Generación de *features* en la búsqueda de estrellas variables en el relevamiento astronómico *VVV*

Juan B Cabral[1,2], Pablo Granitto[3], and Sebastián Gurovich[1] y Dante Minniti[4,5,6]

[1] IATE - Instituto De Astronomía Teórica Y Experimental - Observatoio Astronómico Cordoba, UNC - CONICET Laprida 854, X5000BGR, Córdoba, Argentina Email: jbc.develop@gmail.com
[2] Facultad de Ciencias Exactas, Ingeniería y Agrimensura - UNR, Pellegrini 250 - S2000BTP Rosario, Argentina
[3] CIFASIS - Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas - CONICET - UNR Ocampo y Esmeralda, S2000EZP Rosario, Argentina
[4] The Millennium Institute of Astrophysics (MAS), Santiago, Chile
[5] Departamento de Ciencias Fisicas, Universidad Andres Bello, Republica 220, Santiago, Chile
[6] Vatican Observatory, V-00120 Vatican City State, Italy

Resumen Frente al desarrollo de telescopios terrestres y satelitales que generan relevamientos astronómicos del orden de los *Peta-Bytes*, se expone en este trabajo la metodología a seguir para la generación de *features* de series temporales para el descubrimiento de estrellas variables periódicas en el núcleo, bulbo y una parte del disco de nuestra galaxia utilizando datos del VVV-Survey. A lo largo del trabajo se presenta los datos de dicho relevamiento, la forma de regenerar series temporales a partir de ellos y extraer *features* importantes como el período o diferentes estadísticas de magnitud. Finalmente se proyecta el plan a futuro para utilizar el conocimiento extraído para la creación de catálogos de estrellas variables utilizando aprendizaje automático.

Keywords: features, astronomía, catálogos, vvv

1. Introducción

El desarrollo de modernos telescopios terrestres y satelitales ha impulsado la realización de grandes relevamientos astronómicos, los cuales, a su vez, han generando un crecimiento gigantesco en la dimensionalidad, cantidad y calidad de datos a ser procesados, almacenados y analizados. Actualmente se destaca el relevamiento terrestre llamado "Vista Variables in the Via Lactea" (VVV) [9], cuyo objetivo es producir un mapa tridimensional de una gran parte del centro galáctico (Bulbo) de la Vía Láctea y de una fracción del Disco Galáctico interno.

Este mapa tridimensional se obtuvo a partir de un censo de estrellas, posibilitado gracias a un monitoreo sistemático de estas regiones de la Vía Láctea, que completó un total de 1.929 horas de observación realizadas durante un período

de 5 años (iniciados en el 2010). Para ello, el equipo de astrónomos a cargo del proyecto VVV utilizó el moderno telescopio VISTA, ubicado en Cerro Paranal, II Región, Chile; el cual generó aproximadamente 300 GB de datos por noche.

Los datos del VVV se presentan en una unidad llamada "baldosa" (*tile*, en inglés), la cual es una zona rectangular del cielo relevada a través del tiempo. Cada baldosa se compone de varias imágenes en alta resolución para diferentes tipos de filtros de frecuencias lumínicas en el infrarrojo cercano (cinco en total). Asimismo, por cada imagen existe una base de datos de archivos binarios numérica con los valores de posición, magnitud y color de las fuentes de luz presentes en la imagen, llamada "catálogo fotométrico". La totalidad de las baldosas que constituyen el relevamiento (Área del disco de longitudes galácticas de 250 a 297 grados y el área del bulbo galáctico de 350 a 10 grados) puede apreciarse en la Figura 1.

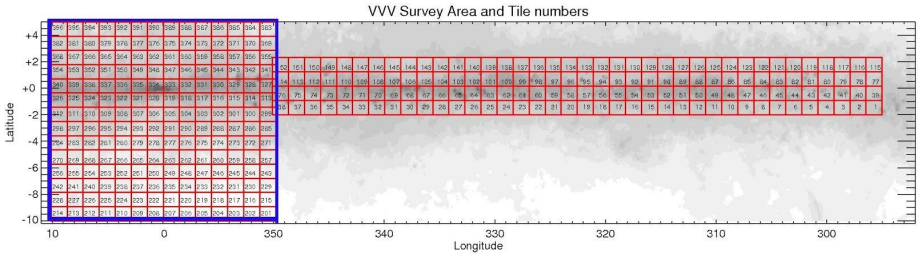

Figura 1. Mapa de enrojecimiento de schlegel adaptado de Minniti et al. 2010 [9], mostrando el área del relevamiento en el cielo en coordenadas galácticas y las baldosas que conforman el VVV. En escala de grises esta la densidad estelar proyectada de este mismo campo tomado del relevamiento astronómico "Two Micron All Sky Survey" [12].

