Clasificación de Series de Tiempo Astronómicas

Muriel Pérez
201011755

8 de abril de 2015

Índice general

Capítulo 1

Introducción

Con los avances en técnicas de observación astronómica que han sucedido en los últimos años, hay grandes cantidades de datos disponibles. Por ejemplo se espera que el *VISTA Variables in the Via Lactea* (VVV) del *European Southern Observatory* (ESO) produzca del orden de 10^9 curvas de luz[1] de fuentes puntuales en el infrarojo cercano con hasta 100 observaciones en diferentes épocas de alta calidad. De la misma forma estudios como la misión Kepler de la *National Aeronautics and Space Administration* (NASA), cuyo objetivo principal es la detección de exoplanetas, tienen como subproductos gran cantidad de curvas de luz.

Para que estos datos sean útiles para la comunidad científica es necesario clasificarlos y extraer sus características. Aunque los métodos automáticos muchas veces no pueden igualar la inspección manual por parte de un experto, la cantidad de datos disponible hace que esta tarea no sea posible en corto tiempo y hace necesario utilizar técnicas de minería de datos. Este interés se manifiesta en proyectos como el *VVV Templates Project* que tiene como objetivo consolidar una base bien definida de curvas de luz de estrellas variables en el infrarojo cercano para ser utilzadas como referencia para la clasificación automática de curvas de luz.

Debido a limitaciones en el tiempo de observación, fallas técnicas, periodos de matenimiento de los instrumentos utilizados y la imposibilidad de abservar todas las regiones del cielo durante todo el año, las curvas de luz no

[1] La curva de luz de una estrella es el resultado de medir su magnitud como función del tiempo. La magnitud de una estrella es el flujo de energía observado en una parte del espectro electromagnético (una banda), delimitada por un filtro, en escala logarítmica (ver el capítulo 4 de [4]).

constan del mismo número de observaciones y éstas no son hechas en intervalos regulares. Como el tiempo durante el cual cada estrella no es observada no es predecible y en ellos no se observan características importantes de las curvas de luz, estas no pueden ser analizadas con técnicas clásicas de análisis de series de tiempo.

En estudios previos [2, 10, 7] se le ha asignado a cada curva de luz un vector, llamado vector de características, y, basado en este vector, se ha hecho la clasificación automática. La escogencia de el vector de características es crucial para el proceso de clasificación porque con él se debe poder clasificar cada curva de luz, es decir, debe logar que, en el espacio de características las clases se superpongan lo menos posible. Para la conformación de este vector se han elegido coeficientes de Fourier de la curva de luz[2, 10, 7], que son calculados mediante métodos como el periodograma de Lomb-Scargle [11] o la minimización de la entropía de Shannon de la gráfica de la curva [1].

Esta elección de caracterísicas no es del todo conveniente porque requiere de gran poder computacional y limita el tipo de objetos que pueden ser clasificados. El cálculo del periodogramas como el de Lomb-Scargle para curvas de luz, y en general el de los métodos utilizados en la literatura, requiere de intentar una gran cantidad de periodos candidatos a ser el periodo de la curva de luz para luego elegir el mejor. Los periodos de los objetos observados varía entre desde unos pocos minutos y varios años por lo cual se requiere probar una gran cantidad de periodos. Por un lado este es un proceso es computacionalmente intensivo, lo que limita su uso en conjuntos grandes de curvas de luz; y por otro lado no es seguro que dé como resultado el periodo real de una curva de luz, por lo que a menudo éste debe ser revisado manualmente. Además el resultado de la clasificación puede ser sensible a la calidad de las curvas de luz que sean elegidas como muestra de entrenamiento [2] y limita el estudio a fuentes periódicas.

