Search of the time-variation of the interstellar extinction with a machine learning method

Application to the variability analysis for future LSST data

J. ITAM-PASQUET

Directors : G.Jasniewicz et D.Puy (LUPM)
Supervisor : N. Mauron (LUPM)
Collaborator : D.Pfenniger (Geneva Observatory)

johanna.pasquet@umontpellier.fr

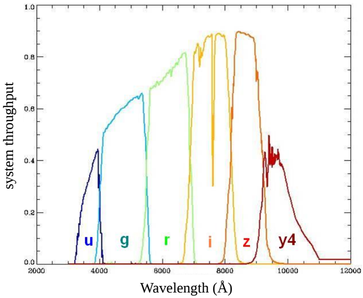
Outline

Magnitudes and filters

- **Apparent magnitude (m)** : logarithmic measure of the intensity of light from an object (I_1), measured in a specific wavelength relative to the intensity of the light from a reference star I_{ref} :

$$m_1 - m_{ref} = -2.5\, log(\frac{I_1}{I_{ref}}) \tag{1}$$

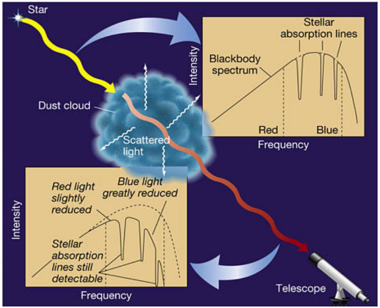
- **Filters** : measure the light flux from a star only in restricted wavelength ranges.



The SLOAN and LSST filter bands, showing total system throughput

Dust Effects on Starlight

- **Extinction** (dimming of the light from stars) : Scattering + absorption
  $\rightarrow A_\lambda$ is the total extinction at wavelength λ
- **Reddening** : the shorter the wavelength, the higher the extinction : blue light is affected more strongly than red light.
  $\rightarrow$ the most common measure of reddening is the color excess : $E(B-V)$



Context : The dark matter

Pfenniger, Combes & Martinet (A&A, 285, 1994) :

- They proposed that a part of the baryonic component of dark matter around spiral galaxies could be in the form of cloud gas, essentially in molecular form and rotationnaly supported
- A factor 10 or more of hydrogen mass underestimate is enough to remove the need of exotic matter in disc galaxies
- Observations of the interstellar medium show that the cold gas is fractal and essentially clumpy down to very small scales (few tens of AU [1])

  → "clumps" with the characteristics at $T = 3\,K$:
  $l = 30\,AU,\ M = 10^{-3} M_{\odot}$

- main difficulty to detect the cold gas in emission because of its low temperature

1. $1\,\mathrm{AU} \sim 1.49 * 10^{11}$ m

Context : The small structures

Evidence of small scale structures

- Interstellar line spectroscopy (Boissé et al., A&A, 559, 2013) show that the column density vary of 11% over 3 years, in some cases, in agreement with some quasars scintillation

Drake and Cook (2003) :

- Search for stellar obscuration events due to dark clouds
- MACHO project light curve of $48 * 10^6$ stars towards the Galactic bulge, Large Magellanic Cloud, and Small Magellanic Cloud
- Such events are expected to be very rare, with much less than 1% of stars in any given direction being obscured at any time
- Clouds occupy the disk and in the halo
- No evidence for a population of dark clouds in either the disk or halo of our Galaxy

Aim of the thesis

- **Goal** : Search for time-variations of the interstellar extinction constrained by baryonic dark matter
- **Methodology** : Develop a search for tempoal variability method, applicable to others variable objects and big future databases

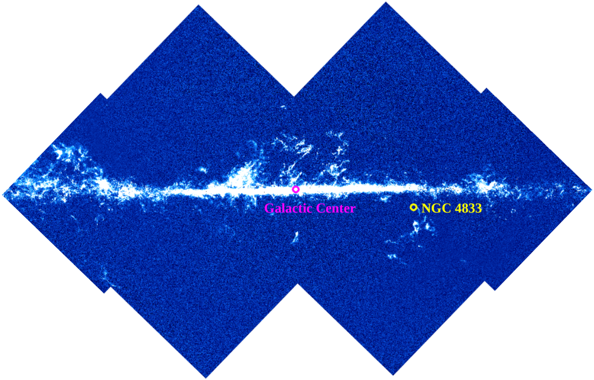
Definition of a globular cluster

- Globular clusters are very massive objects that contain $\sim$ 100 000 stars
- They formed $10 - 13 * 10^9$ years ago
- $\sim$ 180 globular clusters in our Galaxy

NGC 4833 by Hubble Space Telescope (field of view : 3.5')

Advantages of studying a globular cluster

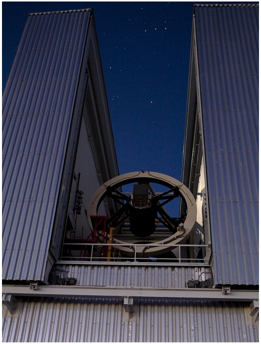
- includes a large number of stars
- A common kinematic system for all stars
- lies behind dusty regions at a latitude of -8°



Planck CO-map

Observations

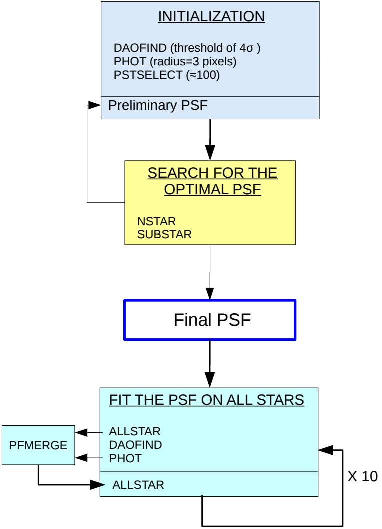
- Carried out at the NTT in Chile (PUY D., PFENNIGER D., DESSAUGES-ZAVADSKY M., 2006, ESO)
- Optical observations in (B,V,I) filters
- 2 observation sets separated by a 6 months period : the **January 2006** and the **July 2006**



