**MASARYKOVA UNIVERZITA**
**PŘÍRODOVĚDECKÁ FAKULTA**
**ÚSTAV TEORETICKÉ FYZIKY A ASTROFYZIKY**

# Bakalářská práce

**BRNO 2016** **MARTIN VO**


**MASARYKOVA UNIVERZITA**
**PŘÍRODOVĚDECKÁ FAKULTA**
**ÚSTAV TEORETICKÉ FYZIKY A ASTROFYZIKY**


# Detekce QSO ve vybraných přehlídkových databázích s pomocí metod data-miningu

Bakalářská práce

**Martin Vo**

**Vedoucí práce: Mgr. Viktor Votruba, Ph.D** **Brno 2016**

# Bibliografický záznam

| | |
|---|---|
| **Autor:** | Martin Vo<br>Přírodovědecká fakulta, Masarykova univerzita<br>Ústav teoretické fyziky a astrofyziky |
| **Název práce:** | Detekce QSO ve vybraných přehlídkových databázích s pomocí metod data-miningu |
| **Studijní program:** | Fyzika |
| **Studijní obor:** | Astrofyzika |
| **Vedoucí práce:** | Mgr. Viktor Votruba, Ph.D |
| **Akademický rok:** | 2015/2016 |
| **Počet stran:** | ix+39 |
| **Klíčová slova:** | Kvazary, data mining, big data, světelné křivky |

# Bibliographic Entry


| | |
|---|---|
| **Author:** | Martin Vo<br>Faculty of Science, Masaryk University<br>Department of Theoretical Physics and Astrophysics |


| | |
|---|---|
| **Title of Thesis:** | Detection of QSO in sky surveys using data-mining methods |
| **Degree Programme:** | Physics |
| **Field of Study:** | Astrophysics |
| **Supervisor:** | Mgr. Viktor Votruba, Ph.D |
| **Academic Year:** | 2015/2016 |
| **Number of Pages:** | ix+39 |
| **Keywords:** | Quasars, data mining, big data, light curves |

# Abstrakt

Cílem této práce bylo vyvinutí metod pro identifikaci kvazarů podle jejich světelných křivek. K odfiltrování světelných křivek bez trendu bylo použito Abbe kriterium. Dále byly též vyjmuty hvězdy s barevnými indexy $B-V > 1.5$ mag a $V-I > 1.5$ mag. Ze světelných křivek spektroskopicky potvrzených kvazarů byla vytvořena šablona, jejíž histogramy a variogramy byly porovnávány v SAX reprezentaci s testovanými světelnými křivkami. Pro všechny možné kombinace parametrů byly vybrány ty nejvhodnější za pomoci pythonovské knihovny GridSearch. Hranice kvazarů a nekvazarů v symbolickém histogram-variogramovém prostoru byly získány clusteringovými metodami. Metoda byla testována na vzorku proměnných a neproměnných hvězd, kde správně bylo identifikováno jako kvazary 86 % a identifikováno nesprávně bylo 21 % nekvazarů. Bylo otestováno přes 1 milion hvězd v OGLE II databázi, z nichž bylo identifikováno jako kvazary téměř 4000 kandidátů. Pro studium světelných křivek jsem vyvinul program v jazyku Python, který je díky objektovému přístupu a způsobu napsání kosterních tříd možné dále rozvíjet v unikátní nástroj pro analýzu a klasifikaci světelných křivek.

# Abstract

The aim of this thesis was to develop the methods for identifying quasars by their light curves. Abbe criterion was used to remove the light curves without trend. Furthermore the stars with color index $B-V > 1.5$ mag and $V-I > 1.5$ mag were also excluded. The spectroscopically confirmed quasars template has been created from the light curves which histograms and variograms were compared in SAX representation with the tested light curves. For all possible combinations of the parameters were selected as the most appropriate those I obtained by using Python library GridSearch. Boundaries of quasars and non-quasars in symbolic histogram-variogram space were obtained by clustering's methods. The method was tested on a sample of variable and non-variable stars from which 86 % were correctly identified as quasars and 21 % incorrectly. It was tested about one million stars in the OGLE database II from which were identified as quasars almost 4000 candidates. For the study of the light curves I developed a program in Python which is due to object oriented approach and the way of writing main classes possible to further develop in a unique tool for analyzing and classifying light curves.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

# ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2015/2016

**Ústav:** Ústav teoretické fyziky a astrofyziky

**Student:** Martin Vo

**Program:** Fyzika

**Obor:** Astrofyzika

Ředitel *Ústavu teoretické fyziky a astrofyziky* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s tématem:

**Téma práce:** Detekce QSO ve vybraných přehlídkových databázích s pomocí metod data-miningu

**Téma práce anglicky:** Detection of QSO in sky surveys using data-mining methods

**Oficiální zadání:**
Úkolem studenta bude osvojení základních dovedností v oblasti analýzy velkých dat a jejich aplikace na problém klasifikace a vyhledávání podobností resp. anomálií ve světelných křivkách. Specifickým úkolem pak bude vyhledání potencionálních kandidátů QSO a porovnání výstupu vlastních metod s výsledky práce Pichara et al.(2012).

**Jazyk závěrečné práce:** ČEŠTINA

**Vedoucí práce:** Mgr. Viktor Votruba, Ph.D.

**Datum zadání práce:** 24. 11. 2015

**V Brně dne:** 1. 12. 2015

Souhlasím se zadáním (podpis, datum):

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Martin Vo | Mgr. Viktor Votruba, Ph.D. | prof. Rikard von Unge, Ph.D. |
| **student** | **vedoucí práce** | **ředitel Ústavu teoretické fyziky a astrofyziky** |

# Poděkování

Na tomto místě bych chtěl poděkovat především vedoucímu své bakalářské práce Viktoru Votrubovi za odborné vedení, cenné rady a revizi samotného textu. Další poděkování patří MetaCentru za poskytnutí svých výpočetních prostředků. Na závěr bych chtěl věnovat své velké díky přítelkyni – Lence, která mi byla oporou a laskavým dozorčím nad mou gramatikou.

# Prohlášení

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

Brno 19.května 2016 .....

Martin Vo

# Obsah

# Úvod

Při pohledu na noční oblohu se nad námi začnou třpitit tisíce hvězd, jejichž počet se ještě mnohonásobně zvětší, nahlédneme-li do dalekohledu. Ještě před pár desítkami let astronomové netušili, že při pohledu na nekonečná hvězdná pole pozorují i exotické objekty hlubokého vesmíru, které několikasetkrát přezařují celé galaxie obsahující stovky miliard hvězd, ale přesto tak vzdálené, že se jeví jako obyčejné hvězdy. Světlo z nich k nám putuje miliardy let, a přestože mnohé z nich už dávno vyhasly, stále nám přináší cenné informace o tom, jak vesmír vypadal, když byl ještě velice mladý. Dnes už máme relativně mnoho pozorování těchto objektů, ale přesto o nich stále moc nevíme.

S přibývajícím množstvím vesmírných a pozemních teleskopů máme k dispozici nesčetné množství astronomických dat. Astronomické přehlídky chrlý každým dnem nová a nová pozorování, které je možné využít i ke studiu zcela jiných objektů, než bylo původně zamýšleno. V následujících kapitolách si ukážeme, jak můžeme tyto tajemné objekty identifikovat a co jsou vůbec zač?

# Kapitola 1

# Objevení

## 1.1 Začátky radioastronomie

Psal se rok 1931 a americký radioinženýr Karl Jansky, jako zástupce Bellových laboratoří, byl pověřen hledáním zdroje rádiového rušení. Mimo šumu zemského původu zaznamenal také rádiový signál přicházející z oblohy. Jeho pozice se měnila s rotací Země, z čehož usoudil, že šum pochází z vesmíru. Zdrojem záření se ukázal být střed naší Galaxie. Jansky byl později přeložen na jiný projekt, ale na jeho práci následně navázali další vědci a dali tak vzniknout novému oboru – radioastronomii. Posupně byly objeveny další zdroje rádiového záření ve vesmíru (např. výbuchy supernov), ale k většině z nich nebyl nalezen vizuální protějšek. Důvodem bylo nepříliš přesné určení polohy objektu radioteleskopy, a proto bylo obtížné rozhodnout, kde přesně na obloze se zdroje rádiového signálu nachází.

Astronomové proto vyvinuli důmyslné metody k přesnějšímu určení jejich pozic. Jednou z nejvýznamějších metod bylo využití zákrytu zdroje Měsícem, kdy se při zakrývání zmenšuje rádiový tok. Na začátku 70. let 20. století byly touto metodou určeny pozice několika silných rádiových zdrojů, které byly zaznamenány do 3C katalogu (3rd Cambridge Catalog).

## 1.2 Tajemné zdroje

Roku 1963 se o dva z těchto silných rádiových zdrojů začal zajímat Maarten Schmidt [1]. 3C 48 a 3C 273 byly velice neobvyklé objekty. Zdálo se, že jejich signál přichází z hvězdy. Z našeho Slunce slabé rádiové signály detekujeme, ale to pouze proto že je od nás na astronomické poměry velice blízko.

Spektrum této hvězdy bylo velkou záhadou. Schmidt zpočátku nemohl určit, jaký prvek vytváří tak silné spektrální čáry jež pozorovali. Potom si všiml, že ony neznámé čáry jsou emisní čáry vodíku, jen jsou velmi výrazně posunuté k červené části spektra. Předpokládáme-li platnost Dopplerova posuvu

$$z = \frac{\Delta\lambda}{\lambda} = \sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}} - 1, \tag{1.1}$$

pak pro 3C 273 ($z = 0.158$) to znamená, že se od nás vzdaluje rychlostí 14.6 % rychlostí světla (!). Podle Hubblova zákona [2]

$$v = H_0 r, \tag{1.2}$$

by takový objekt měl být asi 2 miliardy světelných let daleko. Ze znalosti pozorované hvězdné velikosti $m = 12.8$ mag a vzdálenosti $d$ (dosazujeme v pc) můžeme určit absolutní hvězdnou velikost $M$ (hvězdnou velikost ve vzdálenosti 10 parseků, která vypovídá o skutečnému zářivému výkonu)

$$M = m - 5\log d + 5. \tag{1.3}$$

Dostáváme absolutní hvězdnou velikost -26.7 mag. To znamená, že ve vzdálenosti asi 33 světelných let by zářil jako Slunce (které je asi 2 milionkrát blíže!). Tato hodnota může být použita k odhadu výkonu, kterou určíme kombinací Pogsonovy rovnice:

$$M - M_\odot = -2.5\log\frac{F}{F_\odot} \tag{1.4}$$

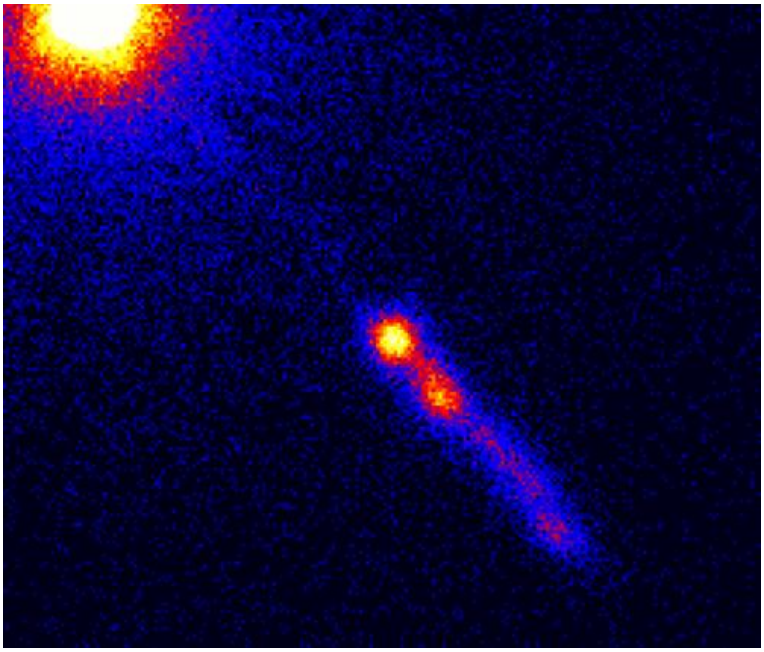
a vztahu pro zářivý výkon:

$$\frac{L}{L_\odot} = \frac{d^2}{d_\odot^2}\,\frac{F}{F_\odot}, \tag{1.5}$$

kde $F$ je zářivý tok, $L$ zářivý výkon a veličiny s indexem $\odot$ náleží Slunci. Dostáváme vztah pro odhad zářivého výkonu:

$$L = 100^{\frac{M_\odot - M}{5}}\, L_\odot = 2.6\cdot 10^{12} L_\odot = 1\cdot 10^{39} W. \tag{1.6}$$

Tento extrémní zářivý výkon řádově odpovídá 100 násobku zářivého výkonu celé naší Galaxie.



Obrázek 1.1: Rentgenový snímek 3C 273 z teleskopu Chandra (s výtryskem)[3]

Těmto objektům se začalo říkat kvazary (ang. quasars – quasi-stellar radio source), což v překladu znamená rádiové zdroje podobné hvězdám. Ani pojmenování však nepomohlo k rozluštění, o jaká astronomická tělesa se jedná. Na snímcích byly objeveny velice zářivé světelné výtrysky vycházející ze samotných kvazarů. Jádro kvazaru pak musí být ještě mnohem energetičnější než výtrysky, aby je bylo schopno produkovat. O jaký útvar v centrech galaxií se jedná, když je několikasetkrát jasnější než celá galaxie stovek miliard hvězd a emituje velké množství rádiových vln? Astronomové vyřešili jednu záhadu, ale další, ještě snad záhadnější, se objevila.

