Searching for open clusters with clustering algorithms in Gaia era

Mario Morvan

University of Barcelona & Mines Saint-Etienne

End-of-Master Project

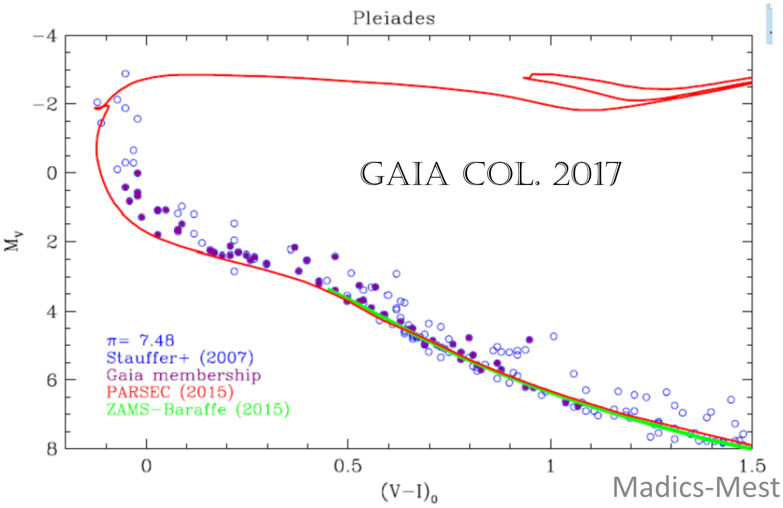
Outline

1. Open Clusters (OCs) → quick intro
2. Astrometry in Gaia era → Big astrometric Data
3. Density-based clustering → kNNs, DBSCAN, OPTICS
4. Searching OCs in TGAS → method & results
5. Scaling-up for next releases

Open clusters

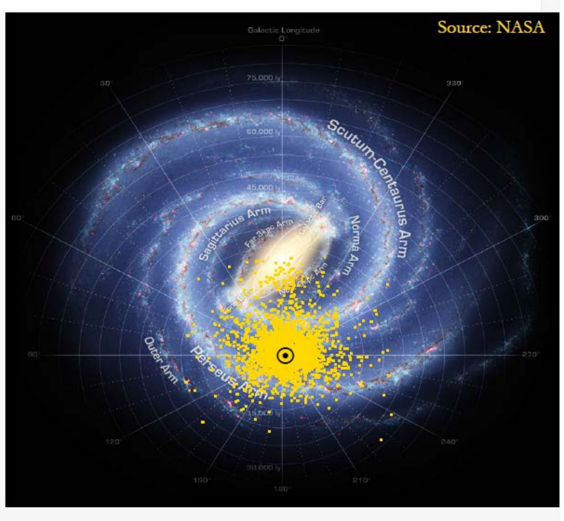
- Bound groups of stars
- Share various properties:
  - Age
  - Metallicity
  - **Position, velocities**
- Photometry:

Pleiades (WEBDA)



GAIA COL. 2017

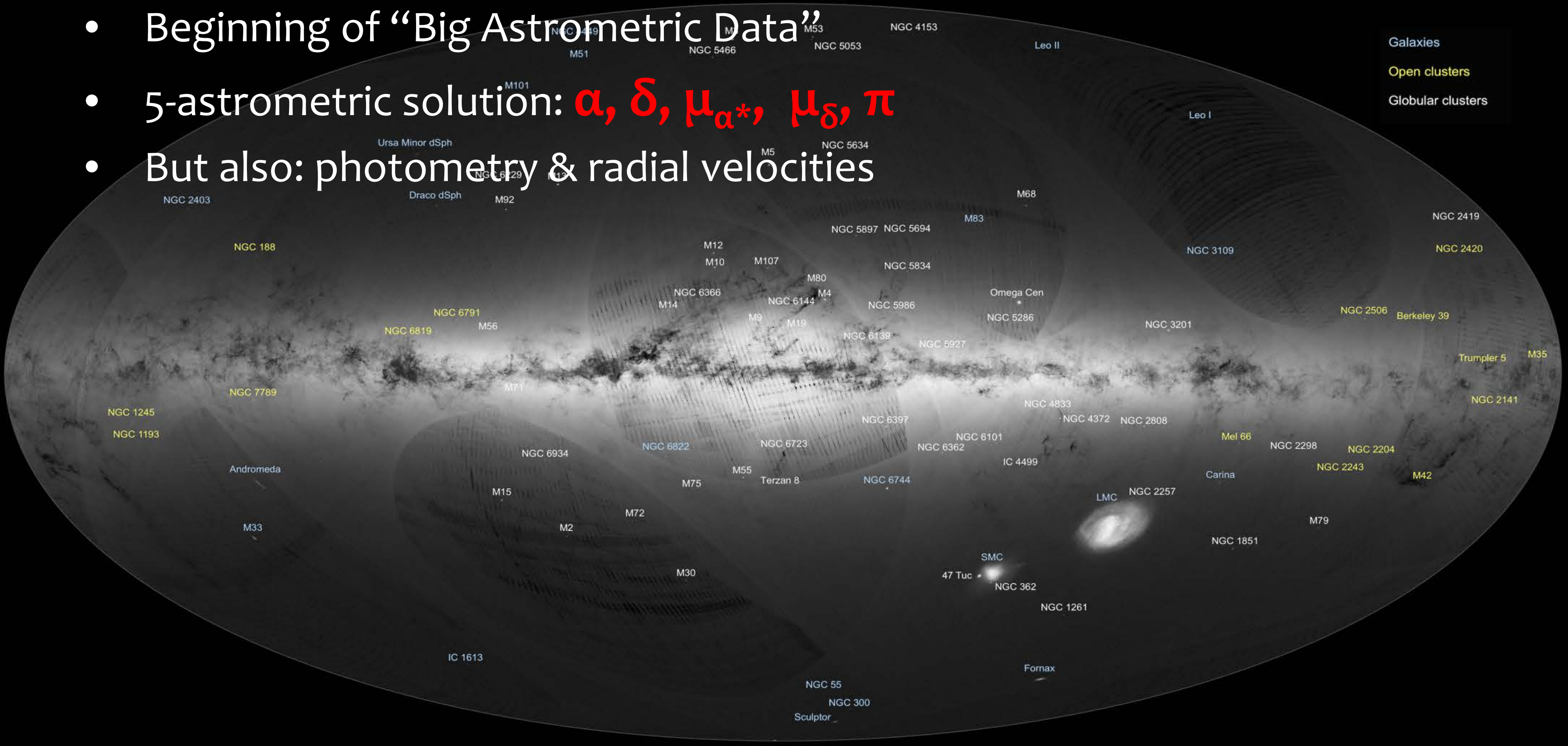
~3000 known OCs



Source: NASA

Astrometry in Gaia Era
Gaia Mission

- ESA’s mission (2013→2022)
- Beginning of “Big Astrometric Data”
- 5-astrometric solution: $\alpha, \delta, \mu_{\alpha*}, \mu_{\delta}, \pi$
- But also: photometry & radial velocities