ESO's New Technology Telescope at La Silla

Photometric data reduction of NGC 4833

Use of DAOPHOT II within IRAF data reduction and analysis environment



Photometric calibration with Standard Stars

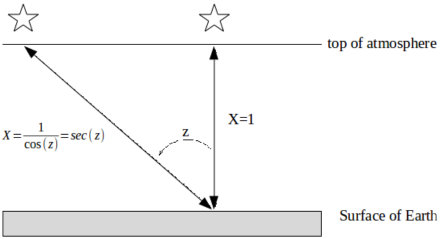
For the January, 2006 data :

- Light of stars is affected by atmospheric extinction :

$$M(\lambda) = m(\lambda) - K_{\lambda} sec\ z \qquad (2)$$

  - m : instrumental magnitude
  - M : calibrated magnitude
  - $sec\ z$: air mass, noted X (=1 at zenith)
  - K : extinction factor

- Selection of standard stars at different air masses



Selecting secondary stars

For the July, 2006 data :

- Use of selected stars from the January photometry as secondary standards
- A weak, linear residual in color and a residual that varied quadratically in X and Y were present :

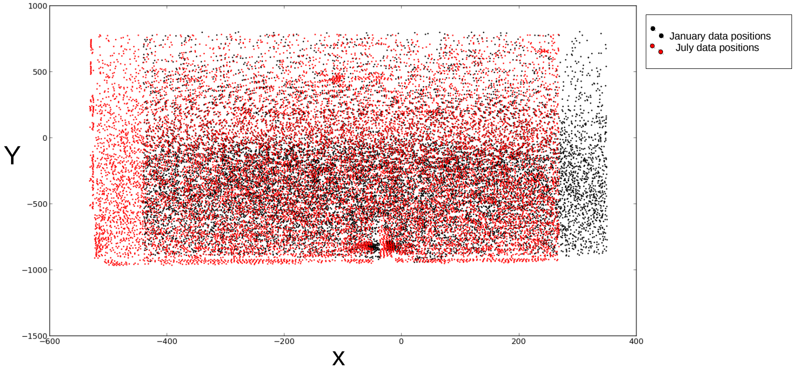
$$m - M = c_0 + c_1(B - V) + c_2X + c_3Y + c_4X^2 + c_5XY + c_6Y^2 \quad (3)$$

coefficients c_i are determined by the method of least squares

---

Final photometric accuracy :
$0.003\,mag \leq mag\ \ uncertainty \leq 0.05\,mag$

Difficulties in the superimposition of images

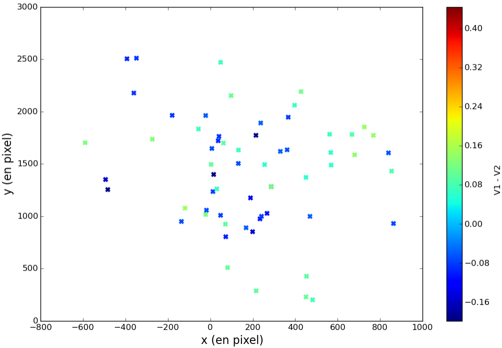


Final superposition accuracy : 0.2 *pixels*

Results on the study of NGC 4833 (I)

ITAM-PASQUET J. et al., The GREAT-ITN final conference, 2014, Barcelona, Spain
ITAM-PASQUET J. et al, Journées de la SF2A 2015, Toulouse, France

- 62 stars vary in a six-month period (beyond a statistical significance of 3σ) and are not known as variables
- Field of view : 0.13 square degree
- Stars number : 5800
- Likelihood of occurring (P) : $\sim 1\%$
- Number of events per square degree (N) : 476



Results on the study if NGC 4833 (II)

Granted 10.4 h of radio observation time with 22m-Mopra telescope (Australia), carried out using VNC remote desktop sofware from Montpellier with the collaboration of Nigel Maxted (Sydney Observatory, Australia)

Goal : Detect CO traces to constrain molecules traces in the line of sight of NGC 4833

- $7'x7'$ map
- a main-beam sensitivity of $\sim 1K$ per channel ($\sim$ 2.5 times higher than existing CO data towards this globular cluster)
- a beam FWHM : 35"($\sim$ 4 times larger than existing CO data)

No CO detection :

- CO is more affected by photodissociation than H_2
- CO may condense into dust grains under 20 K and so be depleted in gas-phase observation

HST globular clusters : problematic and goals

- Expand the field of view→more statistics
- Get more time periods
- wavelengths near UV

→ Study of 20 globular clusters (+27) in *Hubble Space Telescope archive*

Conclusion on the study of HST globular clusters

- 3 interesting globular clusters :
  - **NGC 104** ($E(B-V) = 0.04$)
    - Field of view : 0.29 square degree
    - Likelihood of occurring (P_1) : 6%
    - Number of candidates per square degree (N_1) : 48

      $\rightarrow$ Interest factor$_1$ $= P_1 * N_1 =$ 2.9
  - **NGC 4833** ($E(B-V) = \mathbf{0.32}$)
    - Field of view : 0.007 square degree
    - Likelihood of occurring (P_2) : 11%
    - Number of candidates per square degree (N_2) : 1571

      $\rightarrow$ Interest factor$_2$$= P_2 * N_2 =$ **172**
  - **NGC 7078** ($E(B-V) = 0.1$)
    - Field of view : 0.01 degré carré
    - Likelihood of occurring (P_3) : 4.6%
    - noNumber of candidates per square degree (N_3) : 900