Dentro del barrido total del relevamiento, es de interés astronómico la generación de catálogos de cualquier tipo de objeto, ya que el VVV en la zona bajo estudio identificará como fuentes de luz además de planetas y galaxias un número bastante amplío (Figura 2) de diferentes fenómenos. En este marco de búsqueda de semejante volumen de objetos dentro de un aun mas grande volúmen de datos es destacable el trabajo presentado como tesis doctoral por Cavuoti en el 2013 [2] en el cual explora, entre otras cosas, el uso de aprendizaje automático para el preprosamiento, análisis y presentación de conocimiento.

En este estudio resulta de interes las estrellas de tipo variable RRLyrae, las cuales son del tipo espectral del A al F y su pulsación comprende períodos cortos de entre $\sim 0{,}2$ y $\sim 1{,}2$ días y con variaciones de brillo desde $\sim 0{,}1$ hasta $\sim 0{,}5$ en infrarrojo. Estas fuentes fueron elegidas por existencia de publicaciones que identifican algunas cientos de ellas dentro de relevamiento [5] [6], lo cual facilita la creación de los primeros conjuntos de entrenamiento y prueba.

Ya han habido intentos de utilizar aprendizaje automático sobre VVV, siendo el mas relevante la infraestructura llamada *VVV Templates Project* [1]. En ese trabajo se describe, entre otras cosas, cómo se pueden utilizar fuentes del mismo VVV ya identificadas como de algún tipo en particular (RRLyrae, Cepheids, etc.) en algún otro relevamiento superpuesto en nuestra zona de observación,

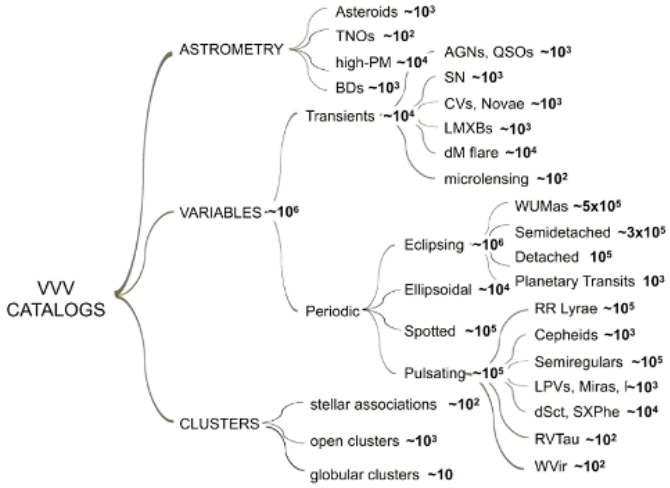

Figura 2. Diagrama adaptado de Minniti et al 2010 [9] que expone el numero esperado de fenómenos astrofísicos que espera detectar VVV en sus catálogos.

como forma sencilla de aumentar nuestro set de entrenamiento y prueba de las clases objetivo.

En este trabajo haremos un resumen del análisis que estamos llevando adelante para lograr, en primera instancia, la extracción de *features* para la generación de catálogos de estrellas tipo RRLyrae dentro del VVV.

2. Catálogos fotométricos múltiple-época

El VVV es un relevamiento público con un tiempo propietario de un año, es decir, únicamente los miembros científicos del VVV y sus estudiantes podrán utilizar los datos dentro del tiempo propietario, posterior al cual los datos preprocesados se dispondrán para todos. Algunos autores de este trabajo pertenecen al grupo científico del VVV, por lo cual se dispone de los datos inmediatamente luego de ser procesados por el pipeline de CASU (*Cambridge Astronomical Survey Unit*) [7].