En [8, 9], los autores notaron que algunas variables descriptivas de la serie de magnitudes de una curva de luz (como su sesgo o su curtosis) sirven para clasificar ciertos tipos de estrellas con clasificadores lineales. En este trabajo retomamos esa idea y construimos un vector de características basadas en variables tomadas de estadística descriptiva. El uso de este tipo de variables tiene las ventajas de que puede ser calculadas de manera rápida y dan como resultado un vector de caracterísicas que sive para realizar clasificación con una taza de éxito alta. Para evaluar esta aproximación al problema utilizamos una parte del Catálogo de Estrellas Variables de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III)[12, 17, 19, 15, 14, 13, 23, 22,

20, 5, 3, 6, 16, 21, 18] que contiene curvas de luz de estrellas previamente clasificadas en seis tipos de variabilidad estelar y curvas de luz de estrellas candidatas a ser clasificadas como Be (ver cuadro 2.1).

En este trabajo utilizamos k-vecinos más cercanos, árboles de clasificación, máquinas de soporte vectorial y bosques aleatorios para realizar la clasificación automática de las curvas de luz basada en nuestra elección de características. Estos clasificadores fueron elegidos porque son aproximaciones muy distintas al problema de clasificación, por su naturaleza no lineal y no paramétrica; y por el hecho de que han mostrado ser efectivos en gran cantidad de aplicaciones prácticas.

Este documento está organizado de la siguiente forma. En el capítulo 2 damos una descripción del conjunto de datos utilizado en este trabajo. En el capítulo 3 presentamos y discutimos la elección de atributos y evaluamos el desempeño de los clasificadores mediante validación cruzada. En los apéndices damos una introducción matemática del problema de aprendizaje en general y una descripción de cada uno de los algorítmos utilizados en el trabajo.

Capítulo 2

El conjunto de Datos

Los datos utilizados en este trabajo provienen de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE-III). OGLE es un proyecto de larga duración cuyo objetivo principal es la búsqueda de materia oscura mediante el aprovechamiento de lentes gravitacionales. La tercera fase del protecto comenzó en 2001 y hace uso de un telescopio de 1,3m de diámetro localizado en el observatorio de Las Campanas, Chile[24]. Uno de los principales resultados de OGLE-III es la reducción y publicación [25] de las curvas de luz de objetos en el bulbo de la Galaxia, la Gran Nube de Magallanes y la Pequeña Nube de Magallanes. En este trabajo utilizamos las curvas de luz de 431653 objetos del catálogo de estrellas variables de OGLE-III de seis tipos de variabilidad (ver tabla 2.1) al cual se puede acceder en la página del proyecto [1] y 475 curvas de luz de estrellas candidatas a ser clasificadas como Be (ESCRIBIR DE DÓNDE FUERON TOMADAS ESTAS).

Las curvas de luz tomadas del catálogo de estrellas variables de OGLE-III se encuentran clasificadas por tipo de variabilidad estelar en un proceso que que involucró, en una etapa, la inspección manual de las curvas de luz (ver referencias en la tabla 2.1) por lo cual tomaremos esta clasificación como verdadera. En este trabajo utilizamos únicamente las curvas de luz registradas en la banda I [2] a pesar de que también se encuentra disponible información adicional sobre las curvas de luz como sus periodos y algunos coeficientes

[1] `http://ogle.astrouw.edu.pl/`

[2] Los objetos observados emiten radiación en una parte amplia del espectro electromagnético. Los telescopios utilizan filtros para recoger solo la radiación emitida por estos objetos en ciertas partes del espectro electromagnético. El filtro I (infrarojo) tiene un ancho de banda de 149nm y una longitud de onda efectiva de 797nm (ver [4])

Cuadro 2.1: Conjunto de datos utilizados. BG hace referencia al Bulbo Galáctico; PNM, a la Pequeña Nube de Magallanes y GNM, a la Gran Nube de Magallanes.