      $\rightarrow$ Interest factor$_3$ $= P_3 * N_3 =$ 41

3D Density law of the distribution of clumps

$C++$ implementation

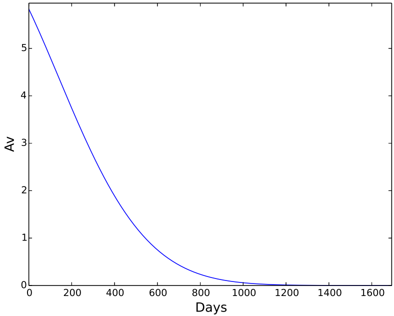
- The centers of the clouds and subclouds are randomly distributed according to a 3D density law
- $N = 6$, $level = 8 \rightarrow 1\ 670\ 616$ clumps

$$\rho(r, r_L) = \begin{cases} \left(\frac{r}{r_L}\right)^{-2} & r < r_L \\ 0 & r \geq r_L \end{cases} ; \quad \rho(r, r_L) = \frac{1}{\left(\left(\frac{r}{r_L}\right)^2 + 1\right)^{\frac{5}{2}}} \quad ; \quad \rho(r, r_L) = e^{-\left(\frac{r}{r_L}\right)^2}$$

Quasi-isothermal profil — Plummer profil — Gaussian profil

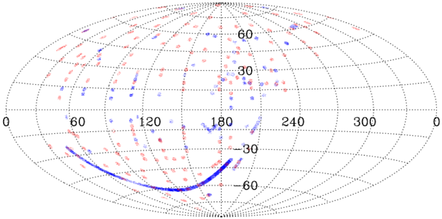
Motions of objects along time

- Galactic velocities of the globular cluster : $U = -100\ km.s^{-1}$; $V = -301\ km.s^{-1}$; $W = -49\ km.s^{-1}$
- Galactic velocities of the Sun and the cloud : $U_\odot = -8\ km.s^{-1}$; $V_\odot = 13\ km.s^{-1}$; $W_\odot = 6\ km.s^{-1}$
- Motion of the clumps inside the cloud : damped gaussian distribution truncated to escape velocity



Defining Stripe 82

- Stripe 82 is a survey of 300 deg^2 equatorial field in the Southern Galactic cap
- Coordinates : -50° $\leq \alpha \leq$ +60° et $-1.26^{\circ} \leq \delta \leq 1.26^{\circ}$
- It was imaged by the Sloan Digital Sky Survey (SDSS) multiple times between 2000 and 2008 with good image quality and low sky background



Stripe 82 is represented by the blue line, red and blue dots are spectroscopic data

Defining Stripe 82

Description

- 67 000 light curves of objects with significant temporal variability in (u, g, r, i et z) magnitudes
- Known objects in the database : $\sim$ 8000 quasars, $\sim$ 500 RR Lyrae and δ scuti, few galaxies, supernovae...

Interests

- $\sim$ 40 000 light curves whose variations are not explained
- $\sim$ 30 observations at different time for each object on average

Defining Stripe 82

Description

- 67 000 light curves of objects with significant temporal variability in (u, g, r, i et z) magnitudes
- Known objects in the database : $\sim$ 8000 quasars, $\sim$ 500 RR Lyrae and δ scuti, few galaxies, supernovae...

Interests

- $\sim$ 40 000 light curves whose variations are not explained
- $\sim$ 30 observations at different time for each object on average

The Challenge

Goals

Short-term : Identify light curves compatible with an obscuration event, ie the passage of a clump in the line of sight
Long-term : Be able to identify them in other databases

Problems

- Big data
- Big uncertainties on characteristics of the variability of the star obscuration by clumps

Solutions

- Synthetize the obscuration event (amplitude ~ 1 *mag* and time scale~ 1 *year*)
- Classification methods to detect obscuration events

The Challenge

Goals

Short-term : Identify light curves compatible with an obscuration event, ie the passage of a clump in the line of sight
Long-term : Be able to identify them in other databases

Problems

- Big data
- Big uncertainties on characteristics of the variability of the star obscuration by clumps

Solutions

- Synthetize the obscuration event (amplitude ~ 1 *mag* and time scale~ 1 *year*)
- Classification methods to detect obscuration events

The Challenge

Goals

Short-term : Identify light curves compatible with an obscuration event, ie the passage of a clump in the line of sight
Long-term : Be able to identify them in other databases

Problems

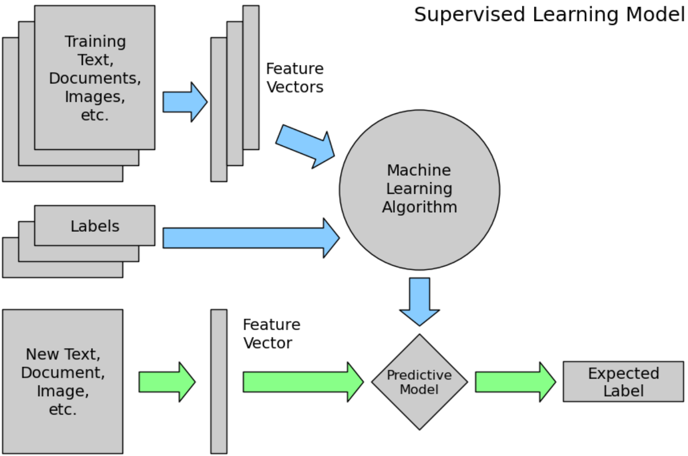
- Big data
- Big uncertainties on characteristics of the variability of the star obscuration by clumps

Solutions

- Synthetize the obscuration event (amplitude ~ 1 *mag* and time scale~ 1 *year*)
- Classification methods to detect obscuration events

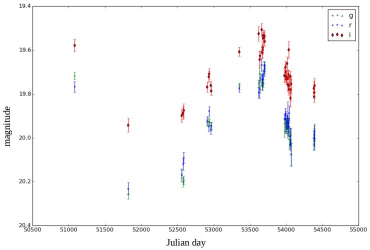
Machine learning

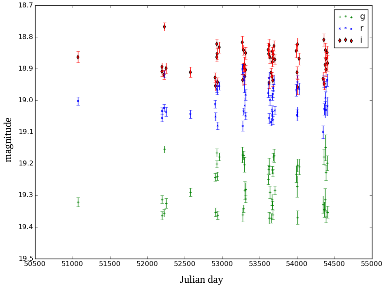
Python implementation (sklearn)



