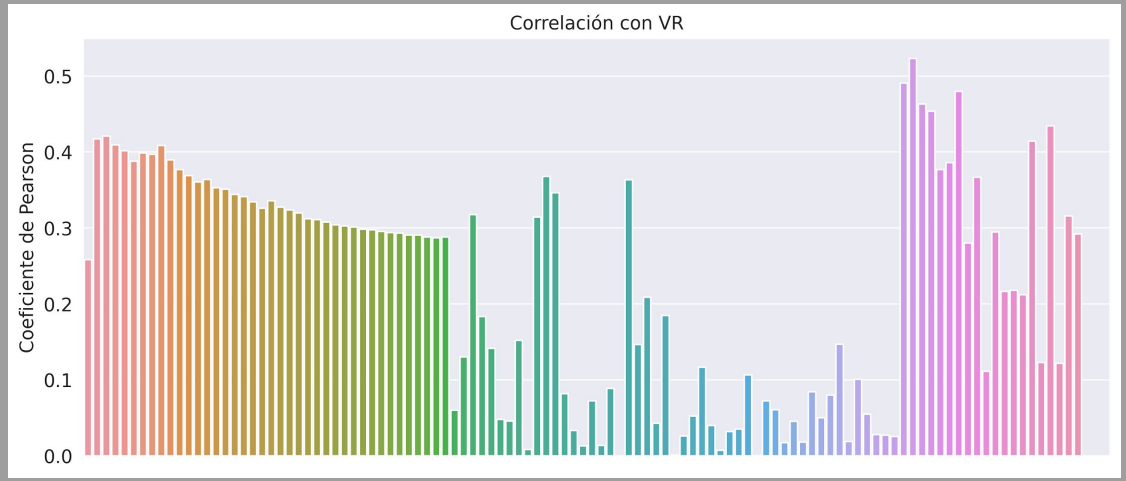

Figura 3: Coeficientes de correlación de cada feature con la VR.

Métodos

- **PREPARACIÓN DE LOS DATOS:** Se hizo limpieza de los datos, y se escalearon todos los features a media cero y varianza unitaria.
- **ENTRENAMIENTO Y AJUSTE FINO:** Probamos 9 técnicas distintas de aprendizaje automático, entrenamos y buscamos los mejores hiperparámetros mediante validación cruzada. El algoritmo final fue un regresor lineal Lasso ajustado con el algoritmo LARS. Se implementó en ScikitLearn sobre los 112 features y todas las combinaciones de sus productos, es decir un total de 6329 features.
- **SELECCIÓN DE FEATURES:** El algoritmo *LassoLars* nos permite ver los coeficientes de cada feature en el modelo, siendo la gran mayoría nulos, con esto hicimos una selección de los más relevantes para la predicción, reduciendo la cantidad de features de 112 a 33 sin afectar la precisión de las predicciones.

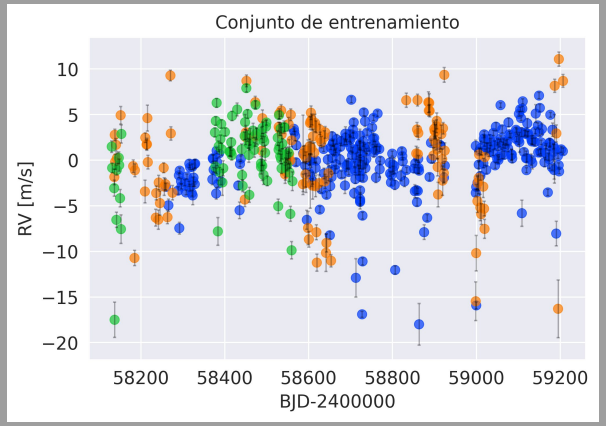

Figura 5: Datos para entrenamiento por estrella. HD 185144 (azul), HD 89269 (naranja), HD 9407 (verde).

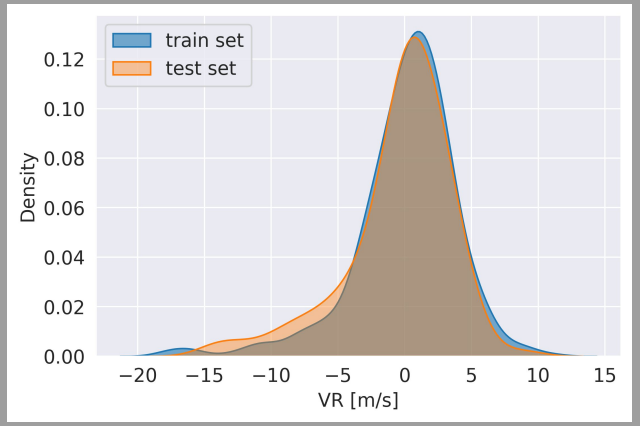

Figura 4: Distribuciones de los conjuntos de entrenamiento y testeo.