Tipo de variabilidad y origen	Número de Objetos
RR Lyrae - BG[12]	16836
RR Lyrae - PNM [17]	2475
RR Lyrae - GNM [19]	24906
Cefeidas - BG [15]	32
Cefeidas - PNM [14]	4630
Cefeidas - GNM [13]	3361
Variables de Largo Periodo - BG [23]	232406
Variables de Largo Periodo - PNM [22]	19384
Variables de Largo Periodo - GNM [20]	91995
Sistema Binario Eclipsante - PNM [5]	6138
Sistema Binario Eclipsante - GNM [3]	26121
δ-Scuti - Nube Mayor de Magallanes[6]	2786
Cefeidas Tipo II - BG [16]	335
Cefeidas Tipo II - PNM [21]	43
Cefeidas Tiplo II - GNM [18]	197
Cadidata a Be - Vía Láctea (cita!)	475

de Fourier (ver referencias en la tabla 2.1). Esta elección se debe a que el cálculo de estas cantidades es computacionalmente intensivo, no siempre se encuentran disponible y proponemos hacer la clasificación utilizando variables tomadas de estadística descriptiva.

Agrupamos los 432128 objetos disponibles en siete clases de variabilidad estelar (ver tabla2.2). Esta elección de clases puede ser refinada puesto que en cada una de estas clases existen subclases. Por ejemplo entre las Cefeidas se puede distinguir entre aquellas que pulsan en su modo fundamental, en su primer sobretono (segundo armónico) o en su segundo sobretono (tercer armónico) (ver figura 2). Sin embargo conocer a qué clase de variabilidad estelar pertence un objeto facilita considerablemente su clasificación en subclases y análisis subsecuentes.

En el Catálogo de Estrellas Variables de OGLE-III, cada curva de luz está disponible en un archivo que contiene tres columnas con los valores de

