Participation in PLAsTiCC

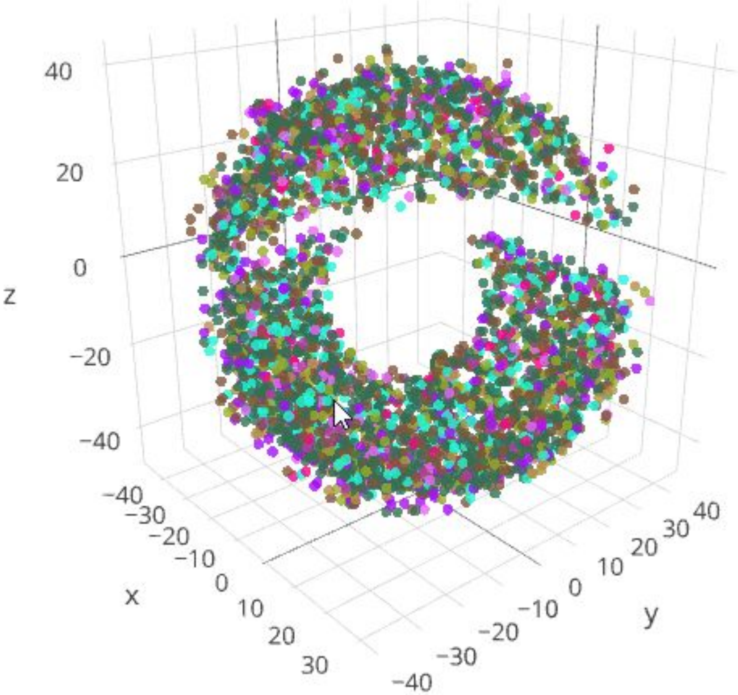
B. Chazelle
Y. Ghernouz
A. Loup
L. Rouquette

Summary

- Context
- Dataset difficulties
- Tools
- Data transformation
- Learning models

Context

- PLAsTiCC Contest
    - Real problem contest powered by LSST
    - Main goal : star classifications based on astrophysical data
- School specific project & Scientific synthesis
    - Mix between scientific monitoring and practical work
    - ~0.5 day a week + extra personal work
    - 4 months duration



https://github.com/Quad-INSA-5IF/Astronomical_photometric_classification

Dataset difficulties

- Very large test dataset
  - Cannot be loaded on RAM with some technologies (Java / Python)
  - Can lead to high computation time with computation consuming models
- Distinct distributions between the training data and the test data
- Unknown "rest of the world" class
- Missing & noised data on time series
- Some classes are poorly represented

Tools

Processing

- Java & Kotlin
  - No Framework
- Python
  - Numpy
  - Pandas
  - Scikit-learn

Visualization

- Usage of MatPlotLib (Python)
- D3.js (Javascript)
- XChart (JVM environment)

Data Transformation

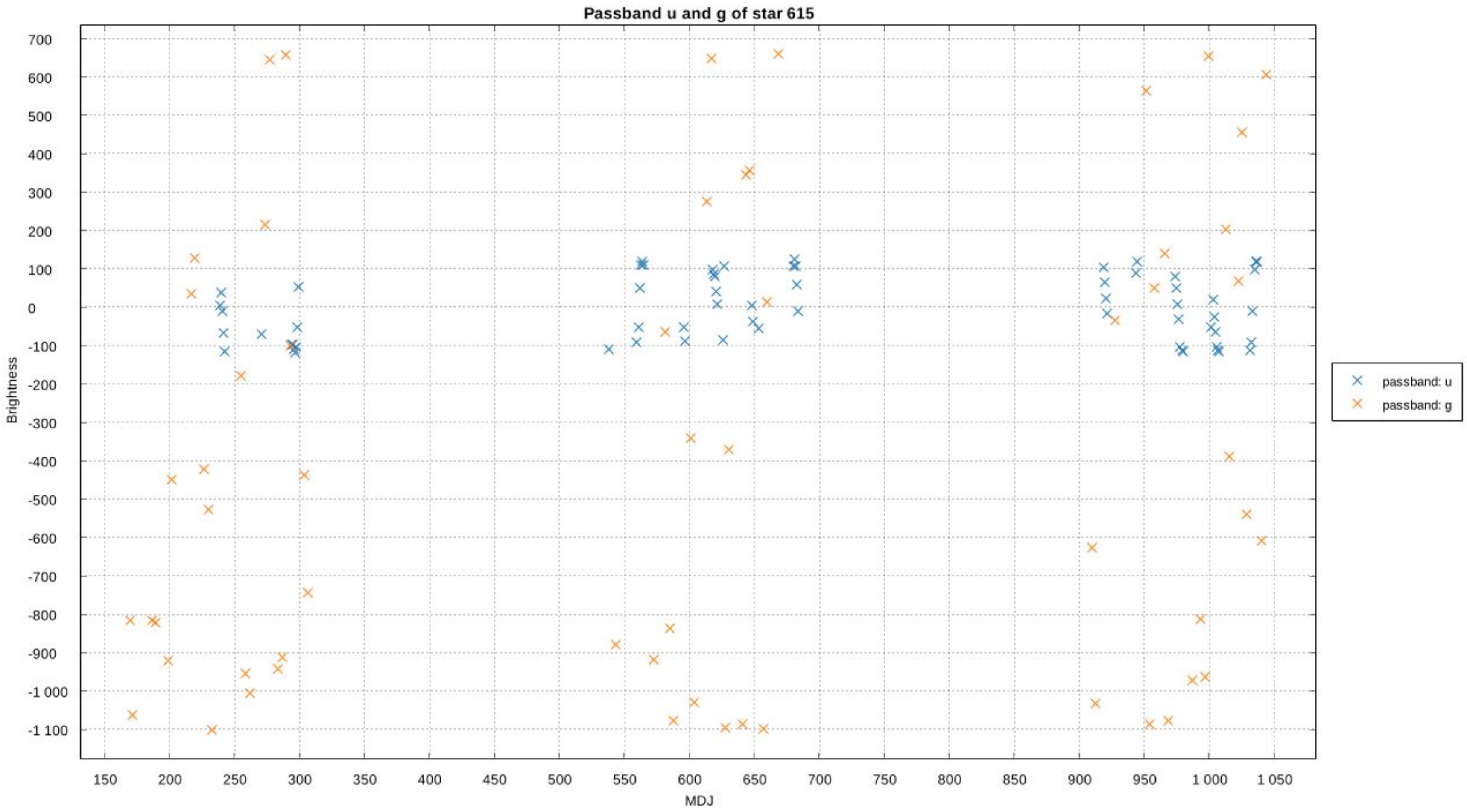
Purposes :

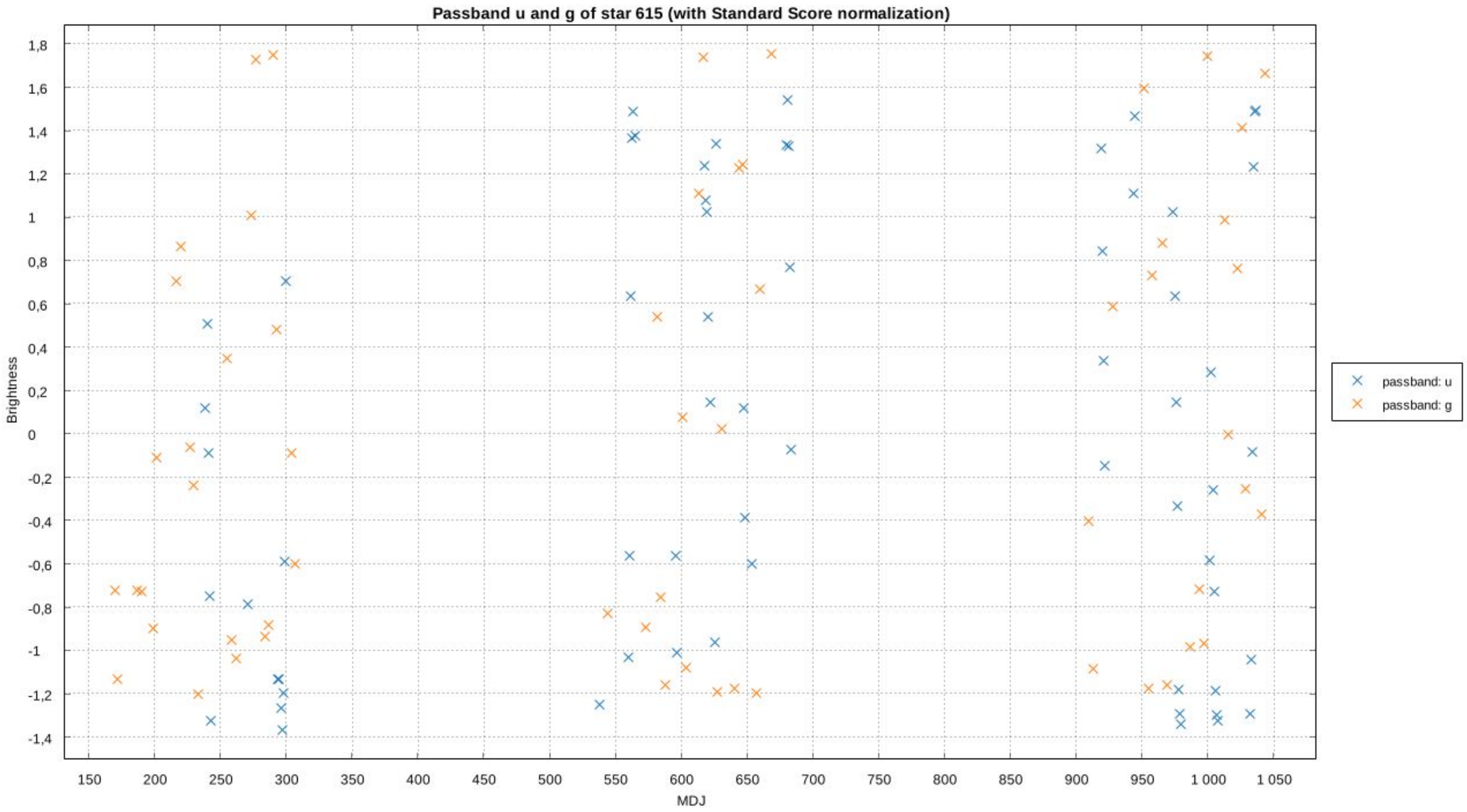
- Reduce the impact of different scales (domains)
- Help the computer to handle NaN values
- Remove outliers (when required)
- Transform data into computable values