# Kapitola 2

# Kvazary

## 2.1 Pozůstatky raného vesmíru

Kvazary patří mezi nejenergetičtější a nejvzdálenější objekty ve vesmíru. Řadí se do třídy aktivních galaktických jader (AGN - Active Galactic Nuclei). Ve středu aktivních galaxií je supermasivní díra, která vyzařuje obrovské množství energie. Podle aktuálních teorií byly kvazary v minulosti běžné, pravděpodobně se nacházely ve středu téměř každé galaxie. Jakmile bylo všechno palivo spotřebováno, jednoduše vyhasly a zbyly po nich jen masivní černé díry [4].

Nyní pozorujeme kvazary jen v nejvzdálenějších koutech vesmíru. Jelikož má světlo konečnou rychlost, tak čím vzdálenější je nějaký objekt, tím delší čas k nám letí a tím více se díváme do minulosti. Právě proto pozorováním kvazarů můžeme zkoumat, jak vesmír vypadal, když byl ještě mladý. Kvazary mohou být zapáleny (popřípadě znovu zapáleny), když se dvě galaxie spojí a supermasivní černá díra je znovu zásobena čerstvým přísunem hmoty. K této situaci by mohlo dojít asi za 3 – 5 miliardy let, když se naše Galaxie „srazí“ s galaxií v Andromédě [5].

Obrázek 2.1: Umělecká představa kvazaru [6]

## 2.2 Vyzařování

### 2.2.1 Akrece

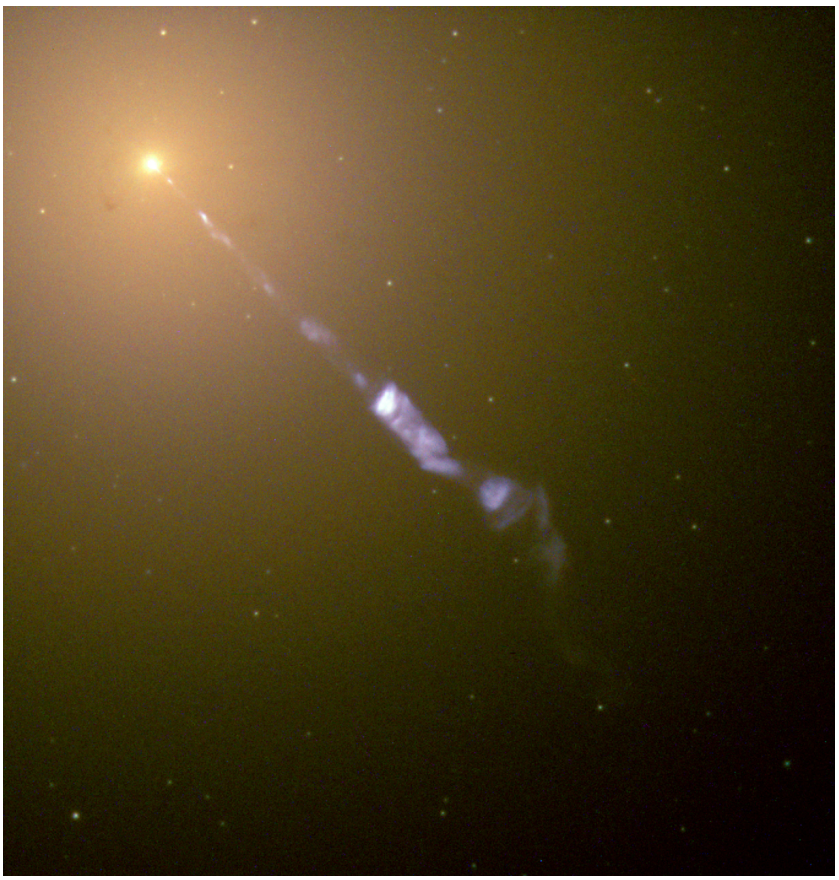
Současný model je založen na představě, že kvazar je tvořen supermasivní černou dírou a akrečním diskem, který ji obklopuje. Černá díra k sobě díky své gravitaci přitahuje okolní hmotná tělesa a vše, co do ní spadne, je nenávratně pryč. Tito obrovští jedlíci si všechny své zásoby „nesní“ najednou, ale uchovávají si je kolem sebe v podobě již zmíněných akrečních disků. Jedná se především o plynnou strukturu z dopadajících těles, kterou černá díra postupně tráví.

Typický výkon kvazaru je $10^{40}$ watů (výkon 100 miliard Sluncí). K tomuto výkonu by supermasivní černá díra musela zkonzumovat 10 hvězd za rok. Nejzářivější kvazary dokonce stráví hmotu o hmotnosti 1000 Sluncí za rok. Změní-li se v nějakém místě kvazaru výkon, změna se začne šířit než zachvátí všechny jeho části. Budeme-li předpokládat, že změna výkonu se šíří téměř rychlostí světla a všechny jeho části jsou ve vzájemném kontaktu, můžeme řádově odhadnout rozměry kvazaru. Například pokud pozorujeme změny jasnosti v řádech týdnů, bude jeho průměr několik světelných týdnů (vzdálenost, za kterou urazí světlo za několik týdnů). [7]

Jasnosti kvazarů se mění v průběhu měsíců až hodin. To znamená, že oblast, ze které tvoří a emitují záření je relativně malá [8]. Vyzařování tak velkého množství energie z tak malého prostoru vyžaduje mnohem efektivnější zdroj, než jadernou fúzi (jako u hvězd). Dostačující výkon může poskytnout exploze hmotných hvězd, ale to jen po dobu pár týdnů. Jediným známým zdrojem tak velkého dlouhodobého výkonu by mohlo být uvolňování potenciální energie při pádu materiálu do masivní černé díry.

### 2.2.2 Jety

Přestože je akreční disk velice zářivý, může být přezářen tzv. jety. Jedná se o dva protilehlé výtrysky z pólů černé díry vyvrhující částice rychlostí blížící se rychlosti světla. O jejich původu a šíření se stále moc neví. Tyto jety jsou extrémně zářivé a mohou být dokonce i jasnější než akreční disk.



Obrázek 2.2: Eliptická galaxie M87 emituje jet, zachyceno Hubblovým teleskopem [9]

# Kapitola 3

# Detekce

## 3.1 Rádiová a rentgenová technika

Objevování kvazarů a AGN (aktivní galaktická jádra) rádiovou či rentgenovou technikou jsou jedny z nejpřímočařejších přístupů, protože v těchto vlnových délkách obyčejné hvězdy a galaxie vyzařují jen velice slabě. K nalezení vizuálního protějšku je třeba rozlišovací schopnosti přibližně $1''$. Úhlové rozlišení $\Theta$ teleskopu o průměru $D$ pozorující ve vlnových délkách $\lambda$ určíme následovně:

$$\Theta = 1.22\frac{\lambda}{D}, \tag{3.1}$$

kde $\Theta$ je v radiánech a $\lambda$ a $D$ v metrech. Například pro optický teleskop o průměru objektivu 1 metr bychom dostali úhlové rozlišení asi $0.1''$, oproti tomu pro radioteleskop (např. pro vlnové délky 5 cm) o průměru 65 metrů dostaneme rozlišení pouhých $192''$.

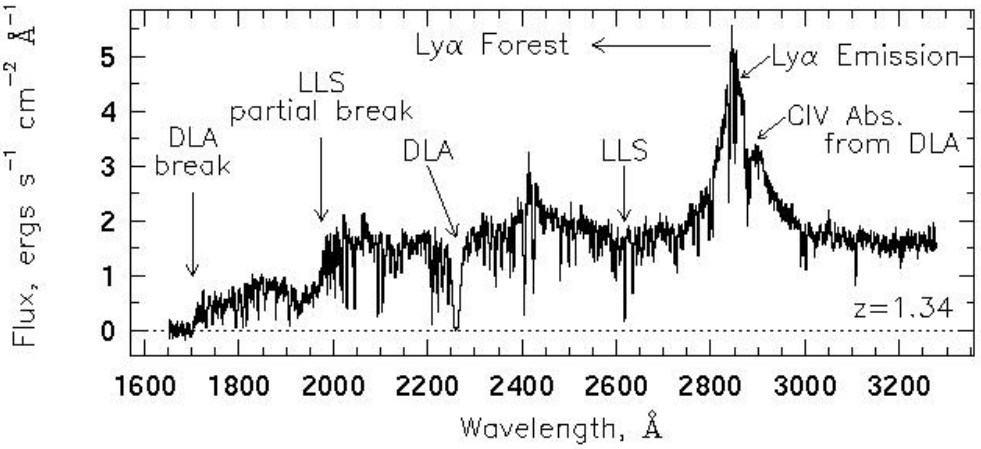
Je zřejmé, že typickými radioteleskopy nelze dosáhnout moc velké přesnosti. K vyřešení této obtíže začali radioastronomové používat větší množství radioteleskopů naskládaných vedle sebe. Této technice se říká interferometrie a je s ní možné dosáhnout rozlišení $0.001''$ nebo i větší. K určení úhlového rozlišení lze opět užít vztah 3.1 s tím rozdílem, že poloměrem teleskopu je nyní vzdálenost mezi nejvzdálenějšími teleskopy.

Například The Very Large Array (VLA) je složen z 27 radioteleskopů o půměru 25 metrů. Teleskopy jsou poskládány do tvaru „Y“.Všech 27 teleskopů vždy míří na stejný objekt a jednotlivá pozorování se poté sečtou dohromady.

Jak jsme si ukázali, úhlové rozlišení už v dnešní době není pro detekci rádiových zdrojů problém, nicméně bylo zjištěno, že pouhých 10 % kvazarů a AGN je „rádiově hlasitých“[10].

## 3.2 Spektrum

Spektroskopické pozorování je jeden z nejpřesvědčivějších důkazů identifikace kvazarů. Z posuvu spektrálních čar lze určit vzdálenost kvazarů. Ze vzdálenosti a pozorované jasnosti už lehce určíme zářivý výkon, jak jsme si již ukázali v kapitole 1.2. Vzhledem k tomu, že ve vesmíru zatím neznáme žádný jiný tak extrémně zářivý zdroj, můžeme jej označit za kvazar. Ze spektra však můžeme vyčíst mnohem více.




Obrázek 3.1: Typické spektrum kvazaru (PKS0454+039, $z = 1.34$). [11]

Spektrum kvazaru je tvořeno poměrně plochým zářením kontinua, které je protkáno absorbčními a emisními čárami. Široké emisní čáry jsou produkovány samotným kvazarem (poblíž černé díry a akrečního disku). Na druhou stranu, drtivá většina jeho absorbčních čar je způsobena plynem mezi kvazarem a Zemí.

## 3.3 UV exces

Jednou z historicky nejběžnějších metod rozeznávání kvazarů je podle UV excesu. Nejlépe prostudovaný vzorek jasných kvazarů, Bright Quasar Survey (součást Palomar-Green přehlídky [12]), rozlišoval kvazary podle hvězdné velikosti ($B < 16.2, M_V < -23$) a barevného indexu ($U - B < -0.4$). Toto kritérium dokáže se značnou jistotou určit, jaké objekty jsou kvazary, nicméně nedokáže identifikovat všechny. Přestože UV exces není typickým rysem kvazarů, který by se dal samostatně použít k jejich rozpoznání, lze jej použít k vytřídění velkého množství hvězd, které s nějvětší pravděpodobností kvazary nejsou.