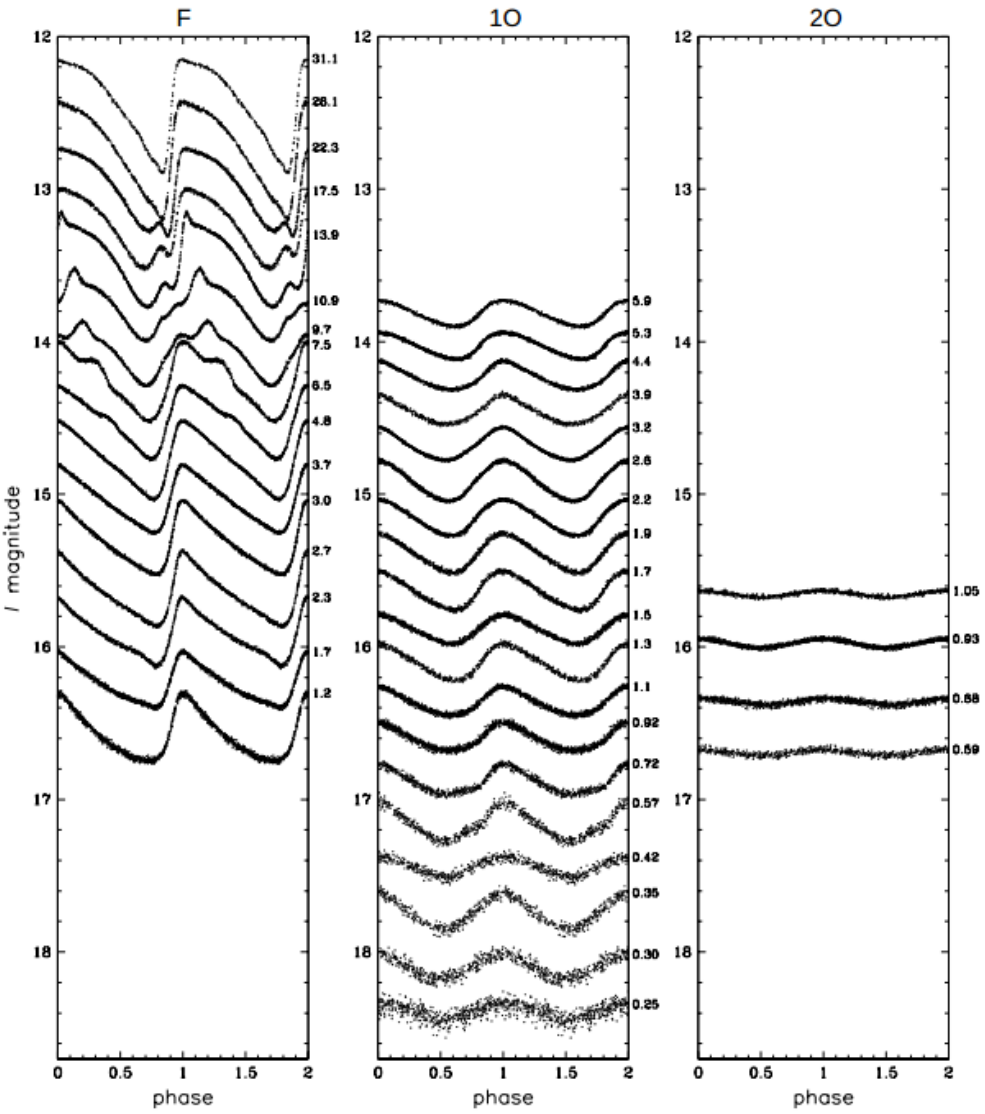

Figura 2.1: Curvas de luz ilustrativas de Cefeidas en modo fundamental (izquierda), primer sobretono (mitad), segundo sobretono (derecha). Los números pequeños a la derecha de cada recuadro muestran los periodos redondeados de las curvas de luz presentadas en los recuadros. Tomado de [15]

Cuadro 2.2: Cantidad de datos por tipo de variabilidad

Tipo de Variabilidad	Cantidad
Variables de Largo Periodo (VLP)	343782
RR Lyrae (RRLyr)	44217
Cefeida (Cef)	8004
Sistema Binario Eclipsante (SBE)	32259
δ-Scuti (δSct)	2788
Cefeida Tipo II (CefT2)	603
Candidata a Be (BeEC)	475
Total	432128

Figura 2.2: (figura pendiente de imagen) Curva de luz (nombre del archivo) del catálogo OGLE-III. Los periodos en los que no hay mediciones corresponden a los momentos del año en los que la zona en la que se encuentra el objeto no puede ser observada debido a la posición relativa entre el Sol y la Tierra.

magnitud, fecha juliana [3] en la que fue tomada cada medida y error en la medida de la magnitud. El número de medidas para cada objeto y la separación temporal varía ampliamente. La separación mínima dos mediciones en toda la muestra es de 0.00147d, la máxima es 2156.9d y en promedio están separadas por 5.1d; por su parte el número promedio de observaciones por objeto es 759; el máximo, 5173; y el mínimo, 11. El 75 % de los objetos cuenta con más de 386 observaciones. Para todos los objetos estas observaciones están repartidas en los (número de años) años en que estuvo activo OGLE-III. En la figura 2 se puede observar una curva de luz del catálogo de estrellas variables de OGLE-III.

[3] La fecha Juliana es el tiempo medido en días desde el 1 de enero de 4713 a. C.

Capítulo 3

Clasificación

3.1. Atributos Seleccionados

3.2. K Vecinos Más Cercanos

Cuadro 3.1: k = 1

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	381	1	0	50	41	0	0
cep	1	6595	4	172	54	986	87
dcst	0	7	2064	340	13	151	0
ebs	51	184	536	30129	573	668	29
lpv	42	187	19	944	342740	371	158
rrlyr	0	971	165	595	308	41911	217
t2cep	0	59	0	29	53	130	112