Data Transformation

We may use different normalizations:

Standard Score	value = (value - mean) / std
Feature scaling	value = (value - min) / (max - min)


Passband u and g of star 615
Brightness
700
600
500
400
300
200
100
0
-100
-200
-300
-400
-500
-600
-700
-800
-900
-1 000
-1 100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
MDJ
passband: u
passband: g



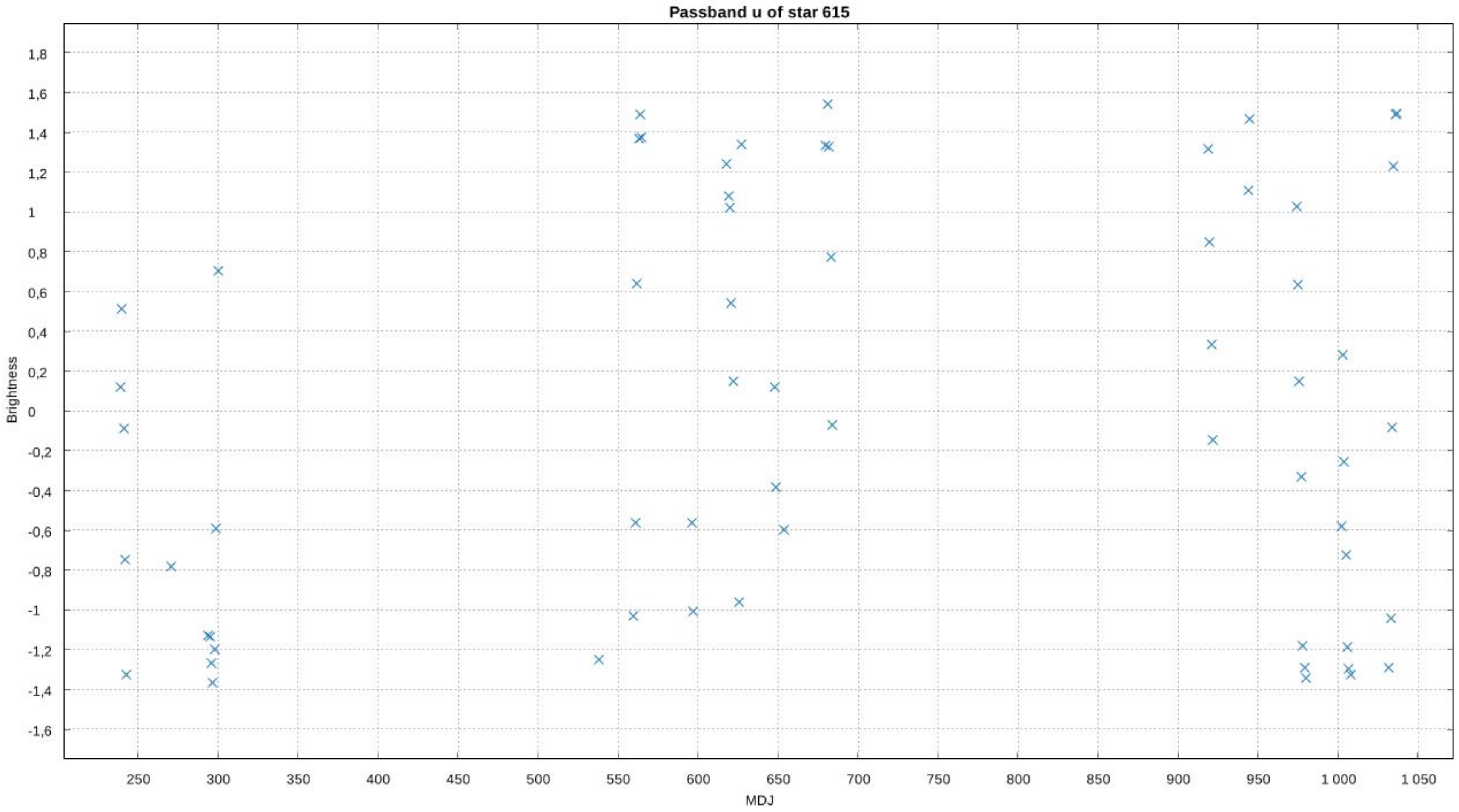
Passband u and g of star 615 (with Standard Score normalization)
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
MDJ
passband: u
passband: g