## 3.4 Světelné křivky

### 3.4.1 Změny jasnosti

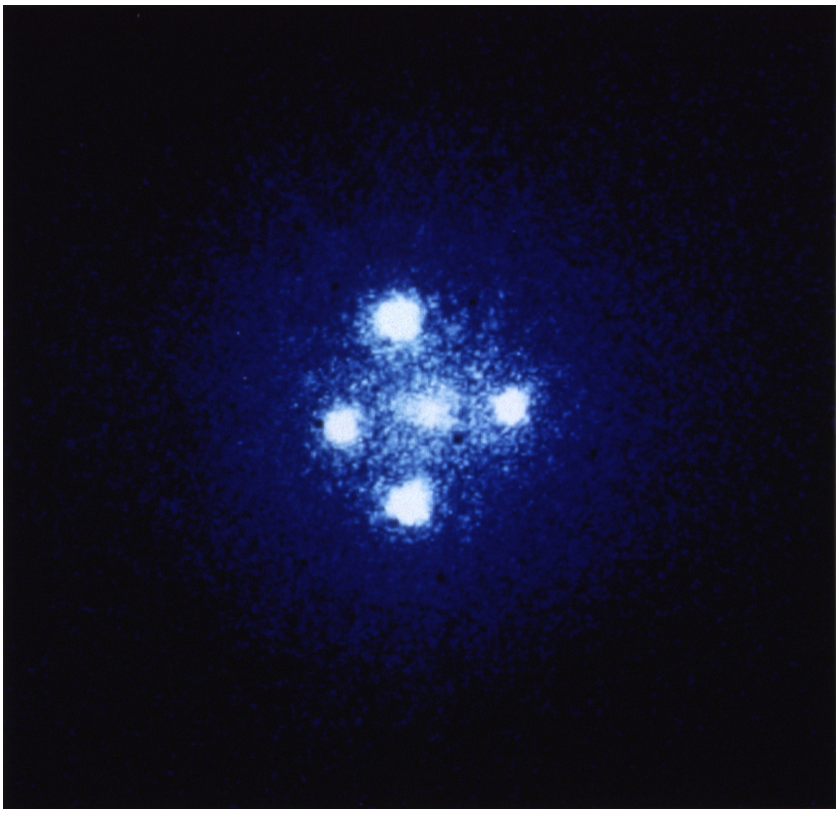
Změny jasnosti jsou pozorovány téměř u všech kvazarů, a to v průměru 10% až 15% jejich světelného toku. Bylo zjištěno, že proměnnost kvazaru je závislá jak na hmotnosti masivní černé díry v jeho středu, tak na efektivitě kvazaru při přeměně potenciální energie na světelnou energii. Změny jasnosti jsou pozorovány napříč všemi vlnovými délkami, jsou aperiodické a jejich amplituda je proměnlivá. [13]

### 3.4.2 Gravitační čočkování

Gravitační čočkování je jev, při kterém dojde k zjasnění zářivého zdroje z důvodu průchodu světla kolem objektu se silným gravitačním polem. Ve světelných křivkách potom pozo-

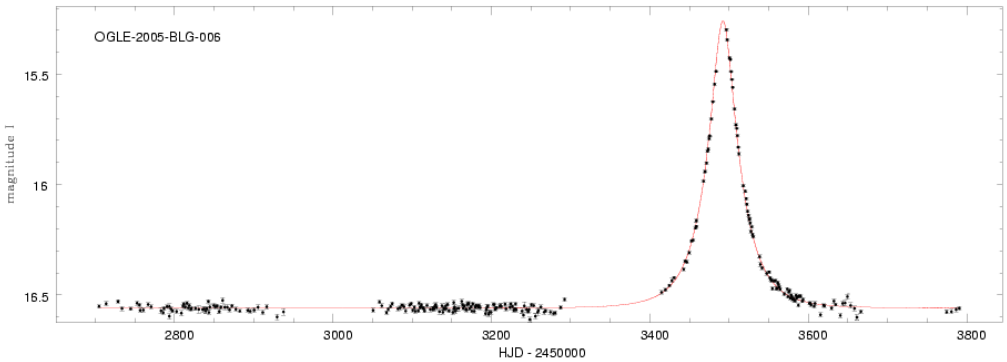
rujeme symetrické zjasňování a pohasínání. Tento jev byl již na počátku 20. století matematicky popsán Albertem Einstainem v jeho Obecné teorii relativity. Díky tomu jsme ze světelných křivek s gravitačním čočkováním schopni určit některé jeho parametry (např. vzdálenost).

Na velice dlouhé cestě, kterou záření kvazarů urazí, je poměrně značná pravděpodobnost, že se paprsky střetnou s hmotným gravitačním objektem (galaxie, kupy galaxií atd.), na kterém „zčočkují“. Jako příklad můžeme uvést jednu z nejznámějších gravitačních čoček pořízená Hubbleovým teleskopem – Einstainův kříž (obr. 3.2).



Obrázek 3.2: Einstainův kříž. Obraz vdáleného kvazaru QSO 2237+0305 je „zčočkován“ na nedaleké galaxii, proto jeden a ten samý kvazar vidíme čtyřikrát. [14]

Ne vždy mají parametry čočkování (hmotnost, vzdálenosti, poloměr složek atd.) ideální hodnoty, aby byl tento jev tak nápadný jako je tomu na snímku 3.2. Přesto jsme však schopni tento úkaz detekovat jako zvýšení jasnosti, kdy maximum světelného toku odpovídá situaci nejvhodnějšího uspořádání systému (zdroj – gravitační čočka – pozorovatel), což za ideálních podmínek nastane, když jsou všechny komponenty čočkování v jedné přímce. Takový průběh světelné křivky je jedinečný a nelze jej pozorovat při žádném jiném ději ve vesmíru.




Obrázek 3.3: Gravitační mikročočkování ve světelné křivce. [15]

# Kapitola 4

# Klasifikace světelných křivek

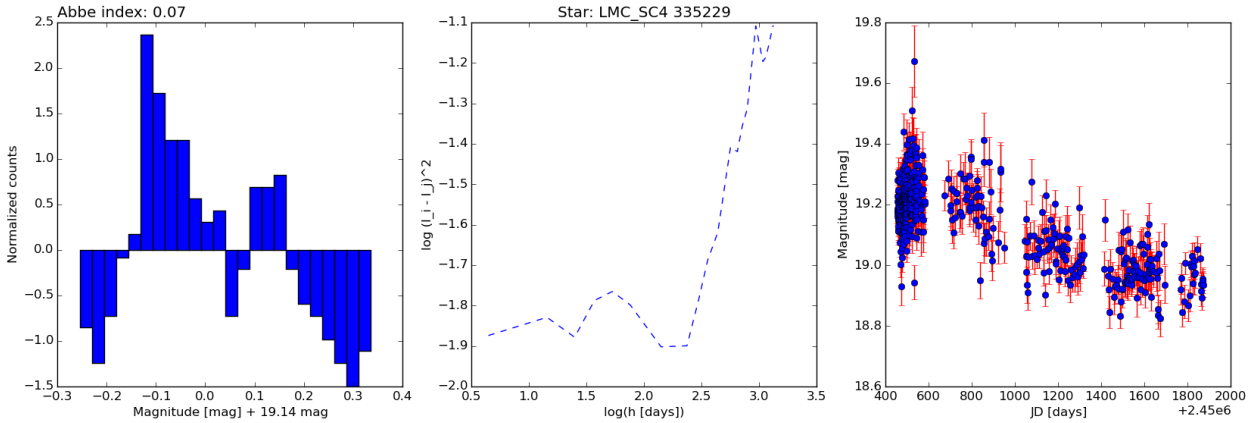
## 4.1 Zpracování dat

Jedním ze základních rysů astronomických objektů je světelná křivka, která popisuje, jak se mění světelné toky v čase. Jejich tvar lze pro různé třídy hvězd klasifikovat a tak blíže pochopit, jaké fyzikální děje odpovídají za pozorované změny jasnosti.

Za posledních pár desítek let bylo pořízeno obrovské množství astronomických dat, které už prakticky nelze zpracovávat jinak něž automaticky s pomocí počítačů. Výstupem pozorování jsou astronomické snímky, které jsou pro vybrané hvězdy zpracovány na světelné křivky. Tím se myslí, že se pro jednotlivé hvězdy spočítá množství světla dopadajícího na čip pro jednotlivé snímky v daných časech. Další práce s takto získanými časovými řadami se významně usnadní, ale přesto je zpracovávání nesčetného množství světelných křivek stále poměrně výpočetně náročné. Naštěstí byly vyvinuty efektivní a důmyslné Data Miningové metody, které ke zpracování dat přistupují z jiného úhlu, čímž se zpracování může mnohonásobně zefektivnit, a dokonce přinést přesnější výsledky, či úplně nové.

## 4.2 Statistické informace

Ačkoliv tvar světelné křivky je pro mnoho astronomických objektů charakteristický (typicky pro proměnné hvědy), u kvazarů pozorujeme různorodé tvary světelných křivek, kterými nelze kompletně reprezentovat celou tuto skupinu. Naštěstí lze ze světelných křivek vyčíst mnohem více, než by se na první pohled mohlo zdát.




Obrázek 4.1: Některé charakteristiky světelné křivky náhodně vybraného kvazaru: Abbe hodnota, histogram, variogram a světelná křivka

### 4.2.1 Variogram

Variogram popisuje, jak moc se mění měřená veličina na různých škálách. Ve světelné křivce nám reprezentuje, jak moc se mění jasnost hvězdy v různě dlouhých časových intervalech. Pro světelnou křivku $I(t)$ spočítáme všechny rozdíly časů $\Delta t(ij) = t_i - t_j$, kde $h$ je časový krok, a všechny páry kvadrátu rozdílů hvězdných velikostí $\Delta I(ij) = \left(I_i - I_j\right)^2$. Variogramem označíme funkci $\Delta I(\Delta t)$.

Tato metoda je pro klasifikaci mnoha kvazarů velice efektivní, protože jejich proměnnost mívá často dlouhodobý charakter (viz 4.1). Například Eyer [16] ve své práci považoval za jedno z kritérií směrnici variogramu, u které vzal v potaz jen křivky s větší směrnicí variogramu. Tato metoda je však vhodná pro rozpoznávání trendů jen v rámci stovek dnů. Níže si ukážeme, že srovnáváním celých tvarů variogramů můžeme dosáhnout pozoruhodných výsledků i pro kratší periody.

### 4.2.2 Histogram

Histogram nám dává informaci o statistickém rozdělení naměřených dat. O světelných křivkách vypovídá, s jakou četností jsou zastoupeny jednotlivé jasnosti (respektive jejich intervaly). Rozpoznávání hvězd pomocí jejich histogramů může být pro určité typy hvězd efektivní (například náhlá zjasnění u Be hvězd). Pro lepší zpracování budeme pracovat s normalizovaným histogramem tak, aby průměrná jasnost (hvězdná velikost) byla posunuta k nule a směrodatná odchylka rovna jedné.

### 4.2.3 Abbe hodnota

Mějme pozorování $n$ po sobě jdoucích pozorování $x_1, \ldots, x_n$ splňující normální rozdělení. Pro náš vzorek definujeme průměrnou hodnotu

$$\bar{x} = \frac{1}{n}\sum_{\mu=1}^{n} x_\mu, \tag{4.1}$$

varianci (rozptyl)

$$s^2 = \frac{1}{n}\sum_{\mu=1}^{n} \left(x_\mu - \bar{x}\right)^2, \tag{4.2}$$

a také průměrný kvadrát rozdílu dvou po sobě jdoucích pozorování

$$\delta^2 = \frac{1}{n-1}\sum_{\mu=1}^{n-1} \left(x_{\mu+1} - x_\mu\right)^2. \tag{4.3}$$

Ukazuje se, že poměr

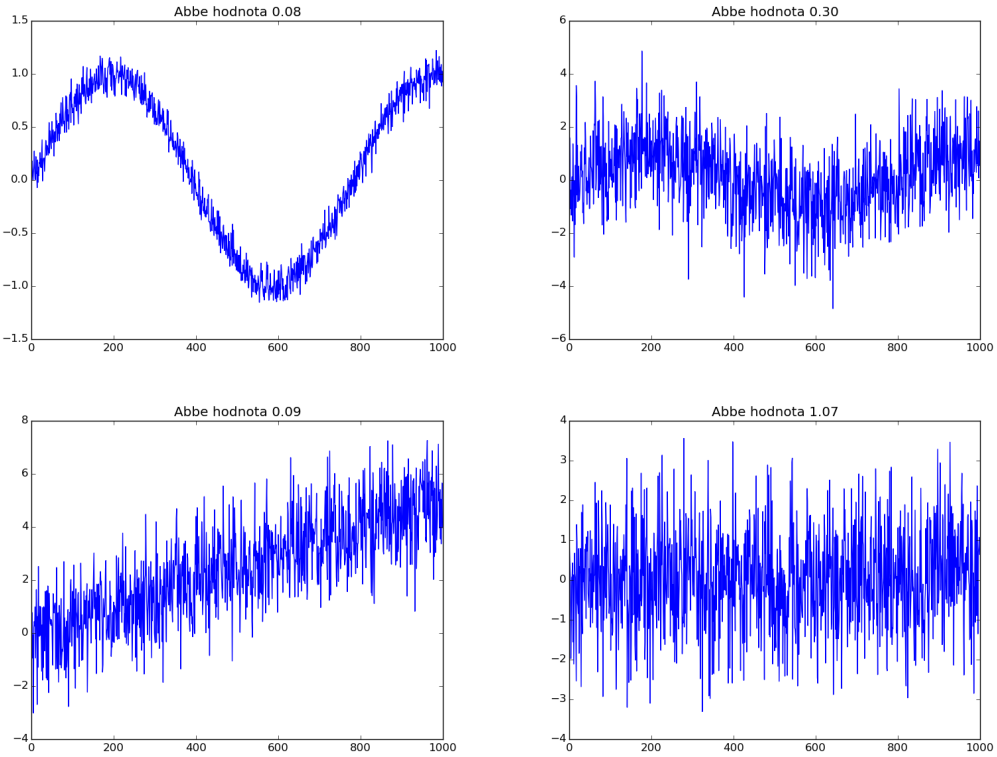
$$\eta = \frac{\delta^2}{s^2} \tag{4.4}$$

je vhodné kritérium k určení, zda je rozdělení vzorku $x_i, \ldots, x_n$ náhodné nebo zda existuje nějaký trend. [17]

Poměr $\mathscr{A} = \frac{\eta}{2}$ se označuje jako Abbe hodnota [18]. Dosazením rovnic 4.2 a 4.3 do 4.4 získáme předpis pro Abbe hodnotu:

$$\mathscr{A}(x) = \frac{n}{2(n-1)} \frac{\sum_{\mu=1}^{n-1}(x_{\mu+1} - x_\mu)^2}{\sum_{\mu=1}^{n}(x_\mu - \bar{x})^2}. \tag{4.5}$$

Konstantní křivky mají Abbe hodnotu blízkou jedné, oproti tomu hladké měnící se vzory mají Abbe hodnotu blížící se nule.