The training database

- 80% of known quasar list[1] ($\sim$ 6500)
- 80% of known RR Lyrae et δ scuti list[2] ($\sim$ 400)



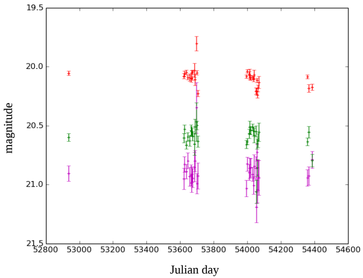


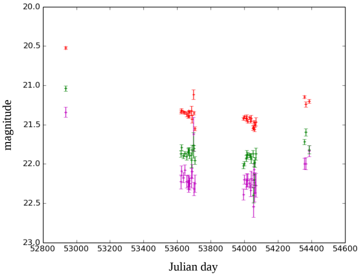
- 10 000 synthetized light curves
- 10 000 light curves whose variations are not identified

[1]Meusinger et al., 2011, Flesch et al., 2015
[2]Surveges et al., 2012 and Ivezic et al. 2007

Obscuration event synthesis

1. Random selection of a star with an amplitude< 0.25
2. Addition of a gaussian with $\sigma \in [400, 800]$, $\mu \in [51000, 55000]$, $Amplitude \in [0.25, 1.25]$
3. Addition of a gaussian noise $\mathcal{N}(0, 0.05)$
4. Repeat the steps to get 10 000 synthetic light curves compatible with an obscuration event





The testing database

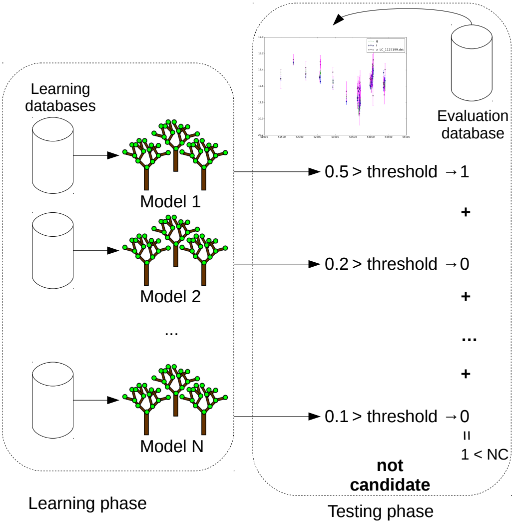
- 20% of known quasar database ($\sim$ 1500) that had not been used in training database
- 20% of known RR Lyrae et δ scuti database ($\sim$ 80) that had not been used in training database
- All light curves whose variations are not identified

Classification method

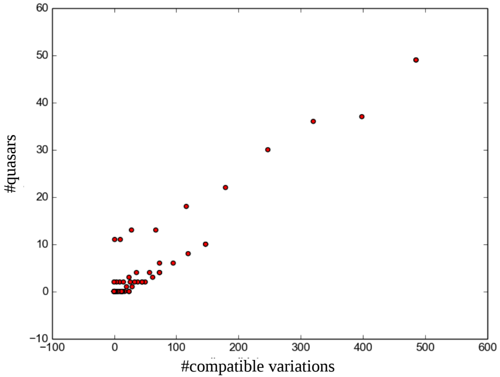
A simplified extraction of features :

Classifier : a set of Random Forests

Fusion of results (I)



Fusion of results (II)



Results and performances (I)

	Recall_{Quasar}	$\text{Precision}_{Quasar}$	$\text{Recall}_{Lyr+\dots}$	$\text{Precision}_{Lyr+\dots}$
My work	95.3%	95%	52.4%	88.4%

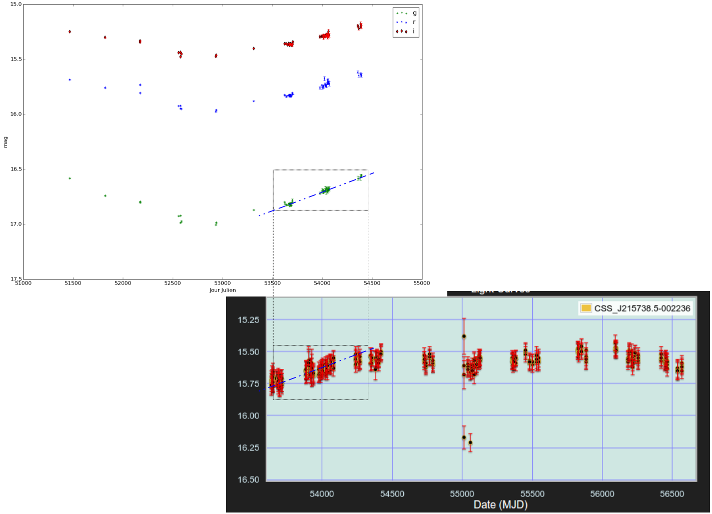
ITAM-PASQUET J., et al., in preparation for A&A

Results and performances (II)

- The algorithm found **43** light curves compatible with an obscuration event
- It did not make a mistake by considering that a quasar light curve is an obscuration event
- It did not make a mistake by considering that a pulsating star light curve is an obscuration event
- It did a mistake by considering that a galaxy light curve is an obscuration event...but it did not learn what looks like a light curve of galaxy

ITAM-PASQUET J., et al., in preparation for A&A

Example of discoveries



Conclusions of my work up to now

My results :

- Variable object detection has till now identified and compatible with the passage through a clump in the line of sight in different directions in the Galaxy
- Great performances of my classification method
- Modelling of the problem and estimation of the likelihood of occuring of the phenomenon consistent with observations

What I learned :

- Increased my knowledge of the small structures of the Interstellar medium
- Gained photometric expertise
- developped skills in the classification of variable objects → **big data**

Conclusions of my work up to now

My results :