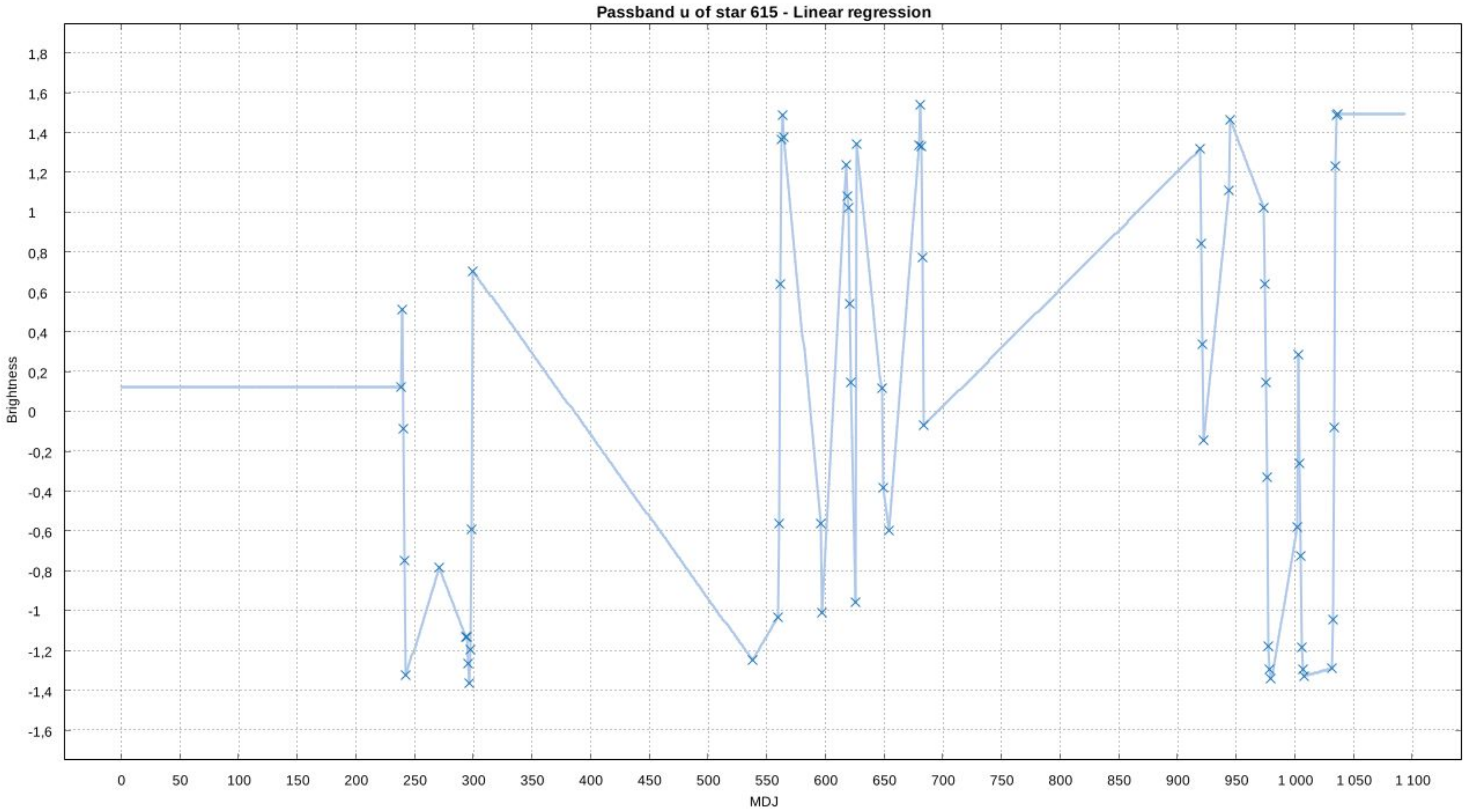
Data Transformation

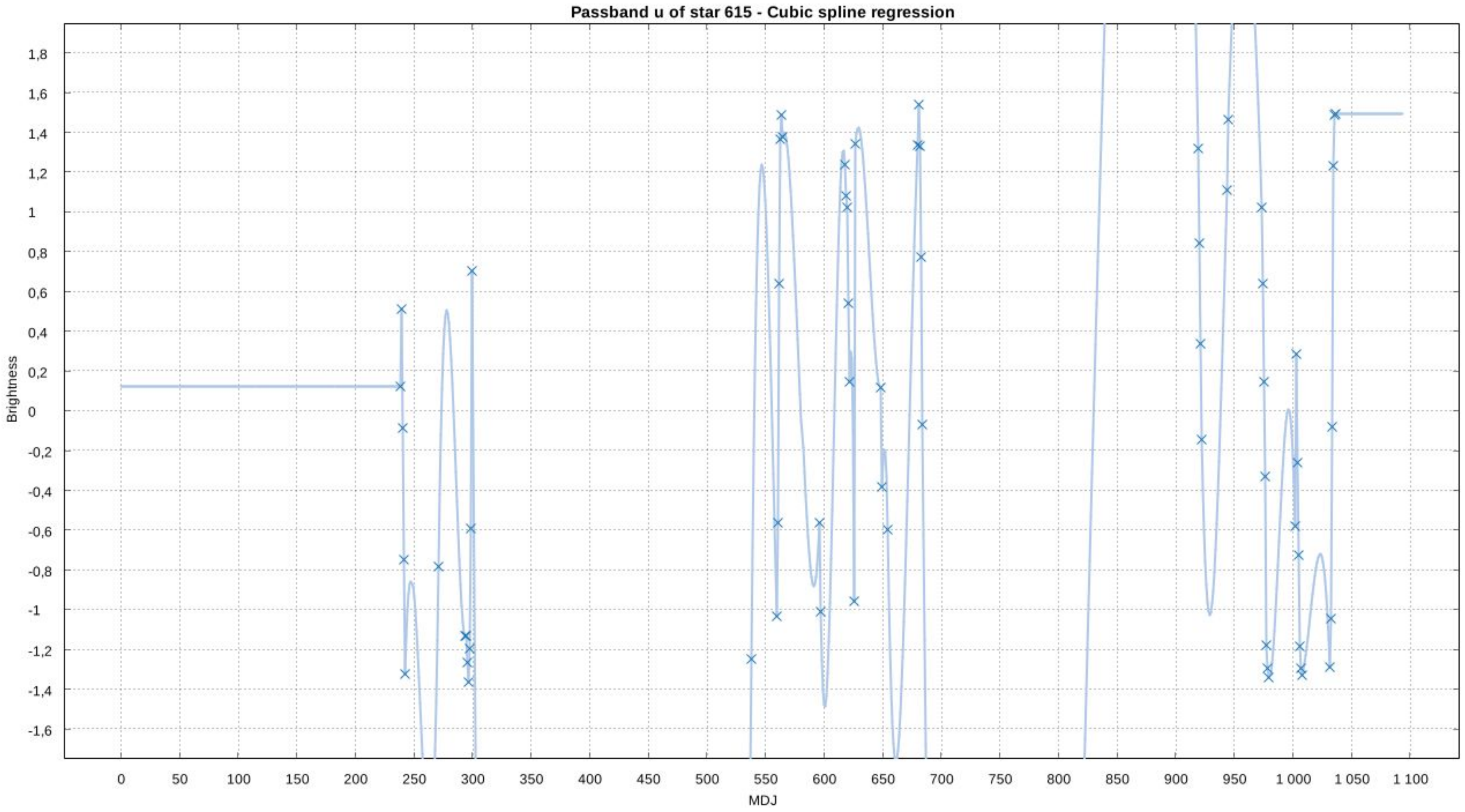
Replacing missing values

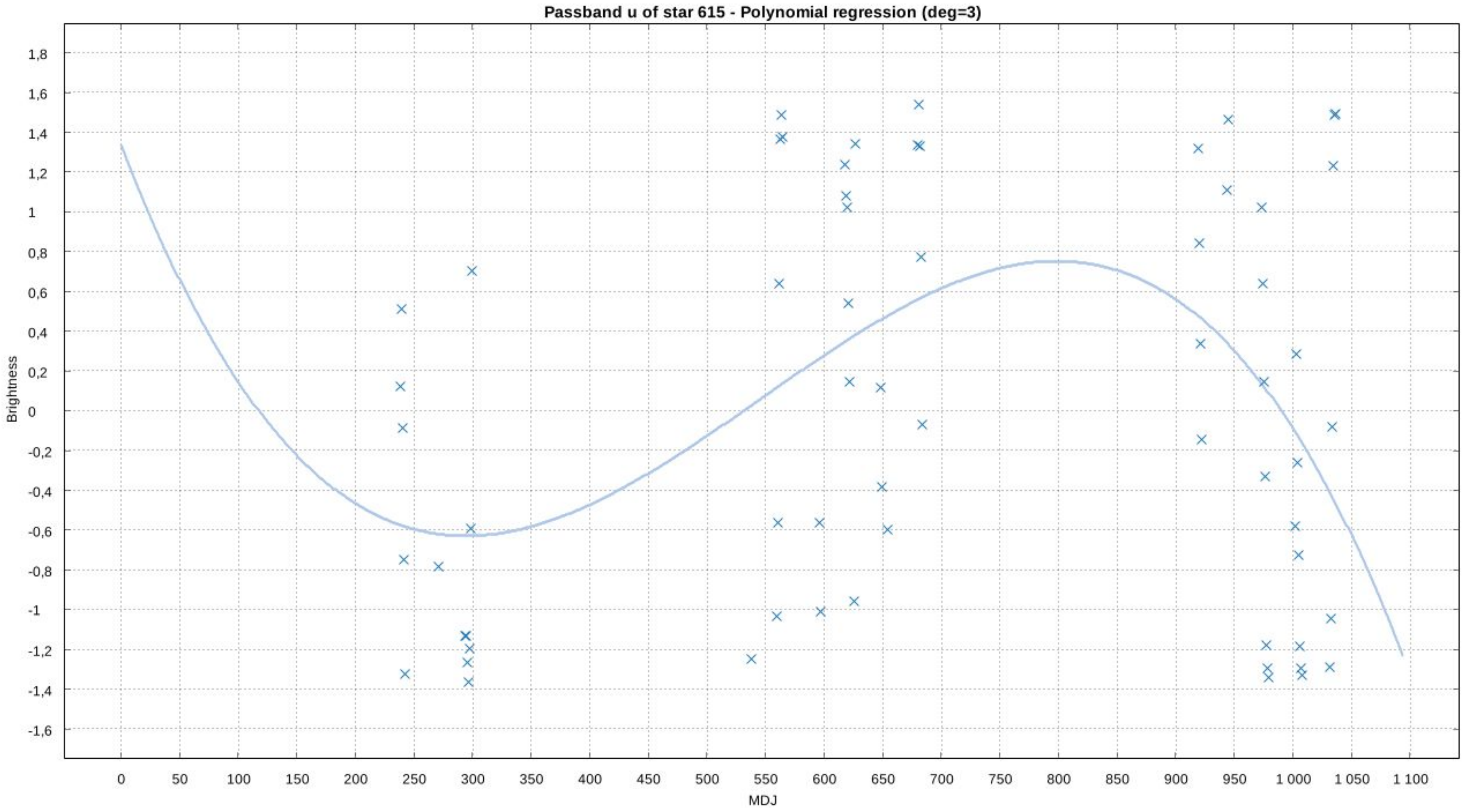
- Linear regression
- Cubic spline
- Polynomial regression
- Random point generation [1]
- Normalization

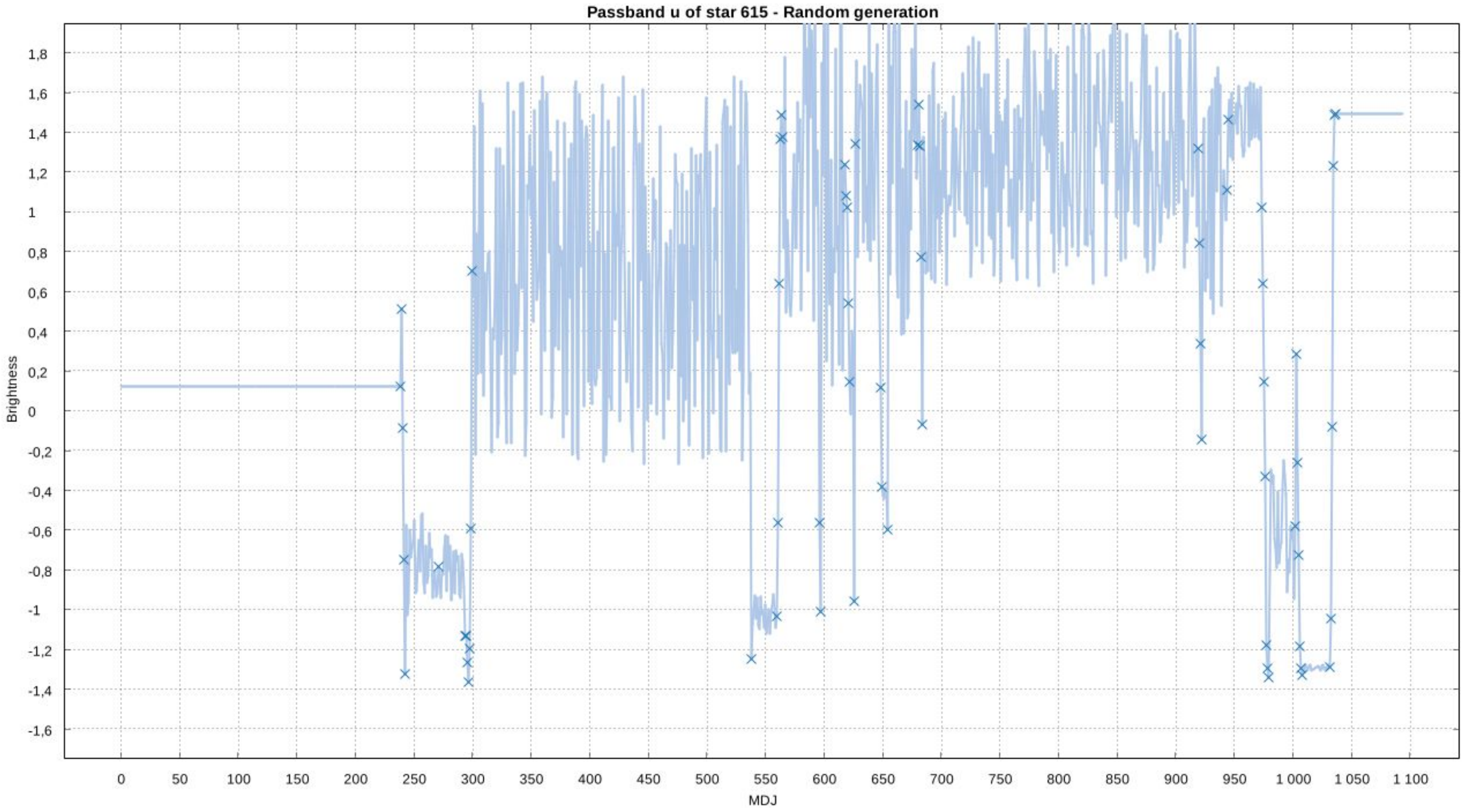
[1] T. A. Hinners, K. Tat, et R. Thorp, « Machine Learning Techniques for Stellar Light Curve Classification », *The Astronomical Journal*, vol. 156, n^{o} 1, p. 7, juin 2018. Disponible à https://arxiv.org/pdf/1710.06804.pdf

