Efficient Catalog Cross-matching Sky Mining Hackastron 2018

Team members:Kristof, Karl & Casey

Our Goal

Take the 50 catalogues

- Cross match them (49 comparisons) and create light curves with the unique measurement in each epoch
- Report the row ID for each source in each epoch
- Sort the answer key by epoch00 row ID

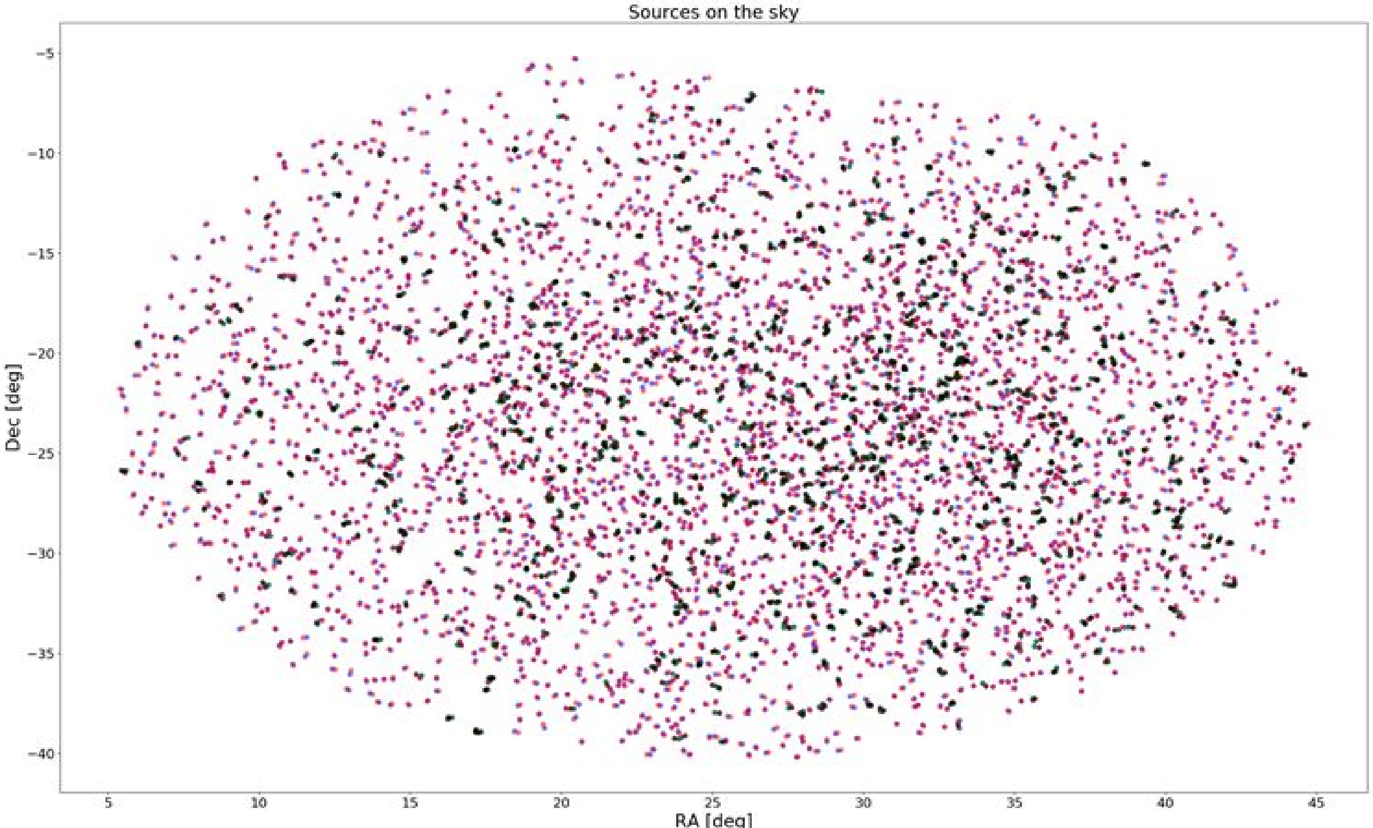
Our Approach

- Simple check: if have ‘Clear Neighbour’, no further analysis

Supplemented with

- Complex check: Hungarian Tinder

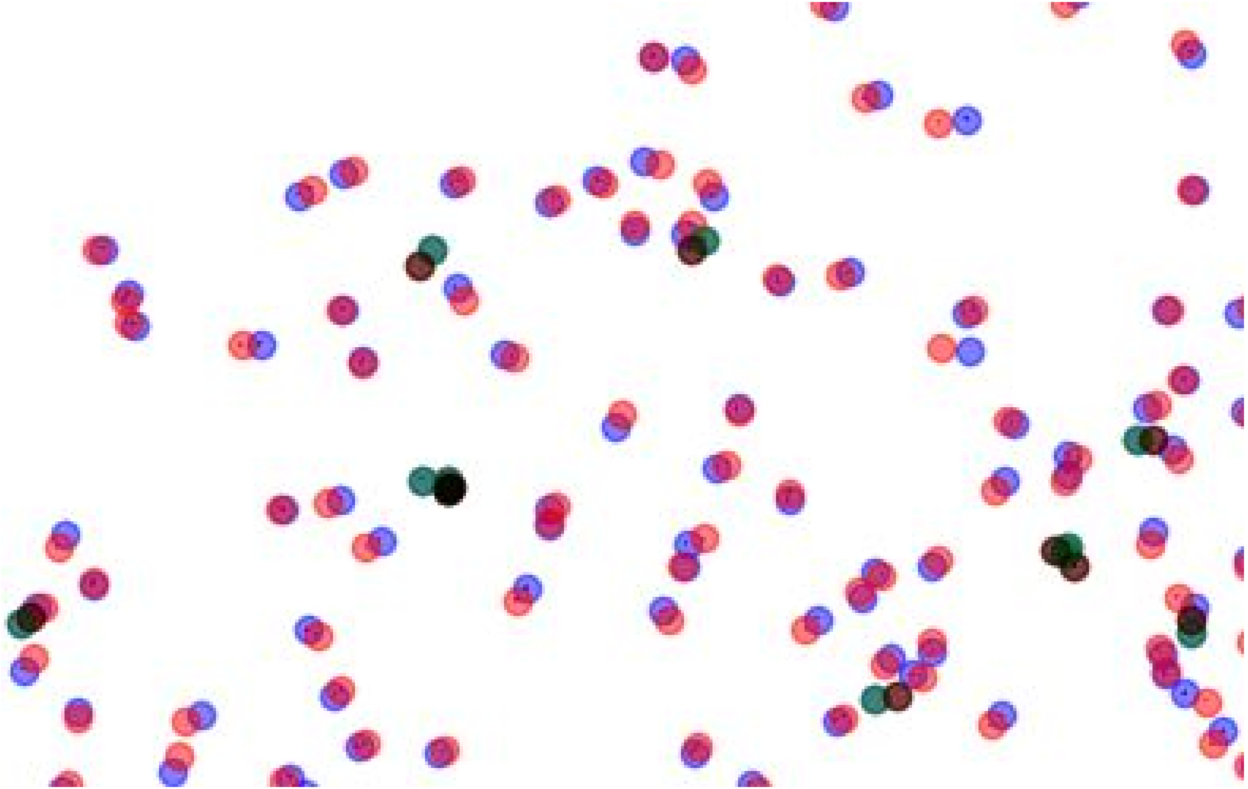
Simple check: ‘Clear neighbour’, Unoptimised= 77%.