- Variable object detection has till now identified and compatible with the passage through a clump in the line of sight in different directions in the Galaxy
- Great performances of my classification method
- Modelling of the problem and estimation of the likelihood of occuring of the phenomenon consistent with observations

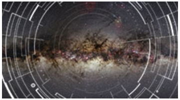
What I learned :

- Increased my knowledge of the small structures of the Interstellar medium
- Gained photometric expertise
- developped skills in the classification of variable objects → **big data**

The Large Synoptic Survey Telescope

- LSST can map the entire visible sky in just a few nights ; each panoramic snapshot with the 3200-megapixel camera covers an area 40 times the size of the full moon.
- 10 years survey of the sky
- 37 billion stars and galaxies
- Three central considerations dictated the design of LSST :

Wide

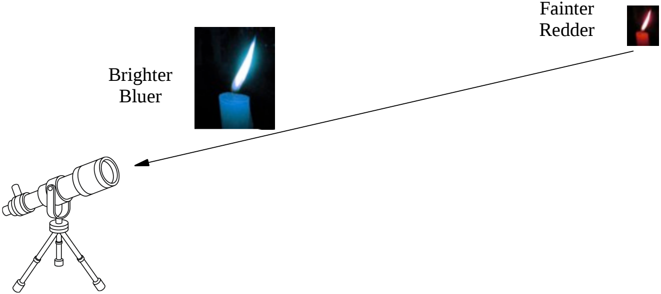


Fast

Deep

Supernovae Ia

- Type Ia supernovae can be used as well-calibrated standard candles
- measurements of the relation between cosmological distance and redshift → the strongest contemporary evidence for an accelerating cosmological expansion



Transient

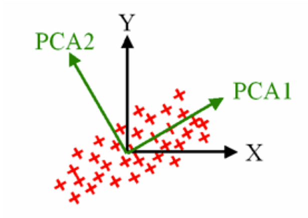

SN 2011fe
Byrne Observatory at Sedgwick Reserve and the Palomar Transient Factory | LCOGT | BJ FULTON

Automatic detection of transients I

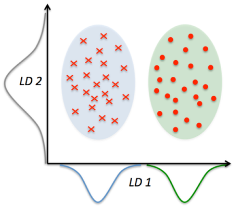
Du Buisson et al. (2015) :

- machine learning detection of SDSS (Sloan) Transient Survey Images to detect supernovae
- Concate to a single vector with 51x51x3 dimension (size of the images*three filters)

- Feature extraction : Principal Component Analysis



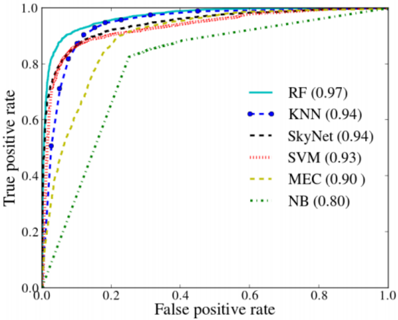
- Feature extraction : Linear Discriminant Analysis



Automatic detection of transients II

They used different classifiers :

- **Random Forest (RF)**
- Minimum Error Classification (**MEC**)
- Naive Bayes (**NB**)
- K-Nearest Neighbours (**KNN**)
- Support Vector Machine (**SVM**)
- Artificial Neural Network (**ANN**)



Discussion about current methods

Problems with the features

- Features are not specific for image processing (Du Buisson et al. (2015))
- Extraction could be incomplete with high level features (Goldstein et al., 2015)

Problems with the classifier

- Features normalization
- Feature extraction step and classification step are separated

What could be the solution ?

Discussion about current methods

Problems with the features

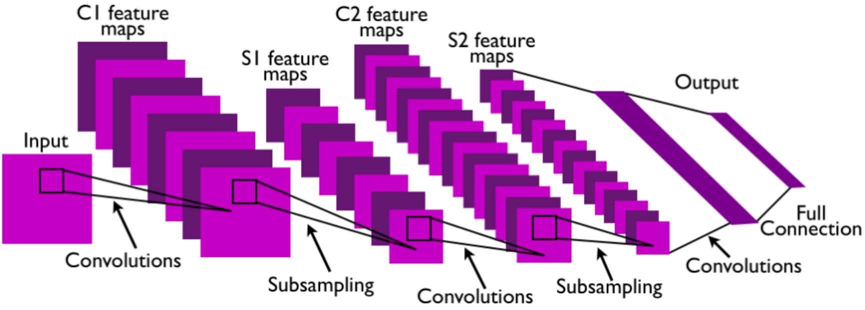
- Features are not specific for image processing (Du Buisson et al. (2015))
- Extraction could be incomplete with high level features (Goldstein et al., 2015)

Problems with the classifier

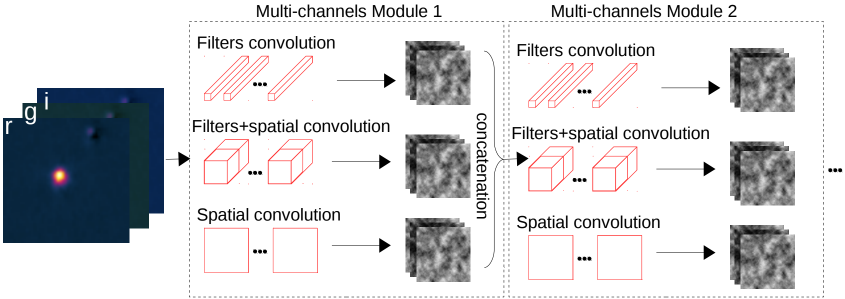
- Features normalization
- Feature extraction step and classification step are separated

What could be the solution ?