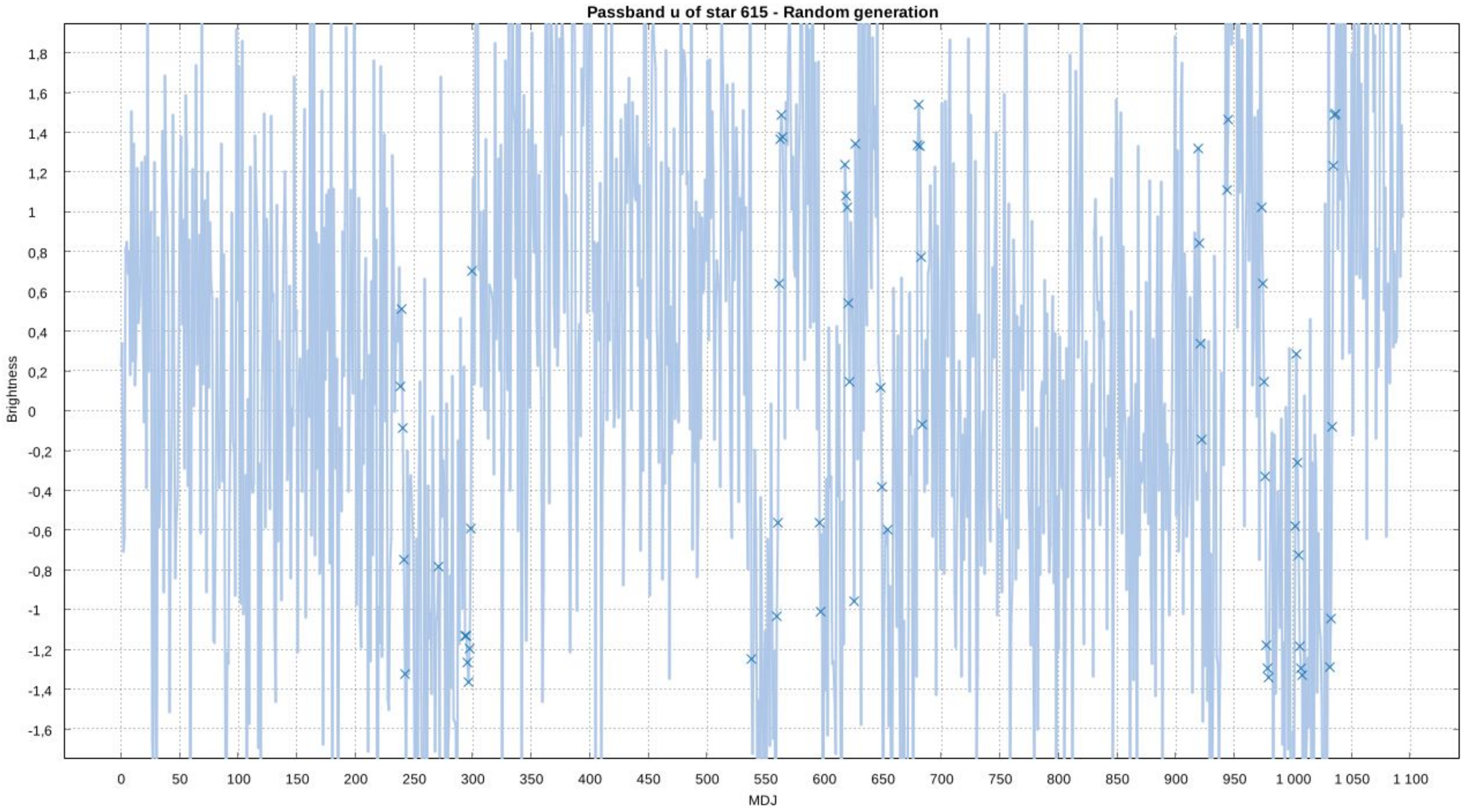

Passband u of star 615
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
MDJ

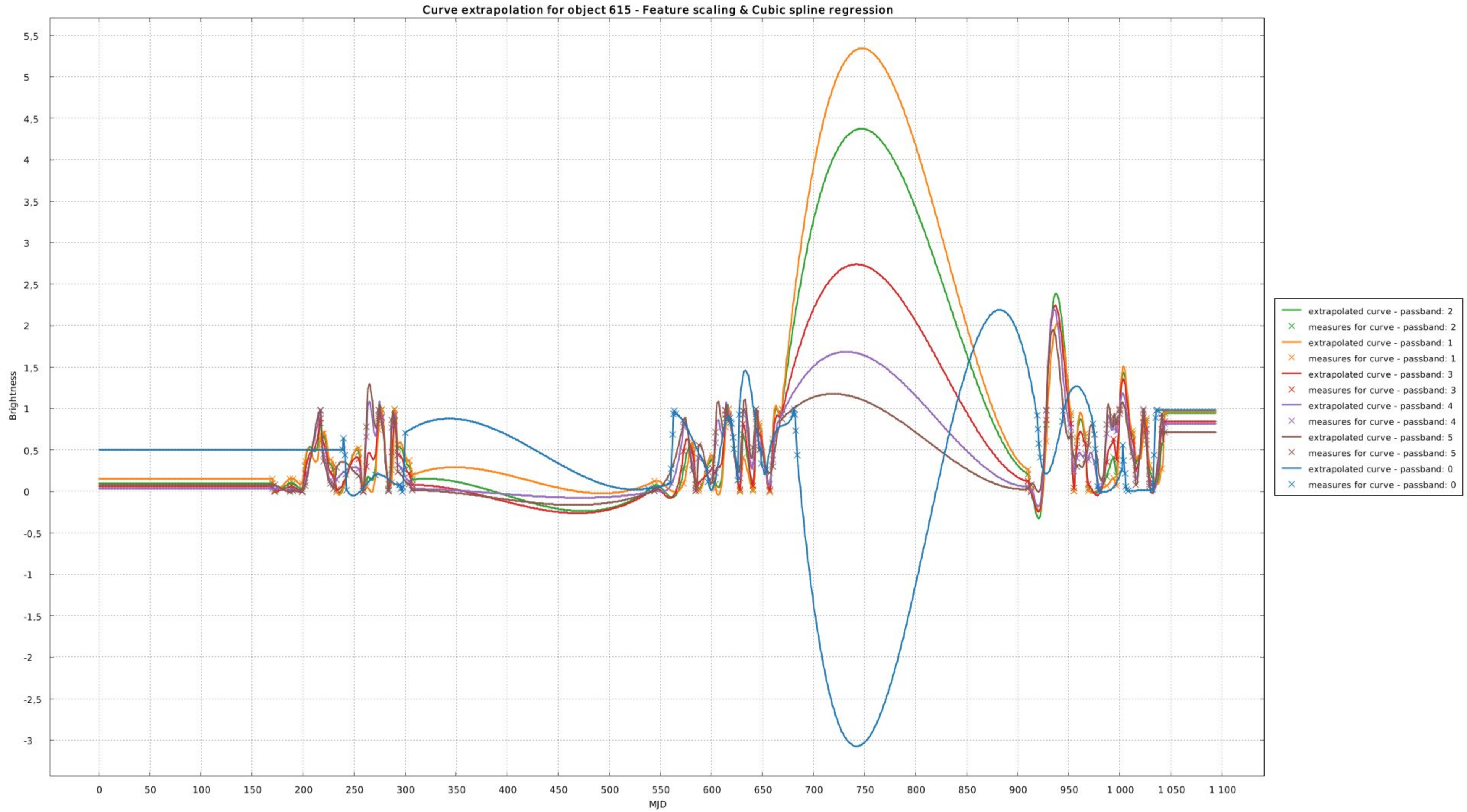

Passband u of star 615 - Linear regression
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MDJ

