

# Clasificación de Series de Tiempo Astronómicas

Muriel Pérez  
201011755

7 de abril de 2015



# Índice general

<b>1. Introducción</b>	<b>5</b>
<b>2. El conjunto de Datos</b>	<b>9</b>
<b>3. Clasificación</b>	<b>13</b>
3.1. Atributos Seleccionados . . . . .	13
3.2. K Vecinos Más Cercanos . . . . .	13
3.3. Árboles de clasificación y regresión . . . . .	13
3.4. Máquinas de Soporte Vectorial . . . . .	13
3.5. Bosques Aleatorios . . . . .	13
<b>4. Apéndices</b>	<b>17</b>
4.1. El Problema del Aprendizaje . . . . .	17
4.2. K Vecinos Más cercanos . . . . .	17
4.3. Árboles de Clasificación y Regresión . . . . .	17
4.4. Máquinas de Soporte Vectorial . . . . .	17
4.5. Bosques Aleatorios . . . . .	17
4.6. Estimación del Error de Clasificación . . . . .	17
<b>5. Cosas que la evolución se llevó</b>	<b>19</b>
5.1. vieja introducción . . . . .	19



# 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<sup>1</sup> 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 escalas de tiempo razonables 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 utilizadas como referencia para la clasificación automática de curvas de luz.

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

---

<sup>1</sup> 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 [2]).

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(citas!) se le ha asignado a cada curva de luz un 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 capturar la información que hace a cada clase de variabilidad diferente de las demás. Usualmente este vector de características está conformado por coeficientes de Fourier de la curva de luz, que son calculados mediante métodos como el periodograma de Lomb-Scargle (cita) o minimización de entropía (cita).

Esta elección de características no es del todo conveniente porque presenta problemas computacionales 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 todos 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. Por un lado estos 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 (citar) y limita el estudio a fuentes periódicas.

En (cita Bayron) los autores notaron que algunas variables descriptivas de la serie de magnitudes de una curva de luz (como su sesgo o su curtosis) servían para clasificar ciertos tipos de estrellas con clasificadores lineales. En este trabajo retomamos esa idea y construimos un vector de características basadas en variables descriptivas. El uso de este tipo de variables tiene las ventajas de que se pueden calcular de manera rápida y da como resultado un vector de características que sirve para realizar clasificación con una tasa de éxito alta. Para evaluar esta aproximación al problema utilizamos una parte de los resultados de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III) 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 atributos. Estos clasificadores fueron elegidos porque representan 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 los resultados de la clasificación automática utilizando los diferentes algoritmos así como la elección de atributos. En los apéndices damos una descripción matemática del problema de aprendizaje en general y una descripción de cada uno de los algoritmos 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 proyecto comenzó en 2001 y hace uso de un telescopio de 1,3m de diámetro localizado en observatorio de Las Campanas, Chile[17]. Uno de los principales resultados de OGLE-III es la reducción[18] y publicación 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 <sup>1</sup> 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 fueron originalmente clasificadas por tipo de variabilidad estelar en un proceso 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. También se encuentra disponible información adicional sobre las curvas de luz como sus periodos y algunos coeficientes de Fourier (ver referencias en la tabla 2.1) pero dado que para futuras bases de datos sin clasificar el cálculo de estas cantidades es computacionalmente intensivo y proponemos hacer la clasificación utilizando variables tomadas de estadística descriptiva, no utilizamos esta información. En este trabajo utilizamos las

---

<sup>1</sup><http://ogle.astrouw.edu.pl/>

Cuadro 2.1: Conjunto de datos utilizados (faltan las citas de los catálogos). 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 [5]	16836
RR Lyrae - PNM [10]	2475
RR Lyrae - GNM [12]	24906
Cefeidas - BG [8]	32
Cefeidas - PNM [7]	4630
Cefeidas - GNM [6]	3361
Variables de Largo Periodo - BG [16]	232406
Variables de Largo Periodo - PNM [15]	19384
Variables de Largo Periodo - GNM [13]	91995
Sistema Binario Eclipsante - PNM [3]	6138
Sistema Binario Eclipsante - GNM [1]	26121
$\delta$ -Scuti - Nube Mayor de Magallanes [4]	2786
Cefeidas Tipo II - BG [9]	335
Cefeidas Tipo II - PNM [14]	43
Cefeidas Tipo II - GNM [11]	197
Candidata a Be - Vía Láctea (cita!)	475

curvas de luz registradas en el filtro I <sup>2</sup> (¿POR QUÉ ELEGIMOS ESTA BANDA?)

Agrupamos los 432128 objetos disponibles en siete clases de variabilidad estelar (ver tabla 2.2). 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 pertenece un objeto facilita considerablemente su clasificación en subclases y análisis subsecuentes.