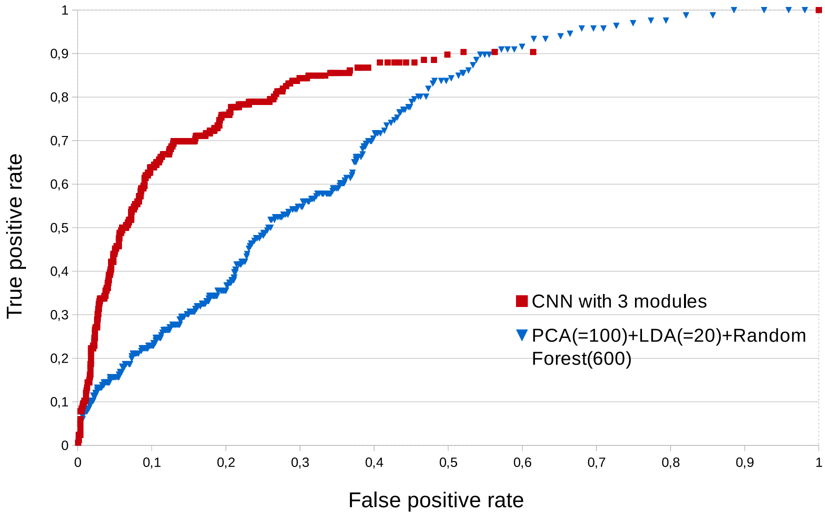
What I propose : Deep learning



My architecture

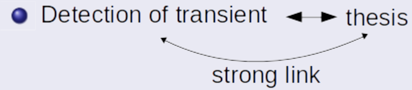


Preliminary results



Conclusions

LSST :



My expertises :

- Expertise photometry
- Automatic Classification methods
- **Future : Deep learning**

Conclusions

LSST :

- Detection of transient ↔ thesis

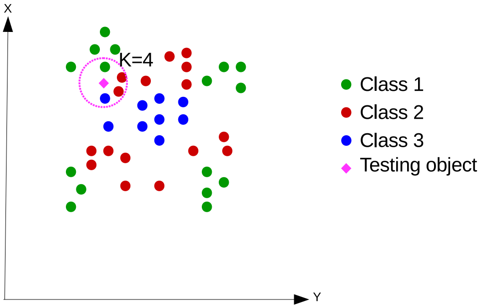
strong link

My expertises :

- Expertise photometry
- Automatic Classification methods
- **Future : Deep learning**

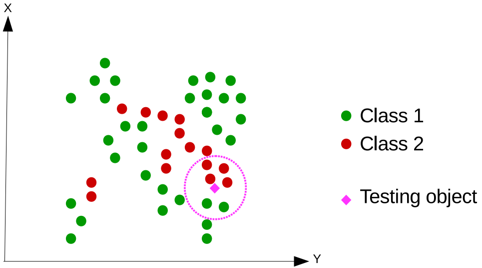
Thank you for your attention !

KNN Classifier



In this example, the KNN algorithm predicts that the testing object belongs to a class 2.

NaiveBayes Classifier



$$P^{prior}_{red} = \frac{15}{46} \qquad \text{Likelihood to give red} \propto \frac{4}{15}$$

$$P^{prior}_{green} = \frac{31}{46} \qquad \text{Likelihood to give green} \propto \frac{2}{31}$$

$$P^{posterior}_{red} = \frac{15}{46} \cdot \frac{4}{15} \simeq 0.08 \qquad P^{posterior}_{green} = \frac{31}{46} \cdot \frac{2}{31} \simeq 0.04$$

In this example, the NaiveBayes algorithm predicts that the testing object belongs to a class red.

Convolution

j

i

35	40	41	45	50
40	40	42	46	32
51	39	29	32	35
25	43	44	30	39
55	35	30	28	41

*

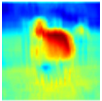
0	0	0
0	0	1
0	1	0

=

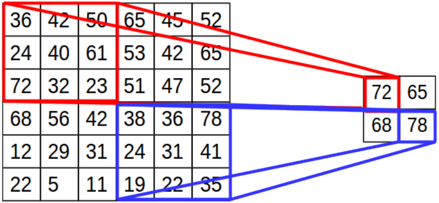
0	0	0	0	0
0	81	75	64	0
0	72	76	65	0
0	79	60	67	0
0	0	0	0	0

*

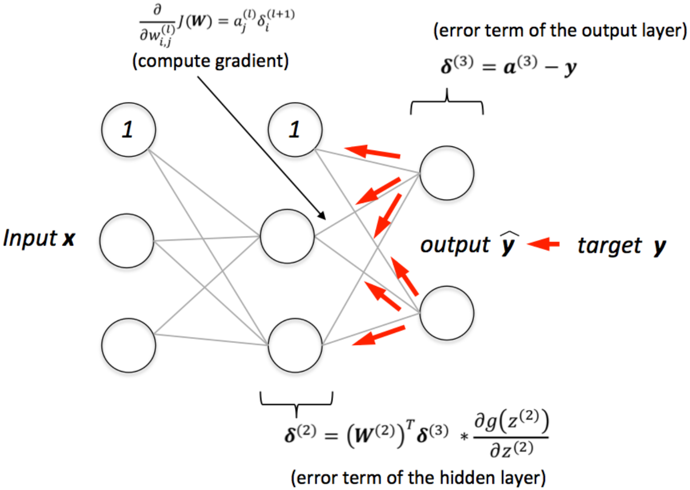
=



Pooling


36 42 50 65 45 52
24 40 61 53 42 65
72 32 23 51 47 52
68 56 42 38 36 78
12 29 31 24 31 41
22 5 11 19 22 35
72 65
68 78

Back-propagation



Physical characteristics of clumps

- Equilibrium of self-gravitating (**virial** equilibrium) :

$$\Omega + 2T = 0 \quad (4)$$

$$R_{vir} = 7au \left(\frac{M}{10^{-3} M_\odot}\right) \left(\frac{T}{10K}\right)^{-1} \quad (5)$$

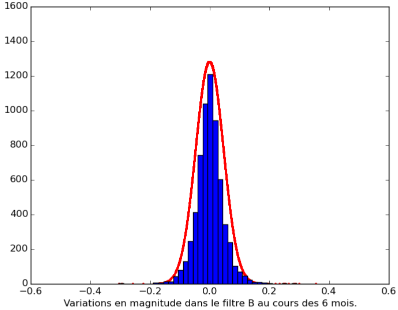
(Lawrence, 2001)