El pipeline de VVV [4] además de prepropocesar cada imagen, brinda por cada una de ellas una base de datos de archivos con los valores de posición, magnitud y color de las fuentes de luz presentes en la imagen, llamada "catálogo fotométrico". Es sobre estos catálogos donde enfocamos este trabajo.

Hay que tener en cuenta que existen diferentes diferentes tipos de catálogos. En este trabajo son importantes dos de ellos:

1. **Pawprint Stack** los sensores infrarrojos de la cama VIR-CAM (Figura 3) del telescopio VISTA tienen márgenes entre ellos. Es por esto que se necesita desplazar el telescopio tres veces variando el eje X y 3 veces el eje Y para completar la imagen de un *tile*. Cada exposición es llamada *Pawprint* y la

[7] `http://casu.ast.cam.ac.uk`

suma de las 6 es llamada *Pawprint Stack* o *época* por ser la observación de un Tile en una fecha dada.

2. **Band-Merge** este catálogo consolida todas las fuentes de un Tile a lo largo de todas las épocas.

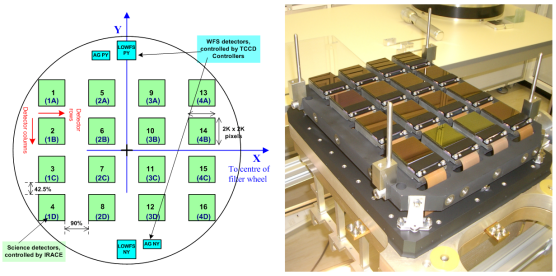

Figura 3. A la izquierda el diagrama de los sensores infrarrojos de VIRCAM y a la derecha una foto de la VIRCAM. Adaptado de `http://www.vista.ac.uk`

Dado las características técnicas del relevamiento, todas las fuentes en los *Band-Merge* de magnitud[8] $\lesssim 12.$ (muy brillantes que saturan el telescopio) y $\gtrsim 16{,}5$ (muy débiles) son ignoradas [5].

3. Reconstrucción de la serie temporal para la generación de features

Como se mencionó antes, para identificar y tipificar una estrella variable *RR Lyrae* es necesario determinar su variación de magnitud en un período dado. Es entonces una necesidad identificar cada fuente presente en un *Band-Merge* de un *Tile* con todas las observaciones existentes en todos los *Pawprint Stack* disponibles para dicho *Tile*, y así reconstruir una serie temporal.

3.1. Emparejamiento por proximidad (*Cross-Matching*)

De cada *Pawprint Stack* se conoce: a que Tile pertenece, a que fecha corresponde la medición, posición (en coordenadas esféricas) y magnitud para cada fuente. Lamentablemente no hay una forma unívoca de determinar que objeto observado en un *Band-Merge* es el mismo que uno observado en el *Pawprint-Stack* ya que no comparten ningún identificador, y si bien las posiciones medidas son cercanas, no son iguales para la misma fuente.

Para superar esta dificultad se identificaron las fuentes a través del método *cross-matching*. Este método consiste en verificar cuales son las fuentes más cercanas en posición catalogo A a las de un Catálogo B, y viceversa dentro de un intervalo D. Solo se asume que la fuente a del catálogo A y la fuente b del catálogo B son las mismas si a es la más cercana a b, b es la más cercana a a y además están dentro del intervalo de D.

El intervalo D elegido en nuestro estudio fue $1/3$ *de arco segundo* [7].

[8] Las fuentes son mas brillantes mientras menor es la magnitud

3.2. Corrección de Fechas

Otro punto importante, a tener en cuenta es que las fechas de los *Pawprint-Stack* están definidas como Días Julianos Medios (MJD del inglés *Mean Julian Date*); que son el promedio de días julianos de todos los *Pawprints* involucrados en el *stack*. Asimismo, un día juliano es la cantidad de días transcurridos desde el mediodía del 1º de enero del año 4713 a. C. En nuestro estudio, para definir las series temporales, el MJD acarrea el problema de que es un formato de fecha *geocéntrico*. Esto lleva a que la misma fuente (observada en dos épocas distintas), dado que la velocidad de la luz es finita, dependa de la posición del observador en el sistema solar cuando es realizada.