1. Neighbr1 vs Nbrs 2 distance ratio
2. Checks for duplicate neighbrs

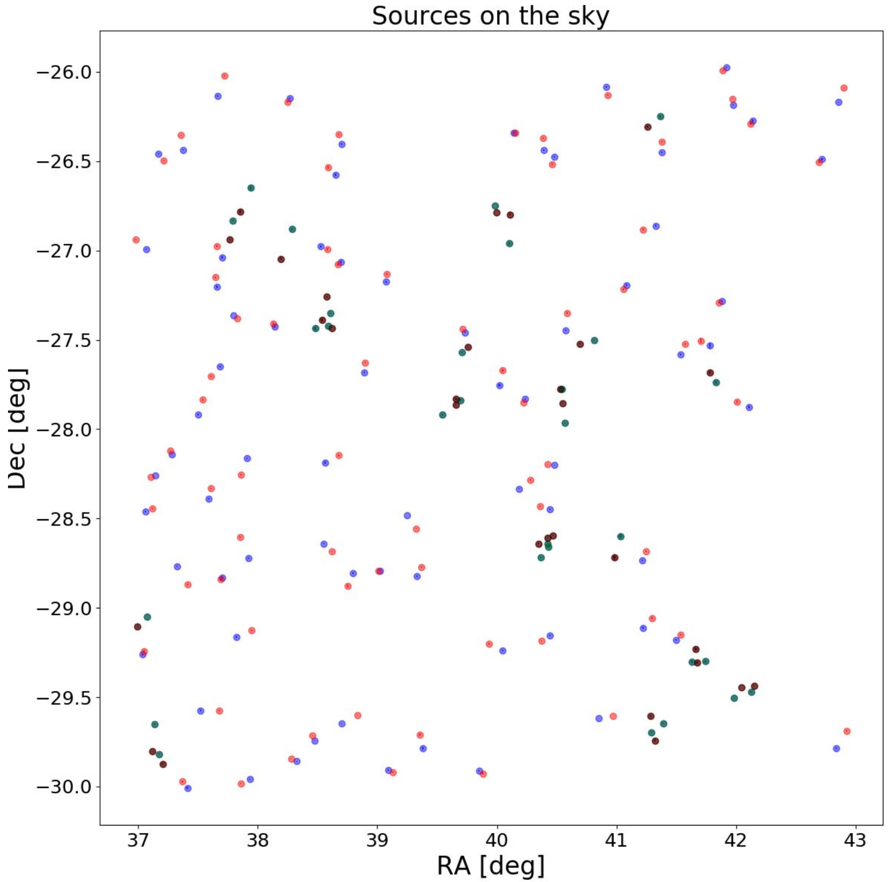
Ambiguous = green & black.

Clear neighbour - zoomed in



Ambiguous = green & black.

Clear neighbour filter on small dataset of 100.



Epoch 00 vs Epoch 01 = 72 unambiguous.

Epoch 00 vs 01 vs… vs 48 vs 49 all unambiguous = 16.

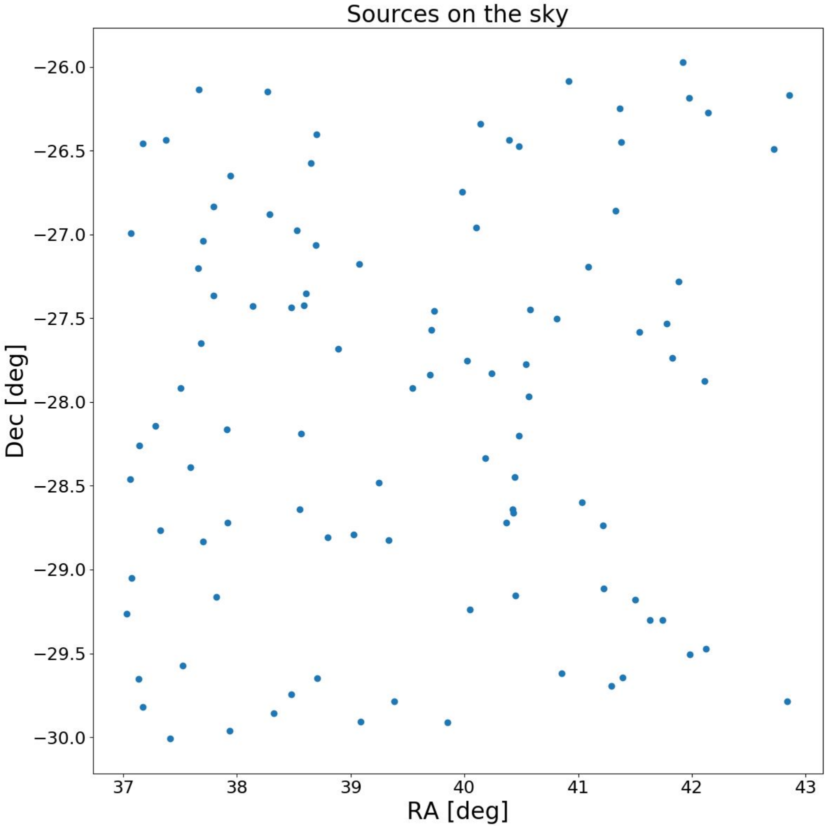
Complex Check
Tinder for galaxy positions: the Hungarian algorithm