<sup>2</sup>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 [2])

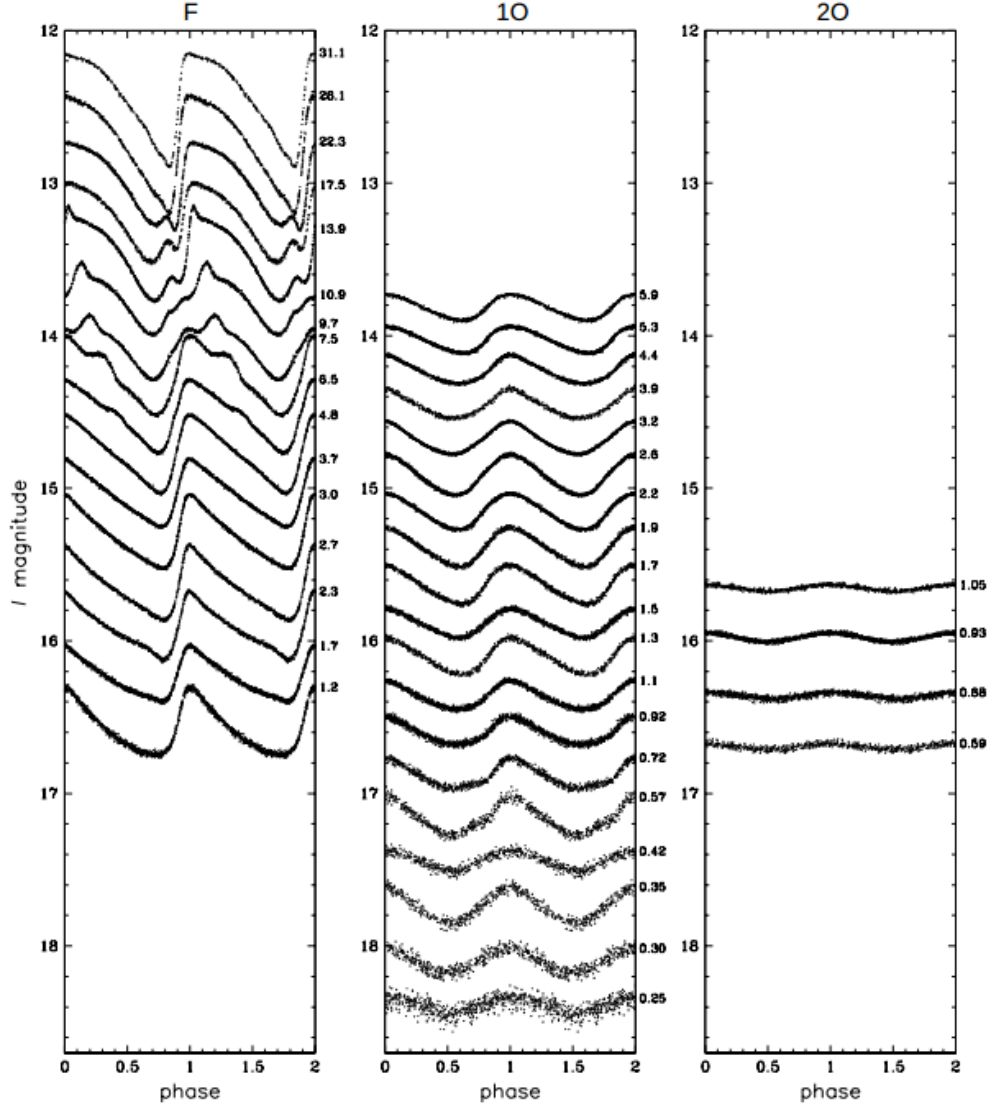


Figura 2.1: Curvas de luz ilustrativas de Cefeidas en modo fundamental (izquierda), primer sobretono (mitad), segundo sobretono (derecha). Los números pequeños a la derecha de cada recuadro muestran los periodos redondeados de las curvas de luz presentadas en los recuadros. Tomado de [8]

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
$\delta$ -Scuti ( $\delta$ Sct)	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.

Cada curva de luz está guarada en un archivo que contiene tres columnas con los valores de magnitud, fecha juliana 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 y 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 activa la tercera fase de OGLE. En la figura (figura pendiente) se puede observar una curva de luz del catálogo de estrellas variables de OGLE-III.

# 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







# 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



# Bibliografia

- [1] D. Graczyk, I. Soszyński, R. Poleski, G. Pietrzyński, A. Udalski, M. K. Szymański, M. Kubiak, L. 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.
- [2] Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, and Karl Johan Donner, editors. *Fundamental Astronomy*. Springer, Berlin ; New York, 5th edition edition, August 2007.
- [3] M. Pawlak, D. Graczyk, I. Soszyński, P. Pietrukowicz, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, L. 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.
- [4] R. Poleski, I. Soszyński, 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. VI. Delta Scuti Stars in the Large Magellanic Cloud. *Acta Astronomica*, 60:1–16, March 2010.
- [5] I. Soszyński, W. A. Dziembowski, A. Udalski, R. Poleski, M. K. Szymański, M. Kubiak, G. Pietrzyński, L. 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.
- [6] 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.
- [7] 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.
- [8] 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 Type II Cepheids in the Galactic Bulge. *Acta Astronomica*, 61:285–301, December 2011.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.

- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] A. Udalski. The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey. *Acta Astron.*, 53(astroph/0401123):291, 2004.
- [18] 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.