Obrázek 4.2: Abbe hodnoty vybraných funkcí s Gaussovským šumem

## 4.3 SAX

Výše bylo zmíněno, že je možné srovnávat histogramy, variogramy či vlastní tvary světelných křivek, ale přichází otázka „Jak kvantifikovat míru podobnosti dvou křivek?“. Nejpřímočařejší metodou je použití euklidovské metriky, sečíst vzdálenosti jednotlivých odpovídajících si bodů obou křivek převedených do normalizovaného tvaru. To se však neukazuje být tím nejvhodnějším přístupem.

Symbolic Aggregate approXimation (SAX) je jedna z Data Miningových metod, která převádí numerická data na symbolický řetězec – na slovo. Hledání dvou podobných tvarů datových řad (světelné křivky, histogramu, variogramu atd.) potom spočívá v porovnávání dvou slov v námi definovaném vzdálenostním prostoru.

Můžeme si například představit automatickou opravu při psaní na počítači, která při překlepu navrhne, jaké slovo jsme pravděpodobně chtěli napsat. Vše co je potřeba, je definovat metriku, která bude určovat vzdálenost mezi písmeny a bude nejspíše souviset s fyzickou vzdáleností kláves na klávesnici. Vyhledávání může vypadat následovně: Máme objekt, který je v symbolickém prostoru reprezentován slovem „FNRK“. Dále máme vzorek objektů, které jsou reprezentovány těmito slovy: „SMRK“ a „POLE“. Máme-li definovanou klávesnicovou metriku tak, že vzdálenost dvou písmen je rovna počtu kláves mezi písmeny, potom můžeme určit vzdálenost dvou slov jako součet vzdáleností jednotlivých písmen.

| první slovo | druhé slovo | počet kláves mezi písmeny |
|---|---|---|
| F | S | 1 |
| N | M | 0 |
| R | R | 0 |
| K | K | 0 |
| Součet vzdáleností | | 1 |

Tabulka 4.1: Vzdálenost dvou blízkých slov v klávesnicové metrice

| první slovo | druhé slovo | počet kláves mezi písmeny |
|---|---|---|
| F | P | 5 |
| N | O | 1 |
| R | L | 5 |
| K | E | 5 |
| Součet vzdáleností | | 16 |

Tabulka 4.2: Vzdálenost dvou vzdálených slov v klávesnicové metrice

Výsledné vzdálenosti mezi slovy jsou $d_{FNRK,SMRK} = 1$ a $d_{FNRK,POLE} = 16$. Pokud jsme napsali slovo „FNRK“ a ve slovníku možných slov byly jen možnosti „SMRK“ a „POLE“, metoda by nám doporučila jako slovo, které jsme nejspíše mysleli, „SMRK“.

Princip pro srovnávání křivek hvězd bude velice podobný, jen budeme muset určit, jak je převedeme na slovo a vhodně si definovat metriku, která nám bude říkat, jak moc jsou si dva objekty podobné.

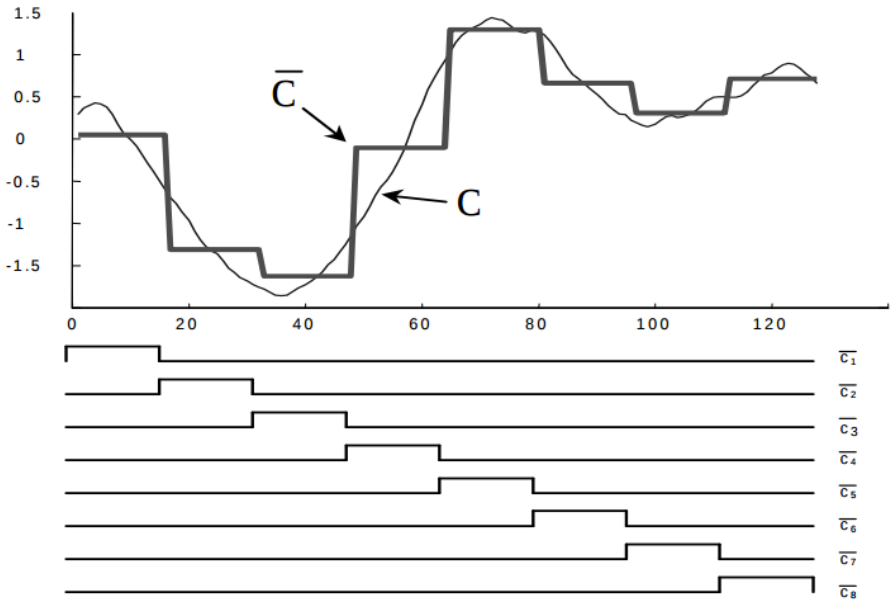
### 4.3.1 Redukce rozměrů – PAA

Než převedeme křivku na řetězec písmen, bude vhodné nejprve snížit její rozměr a rozdělit ji na menší počet úseků. Časová řada $C$ o délce $n$ v $w$–dimensionálním prostoru může být reprezentována jako vektor $\bar{C} = \bar{c}_1, ..., \bar{c}_w$, kde pro prvek $i$ v $\bar{C}$ platí:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}j} c_j. \tag{4.6}$$

.

PAA reprezentace může být vizualizovaná jako aproximace původní časové řady lineární kombinací bázových funkcí, jak je ukázáno na obrázku 4.3 .




Obrázek 4.3: Převedení původní časové řady $C$ na $\bar{C}$ o redukované dimenzi [19]

Nakonec časovou řadu normalizujeme, aby průměrná hodnota byla 0 a směrodatná odchylka byla 1.

### 4.3.2 Diskretizace

Nyní už máme vše co potřebujeme k diskretizaci našich dat na symbolický řetězec. Pro naše data můžeme předpokládat gaussovské rozdělení (světelné křivky tento předpoklad docela dobře splňují [19]), a potom už bude jednoduché stanovit hranice, které budou produkovat $a$ stejně velkých oblastí pod Gaussovou křivkou.

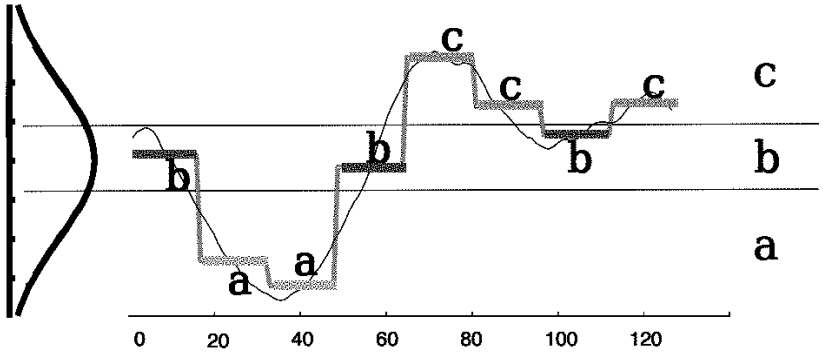
**Definice 1** *Zlomové body*: Jsou prvky setříděného listu čísel $B = \beta_1, ..., \beta_{a-1}$, které ohraničují oblasti, jejichž pravděpodobnost obsazení je pro data s Gaussovským rozdělením stejná (viz. obrázek 4.5)

Tyto hodnoty lze určit ze statistické tabulky, jako například pro hodnoty $a$ od 3 do 10 ukazuje tabulka 4.4.

| $\beta_i$ \ $a$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $\beta_1$ | -0.43 | -0.67 | -0.84 | -0.97 | -1.07 | -1.15 | -1.22 | -1.28 |
| $\beta_2$ | 0.43 | 0 | -0.25 | -0.43 | -0.57 | -0.67 | -0.76 | -0.84 |
| $\beta_3$ | | 0.67 | 0.25 | 0 | -0.18 | -0.32 | -0.43 | -0.52 |
| $\beta_4$ | | | 0.84 | 0.43 | 0.18 | 0 | -0.14 | -0.25 |
| $\beta_5$ | | | | 0.97 | 0.57 | 0.32 | 0.14 | 0 |
| $\beta_6$ | | | | | 1.07 | 0.67 | 0.43 | 0.25 |
| $\beta_7$ | | | | | | 1.15 | 0.76 | 0.52 |
| $\beta_8$ | | | | | | | 1.22 | 0.84 |
| $\beta_9$ | | | | | | | | 1.28 |

Obrázek 4.4: Tabulka obsahující zlomové body, které rozčleňují Gaussovské rozdělení na oblasti se stejnou pravděpodobností (pro $a$ od 3 do 10) [19]

Pomocí této tabulky už můžeme diskretizovat naše normalizovaná a redukovaná data následujícím způsobem: Všem hodnotám menším než první zlomový bod přiřadíme symbol „a“, všem hodnotám větším než první hraniční bod nebo stejný a zároveň menším než druhý zlomový bod přiřadíme symbol „b“ atd. Tento postup ilustruje obrázek 4.5.




Obrázek 4.5: Časová řada je pomocí předdefinovaných zlomových bodů převedena do SAX symbolů. V tomto příkladě, s $n = 128$, $w = 8$ a $a = 3$, je časová řada převedena na slovo **baabccbc** [19]

Všimněme si, že v tomto příkladě je zastoupení symbolů „**a**“ „**b**“ a „**c**“, přibližně stejné (jak bylo požadováno).

**Definice 2** *Slovo*: Sekvence $C$ o délce $n$ může být reprezentovaná jako slovo $\hat{C} = \hat{c}$. Nechť $\alpha_i$ označuje $i$-té písmeno v abecedě ($\alpha_1 = $ a, $\alpha_2 = $ b atd.). Potom převod z PAA reprezentace $\bar{C}$ (tj. normalizovaná časová řada o redukované dimenzi) na slovo $\hat{C}$ provedeme následovně:

$$\beta_{j-1} \leq \bar{c}_i < \beta_j \Rightarrow \hat{c}_i = \alpha_j. \tag{4.7}$$

### 4.3.3 Měření vzdáleností

Nyní, když už máme definovanou naši symbolickou reprezentaci, můžeme definovat měření vzdáleností v naší symbolické metrice.

Jedna z nejjednodušších metod měření vzdáleností $D$ dvou časových řad $Q$ a $C$ o stejné délce je užité Euklidovské vzdálenosti, jak definuje vztah 4.8:

$$D(Q,C) \equiv \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2}. \tag{4.8}$$

V případě redukovaných časových řad (pomocí PAA) $\bar{Q}$ a $\bar{C}$ definujeme jejich vzdálenosti následovně:

$$D_{PAA}(\bar{Q},\bar{C}) \equiv \sqrt{\frac{n}{w}} \cdot \sqrt{\sum_{i=1}^{w} (\bar{q}_i - \bar{c}_i)^2}. \tag{4.9}$$

Pokud dále data transformujeme do symbolické reprezentace, můžeme definovat funkci, které vrací vzdálenost dvou slov:

$$D_{sax}(\hat{Q},\hat{C}) \equiv \sqrt{\frac{n}{w}} \cdot \sqrt{\sum_{i=1}^{w} (dist\,(\hat{q}_i,\hat{c}_i))^2}. \tag{4.10}$$

Oproti předchozí funkci, se 4.10 liší rozdílným určováním vzdáleností jednotlivých prvků časové řady. Byla zavedena funkce $dist()$, která může být implementována pomocí tabulky vzájemných vzdáleností písmen, jak ukazuje obrázek 4.6.

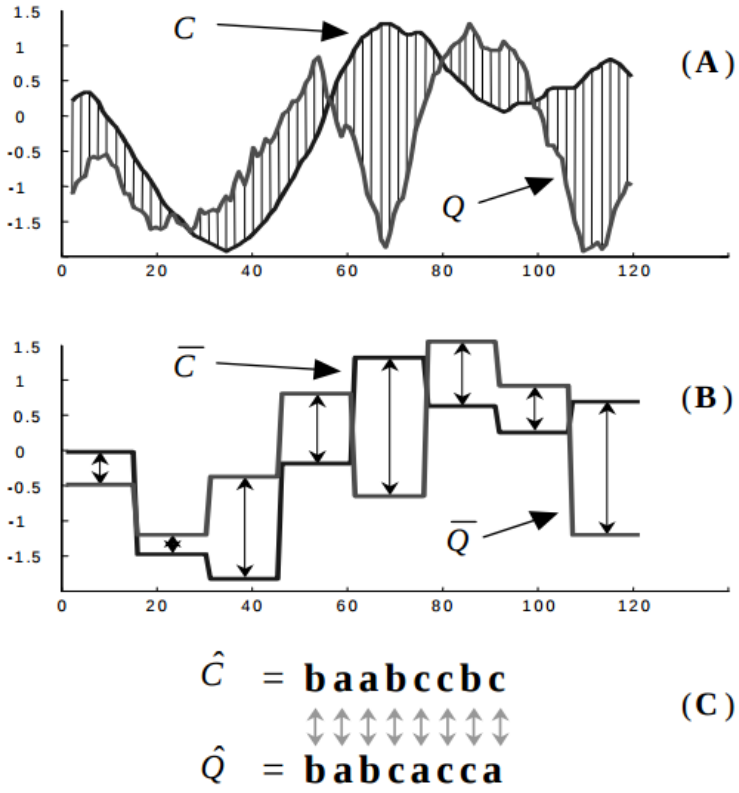
| | **a** | **b** | **c** | **d** |
|---|---|---|---|---|
| **a** | 0 | 0 | 0.67 | 1.34 |
| **b** | 0 | 0 | 0 | 0.67 |
| **c** | 0.67 | 0 | 0 | 0 |
| **d** | 1.34 | 0.67 | 0 | 0 |

Obrázek 4.6: Tabulka vzájemných vzdáleností písmen v symbolické metrice [19]

Hodnoty v jednotlivých buňkách $(r,c)$ mohou být spočítány pomocí následujícího výrazu:

$$bunka_{r,c} = \begin{cases} 0, & \text{pokud } |r-c| \leq 1 \\ \beta_{max(r,c)-1} - \beta_{min(r,c)}, & \text{v ostatních případech} \end{cases} \tag{4.11}$$

Pro danou délku abecedy je třeba tabulku spočítat jen jednou.