http://sci.esa.int/gaia/58281-gaia-s-first-sky-map-annotated/

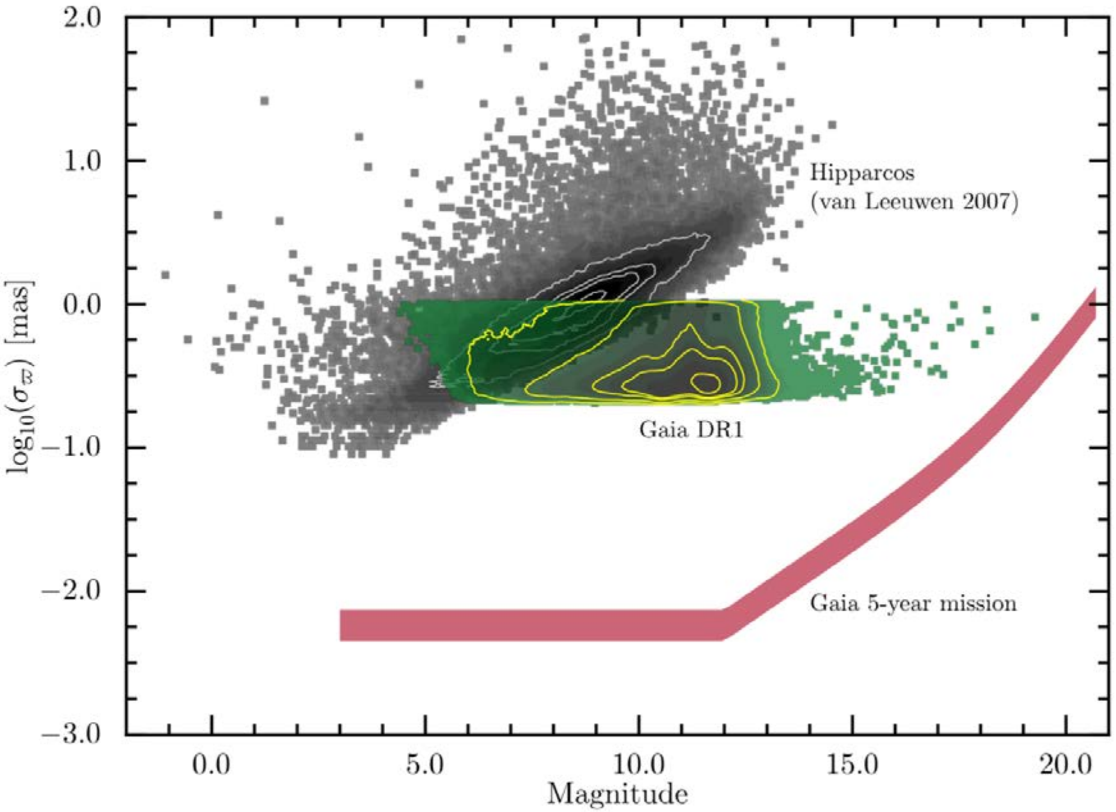
Astrometry in Gaia Era
Astrometric catalogues

Catalogue	Release Date	Nb of Sources with astrometry	Limiting Magnitude	Size	Radial Velocities
Hipparcos	1997	117,955	12	~12 MB	Not available
TGAS	Sept 2016	2,057,050	12	~408 MB	Not available
GDR2	April 2018	$>10^9$	>20	~5 TB	G<12
Final GDR	2022	$>10^9$	>20	~10 TB ?	All

https://www.cosmos.esa.int/web/gaia/release

Astromety in Gaia era
Astrometric Catalogues precision

- Parallax standard error: from mas to 10μas



GAIA COL. 2017

Astrometry in Gaia Era
Our Problem

- Fully-astrometric and unsupervised detection of open clusters
    - Do not adress membership determination
    - Use photometry and OC catalogs (only) for subsequent analyze
- *Idea*: use clustering algorithms to detect OCs as

  5D high astrometric concentrations
- Related works :
    - Gao et al. 2013, Gao 2014 and Bhattacharya et al. 2016 → membership with DBSCAN
    - Gao 2017→ close detection with kNND in TGAS

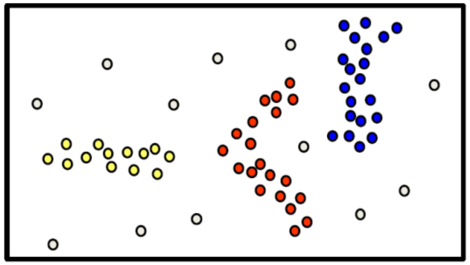
Density-based clustering

- Clustering: find groups of *close* points in data sets
- Types:
  - Partition-based → need to fix a number of cluster
  - Hierarchical → criterion to stop?
  - Distribution-based → requires knowledge and limits scope
  - **Density-based**

    → Clusters = contiguous regions of high-density separated by low density regions

    → Naturally deals with noise

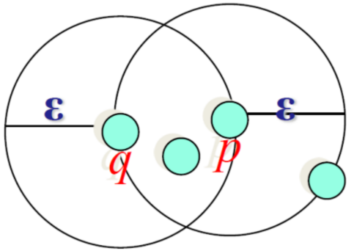
    → Arbitrary number and shapes for its clusters

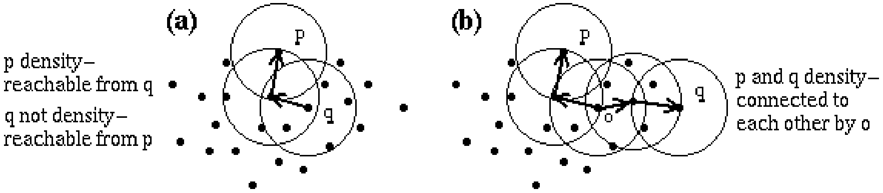


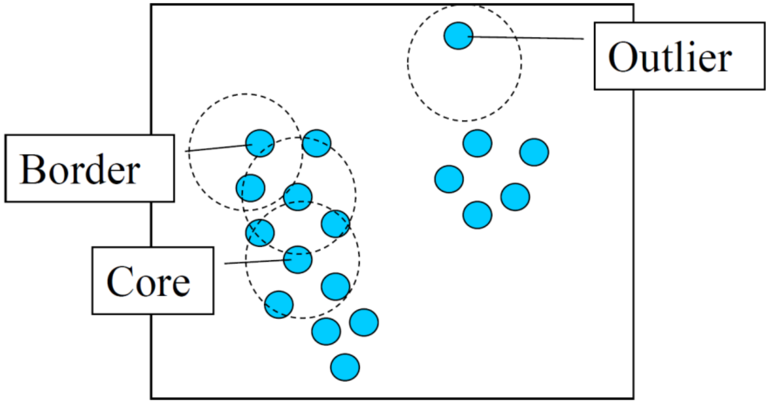
Density-based Clustering DBSCAN

- *'Density-Based Spatial Clustering of Applications with Noise'* (Ester et al. 1996)
- Parameters: (minPts, ε)
- **ε-Neighborhood** –

  Objects within a radius of ε from an object.
- **"High density" region** - ε-Neighborhood of an object contains at least *MinPts* objects