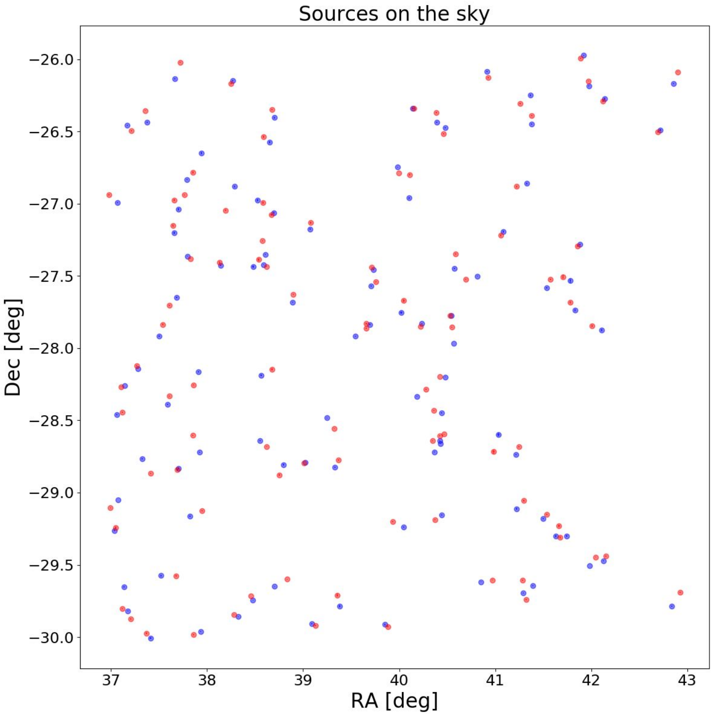
- Minimizes the cost for an NxN matching problem
- Need to introduce a metrics, for matching an observed galaxy position to a galaxy in our sky model.
- **The metrics need to be dimensionless!**
- **Our metics:**

  We measure the probability of a point, that drawn from a 3D distribution defined by our sky model, is closer to the distribution center than the observed position.