Para subsanar la dificultad descrita, se utiliza el Día Juliano Heliocéntrico (HJD del inglés *Heliocentric Julian Date*) el cual corrige el MJD utilizando las diferencias en la posición de la Tierra con respecto al Sol [3].

3.3. Período

Ya identificadas todas las observaciones de la misma estrella y corregidas sus fechas de observación, el siguiente paso consiste en determinar su período. La planificación de observaciones en VVV hace que los *tiles* no se muestreen uniformemente ni en un período dado. Por consiguiente, lo más cómodo es hacer el supuesto de que la frecuencia de observaciones es aleatoria. Si graficamos Los datos directamente como serie temporal no se percibirá ni un período evidente (Figura 4).

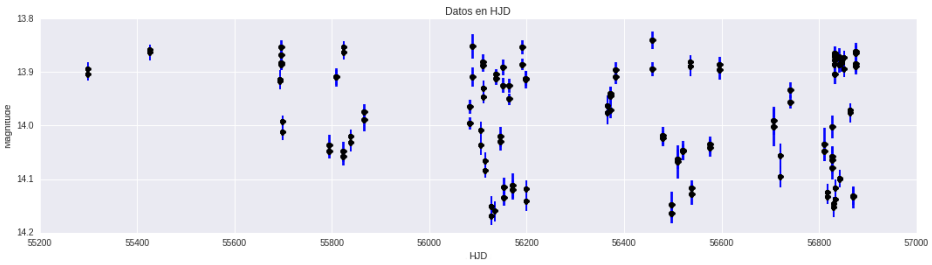

Figura 4. Observaciones de magnitud de una estrella de tipo variable RRLyrae AB del trabajo "Bulge RR Lyrae stars in the VVV tile b201"[5] identificada con el ID `VVV J2703536.01-412829.4`. El eje X representa la fecha de medición y el Y la magnitud en orden inverso.

Para recuperar el período se utiliza el método de Fast Lomb-Scargle[8][11] el cual mide el ajuste de mínimos cuadrados de sinusoides a los datos muestreados. Poniendo en fase los datos sobre este período obtenido queda en evidencia la curva de luz de la fuente periódica (Figura 5)

4. Generación de los conjuntos de datos

Para generar el conjunto de datos de estrellas RRLyrae (Casos positivos), la alternativa elegida fue buscar los catálogos obtenidos por el relevamiento OGLE

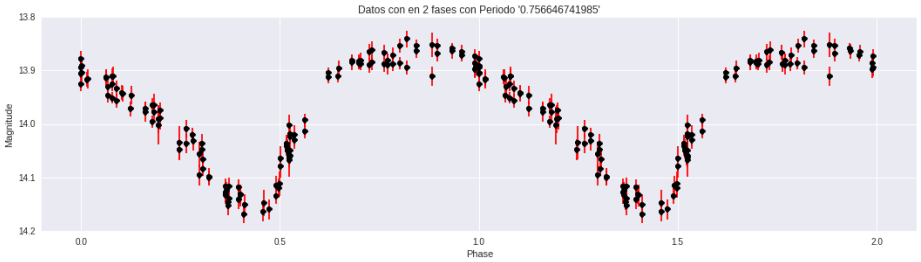

Figura 5. Observaciones de magnitud en fase de la misma estrella correspondiente a la figura 4. El eje X representa la fase de medición y el Y la magnitud en orden inverso. Por comodidad en la observación de la periodicidad de datos se presentan dos períodos de los datos.

3 [13] (estando a espera de datos de OGLE 4 [14]) que superponen su zona de observación del VVV (*Tiles* `b278`, `b261`, `b262`, `b263` y `b264`). De las fuentes de este relevamiento se identificaron las correspondientes en nuestro conjunto de datos con *cross-matching*.

Para los casos de estrellas no variables (Casos negativos) se optó por buscar estrellas que tengan una desviación estándar de magnitud menor a la mediana de las desviaciones estándar de la magnitud de su serie temporal.