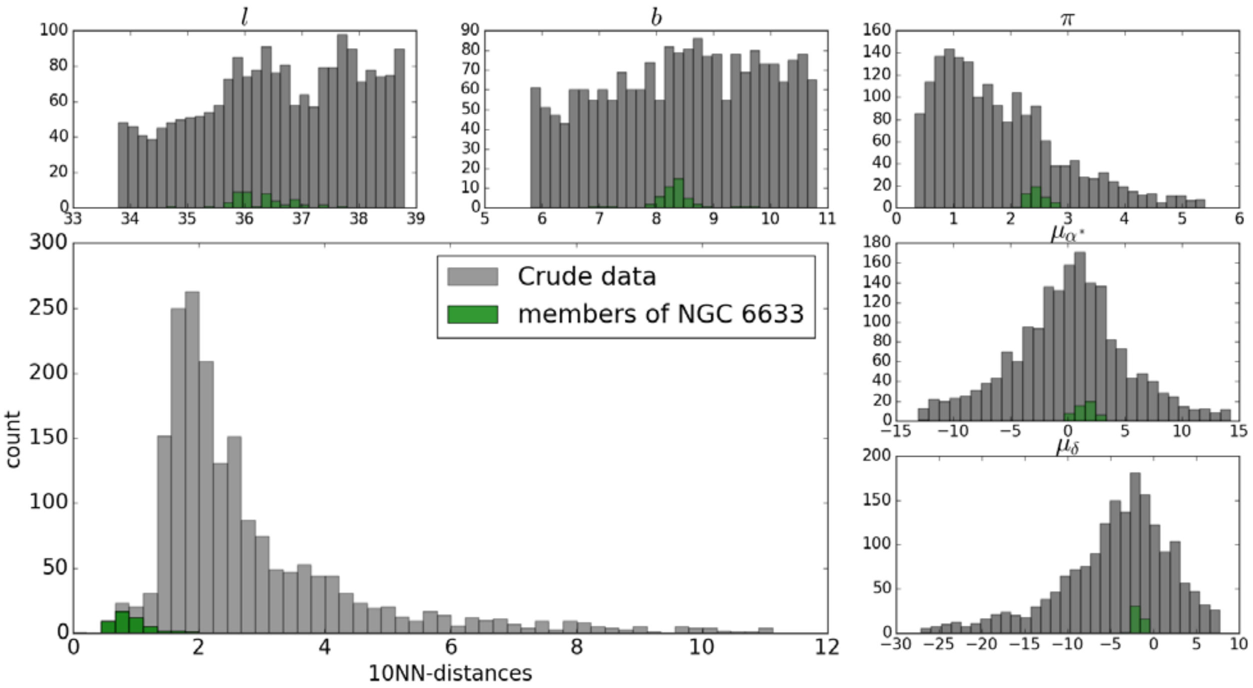



(a)
p
q
p density-reachable from q
q not density-reachable from p
(b)
p
o
q
p and q density-connected to each other by o



→ How to determine ε ?

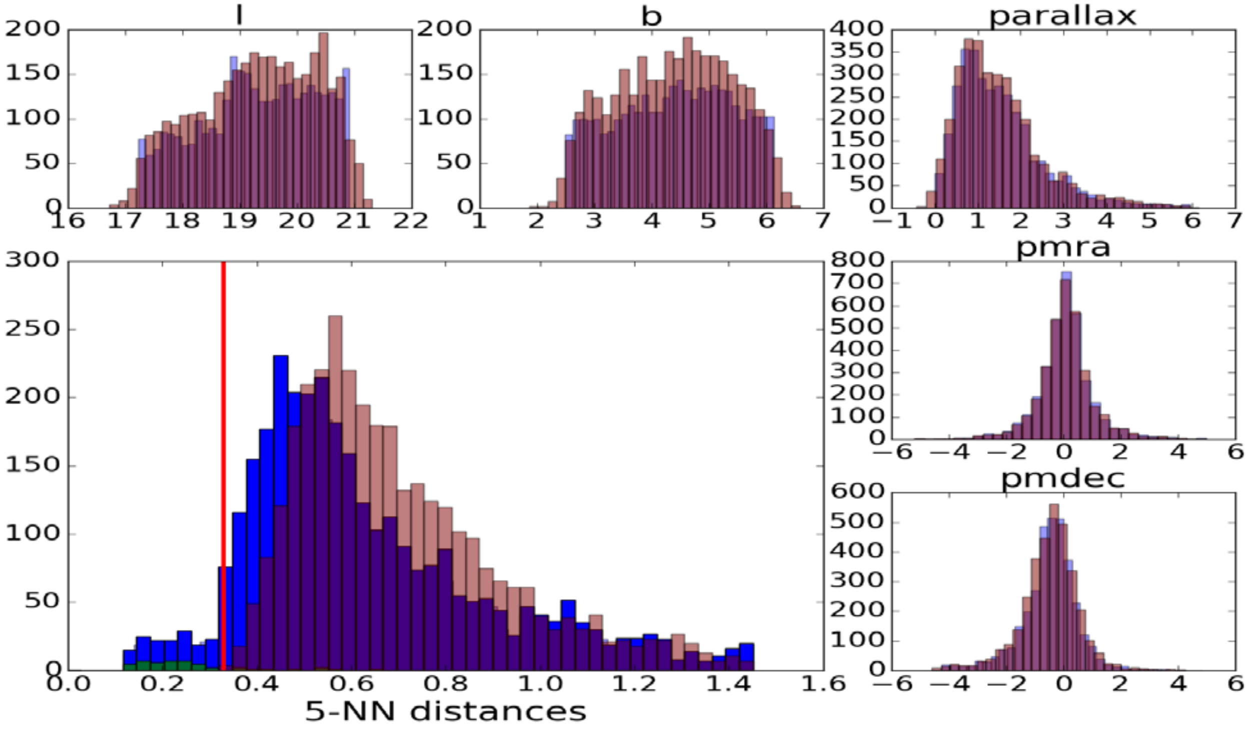
Density-based Clustering
Choice of Epsilon with kNND



→ kNN/DBSCAN correspondance:

p is a DBSCAN core ⟺ *kNND*(p) < ε with k=minPts-1

Density-based Clustering
Automatic choice of Epsilon with kNND



1D-density estimations with Gaussian Kernel

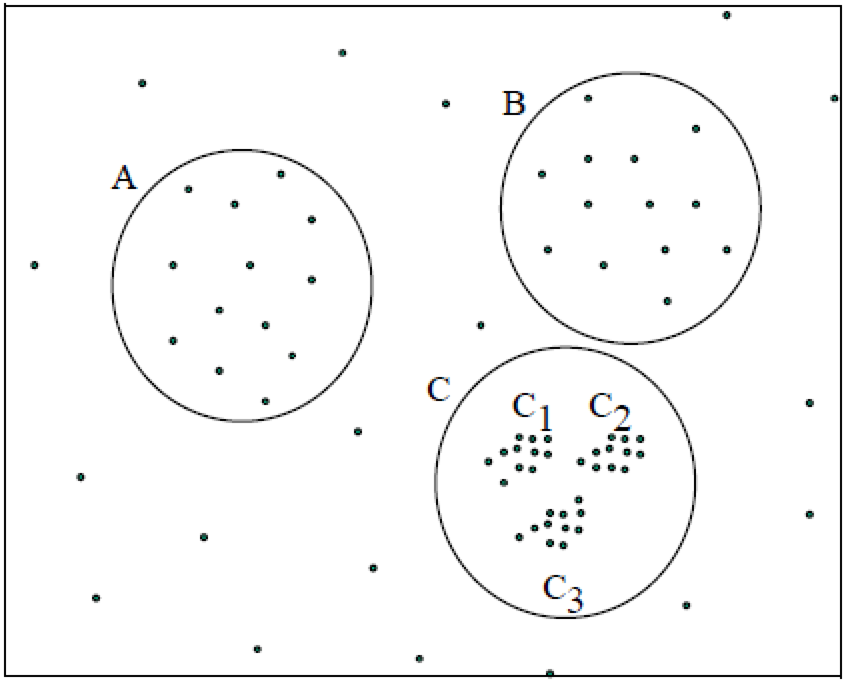
Min of kNN distribution of resampled stars

→ upper limit for ε

Density-based clustering OPTICS

"Ordering Points To Identify the Clustering Structure"
(Ankerst et al. 1999)

What if there are several densities of clusters?