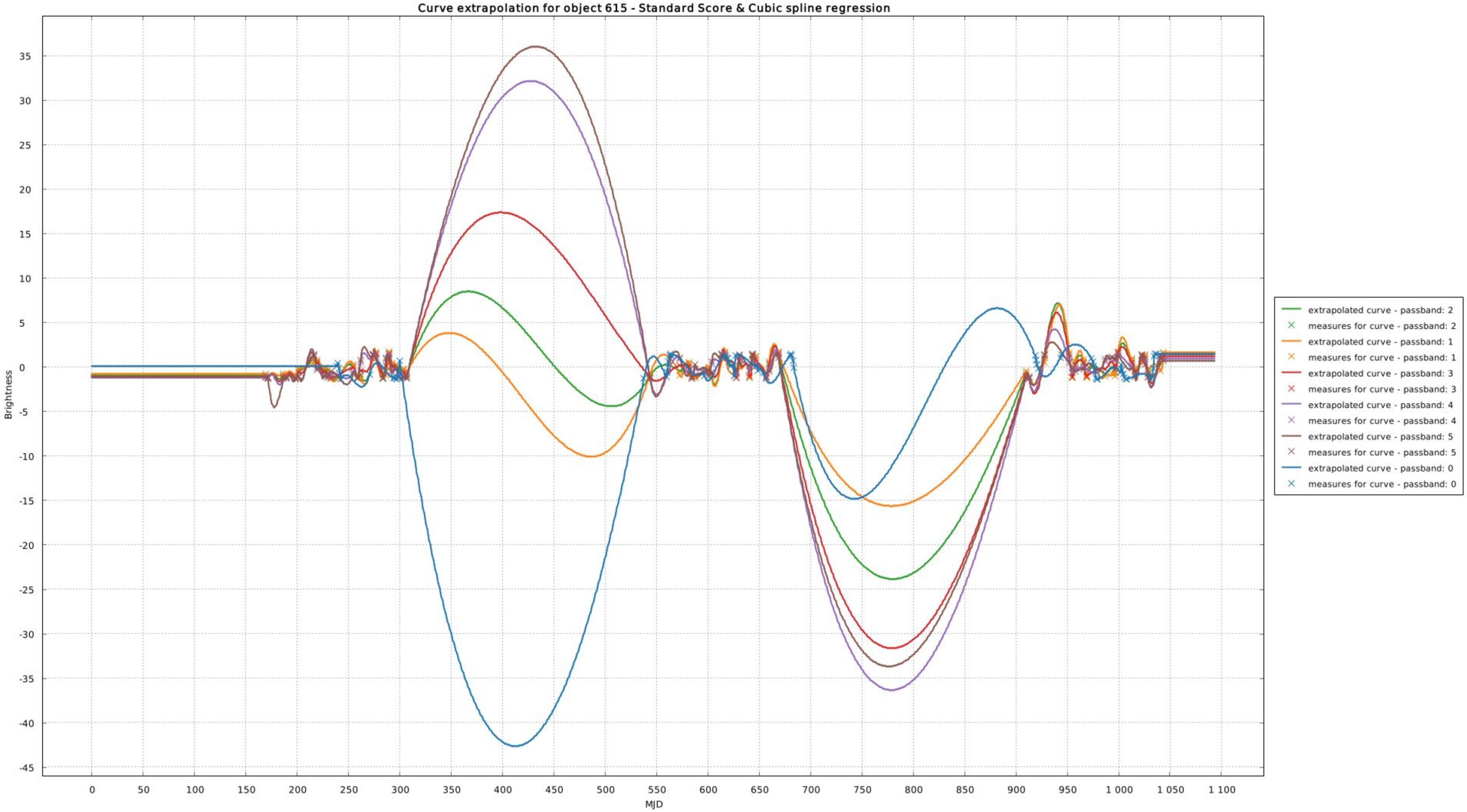

Passband u of star 615 - Cubic spline regression
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
MDJ
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100


Passband u of star 615 - Polynomial regression (deg=3)
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MDJ

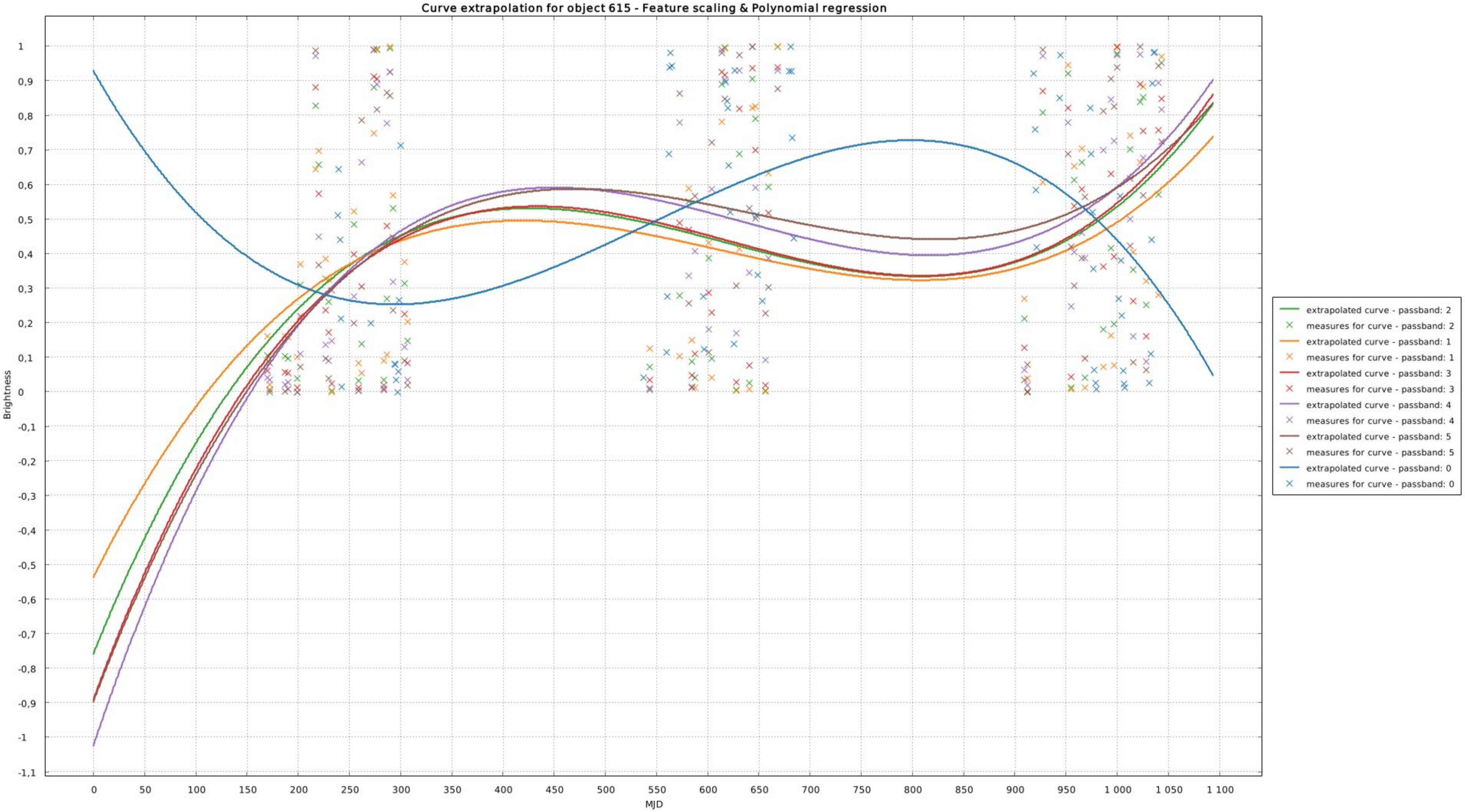
Passband u of star 615 - Random generation
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MDJ

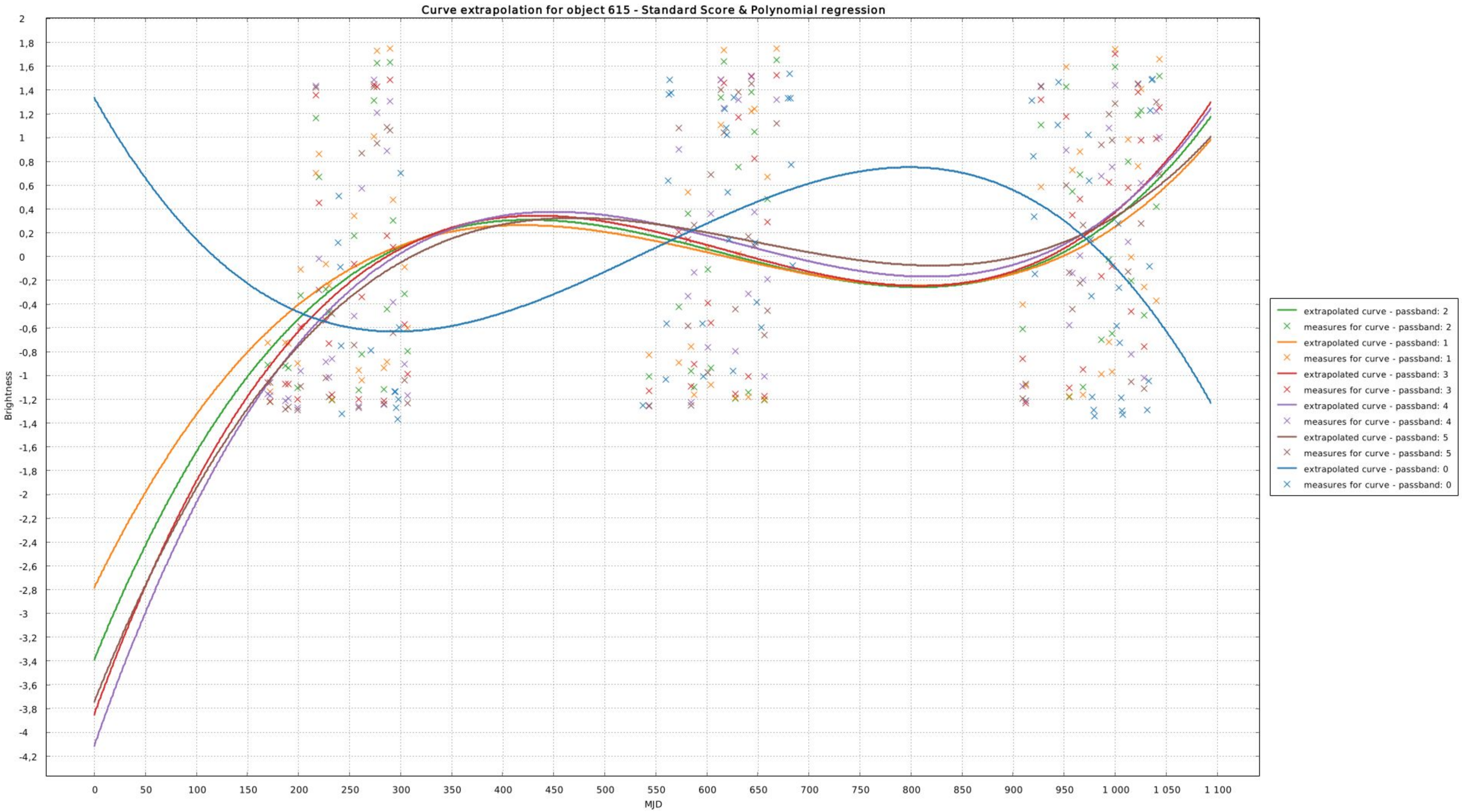

Passband u of star 615 - Random generation
Brightness
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MDJ