5. Trabajo a futuro: Generación de catálogos usando Aprendizaje Automático

Se pretende utilizar los conjuntos de datos obtenidos sobre los *tiles* que superponen a OGLE 3 mencionados en el apartado anterior, para clasificar las estrellas en las clases RRLyrae ($\oplus$) y No Variables ($\ominus$).

Además del *feature* período se utilizarán los promedios pesados, desviaciones estándar, rangos y medianas de las magnitudes [10]. Sobre ellos se realizarán análisis de componentes principales (PCA) y luego se evaluarán los métodos de Árboles Aleatorios, SVM y redes neuronales para elegir el método o métodos que mejores clasificaciones realicen.

Es esperable, dado el volumen de datos de los 4 *tiles* (5 millones de fuentes), que haya una serie de detecciones de falsos positivos (como casos estrellas variables parecidas a las RRLyrae). En esta situación se realizará una inspección visual sobre los resultados obtenidos por el método automático, y de ser un número muy elevado se inspeccionarán las estrellas que más ajusten a los criterios que definen a una RRLyrae. De ser necesario, se propone utilizar técnicas de optimización multi-criterio [15]. Las estrellas que mejor cumplan el set de criterios serán las evaluadas visualmente y de pasar esta prueba serán incorporadas al set de entrenamiento como casos positivos. Como se aprecia, el análisis es iterativo y se considera finalizado cuando el numero de estrellas variables RRLyrae sea igual al porcentaje esperado de este tipo de estrellas dado el total de datos.

Hay que tener en cuenta que VVV es un relevamiento de mucho mas profundidad que OGLE 3 y se espera obtener un número mucho mayor de RRLyrae.

Bibliografía

[1] Angeloni, R., Ramos, R.C., Catelan, M., Dékány, I., Gran, F., Alonso-García, J., Hempel, M., Navarrete, C., Andrews, H., Aparicio, A., Beamín, J.C., Berger, C., Borissova, J., Peña, C.C., Cunial, A., de Grijs, R., Espinoza, N., Eyheramendy, S., Lopes, C.E.F., Fiaschi, M., Hajdu, G., Han, J., Hełminiak, K.G., Hempel, A., Hidalgo, S.L., Ita, Y., Jeon, Y.B., Jordán, A., Kwon, J., Lee, J.T., Martín, E.L., Masetti, N., Matsunaga, N., Milone, A.P., Minniti, D., Morelli, L., Murgas, F., Nagayama, T., Navarro, C., Ochner, P., Pérez, P., Pichara, K., Rojas-Arriagada, A., Roquette, J., Saito, R.K., Siviero, A., Sohn, J., Sung, H.I., Tamura, M., Tata, R., Tomasella, L., Townsend, B., Whitelock, P.: The VVV Templates Project. Towards an Automated Classification of VVV Light-Curves. I. Building a database of stellar variability in the near-infrared. Astronomy & Astrophysics **567** (July 2014) A100 arXiv: 1405.4517.

[2] Cavuoti, S.: Data-rich astronomy: mining synoptic sky surveys. arXiv:1304.6615 [astro-ph] (April 2013) arXiv: 1304.6615.

[3] Eastman, J., Siverd, R., Gaudi, B.S.: Achieving better than 1 minute accuracy in the Heliocentric and Barycentric Julian Dates. Publications of the Astronomical Society of the Pacific **122**(894) (2010) 935

[4] Emerson, J.P., Irwin, M.J., Lewis, J., Hodgkin, S., Evans, D., Bunclark, P., McMahon, R., Hambly, N.C., Mann, R.G., Bond, I., Sutorius, E., Read, M., Williams, P., Lawrence, A., Stewart, M.: VISTA data flow system: overview. Volume 5493. (2004) 401–410

[5] Gran, F., Minniti, D., Saito, R.K., Navarrete, C., Dékány, I., McDonald, I., Ramos, R.C., Catelan, M.: Bulge RR Lyrae stars in the VVV tile $\textit{b201}$. Astronomy & Astrophysics **575** (March 2015) A114 arXiv: 1501.00947.

[6] Gran, F., Minniti, D., Saito, R.K., Zoccali, M., Gonzalez, O.A., Navarrete, C., Catelan, M., Ramos, R.C., Elorrieta, F., Eyheramendy, S., Jordán, A.: Mapping the outer bulge with RRab stars from the VVV Survey. arXiv:1604.01336 [astro-ph] (April 2016) arXiv: 1604.01336.