Density-based clustering OPTICS

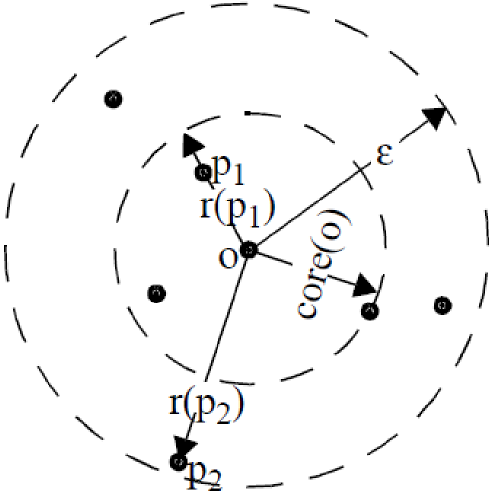
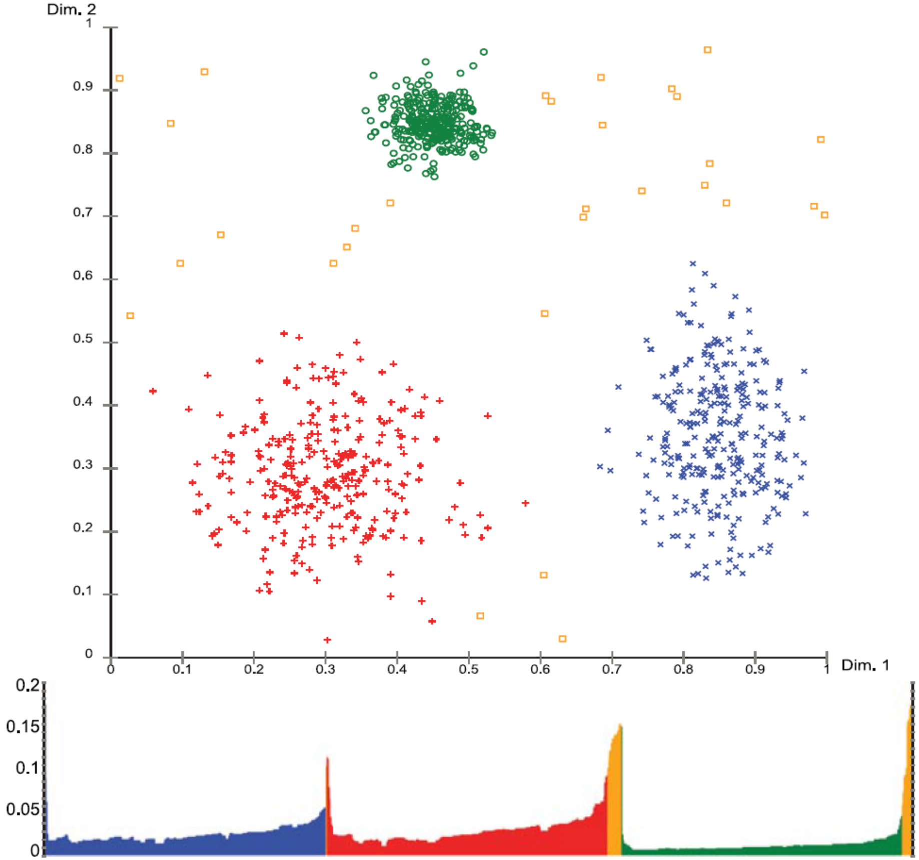


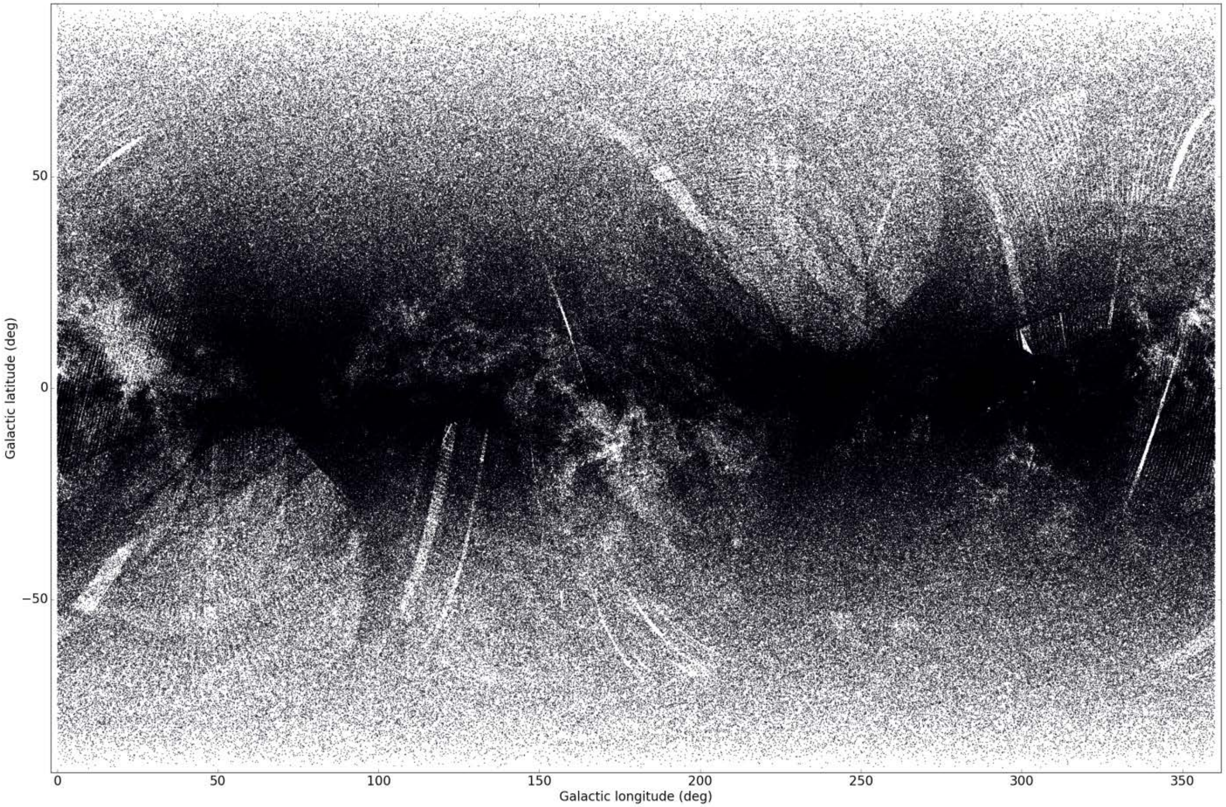
Figure 4. Core-distance(*o*), reachability-distances r(*p*$_1$,*o*), r(*p*$_2$,*o*) for *MinPts*=4