Obrázek 4.7: Vizualizace metod určení vzdáleností dvou časových řad popsaných výše [19]

## 4.4 Data

### 4.4.1 OGLE databáze

K získávání světelných křivek bylo převážně využito OGLE II databáze [20]. Přehlídka OGLE (Optical Gravitational Lensing Experiment) je polský projekt, který se už od roku 1992 zabývá hledáním hmotných objektů na okraji naší Galaxie pomocí gravitačního čočkování (viz kapitola 3.4.2). Tyto objekty by mohly vysvětlit existenci temné hmoty. Hlavními pozorovanými oblastmi jsou galaktická výduť a Magellanova mračna.

### 4.4.2 Světelné křivky

K trénování metod a k zhodnocení jejich výsledků je nezbytné mít k dispozici světelné křivky potvrzených kvazarů. K jejich nalezení bylo využito databáze MQS (The Magellanic Quasars Survey) [21], která obsahuje přes 800 spektroskopicky potvrzených kvazarů. Většina těchto dat je uložena v OGLE III databázi, kde však nejsou zpřístupněna (jen vybrané typy identifikovaných proměnných hvězd). Nicméně některé z nich byly napozorovány i v OGLE II, a proto k nim bylo možné nalézt světelné křivky (okolo 40). Jako další

data byly použity kvazary z MACHO databáze [22], které byly identifikovány a následně spektroskopicky potvrzeny týmem Geha et al. (2003) [23].

K určení přesnosti filtrování byla stažena data neproměnných hvězd, Be hvězdy, hvězdy typu RR Lyrae, cefeidy, dlouhoperiodické proměnné hvězdy, hvězdy s dvojitou periodou a kandidáti na kvazary podle článku Eyera. Identifikované proměnné hvězdy byly staženy z OGLE III databáze [24], kde jsem vybral jen ty, které se nacházely i v OGLE II a díky tomu k nim bylo možné získat barevné indexy [20].

## 4.5 Určování parametrů

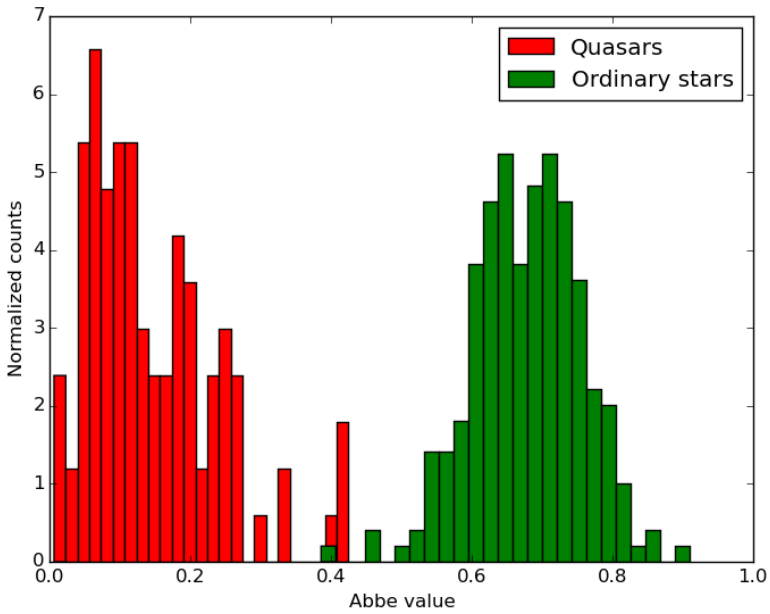
### 4.5.1 Abbe kritérium

Pro efektvitu filtrování je vhodné vytřídit z našeho vzorku co nejvíce neproměnných hvězd pomocí výpočetně nepříliš náročné metody. K tomu nám může poměrně dobře posloužit Abbe hodnota $\mathscr{A}$, kterou jsme si popsali v kapitole 4.2.3. Pro jednotlivé třídy objektů si nejdříve znázorníme rozložení Abbe hodnot.

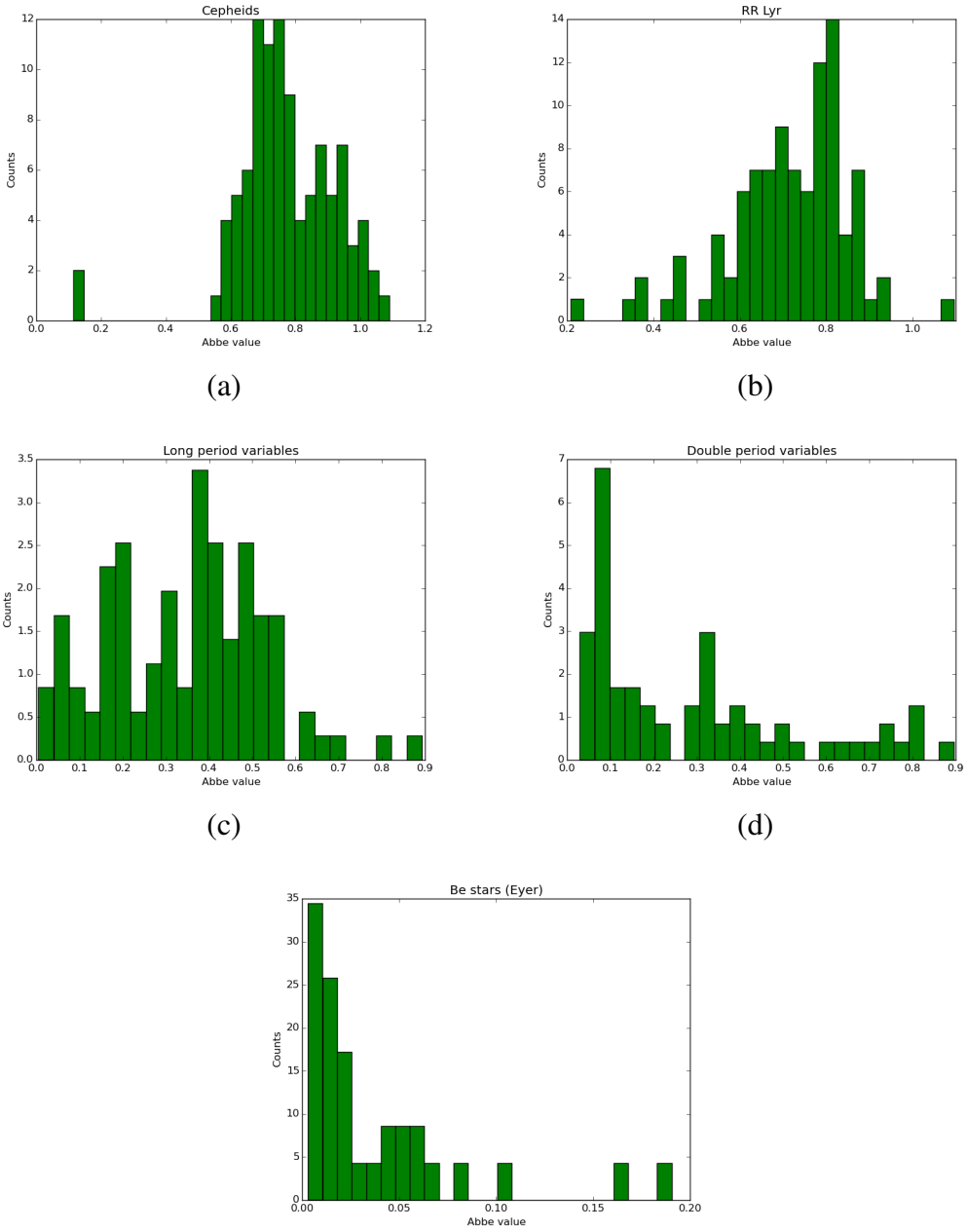
| | $\overline{\mathscr{A}}$ |
|---|---|
| Be-hvězdy (Eyer): | $0.04 \pm 0.04$ |
| Kvazary (MACHO): | $0.14 \pm 0.10$ |
| Kvazary (Eyer): | $0.14 \pm 0.13$ |
| Kvazary (OGLEII): | $0.15 \pm 0.09$ |
| Hvězdy s dvojitou periodou: | $0.29 \pm 0.24$ |
| Dlouho periodické hvězdy: | $0.34 \pm 0.18$ |
| Neproměnné hvězdy: | $0.68 \pm 0.08$ |
| RR Lyrae: | $0.71 \pm 0.14$ |
| Cefeidy: | $0.77 \pm 0.15$ |

Tabulka 4.3: Průměrné Abbe hodnoty pro vybrané typy objektů

Nejmenší Abbe hodnoty mají Be hvězdy a hned po nich kvazary. Podle očekávání na druhém konci tabulky jsou neproměnné hvězdy. Nicméně na první pohled udivující se mohou zdát být poslední dvě políčka – hvězdy typu RR Lyrae a cefeidy, které jsou periodické proměnné hvězdy (čili by měly mít nízkou Abbe hodnotu), ale jejich periody jsou v řádech desítek hodin až maximálně pár desítek dní. Zpracovávané křivky mají však délky typicky 2000 dnů (tj. téměr 6 let) a po sobě jdoucí měření jsou od sebe vzdálena v jednotkách dnů. Z tohoto důvodu není možné detekovat tak krátké trendy, jevící se ve výsledku jako šum.




Obrázek 4.8: Rozdělení četnosti Abbe hodnot pro kvazary a neproměnné hvězdy

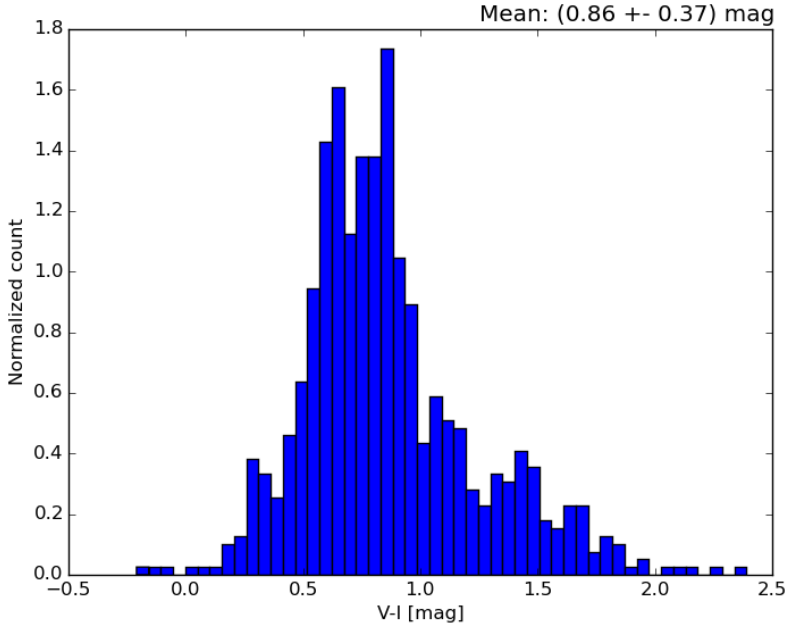



Obrázek 4.9: Rozložení Abbe hodnot pro vybrané typy objektů: (a) Cefeidy, (b) hvězdy typu RR Lyr, (c) dlouhoperiodické proměnné hvězdy, (d) hvězdy s dvojitou periodou a (e) Be hvězdy