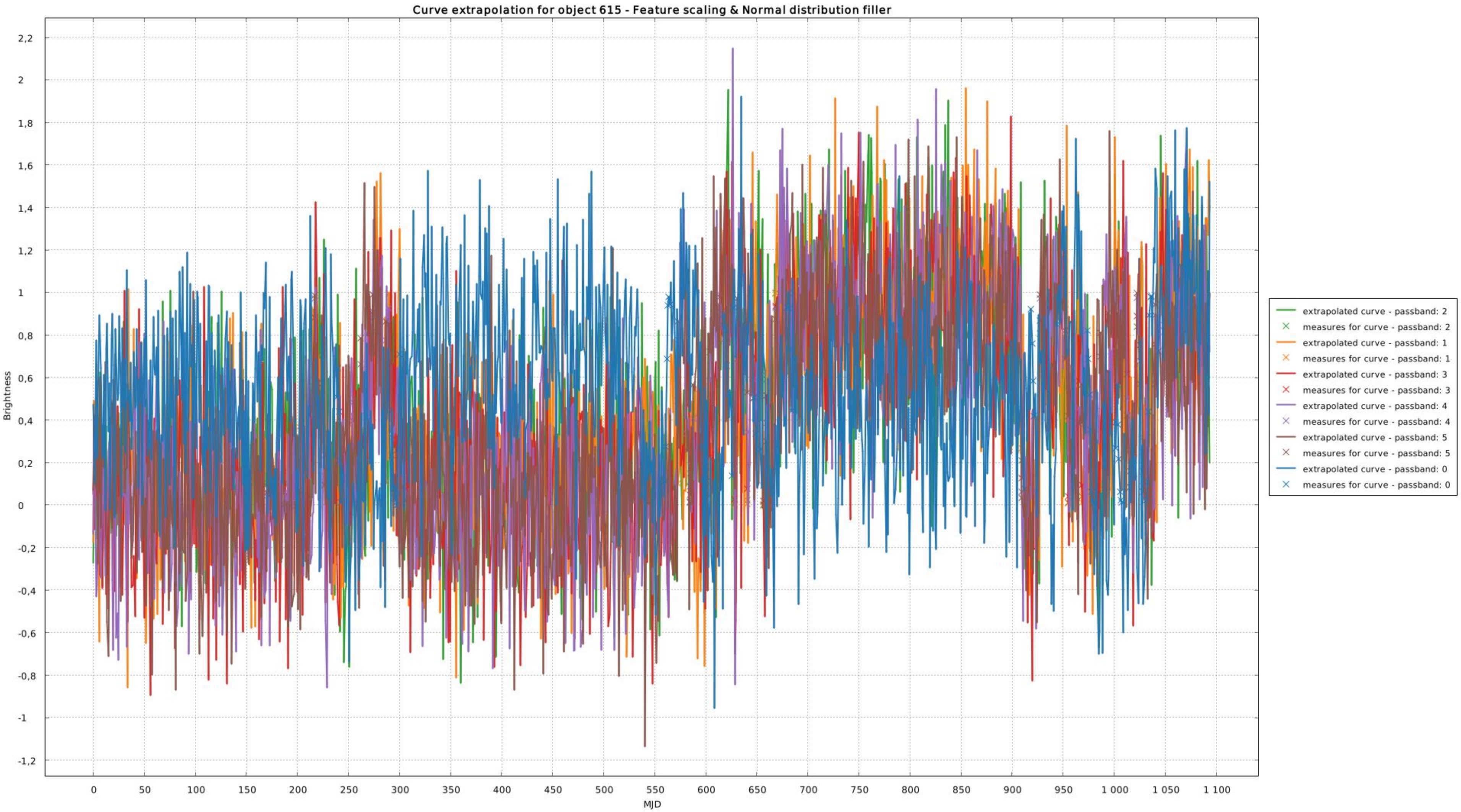

Curve extrapolation for object 615 - Feature scaling & Cubic spline regression
Brightness
MJD
5,5
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
-1
-1,5
-2
-2,5
-3
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
extrapolated curve - passband: 2
measures for curve - passband: 2
extrapolated curve - passband: 1
measures for curve - passband: 1
extrapolated curve - passband: 3
measures for curve - passband: 3
extrapolated curve - passband: 4
measures for curve - passband: 4
extrapolated curve - passband: 5
measures for curve - passband: 5
extrapolated curve - passband: 0
measures for curve - passband: 0

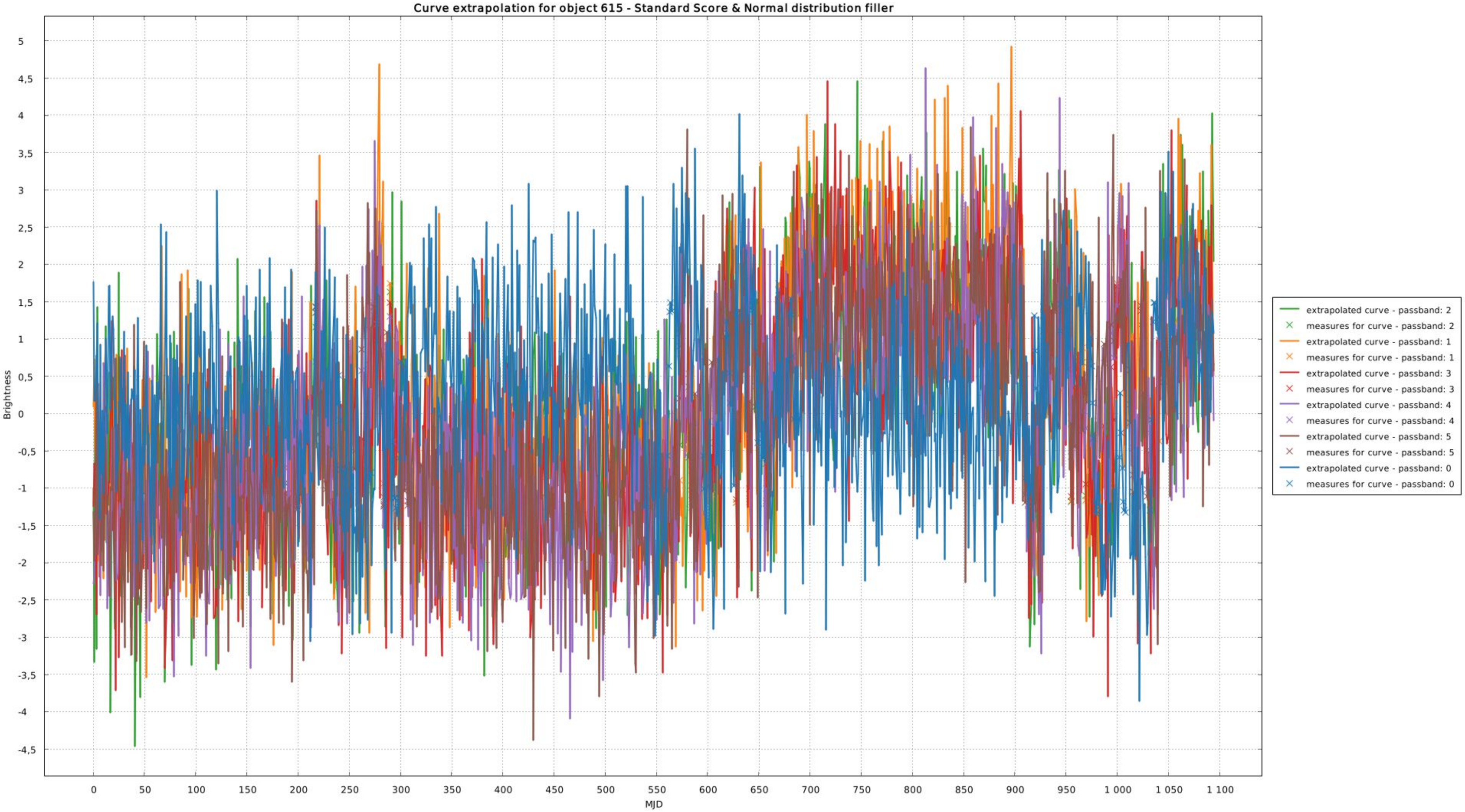


Curve extrapolation for object 615 - Standard Score & Cubic spline regression


Curve extrapolation for object 615 - Feature scaling & Polynomial regression
Brightness
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
-0,2
-0,3
-0,4
-0,5
-0,6
-0,7
-0,8
-0,9
-1
-1,1
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MJD
extrapolated curve - passband: 2
measures for curve - passband: 2
extrapolated curve - passband: 1
measures for curve - passband: 1
extrapolated curve - passband: 3
measures for curve - passband: 3
extrapolated curve - passband: 4
measures for curve - passband: 4
extrapolated curve - passband: 5
measures for curve - passband: 5
extrapolated curve - passband: 0
measures for curve - passband: 0


Curve extrapolation for object 615 - Standard Score & Polynomial regression
Brightness
MJD
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
-1,4
-1,6
-1,8
-2
-2,2
-2,4
-2,6
-2,8
-3
-3,2
-3,4
-3,6
-3,8
-4
-4,2
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
extrapolated curve - passband: 2
measures for curve - passband: 2
extrapolated curve - passband: 1
measures for curve - passband: 1
extrapolated curve - passband: 3
measures for curve - passband: 3
extrapolated curve - passband: 4
measures for curve - passband: 4
extrapolated curve - passband: 5
measures for curve - passband: 5
extrapolated curve - passband: 0
measures for curve - passband: 0


Curve extrapolation for object 615 - Feature scaling & Normal distribution filler
Brightness
MJD
2,2
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
-1,2
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
extrapolated curve - passband: 2
measures for curve - passband: 2
extrapolated curve - passband: 1
measures for curve - passband: 1
extrapolated curve - passband: 3
measures for curve - passband: 3
extrapolated curve - passband: 4
measures for curve - passband: 4
extrapolated curve - passband: 5
measures for curve - passband: 5
extrapolated curve - passband: 0
measures for curve - passband: 0


