A photometric unsupervised candidate quasars selection algorithm

Raffaele D'Abrusco

Department of Physical Sciences - University "Federico II", Naples
Institute of Astronomy - University of Cambridge

NAM 2007 - 18th April 2007
Preston

Quasars

Quasars are peculiar objects. In early years of quasar search broad criteria were used to distinguish them from stars and galaxies ("Quasi-Stellar Objects", Burbridge's):

- Optical star-like objects identified with radio-sources.
- Optical variability.
- Large ultraviolet emission (compared to stars).
- Spectral features: continuous spectra with broad emission lines and with absorption lines in few cases.
- Large redshift.

Tight observational similarities (Schmidt, 1963) and similar energetic requirements are seen in quasars and central compact regions of some galaxies. This has inspired a unified scheme where both types of sources (and other types of active emission from galaxies) are explained as different manifestations of astrophysical objects called AGNs. Differences in orientation and luminosity of these objects mix up the observational scenario.

Quasars are very important for both astrophysics and cosmology as witnesses of the evolution of galaxies and remarkable probes of the far universe. Their selection is also a benchmark for data mining techniques because of their heterogeneous appearance.

Quasars

Quasars are peculiar objects. In early years of quasar search broad criteria were used to distinguish them from stars and galaxies ("`Quasi-Stellar Objects`", Burbridge's):

Today the scenario is more complex then it was in those pioneering days, because new families of quasars have been discovered and observations span over the whole e.m. spectrum.

Tight observational similarities (Schmidt, 1963) and similar energetic requirements are seen in quasars and central compact regions of some galaxies. This has inspired a unified scheme where both types of sources (and other types of active emission from galaxies) are explained as different manifestations of astrophysical objects called AGNs. Differences in orientation and luminosity of these objects mix up the observational scenario.

Quasars are very important for both astrophysics and cosmology as witnesses of the evolution of galaxies and remarkable probes of the far universe. Their selection is also a benchmark for data mining techniques because of their heterogeneous appearance.

Photometric identification of quasars

Algorithms for photometric identification of candidate quasars select sources according to different techniques. Most important of these are:

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of quasars are radio-loud).
- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

Photometric identification of quasars

Algorithms for photometric identification of candidate quasars select sources according to different techniques. Most important of these are:

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of quasars are radio-loud).

Not used anymore

- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

Photometric identification of quasars

Algorithms for photometric identification of candidate quasars select sources according to different techniques. Most important of these are:

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of quasars are radio-loud).
- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Not used anymore

Good

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

Photometric identification of quasars

Algorithms for photometric identification of candidate quasars select sources according to different techniques. Most important of these are:

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of quasars are radio-loud).
- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("stellar locus") occupied by stars.

Not used anymore

Good

Better

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

An example: SDSS quasars targeting algorithm

SDSS quasars candidate selection algorithm (Richards et al, 2002) targets star-like objects as quasars candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:

- they lie more then 4σ far from a cylindrical region containing the "stellar locus" (S.L.), where σ depends on photometric errors.
- they lie inside an inclusion region, even if not meeting the previous requirement.

1. Designated as **inclusion regions** are regions where S.L. meets quasar's area (due to absorption from Lyα forest entering the SDSS filters, which change continuum power spectrum power law spectral index). All objects in these areas are retained in order to sample [2.2, 3.0] redshift range (where quasars density is also declining), but at the cost of a worse efficiency (Richards et al, 2001).
2. Designated as **exclusion regions** are regions outside the main "stellar locus" but clearly populated by stars only (usually WDs). All objects in these regions are discarded.

Overall performance of the algorithm: completeness c = 95%, efficiency e = 65%, but locally (in colours and redshift) much less.

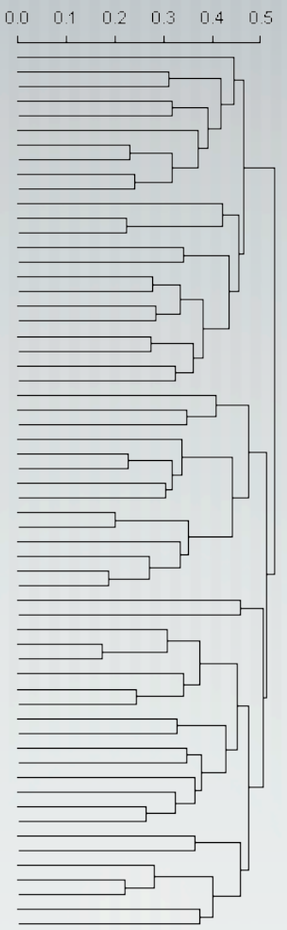
Unsupervised clustering

Our candidate quasar search algorithm is based on unsupervised clustering inside colours space and exploits mixed (spectro + photo) datasets. Once clusters have been formed by the unsupervised algorithms, knowledge-base (spectroscopic types) is used (f.i., “labels” associated to objects within each cluster) in order to understand the nature of correlation between parameters and discover common properties of cluster members.