3.3. Árboles de clasificación y regresión

3.4. Máquinas de Soporte Vectorial

3.5. Bosques Aleatorios

Cuadro 3.2: k=1, validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.80	0.00	0.00	0.00	0.00	0.00	0.00
cep	0.00	0.82	0.00	0.01	0.00	0.02	0.14
dcst	0.00	0.00	0.74	0.01	0.00	0.00	0.00
ebs	0.11	0.02	0.19	0.93	0.00	0.02	0.05
lpv	0.09	0.02	0.01	0.03	1.00	0.01	0.26
rrlyr	0.00	0.12	0.06	0.02	0.00	0.95	0.36
t2cep	0.00	0.01	0.00	0.00	0.00	0.00	0.19

Cuadro 3.3: Matriz de confusión para CART

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	470	12	6	1039	8906	47	18
cep	0	6210	12	18	320	3098	92
dcst	0	35	2511	5538	91	1615	1
ebs	0	1	148	21346	346	64	1
lpv	5	107	24	1762	304810	100	3
rrlyr	0	648	82	552	13880	34313	92
t2cep	0	991	5	2004	15429	4980	396

Cuadro 3.4: Tasas de clasificación estimadas por validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.99	0.00	0.00	0.03	0.03	0.00	0.03
cep	0.00	0.78	0.00	0.00	0.00	0.07	0.15
dcst	0.00	0.00	0.90	0.17	0.00	0.04	0.00
ebs	0.00	0.00	0.05	0.66	0.00	0.00	0.00
lpv	0.01	0.01	0.01	0.05	0.89	0.00	0.00
rrlyr	0.00	0.08	0.03	0.02	0.04	0.78	0.15
t2cep	0.00	0.12	0.00	0.06	0.04	0.11	0.66

Cuadro 3.5: $\gamma = 0{,}1$, costo = 16

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	297	0	2	26	16	0	0
cep	0	4622	0	55	85	788	69
dcst	0	0	1683	163	4	47	0
ebs	89	48	865	29439	198	563	5
lpv	89	490	22	1380	342823	1113	255
rrlyr	0	2844	216	1196	656	41706	274
t2cep	0	0	0	0	0	0	0

Cuadro 3.6: $\gamma = 0{,}1$, costo = 16, validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.63	0.00	0.00	0.00	0.00	0.00	0.00
cep	0.00	0.58	0.00	0.00	0.00	0.02	0.11
dcst	0.00	0.00	0.60	0.01	0.00	0.00	0.00
ebs	0.19	0.01	0.31	0.91	0.00	0.01	0.01
lpv	0.19	0.06	0.01	0.04	1.00	0.03	0.42
rrlyr	0.00	0.36	0.08	0.04	0.00	0.94	0.45
t2cep	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Capítulo 4

Apéndices

4.1. El Problema del Aprendizaje

4.2. K Vecinos Más cercanos

4.3. Árboles de Clasificación y Regresión

4.4. Máquinas de Soporte Vectorial

4.5. Bosques Aleatorios

4.6. Estimación del Error de Clasificación

Capítulo 5

Cosas que la evolución se llevó

5.1. vieja introducción

Bibliografía

[1] Pablo M. Cincotta, Mariano Mendez, and Josue A. Nunez. Astronomical time series analysis. I. A search for periodicity using information entropy. *The Astrophysical Journal*, 449:231, 1995.

[2] Jonas Debosscher, L. M. Sarro, Conny Aerts, J. Cuypers, Bart Vandenbussche, R. Garrido, and E. Solano. Automated supervised classification of variable stars-I. Methodology. *Astronomy & Astrophysics*, 475(3):1159–1183, 2007.

[3] D. Graczyk, I. Soszyński, R. Poleski, G. Pietrzyński, A. Udalski, M. K. Szymański, M. Kubiak, Ł. Wyrzykowski, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XII. Eclipsing Binary Stars in the Large Magellanic Cloud. *Acta Astronomica*, 61:103–122, June 2011.

[4] Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, and Karl Johan Donner, editors. *Fundamental Astronomy*. Springer, Berlin ; New York, 5th edition edition, August 2007.