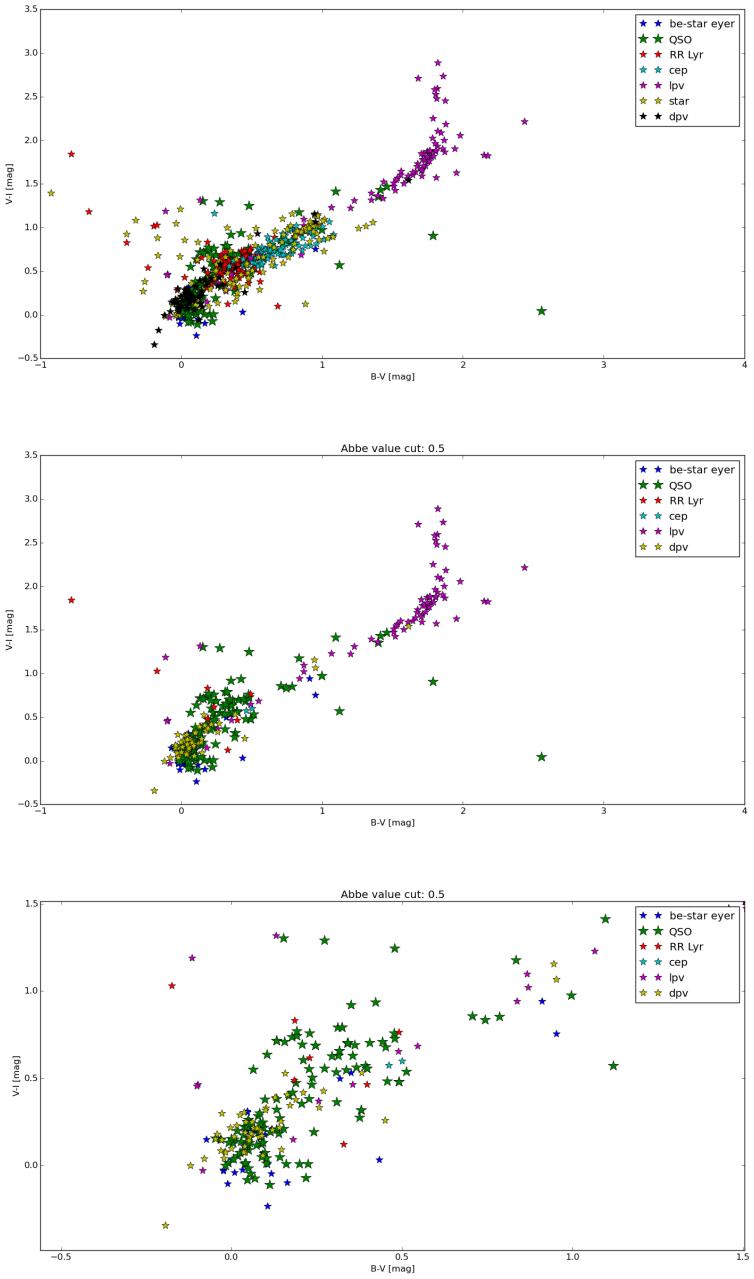
Je patrné, že zvolíme-li vhodně limitní velikost Abbe hodnoty, lze hned na začátku třídícího procesu vyloučit velké množství nežádoucích objektů. Za kritérium vybereme nejvyšší Abbe hodnotu ze vzorku potvrzených kvazarů – $\mathscr{A} = 0.5$.

### 4.5.2 Barevné indexy

V OGLE II databázi jsou kromě světelných křivek dostupné i barevné indexy pro většinu hvězd ve filtrech $B$, $V$ a $I$. V MQS katalogu potvrzených kvazarů jsou uvedeny i barevné indexy $V - I$, jejichž distribuci (pro cca 800 kvazarů) znázorňuje histogram 4.10.




Obrázek 4.10: Distribuce V-I hodnot pro spektroskopicky potvrzené kvazary z MQS databáze




Obrázek 4.11: Barevné indexy tří vybraných hvězd: Be hvězdy, kvazary, hvězdy typu RR Lyrae, cefeidy, dlouhoperidické hvězdy, neproměnné hvězdy a hvězdy s dvojitou periodou. Na druhém obrázku jsou vyjmuty hvězdy s Abbe indexem větším než 0.5 a v posledním obrázku je přiblížení do husté oblasti ve středu.

Jedna z možností, jak vyčlenit některé hvězdy, by byla vzít v potaz jen skupinu hvězd okolo $[B-V=0.2, V-I=0.5]$, kde se nejvíce koncentrují kvazary. Například Eyer ve své práci vyloučil všechny hvězdy s $V-I>0.9$. Tímto bychom však nesprávně vyloučili, jak ukazuje i histogram 4.10, spoustu kvazarů, proto s důvěrou v další filtrovací metody vyloučíme až hvězdy s $B-V>1.5$ a $V-I>1.5$.

| | Počet objektů prošlých jednotlivými filtry [%] | |
|---|---|---|
| | Abbe kritérium | barevný index |
| Kvazary (MACHO) | 100 | 100 |
| Kvazary (OGLEII) | 100 | 100 |
| Kvazary (Eyer) | 96 | 94 |
| Be-hvězdy (Eyer) | 100 | 100 |
| Hvězdy s dvojitou periodou | 82 | 80 |
| Dlouho periodické hvězdy | 82 | 20 |
| Cefeidy | 2 | 2 |
| RR Lyr | 8 | 7 |
| Neproměnné hvězdy | 1 | 0 |

Tabulka 4.4: Počet objektů splňujících jednotlivá kritéria. V prvním sloupci je počet objektů s Abbe hodnotou menší než 0.5, v druhém sloupci počet hvězd, které zároveň splňují podmínku $B-V < 1.5$ mag a $V-I < 1.5$ mag.

### 4.5.3 Symbolický vzdálenostní prostor

#### SAX parametry

Primární metodou k identifikaci kvazarů bude srovnávání variogramů a histogramů vyšetřovaných objektů se známými kvazary pomocí jejich symbolické reprezentace, kterou jsme si blíže popsali v kapitole 4.3. Nejprve ze světelných křivek určíme histogramy a variogramy, které vzápětí převedeme na „slova“. V našem prostoru bude každý hvězdný objekt reprezentován dvojicí slov – histogramové a variogramové slovo.

Než se dostaneme k samotnému srovnávání křivek, je třeba určit 6 volných parametrů: Délka abecedy $a$, poměr počtu písmen ku délce časové řady $r$ a mezní vzdálenost $d_{lim}$ (3 pro histogramové slovo + 3 pro variogramové slovo). Pro urychlení výpočtu byl počet parametrů snížen na 5 zavedením substituce $d_{lim} = d_{lim,hist} + d_{lim,vario}$. Jednotlivé mezní vzdálenostni budeme detailněji určovat v další kapitole. K trénování metody byla využita pythonovská knihovna Grid Search [25], která v celém prostoru parametrů hledá tu nejoptimálnější kombinaci parametrů. Funkci optimálnosti definujeme následovně:

$$f(a_{hist}, r_{hist}, a_{vario}, r_{vario}, d_{lim}) = \frac{n_x}{N_x} - \frac{n_y}{N_y}, \tag{4.12}$$

kde $N_x$ a $N_y$ je celkový počet kvazarů a nekvazarů, $n_x$ a $n_y$ je počet hvězd prošlých filtrováním pro dané parametry. Optimalizace potom spočívá v nalezení takových parametrů, pro které má funkce maximum. Tahle operace je výpočetně velice náročná a na osobním počítači ji prakticky nelze v požadovaném rozsahu provést (doba výpočtu několik týdnů, až měsíců). Pro tyto účely jsem využil výpočetních prostředků, které nabízí superpočítače MetaCentra [26]. Pro první běh s velkými kroky byl rozpyl parametrů široký tak, aby zahrnoval všechny smysluplné hodnoty parametrů. Za cíl měl jen přibližně určit do kterých oblastí jednotlivé hodnoty konvergují. Knihovna GridSearch umožňuje běh více procesů naráz, čímž se doba výpoču velmi zkrátila.

| | od | do | délka kroku |
|---|---|---|---|
| délka abecedy histogramu | 5 | 19 | 3 |
| délka abecedy variogramu | 5 | 19 | 3 |
| počet dnů na jedno písmeno histogramu | 5 | 120 | 30 |
| počet dnů na jedno písmeno variogramu | 5 | 120 | 30 |
| součet mezních vzdáleností | 2 | 40 | 9 |

Tabulka 4.5: První běh přes široký rozptyl parametrů s dlouhými kroky. Výpočet na 3 počítačích s 15 procesory a na každém byla využita vyrovnávací paměť 1 GB během 15 paralelních úloh s dobou výpočtu 13h

Další dva výpočty byly provedeny již s kratšími kroky v oblastech získaných předchozími výpočty. Mějme pro parametr $\alpha$ hodnotu $x_\alpha$ určenou z předchozího výpočtu s krokem $\delta x$, potom interval hodnot $\Delta$ pro další vyhledávání bude následující:

$$\Delta = (x - \delta x; x + \delta x). \tag{4.13}$$

Poslední výpočet proběhl na 3 počítačích s 3 procesory a na každém bylo využito 800 MB RAM během 12 paralelních úloh s dobou výpočtu téměř 35 h. Výsledky jsou uvedeny v tabulce 4.6,

| | délka abecedy | počet dnů na jedno písmeno |
|---|---|---|
| histogram | 7 | 97.5 |
| variogram | 17 | 9 |

Tabulka 4.6: Optimální parametry pro filtrování získané prohledáváním všech kombinací parametrů ve zvolených intervalech. Orientační hodnota mezní vzdálenosti $d_{lim} = 11.5$.

### Hranice

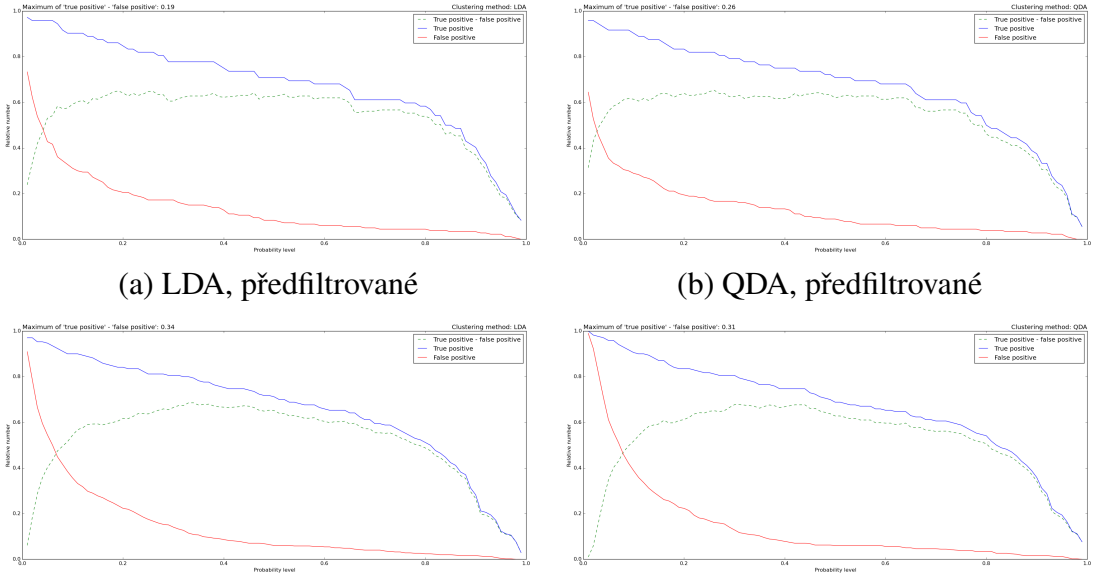
Nyní, když jsme si určili délky slov a abeced pro histogramové i variogramové slovo, se můžeme blíže zabývat určením kritéria podobnosti dvou světelných křivek. Jinými slovy jak moc si musí být dva objekty podobné, abychom je označili za objekty stejného typu. Oproti předchozím kritériím můžeme naší porovnávací metodou určit jen relativní vzdálenosti mezi dvěma objekty (jak moc jsou dva objekty rozdílné). Absolutní souřadnice vyšetřované hvězdy v histogram-variogramovém prostoru určíme následovně:

- Vytvoříme šablonu
  - z $n$ objektů jejichž typ budeme v datech hledat (v našem případě kvazary) vytvoříme šablonovou množinu tak, že vybereme takové světelné křiky, o kterých jsme si jisti, že reprezentují vybraný typ objektů
- Připravíme objekty
  - převedeme světelné křivky šablonových hvězd i vyšetřované hvězdy na histogramová a variogramová slova

- Porovnáme hvězdu s šablonou
  - porovnáme dvojici slov vyšetřované hvězdy se všemi dvojicemi slov v šabloně
  - výsledné vzdálenosti reprezentuje vektor $\hat{d} = ([d_{hist,1}, d_{vario,1}], .., [d_{hist,n}, d_{vario,n}])$
- Najdeme nejkratší vzdálenost
  - hledáme takový prvek $i$ vektoru $\hat{d}$ pro který bude $d_{hist,i} + d_{vario,i}$ minimální.
- Označíme souřadnice vyšetřované hvězdy jako $[d_{hist,i}, d_{vario,i}]$

Kvazary si rozdělíme na trénovací a testovací množinu. Trénovací vzorek bude sloužit ke srovnávání a identifikaci dat. Testovací množinu zařadíme zpět k ostatním objektům a budeme hledat takovou hranici v našem prostoru, která bude vymezovat oblast s co největším počet kvazarů a s nejmenším početem nekvazarů.

K určení hranice využijeme clusterovací metody, které „naučíme“ rozpoznávat kvazary. Ze všech zkoušených metod jsou LDA [27] a QDA [28] metody nejúspěšnější, a proto dále budeme testovat právě tyto dvě, a to jak na předfiltrovaných datech, tak na nepředfiltrovaných. Zadáním souřadnic kvazarů a nekvazarů v histogram-variogramovém prostoru nám naše metody vypočítají pole, které reprezentuje pravděpodobnost s jakou je objekt v daném bodě kvazar (viz. obrázek 4.13). Určení hranice nyní spočívá v nalezení takové pravděpodobnosti, jejíž vrstevnice (křivka, která spojuje body se stejnou hodnotou pravděpodobností) ohraničuje oblast s nejvíce kvazary a nejméně nekvazary. Pro všechny možné hodnoty pravděpodobností určíme relativní hodnoty správně identifikovaných, nesprávně identifikovaných objektů a funkci jejich rozdílu.