- We assumed Gaussian distribution in each marginal direction

Initial model

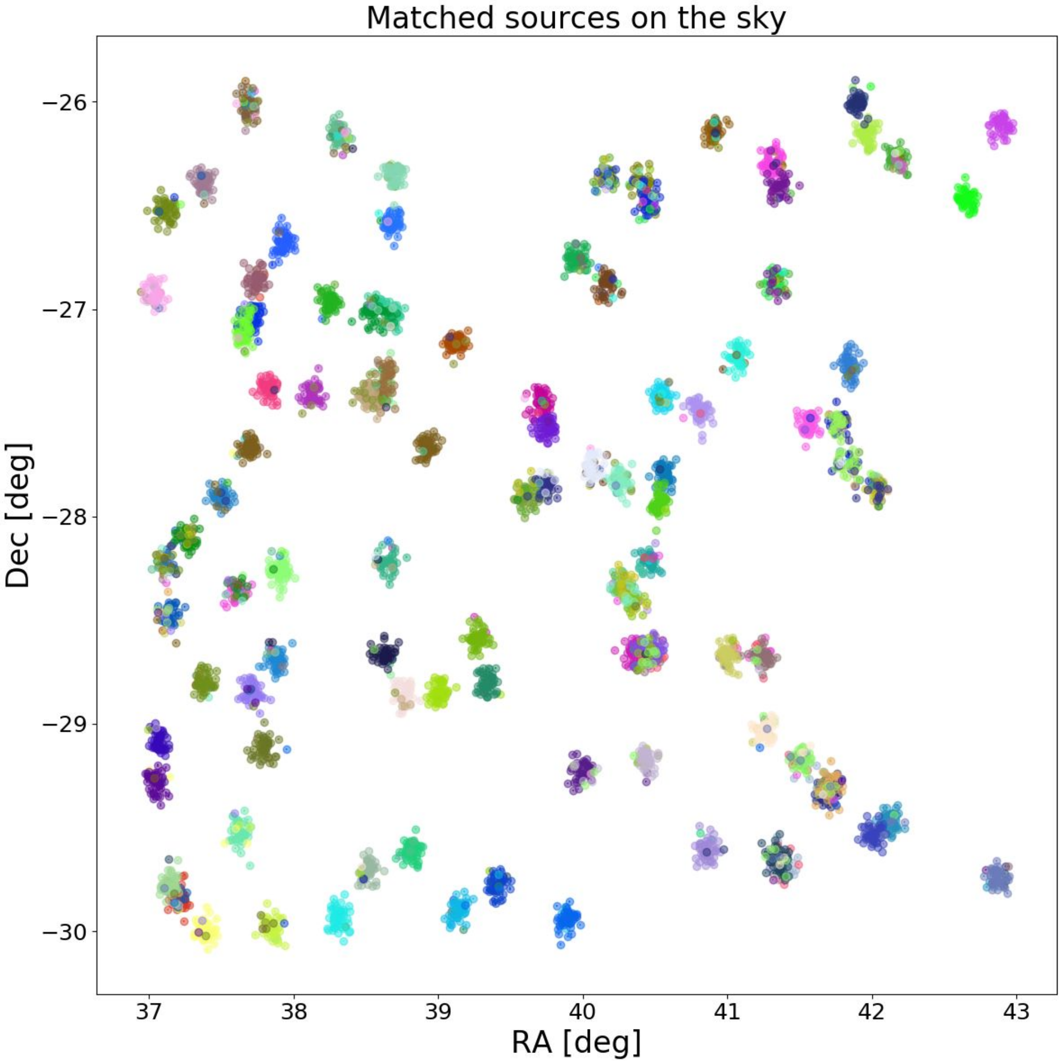


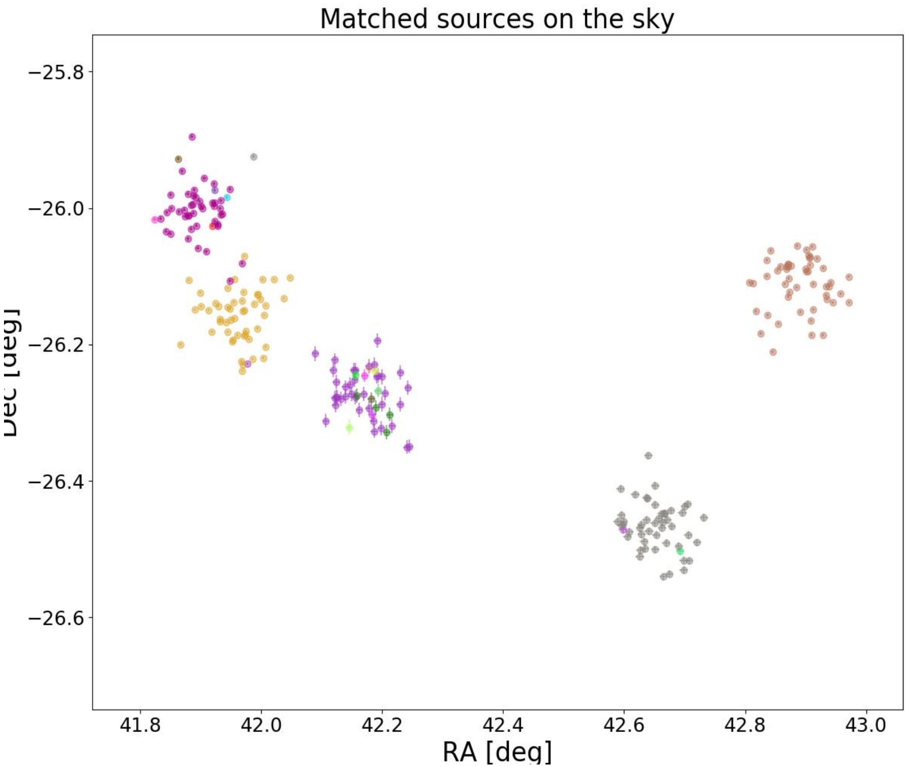


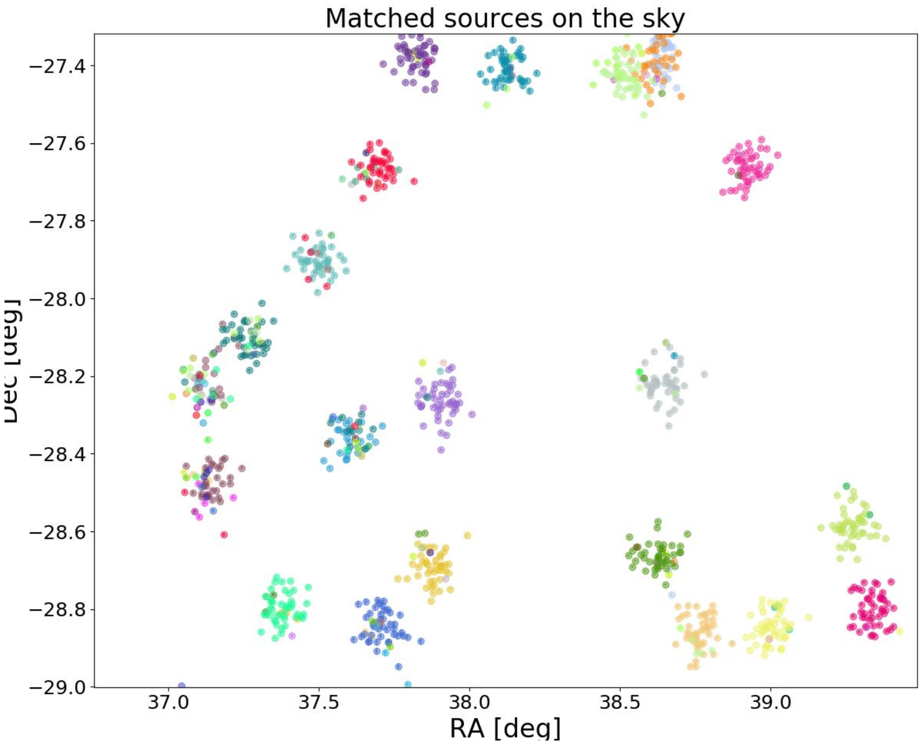
- We set our initial model based on the first observation
- We updated the model in while processing observations

Final results

- We identified the isolated galaxies, but we had some issues with the cluster galaxies.
- Unfortunately, we couldn’t quantify the algorithm preciseness