KRIEGEL ET AL. 2011

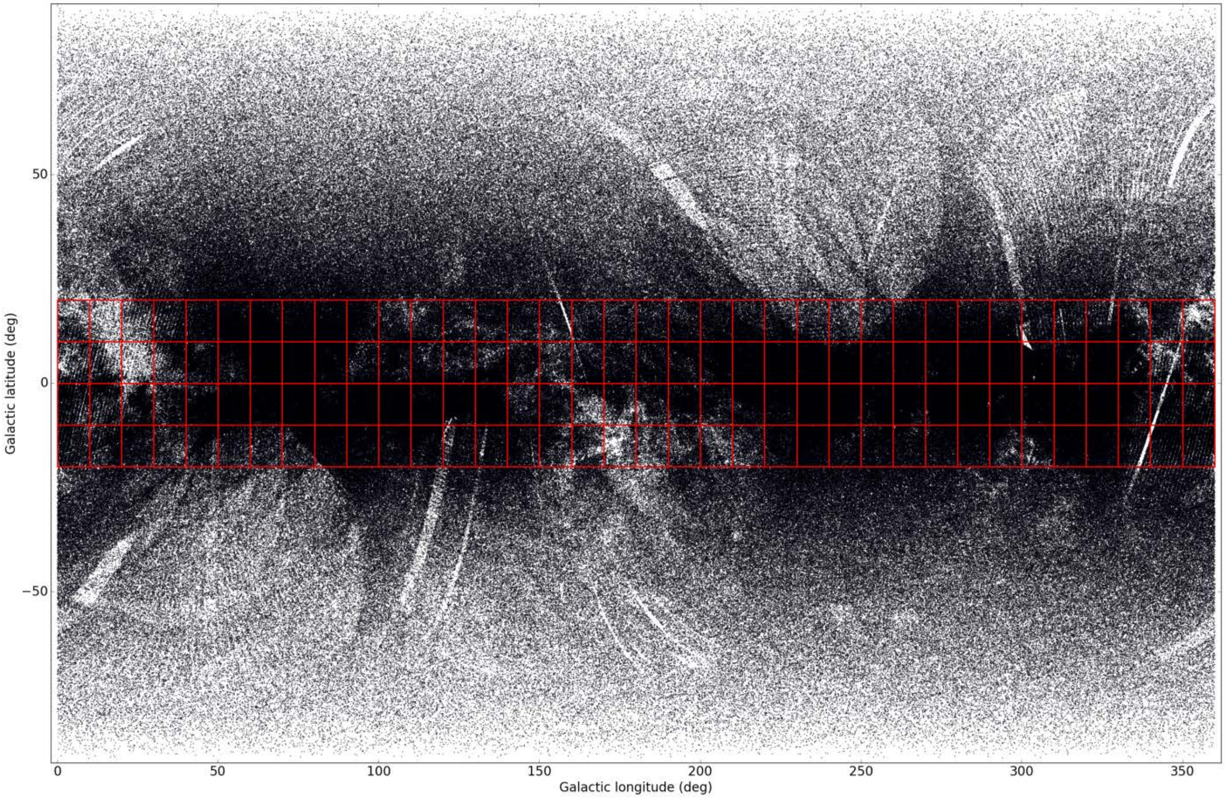
Searching OCs in TGAS

- Start with 2,057,050 stars with (*l*,*b*, $\mu_{\alpha*}$, μ_δ, π)



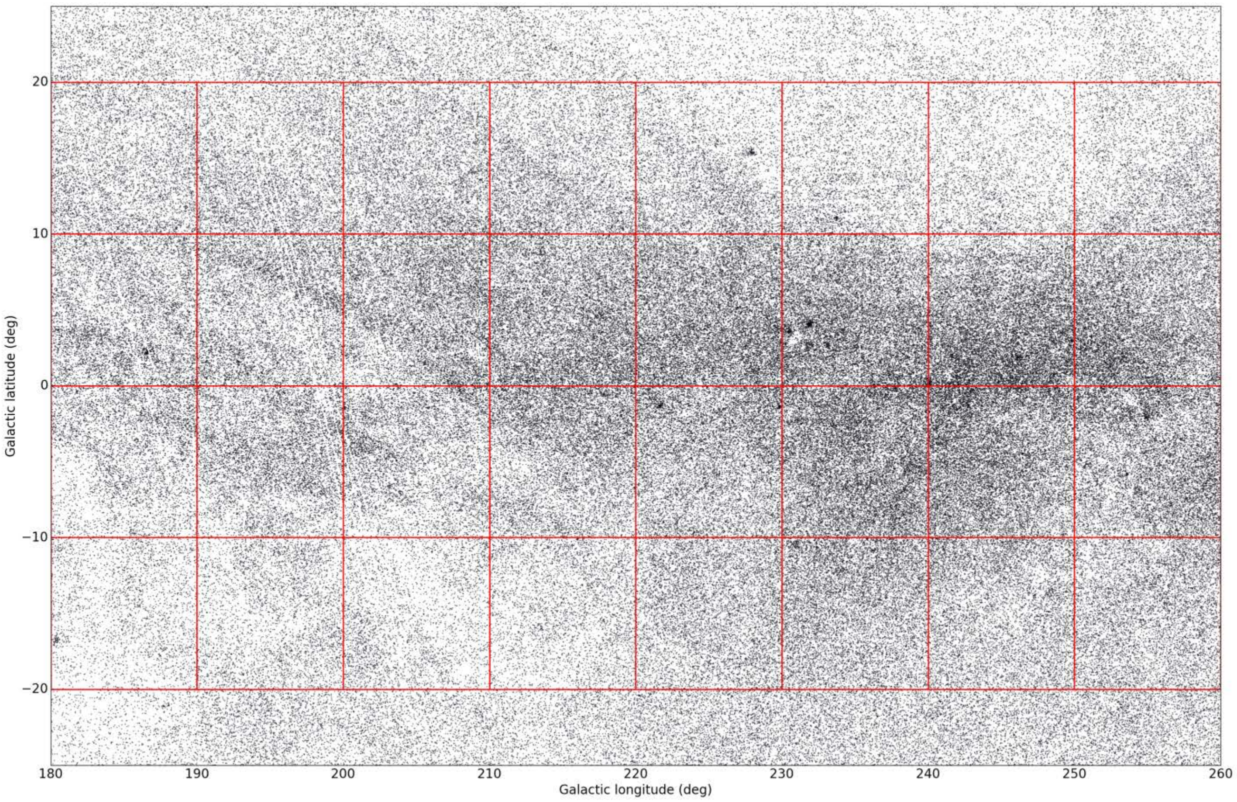
Searching OCs in TGAS Preprocessing

- Galactic disk, cleaning extreme values
- division into rectangles, normalization