[5] M. Pawlak, D. Graczyk, I. Soszyński, P. Pietrukowicz, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, S. Kozłowski, and J. Skowron. Eclipsing Binary Stars in the OGLE-III Fields of the Small Magellanic Cloud. *Acta Astronomica*, 63:323–338, September 2013.

[6] R. Poleski, I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VI. Delta Scuti Stars in the Large Magellanic Cloud. *Acta Astronomica*, 60:1–16, March 2010.

[7] Joseph W. Richards, Dan L. Starr, Nathaniel R. Butler, Joshua S. Bloom, John M. Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *The Astrophysical Journal*, 733(1):10, May 2011.

[8] Bayron Stevenson Rodríguez Feliciano and José Alejandro García Varela. *Análisis estadístico en poblaciones de estrellas variables.* Tesis (Físico). Universidad de los Andes. Bogotá : Uniandes, 2012., 2012.

[9] B. E. Sabogal, A. García-Varela, and R. E. Mennickent. Search for Southern Galactic Be Star Candidates. *Publications of the Astronomical Society of the Pacific*, 126:219–226, 2014.

[10] L. M. Sarro, Jonas Debosscher, M. López, and Conny Aerts. Automated supervised classification of variable stars-II. Application to the OGLE database. *Astronomy & Astrophysics*, 494(2):739–768, 2009.

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

[12] I. Soszyński, W. A. Dziembowski, A. Udalski, R. Poleski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, S. Kozłowski, and P. Pietrukowicz. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XI. RR Lyrae Stars in the Galactic Bulge. *Acta Astronomica*, 61:1–23, March 2011.

[13] I. Soszyński, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, L. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. I. Classical Cepheids in the Large Magellanic Cloud. *Acta Astronomica*, 58:163–185, September 2008.

[14] I. Soszyński, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VII. Classical Cepheids in the Small Magellanic Cloud. *Acta Astronomica*, 60:17–39, March 2010.

[15] I. Soszyński, A. Udalski, P. Pietrukowicz, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, and S. Kozłowski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XIV. Classical and TypeII Cepheids in the Galactic Bulge. *Acta Astronomica*, 61:285–301, December 2011.

[16] I. Soszyński, A. Udalski, P. Pietrukowicz, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, and S. Kozłowski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. Type II Cepheids in the Galactic Bulge - Supplement. *Acta Astronomica*, 63:37–40, March 2013.

[17] I. Soszyński, A. Udalski, M. K. Szymański, J. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IX. RR Lyr Stars in the Small Magellanic Cloud. *Acta Astronomica*, 60:165–178, September 2010.

[18] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. II.Type II Cepheids and Anomalous Cepheids in the Large Magellanic Cloud. *Acta Astronomica*, 58:293, December 2008.

[19] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. III. RR Lyrae Stars in the Large Magellanic Cloud. *Acta Astronomica*, 59:1–18, March 2009.

[20] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IV. Long-Period Variables in the Large Magellanic Cloud. *Acta Astronomica*, 59:239–253, September 2009.

[21] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VIII.

Type II Cepheids in the Small Magellanic Cloud. *Acta Astronomica*, 60:91–107, June 2010.

[22] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, and P. Pietrukowicz. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XIII. Long-Period Variables in the Small Magellanic Cloud. *Acta Astronomica*, 61:217–230, September 2011.

[23] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, P. Pietrukowicz, and J. Skowron. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XV. Long-Period Variables in the Galactic Bulge. *Acta Astronomica*, 63:21–36, March 2013.

[24] A. Udalski. The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey. *Acta Astron.*, 53(astro-ph/0401123):291, 2004.

[25] A. Udalski, M. K. Szymanski, I. Soszynski, and R. Poleski. The Optical Gravitational Lensing Experiment. Final Reductions of the OGLE-III Data. *Acta Astronomica*, 58:69–87, 2008.