[7] Gray, J., Szalay, A., Budavári, T., Thakar, A.R., Nieto-Santisteban, M.A., Lupton, R.: Cross-Matching Multiple Spatial Observations and Dealing with Missing Data. Technical Report MSR-TR-2006-175, Microsoft Research (December 2006)

[8] Lomb, N.R.: Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science **39**(2) (February 1976) 447–462

[9] Minniti, D., Lucas, P.W., Emerson, J.P., Saito, R.K., Hempel, M., Pietrukowicz, P., Ahumada, A.V., Alonso, M.V., Alonso-Garcia, J., Arias, J.I., Bandyopadhyay, R.M., Barbá, R.H., Barbuy, B., Bedin, L.R., Bica, E., Borissova, J., Bronfman, L., Carraro, G., Catelan, M., Clariá, J.J., Cross, N., de Grijs, R., Dékány, I., Drew, J.E., Fariña, C., Feinstein, C., Lajús, E.F., Gamen, R.C., Geisler, D., Gieren, W., Goldman, B., Gonzalez, O.A., Gunthardt, G., Gurovich, S., Hambly, N.C., Irwin, M.J., Ivanov, V.D., Jordán, A., Kerins, E., Kinemuchi, K., Kurtev, R., López-Corredoira, M., Macca-

rone, T., Masetti, N., Merlo, D., Messineo, M., Mirabel, I.F., Monaco, L., Morelli, L., Padilla, N., Palma, T., Parisi, M.C., Pignata, G., Rejkuba, M., Roman-Lopes, A., Sale, S.E., Schreiber, M.R., Schröder, A.C., Smith, M., Jr., L.S., Soto, M., Tamura, M., Tappert, C., Thompson, M.A., Toledo, I., Zoccali, M., Pietrzynski, G.: VISTA Variables in the Via Lactea (VVV): The public ESO near-IR variability survey of the Milky Way. New Astronomy **15**(5) (July 2010) 433–443

[10] Nun, I., Protopapas, P., Sim, B., Zhu, M., Dave, R., Castro, N., Pichara, K.: FATS: Feature Analysis for Time Series. arXiv:1506.00010 [astro-ph] (May 2015) arXiv: 1506.00010.

[11] Scargle, J.D.: Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal **263** (1982) 835–853

[12] Skrutskie, M.F., Cutri, R.M., Stiening, R., Weinberg, M.D., Schneider, S., Carpenter, J.M., Beichman, C., R. Capps, Chester, T., Elias, J., Huchra, J., Liebert, J., Lonsdale, C., Monet, D.G., Price, S., Seitzer, P., T. Jarrett, Kirkpatrick, J.D., Gizis, J.E., Howard, E., Evans, T., Fowler, J., Fullmer, L., Hurt, R., Light, R., Kopan, E.L., Marsh, K.A., McCallon, H.L., Tam, R., Dyk, S.V., Wheelock, S.: The Two Micron All Sky Survey (2mass). The Astronomical Journal **131**(2) (2006) 1163

[13] Soszynski, I., Dziembowski, W.A., Udalski, A., Poleski, R., Szymanski, M.K., Kubiak, M., Pietrzynski, G., Wyrzykowski, L., Ulaczyk, K., Kozlowski, S., Pietrukowicz, P.: The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XI. RR Lyrae Stars in the Galactic Bulge. arXiv:1105.6126 [astro-ph] (May 2011) arXiv: 1105.6126.

[14] Soszynski, I., Udalski, A., Szymanski, M.K., Pietrukowicz, P., Mroz, P., Skowron, J., Kozlowski, S., Poleski, R., Skowron, D., Pietrzynski, G., Wyrzykowski, L., Ulaczyk, K., Kubiak, M.: Over 38000 RR Lyrae Stars in the OGLE Galactic Bulge Fields. arXiv:1410.1542 [astro-ph] (October 2014) arXiv: 1410.1542.

[15] Zopounidis, C., Doumpos, M.: Multicriteria classification and sorting methods: A literature review. European Journal of Operational Research **138**(2) (April 2002) 229–246