Resultados

Como métrica para evaluar el modelo en el conjunto de testeo se usó el WRMSE (*weighted root mean squared error*) y obtuvimos un valor de **~1.47 m/s**. En la tabla 1 mostramos la desviación estándar de los datos de testeo, el WRMSE y el score R2 de las predicciones en el conjunto de testeo completo y para cada estrella por separado. Observamos que el feature con mayor importancia en el modelo fue **“drift rv”**, este es un parámetro que se obtiene al medir cuánto se movió el punto cero de la longitud de onda del instrumento desde la última calibración a la noche de la observación.

	DATASET	HD185144	HD89269	HD9407
SD [m/s]	3.96	2.63	5.62	4.31
WRMSE [m/s]	1.47	1.40	1.90	1.74
R2 Score	0.71	0.63	0.86	0.77

Tabla 1: Evaluación del modelo en el set de testeo completo y separado por estrella.

Resultados

En las figuras 6, 7 y 8 se muestran los datos de testeo y las predicciones para cada una de las estrellas. En HD 89269 y HD 9407 vemos que **el modelo explica el 86 y 77%** de la dispersión de los datos respectivamente, mientras que en HD 185144 cerca del **63%**. Para este último caso sin embargo, si aplicamos nuestro modelo como corrección a los datos de testeo nos permite reducir la dispersión de las velocidades radiales de 2.6 m/s a 1.51 m/s, mientras que Courcol et. al[3] con el método tradicional de las estrellas constantes para la misma estrella obtienen una dispersión de 1.61 m/s.

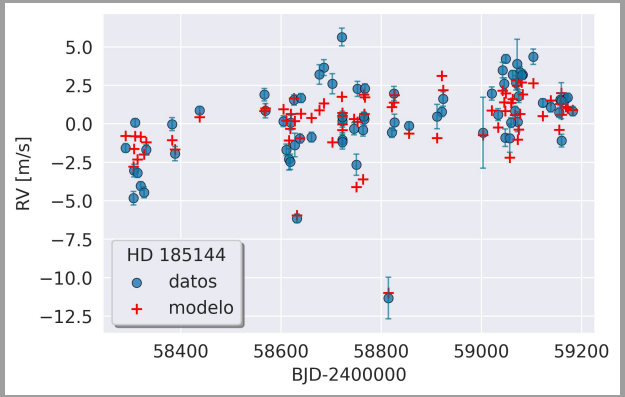

Figura 6: Predicciones paraHD 185144.

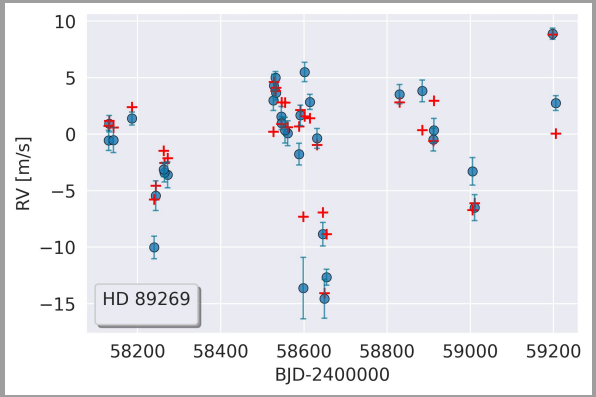

Figura 7: Predicciones para HD 89269.

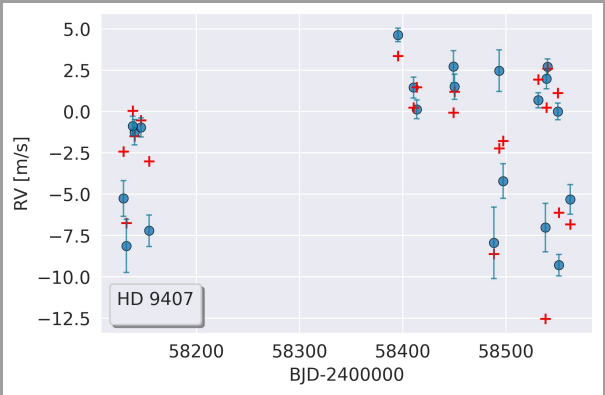

Figura 8: Predicciones para HD 9407.

Conclusiones y trabajo futuro

Pudimos construir un modelo robusto que predice con muy buena precisión la dispersión de las velocidades radiales debida a errores sistemáticos del espectrógrafo SOPHIE en tres estrellas de velocidad radial *constante* de distintos tipos espectrales. El algoritmo final utiliza 33 features de las cuales identificamos a "**drift_rv**" como la más importante para predecir los errores sistemáticos. Como trabajo futuro vamos a probar el modelo en alguna estrella que no haya sido parte del conjunto de entrenamiento.

Referencias

1. *Deriving High-Precision Radial Velocities. Figueira (2018)*
2. *Detecting the spin-orbit misalignment of the super-Earth 55 Cancri e*. Bourrier & Hébrard. (2014)*
3. *The Sophie Search for northern extrasolar planets VII. A warm Neptune orbiting HD 164595. Courcol et al (2015)*

Grupo de exoplanetas ROCKY

contacto: juann.serrano@gmail.com

Facultad de Ciencias Astronómicas y Geofísicas
UNIVERSIDAD NACIONAL DE LA PLATA