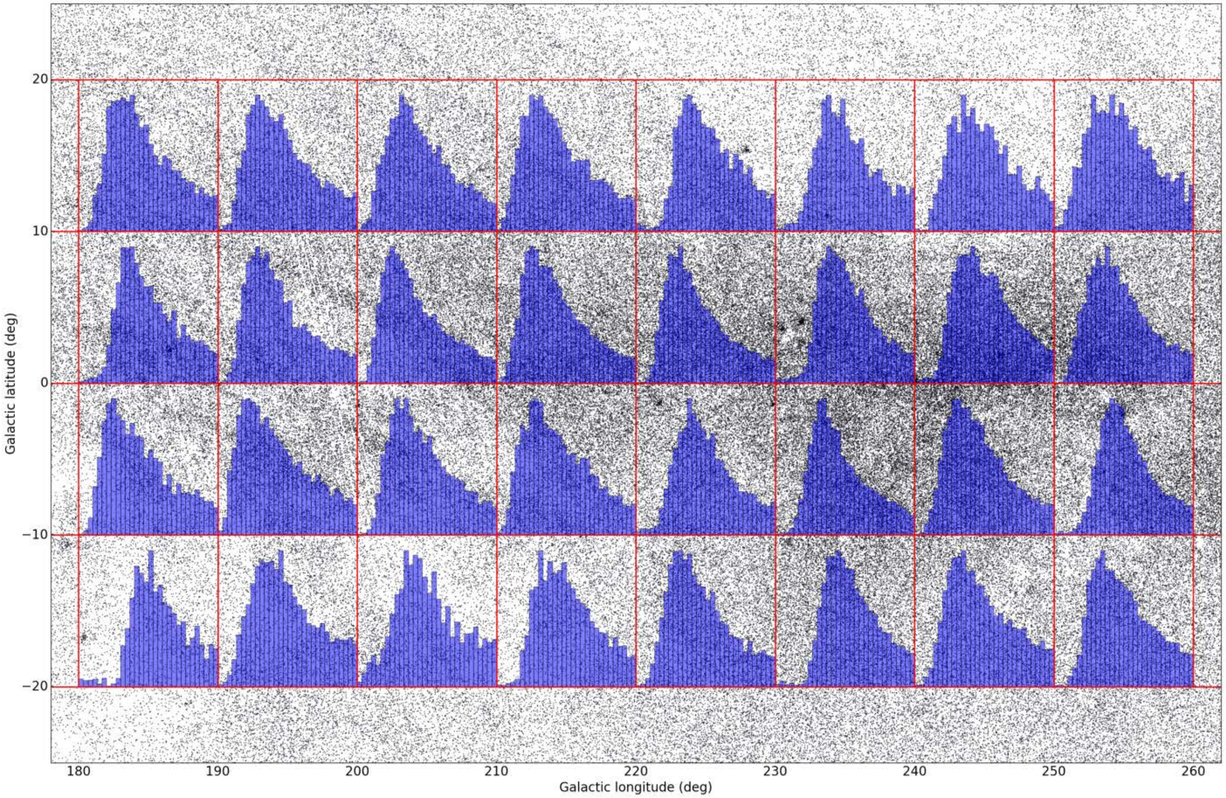
Searching OCs in TGAS Preprocessing

- Galactic disk, cleaning extreme values
- division into rectangles, normalization



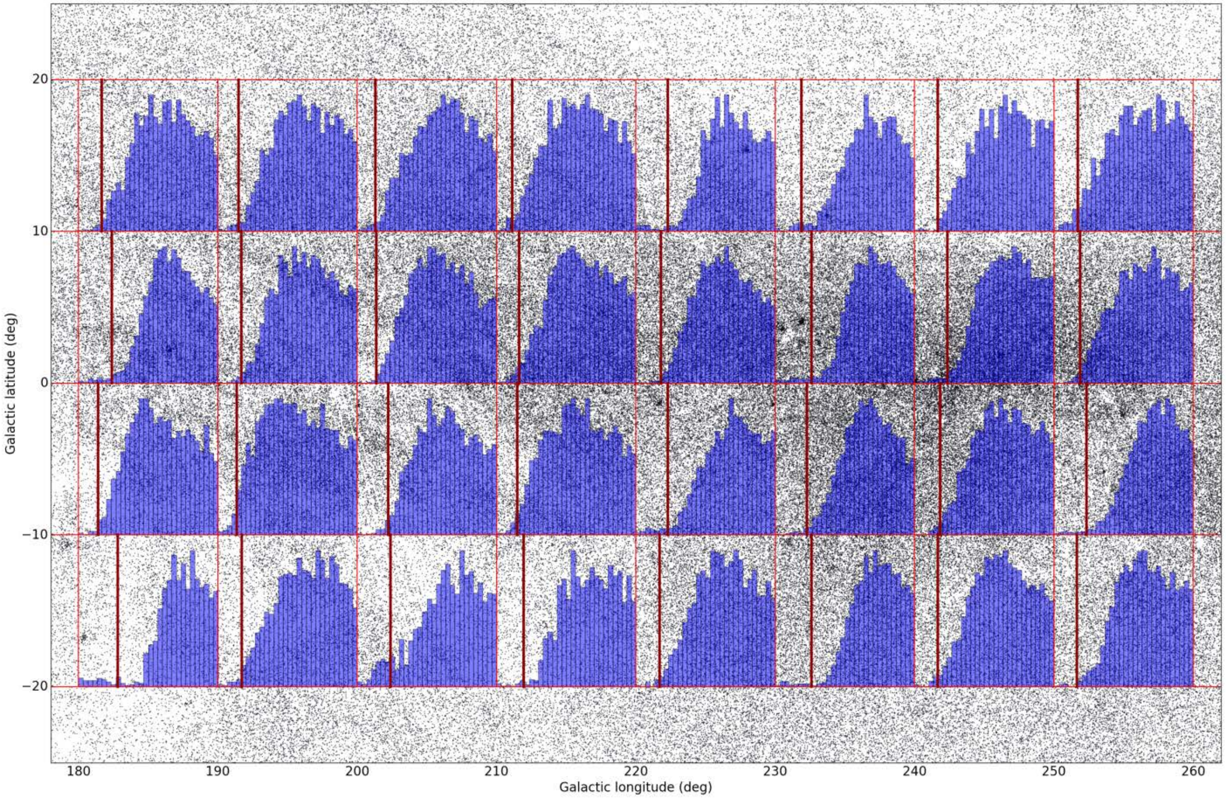
Searching OCs in TGAS Epsilon determination

- KNND computation in each square



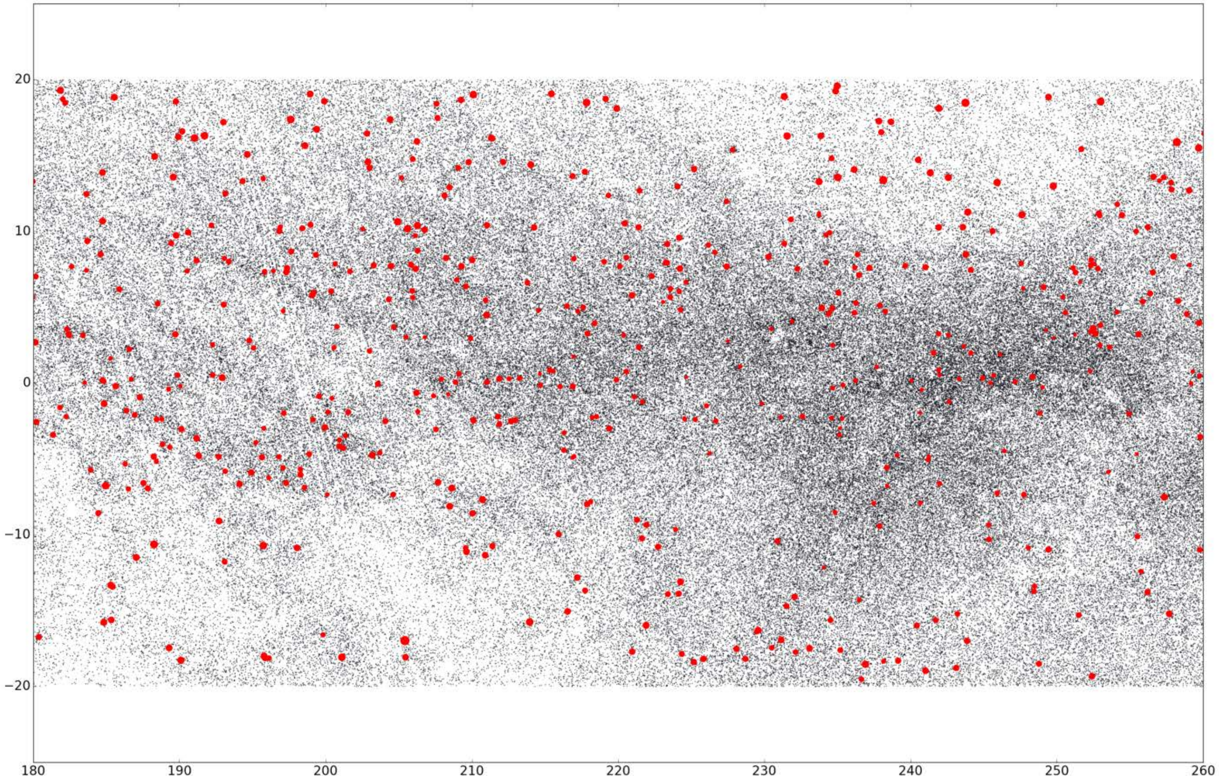
Searching OCs in TGAS
Epsilon determination