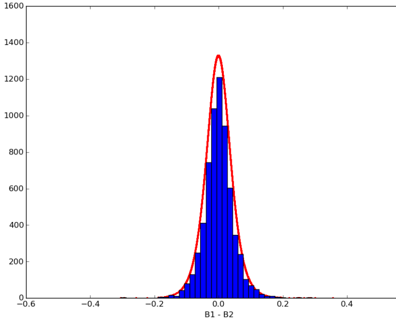
- Once a fragment becomes opaque to its own radiation, it will radiate almost as a blackbody. The mass of the smallest fragment is obtained by considering that the rate of radiation loss $\sim$ the rate of gain in gravitationnal energy.

$$M \sim 0.007 \frac{T^{\frac{1}{4}}}{\mu^{\frac{9}{4}}} M_\odot \quad (6)$$

- If $T \in [3K, 10\,K]$ and $\mu = 2.4$, $M \in [10^{-3} M_\odot\ ,\ 2.10^{-3} M_\odot]$

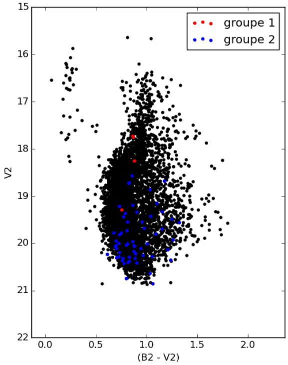
logistique/gaussienne

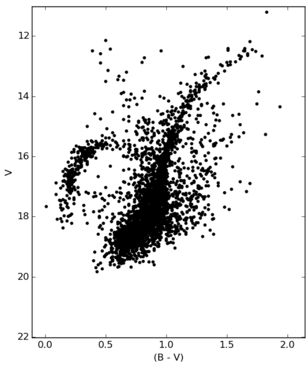




The candidates in the Color Magnitude Diagram

Group 1 : magnitudes with photometric uncertainties < 0.02 mag and beyond $3\,\sigma$ statistical significance
Group 2 : magnitudes with photometric uncertainties ≥ 0.02 mag and beyond a $3\,\sigma$ statistical significance





Structure fractale du nuage : idées générales

- La distribution de taille des sous-structures dans un fractal suit l'équation (Mandelbrot, 1983) :

$$N(\lambda > L) \propto L^{-D} \tag{7}$$

où N est le nombre de structures auto-similaires sur une échelle λ plus grande que L, et D est la dimension fractale.

Structure fractale du nuage : motivations I

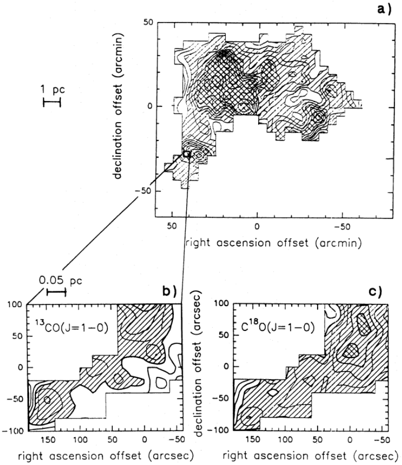


Figure : Carte d'intensité intégrée sur le champ de Persée (Falgarone, 1991)

Structure fractale du nuage : motivations II

- Les nuages interstellaires moléculaires ont une distribution de masse en loi de puissance : $M \propto L^{\kappa}$

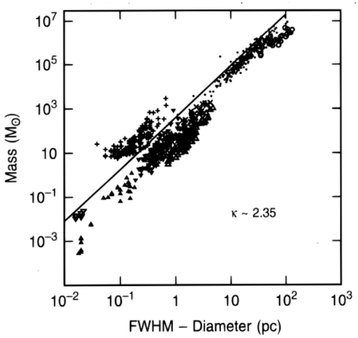


Figure : Masse des nuages versus taille de la FWHM en CO dans l'Ophiuchus, la nébuleuse de la Rosette, le nuage de Maddalena-Thaddeux et des nuages galactiques (Falgarone, 1996)

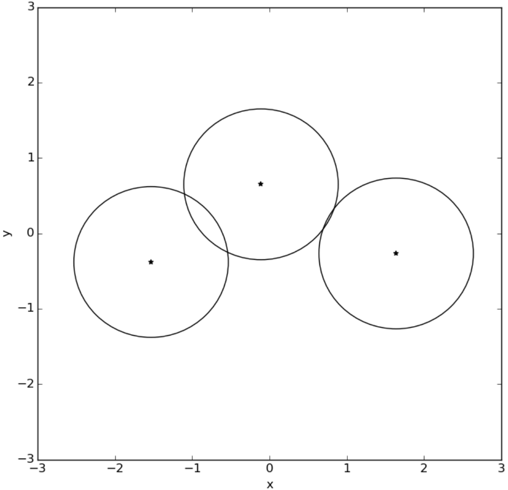
Méthodologie du code fractal

Génération d'un nuage hiérarchique fractal en 3D par fragmentation récursive :

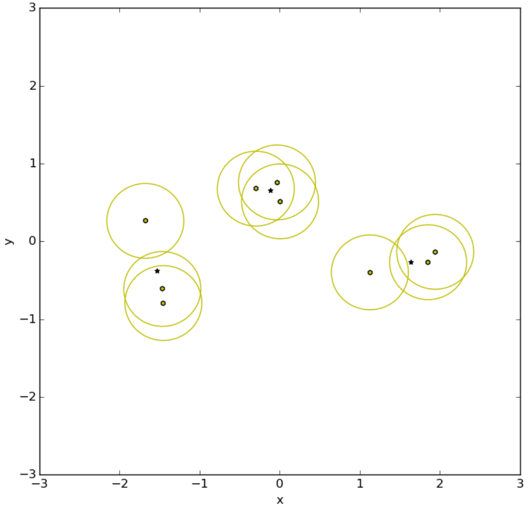
- Distribution aléatoire de N centres de "sous-nuages" dans une sphère de rayon R_{max} selon une loi de densité 3D, $\rho(r, r_L)$ avec r_L l'échelle de longueur
- Réduction de l'échelle de longueur par un facteur spatial de réduction $\alpha = \frac{r_{L-1}}{r_L} = N^{-\frac{1}{D}}$ et redistribution aléatoire de N centres de "sous-sous nuages"
- Au dernier niveau de récursivité : distribution aléatoire de N^L clumps