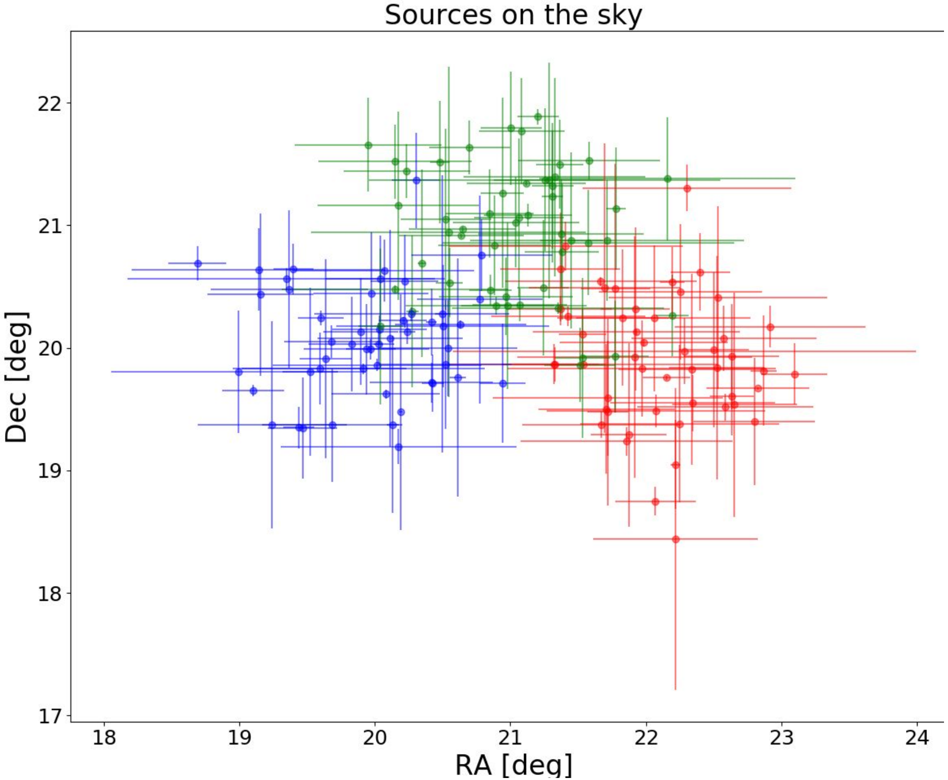


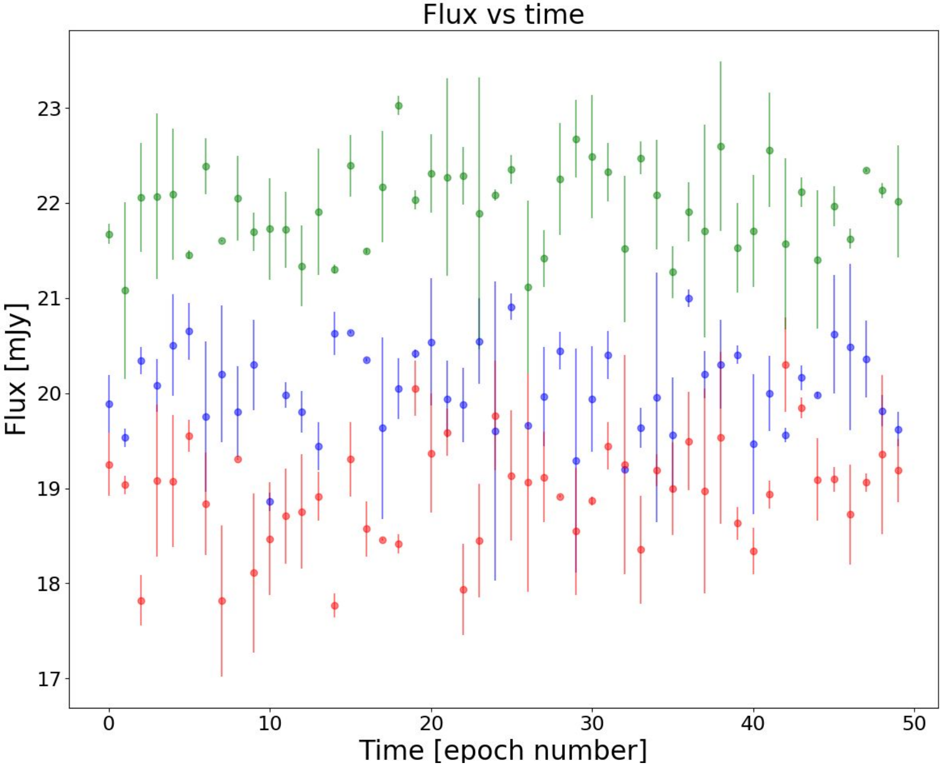

Matched sources on the sky
Dec [deg]
RA [deg]
−25.8
−26.0
−26.2
−26.4
−26.6
41.8
42.0
42.2
42.4
42.6
42.8
43.0