- kNND of 1D-kde of parameters

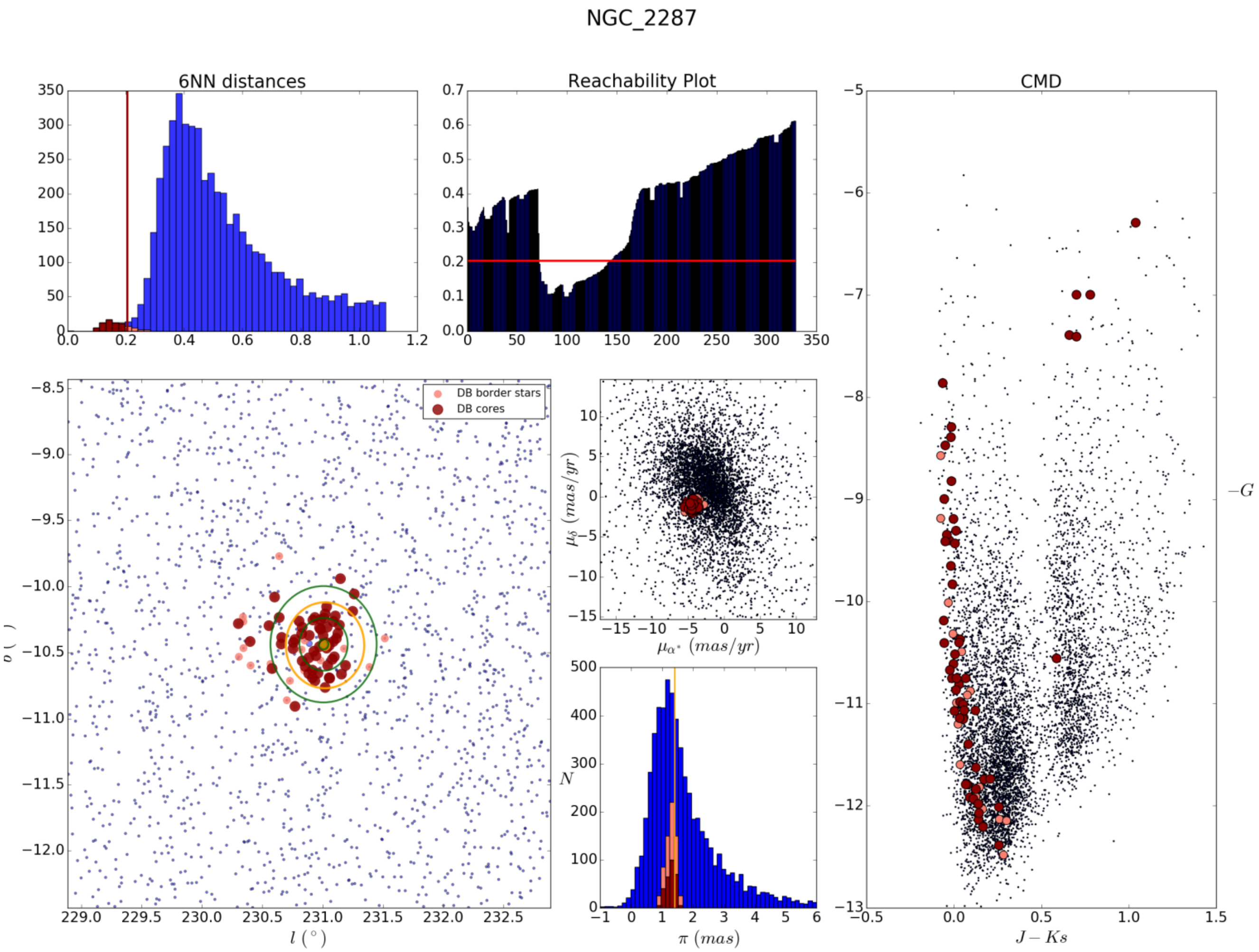


Searching OCs in TGAS
Running DBSCAN

- Remove clusters on the edge
- Merge results with 2 shifted grids



Searching OCs in TGAS
A matched OC



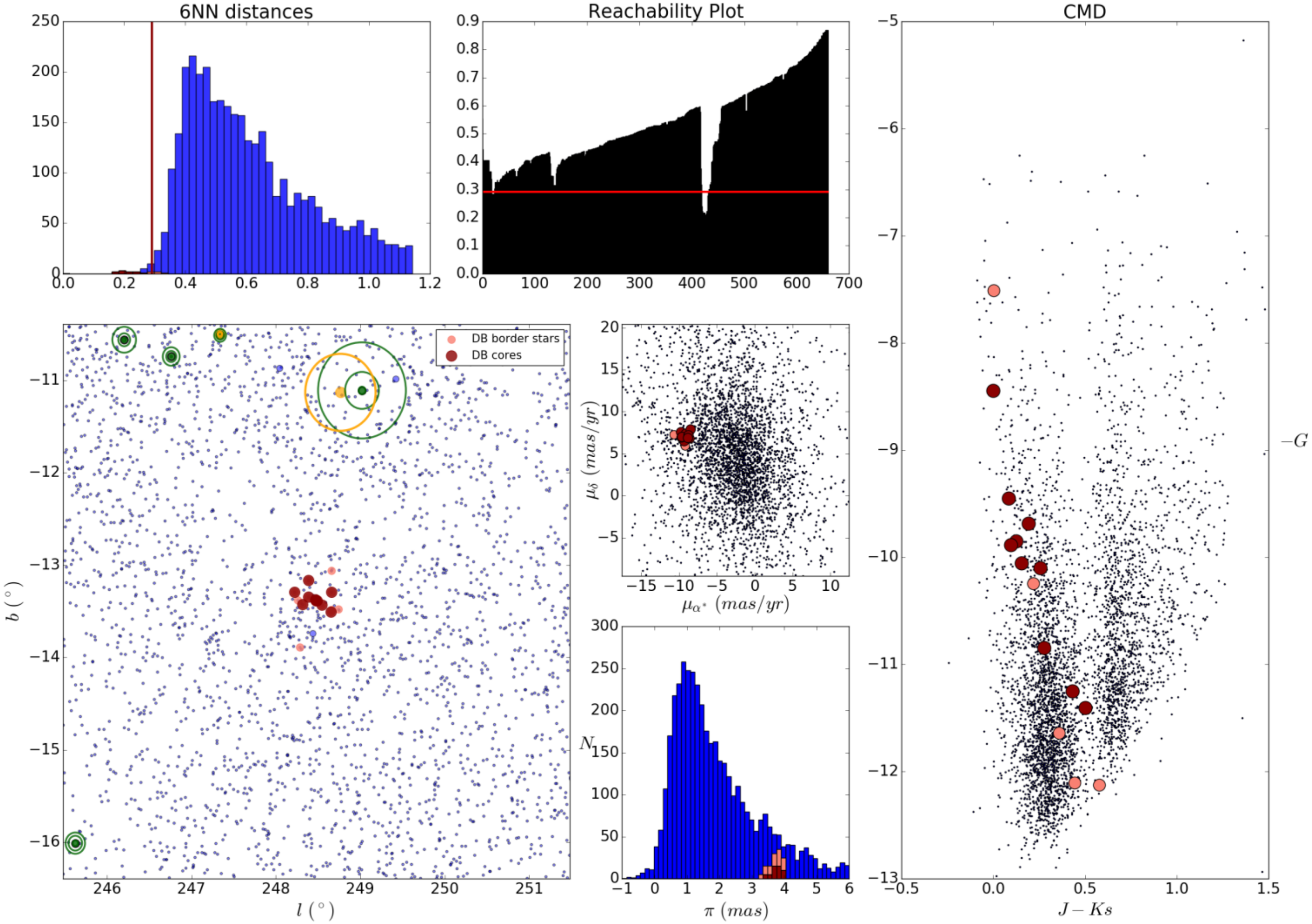
Searching OCs in TGAS
Some results

- Seeing matches while Merging 15 clusterings:
  - – $\lesssim$1000 matches with catalogued OCs
  - – $\lesssim$80 OB associations
  - – $\lesssim$50 removed OCs
- But also a lot of non-matched DB clusters

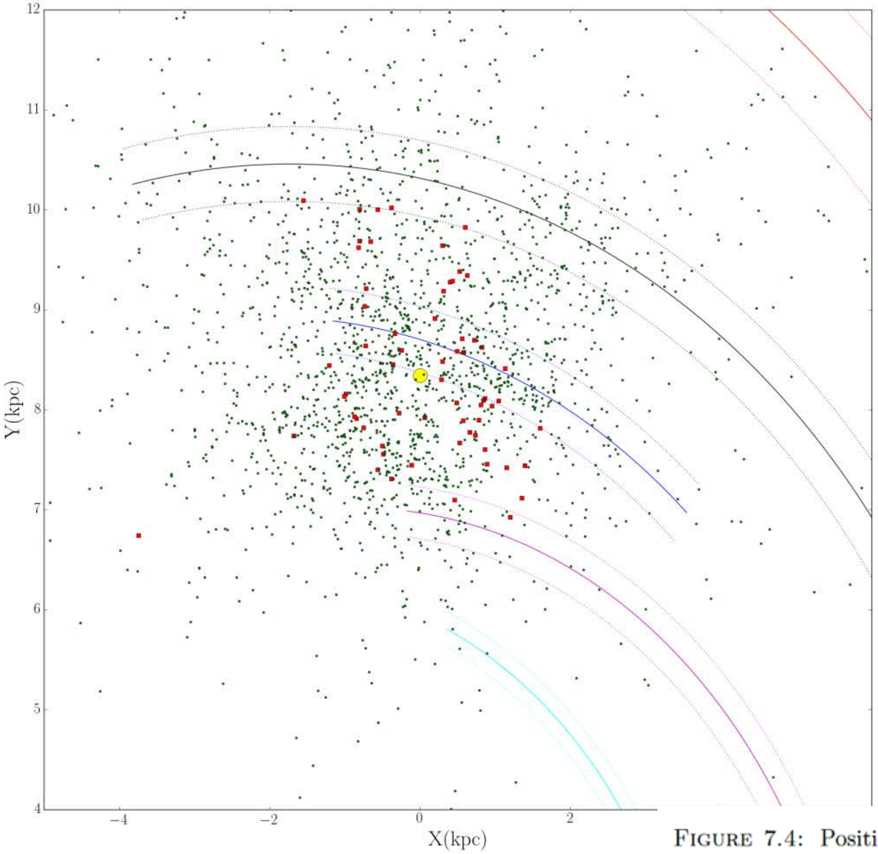
→filters to look at most interesting cases

→60 selected interesting non-matched clusters

Searching OCs in TGAS
A non-matched cluster



Searching OCs in TGAS
Non matched clusters



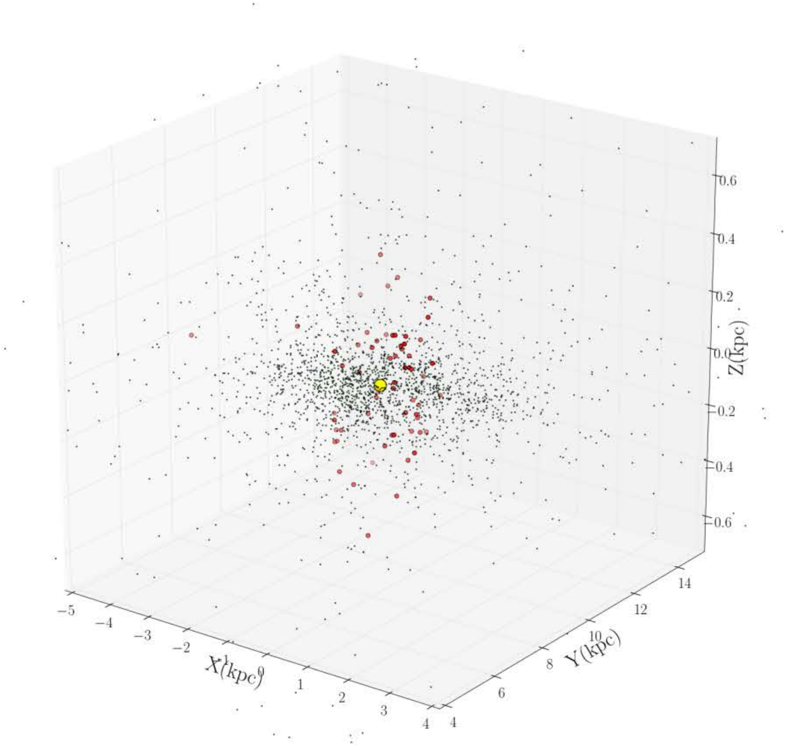


FIGURE 7.4: Position of 60 clusters candidates in red and 2040 Dias catalogued open clusters in green in rectangular Galactic coordinates. On the upper plot, the clusters were projected onto the Galactic XY-plane and the curved lines represent the spiral arms as proposed by Reid et al. (2014). On both plots, yellow circles locate the Sun's position. Dias open clusters extend much more from th Sun, since strong relative errors in parallaxes and proper motions make a fully-astrometric detection of open clusters with TGAS data complicated above $2pc$ from the Sun.

Scaling-up to subsequent GDR

- Gaia Data Release 2 will have to be stored in a computer cluster
- Area density of stars x500
  - →Adapt the grid, minPts…
  - →Intermediate filters
- kNN, DBSCAN and OPTICS are parallelizable
  (ZHANG ET AL. 2012, HE ET AL. 2011)

→ Soon in Spark?

- In the end, include radial velocities?

Conclusions

- Astrometry great for OCs
- Such a method using DB clustering is:
  - Promising
  - Largely improvable
  - Scalable
- GDR2 will
  - Enrich the knowledge of known OCs
  - Decide all the uncertain cases
  - Allow to discover many more

Thank you !

Any question, comment suggestion… ?