Méthodologie du code fractal

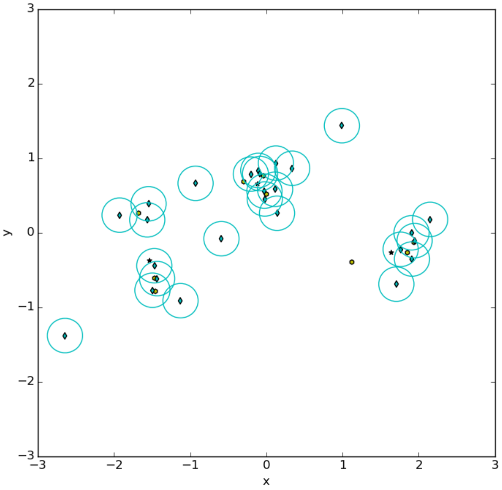
- $N = 3$, $profondeur = 6 \rightarrow 729$ clumps



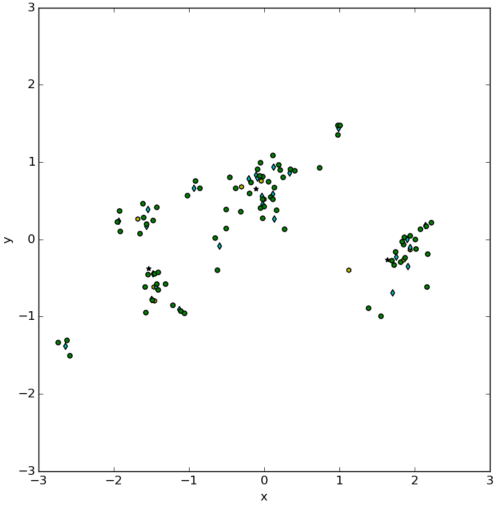
Méthodologie du code fractal



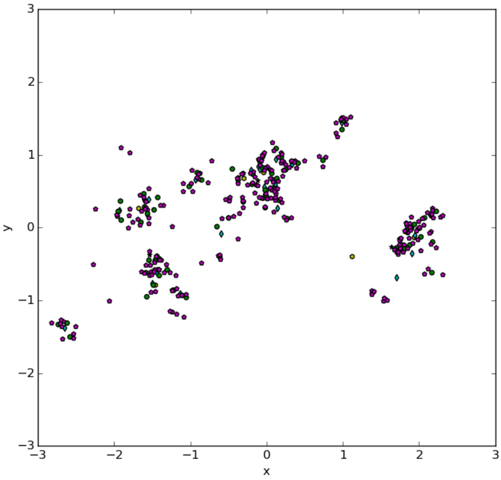
Méthodologie du code fractal



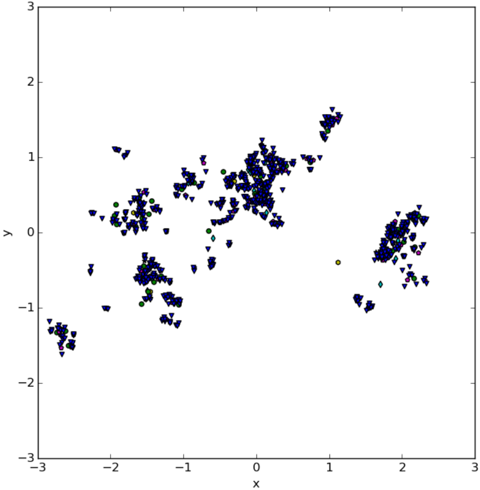
Méthodologie du code fractal



Méthodologie du code fractal



Méthodologie du code fractal

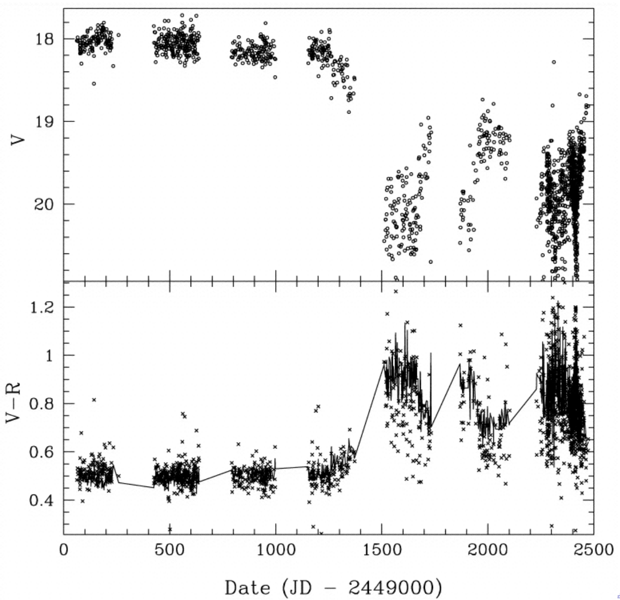


Influence des paramètres de structure

- Simulations : $6 * 100\ k$ étoiles

D=1.5	Plummer	gaussien	pseudo-isotherme
# interactions	13	190	256
$A_{v,max}$ moyen	7.95	5.37	5.63
D=1.8			
# interactions	18	146	46
$A_{v,max}$ moyen	4.15	4.93	3.76
D=2.0			
# interactions	8	21	46
$A_{v,max}$ moyen	4.64	7.68	5.48

Drake et Cook



Arbre de décision