(a) LDA, předfiltrované

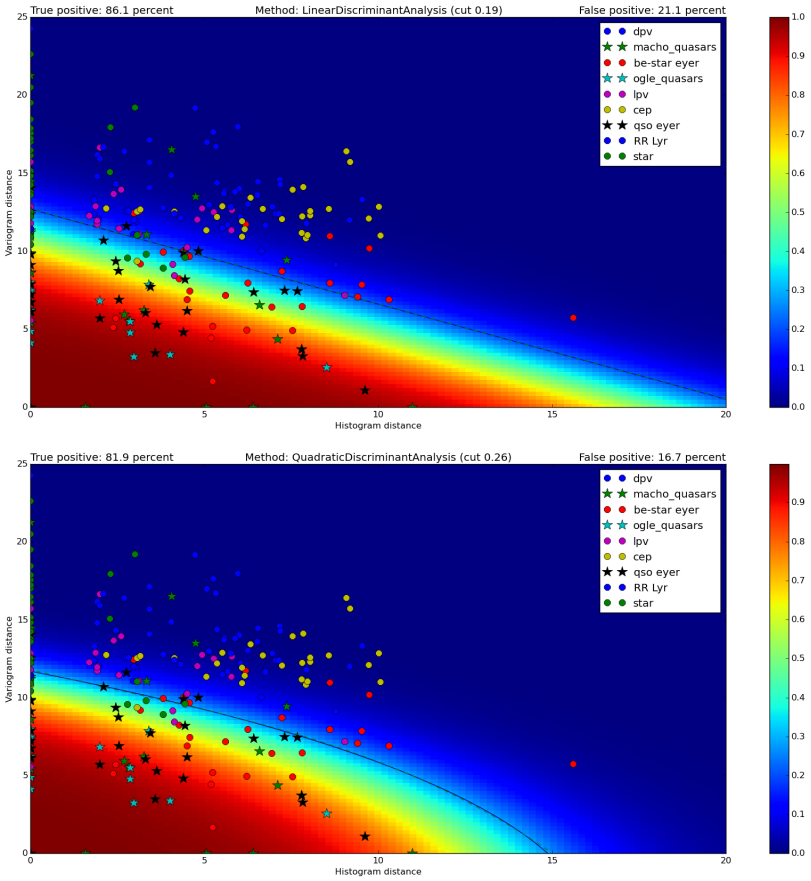
(b) QDA, předfiltrované

(c) LDA, nepředfiltrované

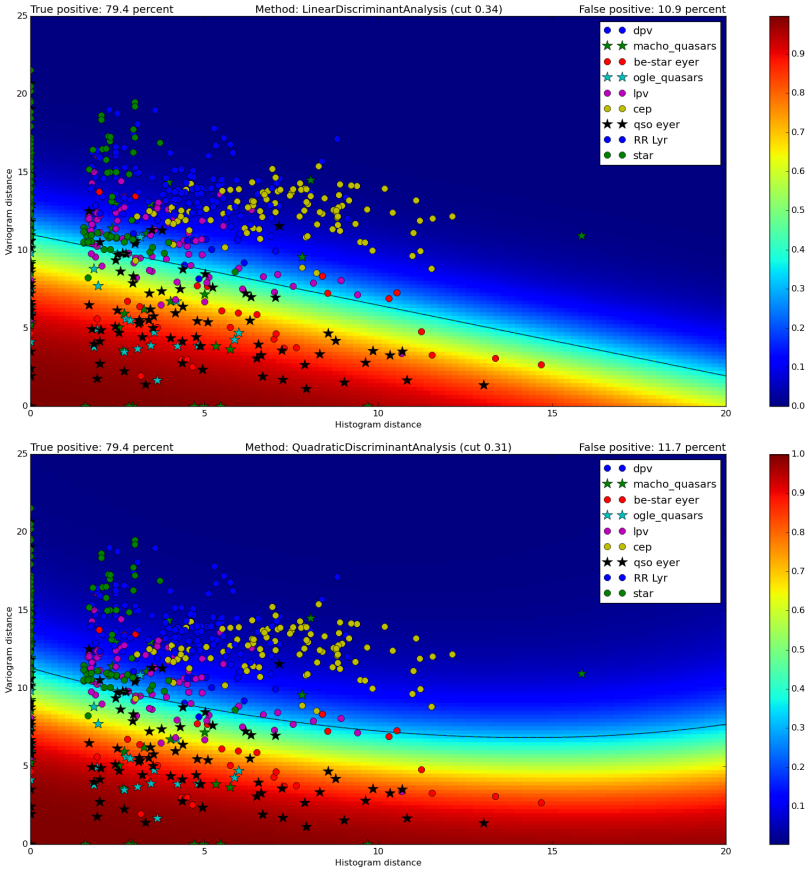
(d) QDA, nepředfiltrované

Obrázek 4.12: Grafy znázorňují množství správně identifikovaných, nesprávně identifikovaných a jejich rozdíl v závislosti na volbě hraniční pravděpodobnosti. Každý graf v levém horním rohu ukazuje optimální hodnotu pravděpodobnosti pro identifikaci.

Pro každý případ byla určena optimální hodnota pravděpodobnosti jako maximum funkce „správně identifikováno“ – „nesprávně identifikováno“. Tyto hodnoty se následně užili jako hraniční kritérium pro jednotlivé případy.




Obrázek 4.13: Histogram-variogramový prostor reprezentující vzdálenost (odlišnost) od trénovacího vzorku kvazarů. Na prvním obrázku je LDA metoda [27] a na druhém QDA metoda [28]. Testovaná data jsou vytříděná o Abbe kritérium a barevný index (viz výše). Barevné pozadí reprezentuje s jakou pravděpodobností je objekt v dané oblasti kvazarem. Body s hvězdicovým symbolem jsou kvazary a puntíky jsou ostatní objekty.




Obrázek 4.14: Histogram-variogramový prostor pro nepředfiltrovaná data

Výsledky identifikace pro vypočítané hranice znázorňuje tabulka 4.14.

| | Předfiltrované | | Nepředfiltrované | |
|---|---|---|---|---|
| | QDA | LDA | QDA | LDA |
| | Množství prošlých hvězd [%] | | | |
| Kvazary (MACHO) | 61 | 58 | 68 | 73 |
| Kvazary (OGLEII) | 91 | 91 | 95 | 95 |
| Kvazary (Eyer) | 82 | 83 | 83 | 90 |
| Be-hvězdy (Eyer) | 87 | 81 | 57 | 63 |
| Hvězdy s dvojitou periodou | 2 | 2 | 3 | 3 |
| Dlouho periodické hvězdy | 24 | 22 | 13 | 23 |
| Cefeidy | 2 | 2 | 7 | 7 |
| RR Lyr | 3 | 1 | 3 | 3 |
| Neproměnné hvězdy | 3 | 3 | 17 | 27 |

Tabulka 4.7: Množství hvězd identifikovaných jako kvazary podle výše určených parametrů. Relativní počty předfiltrovaných hvězd se vztahuje k velikosti vzorku, které už prošli předfiltrováním (tzn. nejde o absolutní relativní počet, jako u nepředfiltrovaných)

O trochu úspěšnější se ukazuje být LDA metoda a proto pro definovaní hranic využijeme právě tuhle metodu. Jako hraniční oblast volíme konturu na hladině pravděpodobnosti 0.34,

kterou definuje lineární funkce:

$$d_{max,vario}(d_{hist}) = -0.45 d_{hist} + 11 \tag{4.14}$$

Vyšetřována hvězda bude klasifikovaná jako kvazar pokud:

| identifikace | |
|---|---|
| kvazar | $d_{vario} < d_{max,vario}(d_{hist})$ |
| nekvazar | $d_{vario} \geq d_{max,vario}(d_{hist})$ |

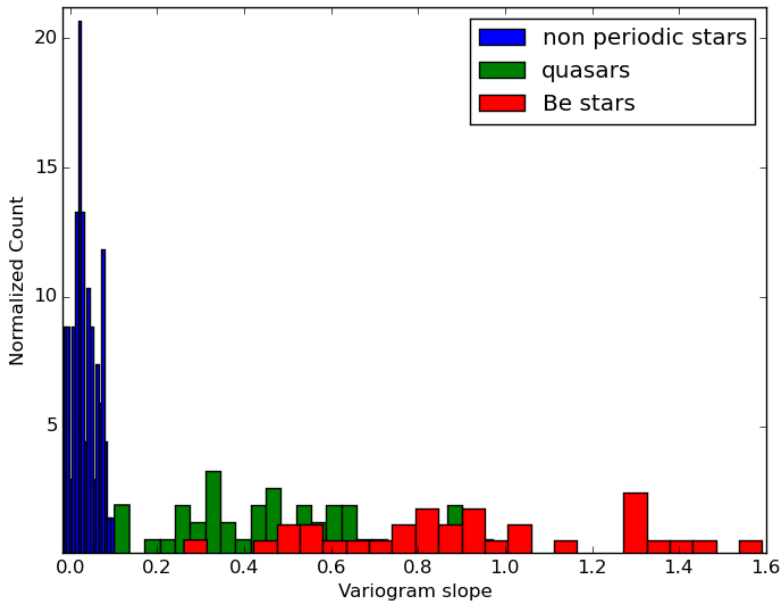
Tabulka 4.8: Rozhodovací tabulka

V tabulce 4.7 lze vidět, že příliš nezáleží, zda jsou data předtříděna Abbe hodnotou a barevným indexem, či nikoliv. Odhadovou přesnost metody popisuje tabulka 4.9.

| | správně identifikováno [%] | nesprávně identifikováno [%] |
|---|---|---|
| kvazary | 86 | 14 |
| nekvazary | 79 | 21 |

Tabulka 4.9: Úspěšnost identifikace kvazarů v symbolickém prostoru histogramu a variogramu. Kritérium bylo určeno LDA metodou podle rozhodovací tabulky 4.8 pro pravděpodobnost 0.19.

Vzorek hvězd, nesprávně identifikovaných jako kvazary, obsahuje především Be hvězdy, jejichž světelné křivky jsou velice podobné světelným křivkám kvazarů. K vyčlenění některých z nich by bylo možné určit směrnice jejich variogramů. Změny jasnosti většiny Be hvězd jsou výraznější a rychlejší, proto by jejich směrnice měli být větší. Nicméně, jak ukazuje histogram 4.15, to neplatí obecně, a proto je toto kritérium pouze orientační.




Obrázek 4.15: Distribuce směrnic variogramů pro neproměnné hvězdy, kvazary a Be hvězdy

# Kapitola 5

# Light Curve Analyzer

K účelům klasifikace hvězd podle jejich světelných křivek jsem vyvinul poměrně rozsáhlý program (téměř 6 tisíc řádků), který má potenciál se stát užitečným nástrojem pro studium a klasifikaci světelných křivek. Program je napsán objektově v jazyce Python. Jednotlivé třídy jsou vytvořeny tak, aby bylo možné doplňovat další segmenty, či přepisovat stávající. Například přidat další filtr pro třídění hvězd nebo dodat konektor pro práci s jinou databází.

Přestože jsem program zatím využíval jen pro identifikaci kvazarů, není problém ho využít ke klasifikaci jiného typu objektů. Stačí jen načíst rozdílné světelné křivky jako předlohu a nalézt jiné parametry, obdobně jak bylo popsáno v kapitole 4.5. Popřípadě zaimplementovat si vlastní filtry a identifikovat hvězdy podle úplně jiných kritérií (fitovat křivky, třídit hvězdy podle směrnice variogramu atd.)

## 5.1 Struktura

Program je rozčleněn do balíků, ve kterých se nachází jednotlivé moduly se svými třídami a metodami.

- entities
  - elementární třídy pro práci s astronomickými objekty jako například: „LightCurve“ (světelná křivka), „Star“ (hvězda) a „AbstractCoordinates“ (společná třída pro souřadnicové třídy), v neposlední řadě jsou zde třídy s výjimkami (chybové třídy)
- db tier
  - třídy které jsou zodpovědné za dodání hvězd (jako objekty typu „Star“) se svými atributy (identifikátor, souřadnice, světelná křivka atd.)
- stars processing
  - třídy, které mají na starost filtrování hvězd
- utils

  - moduly pro analýzu časových řad, pokročilejší implementace prohledávání databází, vizualizace dat a pomocné metody pro práci s hvězdnými objekty

- tests
  - sada testovacích tříd pro ladění programu
- commandline
  - spustitelné moduly

## 5.2 Filozofie programu

### Hvězda

V tomto oddíle si lehce přiblížíme některé třídy a princip funkčnosti programu. Jak bylo naznačeno výše, základním pilířem programu je třída „Star“. Jedná se o základní astronomický objekt, který obsahuje velké množství informací nejen o samotné hvězdě, kterou reprezentuje. Jedním z jeho nejdůležitějších atributů je objekt „LightCurve“, který obsahuje informace o světelné křivce.

Na první pohled by se mohlo zdát zbytečné tvořit třídu na vše, co by mohlo být uloženo jen jako dva řádky hodnot (časy a hvězdné velikosti) v nějaké proměnné. Obrovská výhoda objektově orientovaného přístupu se ukazuje při zpracovávání dat v různých formátech, různými metody atd. Co když v jedné světelné křivce jsou přepálené pixely reprezentovány jako „99“, „N/A“ nebo jako volná mezera? Kouzlo tkví v samotné implementaci třídy, která rozhoduje o přístupu k datům a je možné reagovat na jiný formát připsáním pár řádků v samotné třídě.