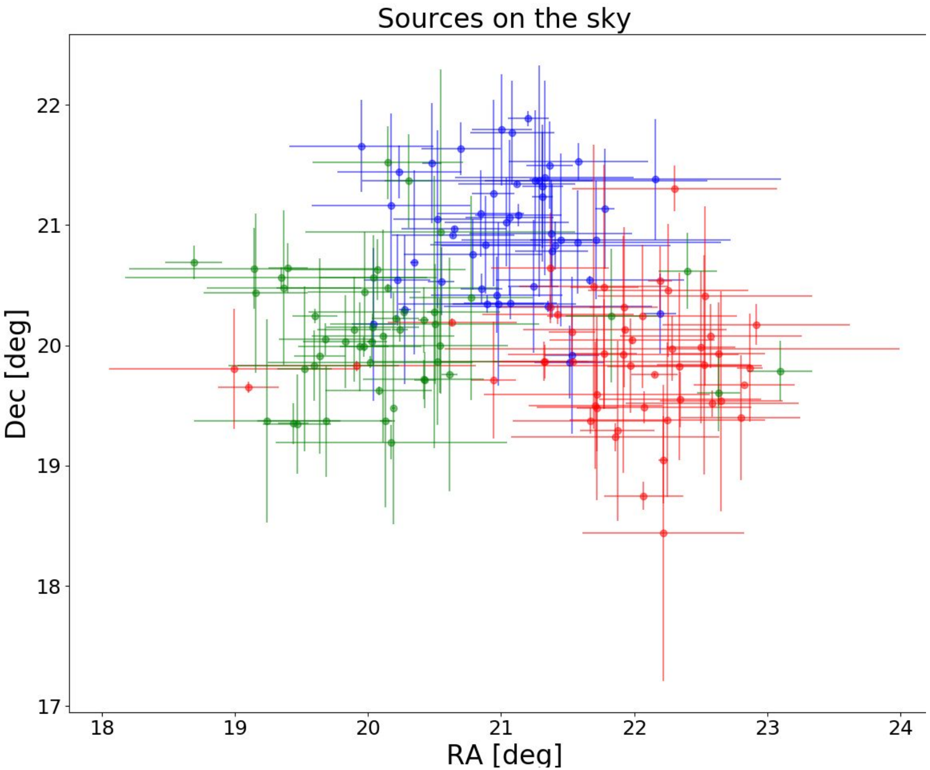

Matched sources on the sky
Dec [deg]
RA [deg]
−27.4
−27.6
−27.8
−28.0
−28.2
−28.4
−28.6
−28.8
−29.0
37.0
37.5
38.0
38.5
39.0

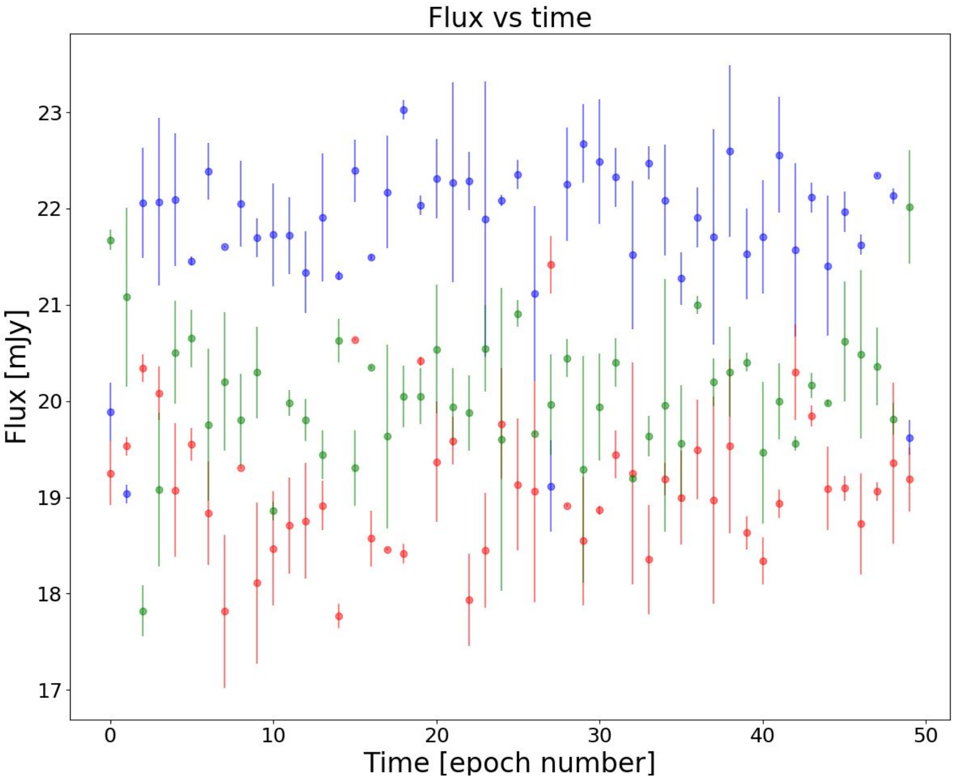
Zoom in simulation, issues & future improvement

- We simulated 3 galaxies close to each other to quantify the algorithm efficiency
- We find that the method worked in ~95% of the data (43 out of 50)
- However the method scales with O(N^2)
- A possible solution would be to use the **Bellmann-Ford algorithm**


Sources on the sky
22
21
20
19
18
17
Dec [deg]
18
19
20
21
22
23
24
RA [deg]


Flux vs time
23
22
21
20
19
18
17
Flux [mJy]
0
10
20
30
40
50
Time [epoch number]


Sources on the sky
Dec [deg]
RA [deg]
22
21
20
19
18
17
18
19
20
21
22
23
24


Flux vs time
Flux [mJy]
Time [epoch number]
23
22
21
20
19
18
17
0
10
20
30
40
50

Future Plans

1. Quantify Tinder comparison

2. Optimise Simple
    a. Already checks for false positives; add check for true negative

3. Combine:
    a. For each comparison, combine output of both algorithms to maintain 90%.