Curve extrapolation for object 615 - Standard Score & Normal distribution filler
Brightness
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
-1
-1,5
-2
-2,5
-3
-3,5
-4
-4,5
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1 000
1 050
1 100
MJD
extrapolated curve - passband: 2
measures for curve - passband: 2
extrapolated curve - passband: 1
measures for curve - passband: 1
extrapolated curve - passband: 3
measures for curve - passband: 3
extrapolated curve - passband: 4
measures for curve - passband: 4
extrapolated curve - passband: 5
measures for curve - passband: 5
extrapolated curve - passband: 0
measures for curve - passband: 0

Learning models

Deep learning

Pros	Cons
● May integrate some noise in the model ● Do not require important data transformation and analysis	● Cannot handle properly the "rest of the world" class ● Sensible to class distribution ● No explainable ● Need very large sets

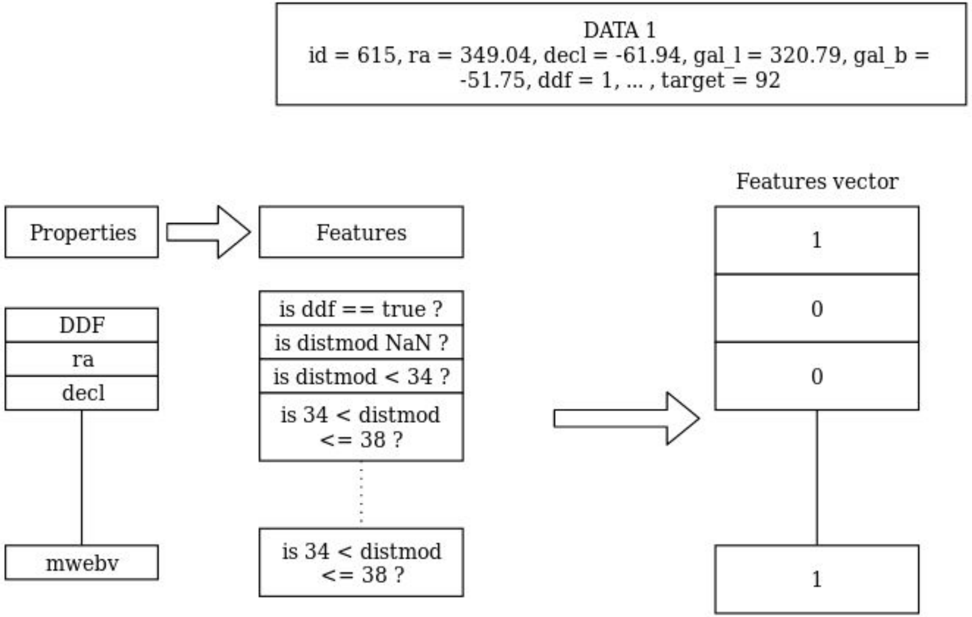
Learning models

Decision Tree

Pros	Cons
● Can handle missing data (axis are not required to be continued) ● Can manage nominal properties ● Explainable if the size is not big	● Overfit quickly ● Cannot manage an unknown class

Learning models

KNN + Features engineering + cos similarity



DATA 1
id = 615, ra = 349.04, decl = -61.94, gal_l = 320.79, gal_b = -51.75, ddf = 1, ... , target = 92

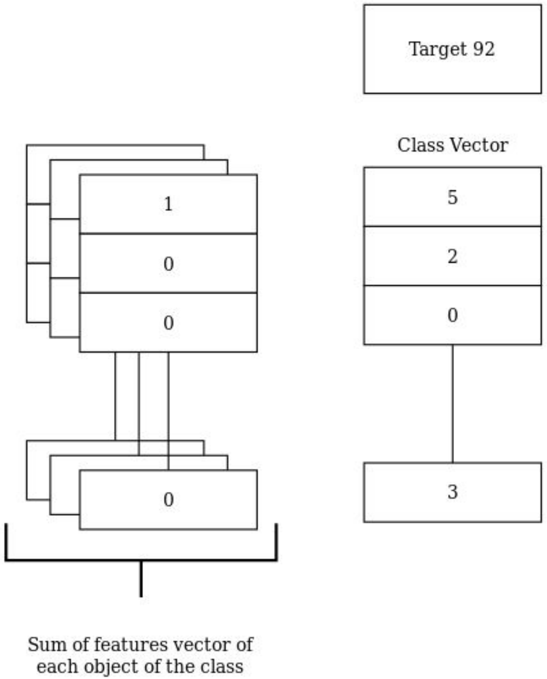
Properties

DDF
ra
decl
mwebv

Features

is ddf == true ?
is distmod NaN ?
is distmod < 34 ?
is 34 < distmod <= 38 ?
is 34 < distmod <= 38 ?