### Získávání dat

Nyní je otázkou, jak získat data k vytvoření hvězdných objektů? K tomu slouží databázové konektory, které implementují základní metody, které by měly mít všechny třídy tohoto typu, jako například „Stáhni světelnou křivku“, „Vyhledej informace o hvězdě v databázi atd.“. Vedle těchto databázových klientů stojí třída načítající hvězdné objekty z datových souborů světelných křivek. Pro třídy stahující data přes TAP protokol se zde nachází abstraktní třída, která má zaimplementované metody pro práci s tímto protokolem. Dále si uvedeme třídu, která pro dva databázové klienty vyhledává hvězdy, které se nachází v obou databázích.

V neposlední řadě balík obsahuje třídu „StarsProvider“, která je prostředníkem mezi klienty získávající data (ať už z databáze nebo ze složky) a okolním světem. To znamená, že na všechny databáze se dotazujeme stejně a vždy dostaneme výstup ve stejném formátu (list objektů typu „Star“). Díky tomu není důležité jak jsou v dotazované databázi označeny data, například zda se rektascence zadává ve stupních, radiánech nebo v hodinách. Všechny tyto náležitosti si zajišťují konektory samy. S problémem s rozdílným formátem souřadnic by nám pomohly objekty „RightAscension“ a „Declination“, které uchovávají samotnou informaci o poloze. Konkrétní databáze si z nich už zavolají souřadnice v požadovaném formátu.

## Filtrování

V tomto bodě už máme k dispozici data v podobě hvězdných objektů a můžeme se pustit do samotného třídění. Hlavní třídou je zde „FilteringManager“, která vyfiltruje hvězdné objekty podle vložených filtrů. Všechny konkrétní implementace filtru musí obsahovat dvě metody: „Připrav hvězdy“ a „Vyfiltruj hvězdy“, které využívají dle potřeby funkce pro zpracovávání dat (například „data analysis“, „SAX“ atd.). Práce s filtrovacím managerem je už potom jednoduchá, jednou metodou se načtou hvězdy k filtrování, druhou se načtou filtry a třetí vrátí hvězdy splňující podmínky jednotlivých filtrů.

```
from db_tier.stars_provider import StarsProvider
from stars_processing.filtering_manager import FilteringManager
from stars_processing.filters_impl.compare import ComparingFilter
from stars_processing.filters_impl.word_filters  import
    HistShapeFilter,\
     VariogramShapeFilter
from stars_processing.filters_impl.abbe_value import AbbeValueFilter
from stars_processing.filters_impl.color_index import ColorIndexFilter
from utils.output_process_modules import saveIntoFile
from utils.stars import plotStarsPicture
from commandline.filtering_parameters import params as p
from entities.right_ascension import RightAscension
from entities.declination import Declination

#=======================================================
'''
EXAMPLE of searching for quasars in certain area in OGLE II database
'''
#=======================================================

obtain_params ={
        "ra":           RightAscension(5.56,ra_format="hours"),
        "dec":          Declination(-69.99, dec_format="degrees"),
        "delta":        3,
        "target":       "lmc"
         }

#------    Get qusars with light curves    ------------

files_prov = StarsProvider().getProvider(path=p.OGLE_QSO_PATH,
                                         obtain_method="file",
                                         star_class="quasar")
quasars =  files_prov.getStarsWithCurves()


#----  Download  stars from OGLE II database    ------

ogle_prov = StarsProvider().getProvider(obtain_method="ogle",
                                        obtain_params=obtain_params)
stars = ogle_prov.getStarsWithCurves()

#Filter which compares two stars according to given subfilters
cf = []
```

```
cf.append(HistShapeFilter(days_per_bin=p.HIST_DAYS_PER_BIN,
                          alphabet_size=p.HIST_ALPHABET_SIZE))

cf.append(VariogramShapeFilter(days_per_bin=p.VARIO_DAYS_PER_BIN,
                               alphabet_size=p.VARIO_ALPHABET_SIZE))


#Decision function which decides whether a star
#will pass thru filtering according to its histogram
#and variogram distance from template
def dec_func_t(distances):
    hist_dist,vario_dist = distances
    return p.VAR_HIST_A * hist_dist + p.VAR_HIST_B > vario_dist

#Load comparative sub filters, template stars
#of quasars and decision function
comp_filt = ComparingFilter(cf, quasars, dec_func_t,
                            search_opt="closest")

#Abbe value limit filter
abbe_filter = AbbeValueFilter(abbe_lim=p.ABBE_LIM)


#Color index filter with its decision function
def dec_func_c(bv,vi): return bv >= p.BV_MIN and vi >= p.VI_MIN

color_filter = ColorIndexFilter(dec_func_c)


#--------------    Perform filtering    ----------------
#Load inspected stars and filters
filteringManager = FilteringManager(stars)
filteringManager.loadFilter(comp_filt)
filteringManager.loadFilter(abbe_filter)
filteringManager.loadFilter(color_filter)

#Perform filtering and return stars passed thru filter
result_stars = filteringManager.performFiltering()


#------   Plot and save stars passed thru filter   ----

#Save passed stars object
saveIntoFile(result_stars, p.STARS_OBJECT_PATH, p.
    RESULT_STARS_FILE_NAME)

#Plot stars
plotStarsPicture(result_stars)
```

Listing 5.1: „Ukázka získání dat z OGLE II databáze a jejich třídění podle zadaných filtrů“

# Kapitola 6

# Hledání v OGLE II databázi

OGLE II datábáze obsahuje přes 40 milionů hvězd ve třech oblastech: Velký Magellanův oblak (LMC), Malý Magellanův oblak (SMC) a galaktická výduť (BUL). Při prvním prohledávání bylo množství identifikovaných objektů poměrně vysoké, a proto jsem se rozhodl snížit pravděpodobností hranici, kterou jsme zjišťovali v 4.5.3, na hladinu 88 %. Na této hladině by mělo být správně identifikováno jako kvazary 43.1 % objektů a nesprávně určeno 3.3 % (2.6 % Be hvězdy a 0.7 % dlouhoperiodické hvězdy).

Při vizuální kontrole klasifikovaných objektů tyto hodnoty orientačně odpovídají. Prozatím bylo provedeno systematické prohledávání v LMC, kde bylo vyšetřeno přes 1 milion hvězdných objektů. Z tohoto množství bylo nalezeno téměr 4 tisíce kandidátů na kvazary. K prohledávání bylo opět využito MetaCentra, kde pro každou z 21 oblastí LMC beží jeden proces, který systematicky prochází všechny hvězdy podle identifikačního čísla.

# Závěr

Ve své práci jsem se zabýval vývojem a testováním metod, které by byly schopny identifikovat světelné křivky kvazarů. Všechny tyto metody byly testovány a laděny na vzorku kvazarů, neproměnných hvězd, Be hvězd, cefeid, hvězd typu RR Lyrae, dlouhoperiodických hvězd a hvězd s dvojitou periodou.

Jako první kritérium jsem uvažoval Abbe hodnoty jednotlivých časových řad. Hraniční hodnotu jsem určil z histogramu jako nejvyšší Abbe hodnotu ze vzorku kvazarů, čímž se vyloučily především neproměnné hvězdy a proměnné hvězdy s periodou menší nebo srovnatelnou s dobou mezi jednotlivými pozorováními. Dalším kritériem byl barevný index, který jsem opět zvolil střídmě tak, aby nevyloučil žádné kvazary. Tímto způsobem byly odstraněny především dlouhoperiodické proměnné hvězdy ležící v pravé horní části barevného diagramu.

Hlavním filtrovacím kritériem byla vzdálenost v histogram-variogramovém prostoru. Jednotlivé souřadnice byly určeny srovnáváním vyšetřovaných hvězd s trénovacími kvazary. Vložením vzorku testovacích kvazarů bylo možné clusteringovými metodami určit pravděpodobností pole, které určovalo s jakou pravděpodobností je objekt na daných souřadnicích kvazar. Hraniční pravděpodobnost jsme určili tak, aby oblast, kterou ohraničuje křivka spojující body o této pravděpodobnosti, obsahovala co největší počet kvazarů a co nejmenší počet nekvazarů. Pro takto zvolenou hladinu je množství správně identifikovaných kvazarů 86 % a pro nesprávně označených jako kvazary 21 %.

Pro identifikaci kvazarů jsem vyvinul program v jazyku Python, který by se mohl stát univerzálním nástrojem pro zpracovávání astronomickcých dat. Program zatím obsahuje především třídy a metody pro získávání a zpracování světelných křivek, ale možnosti jeho využití jsou neomezené. Je možné do něj doplňovat nové prvky (filtry, konektory pro jiné databáze atd.), aniž by se změnila funkčnost programu.

Nakonec jsem pomocí vyvinutých metod začal prohledávat OGLE II databázi. Z důvodu relativně velkého množství kandidátů na kvazary jsem se rozhodl zvýšit pravděpodobnostní hladinu správné identifikace na 88 %, aby byla kontaminace o špatně identifikované objekty minimální. Nalezeno by mělo být 43.1 % kvazarů (ze všech kvazarů, které by tam měly být) s tím, že výsledný vzorek bude obsahovat i 3.3 % jiných objeků (převážně Be hvězd). Zatím byla prohledána asi 1/40 databáze a z 1 milionu hvězd bylo nalezeno téměř 4000 kandidátů. Všichni takto získaní kandidáti musí být dále potvrzeni jinou metodou (např. spektroskopicky).

# Seznam použité literatury


[1] M. Schmidt, "3C 273 : A Star-Like Object with Large Red-Shift," *Nature*, vol. 197, p. 1040, Mar. 1963.

[2] E. Hubble, "3C 273 : A relation between distance and radial velocity among extra-galactic nebulae," *National Academy of Science*, Apr. 1929.

[3] "3c273 chandra." http://en.es-static.us/upl/2005/02/3C273.jpeg.

[4] R. Irion, "A Quasar in Every Galaxy?," *Astrophysical Journal*, Feb. 2014.

[5] T. D. E., "End of the world: You won't feel a thing," *Science News*, vol. 131, no. 25, pp. 391–391, 1987.

[6] "Q is for quasar." http://aisforeducation.pressible.org/daverafe/q-is-for-quasar.

[7] V. Karas, "Podobnosti v astrofyzice: Od kvazarù k pulzarům," 1996.

[8] "Quasar." https://en.wikipedia.org/wiki/Quasar.

[9] "M87 jet." https://upload.wikimedia.org/wikipedia/commons/3/39/M87_jet.jpg.

[10] A. Sandage and J. D. Wyndham, "On the Optical Identification of Eleven New Quasi-Stellar Radio Sources.," *Astrophysical Journal*, vol. 141, p. 328, Jan. 1965.

[11] "Basics of quasar spectra." https://ned.ipac.caltech.edu/level5/Charlton/Charlton1_1.html.

[12] M. Schmidt and R. F. Green, "Quasar evolution derived from the Palomar bright quasar survey and other complete quasar surveys," *Astrophysical Journal*, vol. 269, pp. 352–374, June 1983.

[13] U. of Illinois at Urbana-Champaign, "Quasar Light Variability Linked To Black Hole Mass.," *ScienceDaily*, Jan. 2007.

[14] "Einstain cross." http://hubblesite.org/newscenter/archive/releases/1990/20/image/a/.

[15] "Gravitional microlensing." http://ogle.astrouw.edu.pl.

[16] L. Eyer, "Search for QSO candidates in OGLE-II data," June 2002.

[17] J. Von Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *The Annals of Mathematical Statistics*, vol. 12, no. 4, pp. 367–395, 1941.

[18] N. Mowlavi, "Searching transients in large-scale surveys. A method based on the Abbe value," *Astronomy and Astrophysics*, vol. 568, p. A78, Aug. 2014.

[19] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, ACM, 2003.

[20] "Ogle photometric databases." http://ogledb.astrouw.edu.pl/ ogle/photdb/index.html.

[21] "The magellanic quasars survey." http://www.astrouw.edu.pl/ simkoz/MQS.

[22] "The macho project." http://wwwmacho.anu.edu.au/.

[23] M. Geha, C. Alcock, R. A. Allsman, D. R. Alves, T. S. Axelrod, A. C. Becker, D. P. Bennett, K. H. Cook, A. J. Drake, K. C. Freeman, K. Griest, S. C. Keller, M. J. Lehner, S. L. Marshall, D. Minniti, C. A. Nelson, B. A. Peterson, P. Popowski, M. R. Pratt, P. J. Quinn, C. W. Stubbs, W. Sutherland, A. B. Tomaney, T. Vandehei, and D. L. Welch, "Variability-selected Quasars in MACHO Project Magellanic Cloud Fields," *Astrophysical Journal*, vol. 125, pp. 1–12, Jan. 2003.

[24] "Ogle-iii catalog of variable stars." http://ogledb.astrouw.edu.pl/ ogle/CVS/.

[25] "Grid search: Searching for estimator parameters." http://scikit-learn.org/stable/modules/grid_search.html.

[26] "Metacentrum vo." https://metavo.metacentrum.cz/.

[27] "Linear discriminant analysis (lda)." http://scikit-learn.org/0.16/modules/generated/sklearn.lda.LDA.html.

[28] "Quadratic discriminant analysis (qda)." http://scikit-learn.org/0.16/modules/generated/sklearn.qda.QDA.html.