Features vector



Target 92

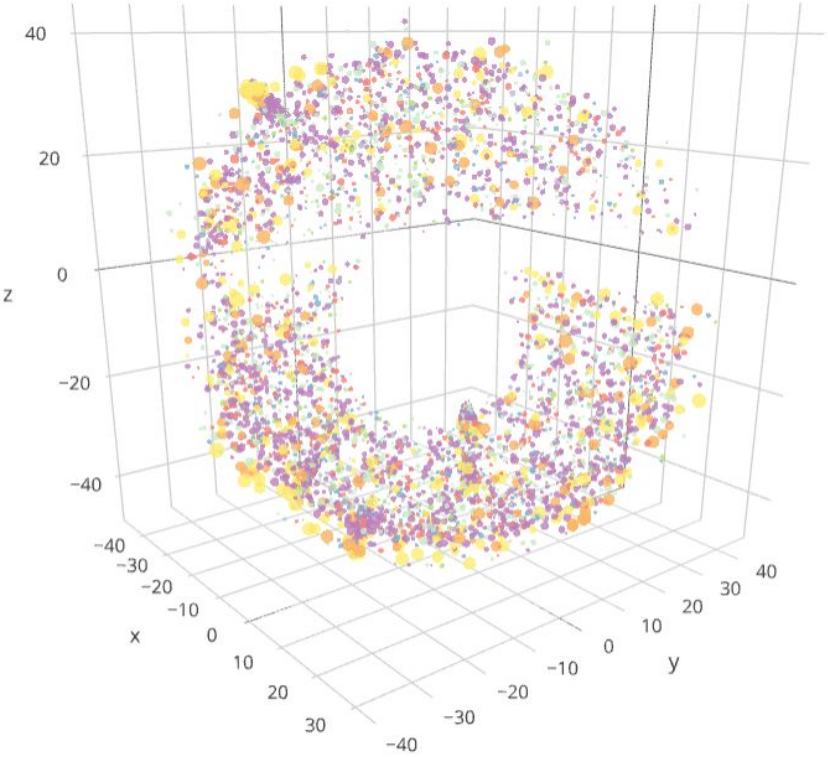
Class Vector

Sum of features vector of each object of the class

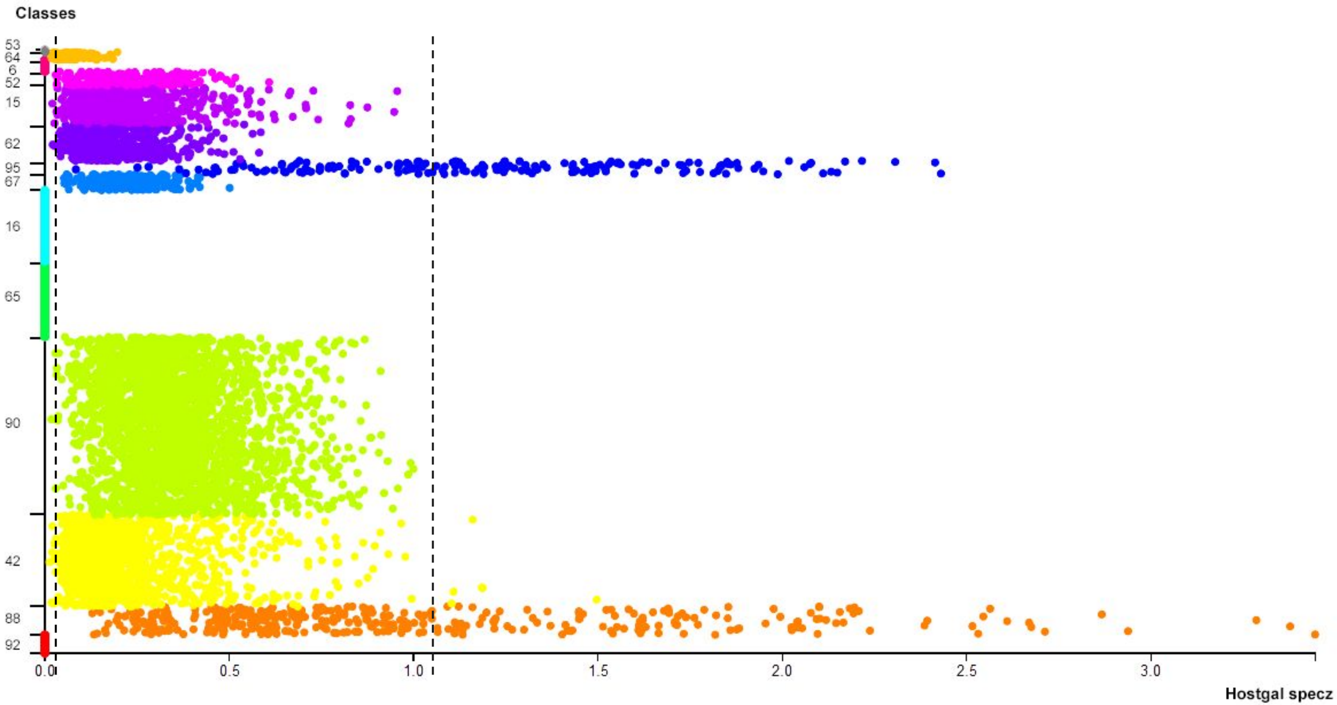
Learning models

KNN + Features engineering + cos similarity

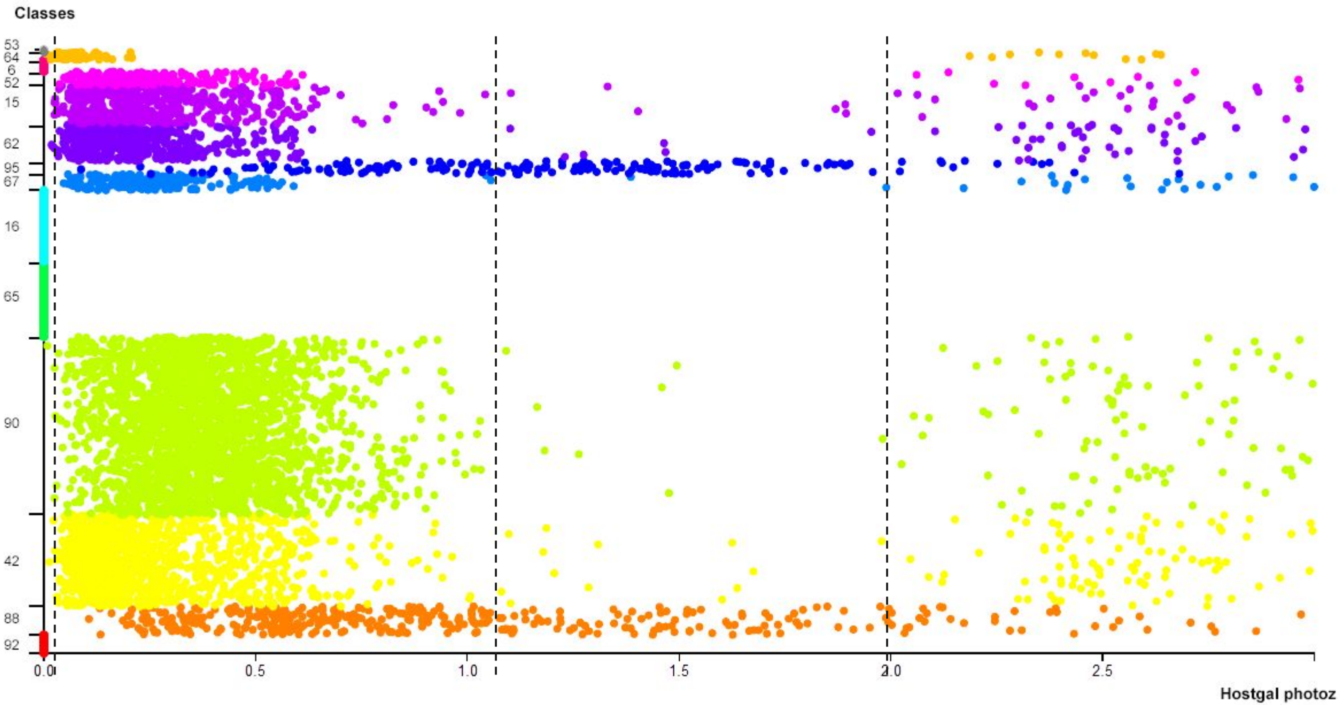
Pros	Cons
● Easy to create ● Can manage the unknown class	● Required to extract meaningful features (very complex task)


40
20
0
z
−20
−40
−40
−30
−20
−10
0
x
10
20
30
40
30
20
10
0
y
−10
−20
−30
−40

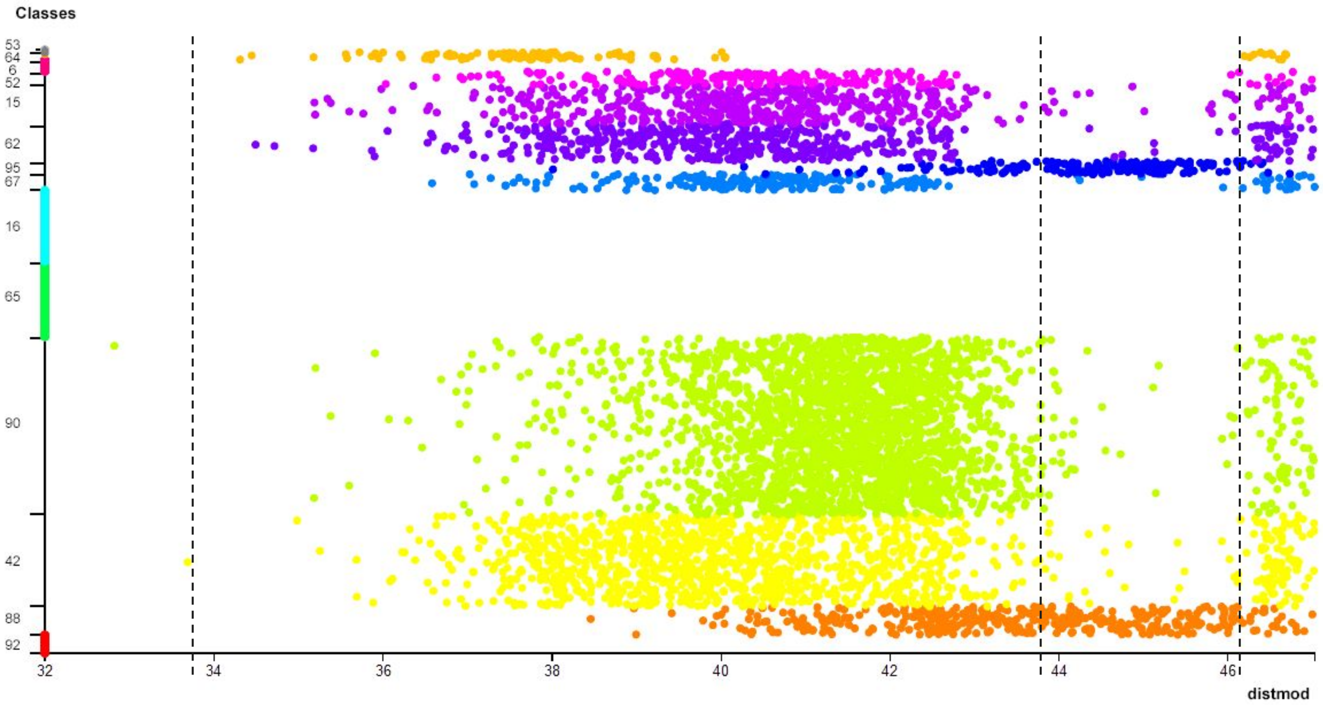
Features engineering



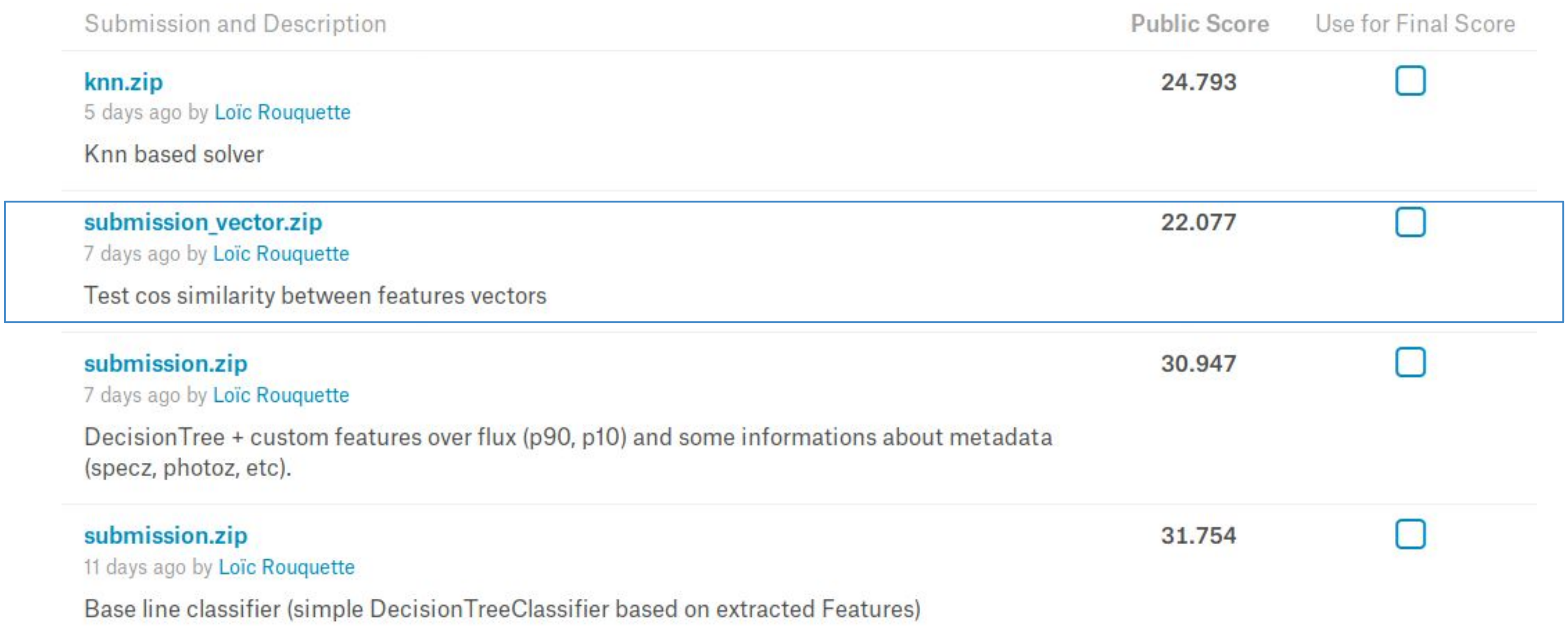
Features engineering



Features engineering



Our results on PLAsTiCC



Submission and Description	Public Score	Use for Final Score
knn.zip 5 days ago by Loïc Rouquette Knn based solver	24.793	□
submission_vector.zip 7 days ago by Loïc Rouquette Test cos similarity between features vectors	22.077	□
submission.zip 7 days ago by Loïc Rouquette DecisionTree + custom features over flux (p90, p10) and some informations about metadata (specz, photoz, etc).	30.947	□
submission.zip 11 days ago by Loïc Rouquette Base line classifier (simple DecisionTreeClassifier based on extracted Features)	31.754	□

Thanks for your attention