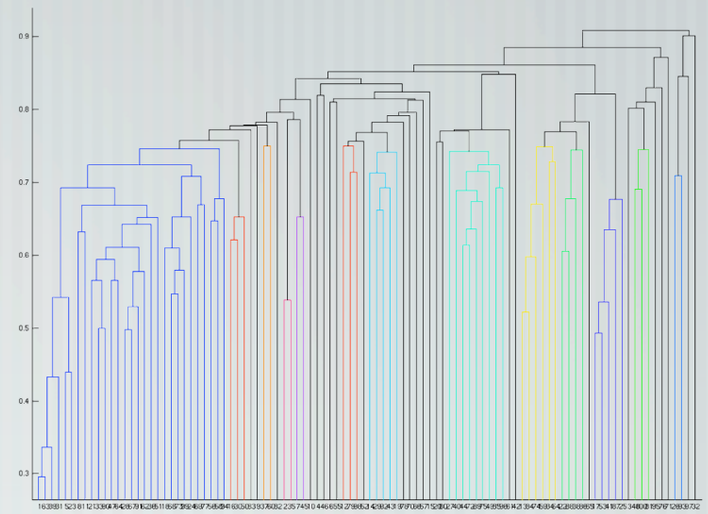
Parameter space

Clustering algorithms: (PPS + NEC)

Parameter space with clusters



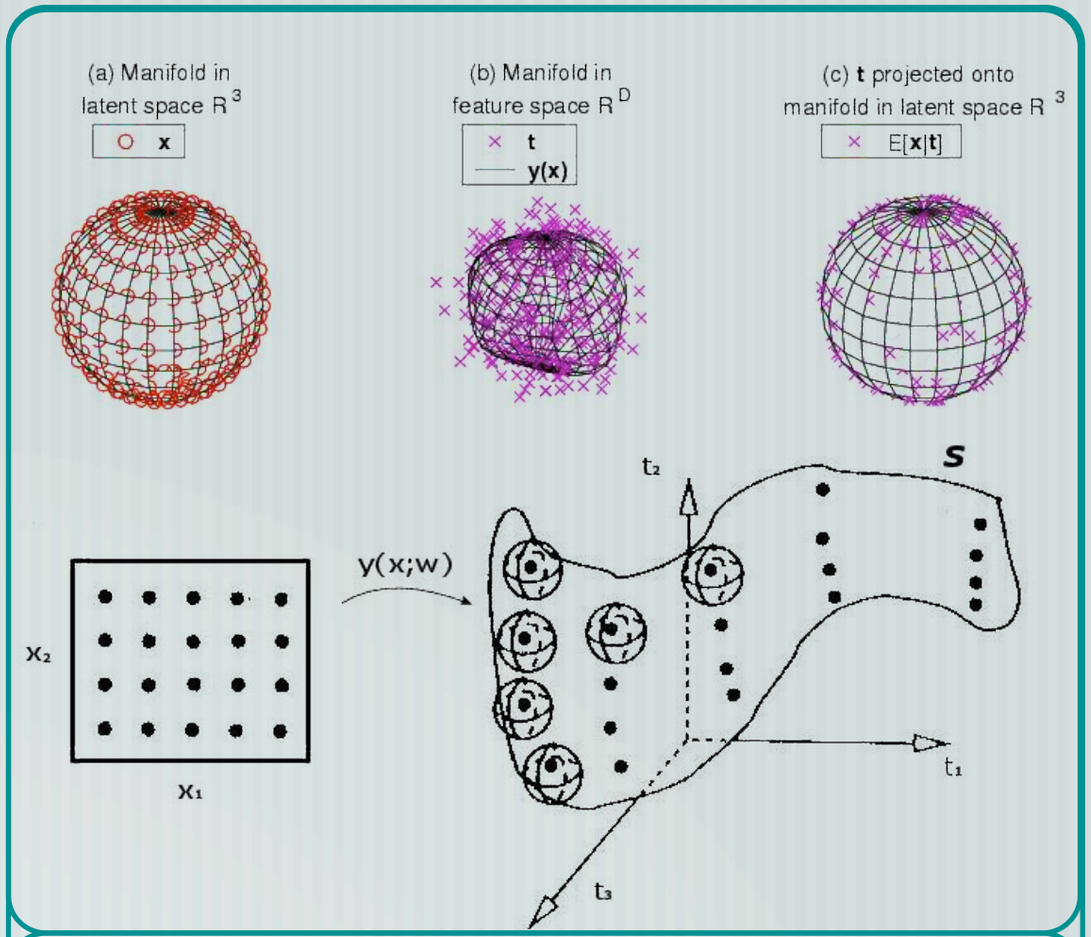
Parameter space with clusters and labels association



2 algorithms for unsupervised clustering

PPS

Probabilistic Principal Surfaces are a non-linear extension of principal components which defines a non linear parametric mapping from a Q-dimensional to D-dimensional space (usually Q << D), in our case the surface of a 2-sphere (Chang, 2001), where the “latent variables” are the knots of the grid on the sphere. In other words, a clusterization of pts in paramater space is produced from the scratch.



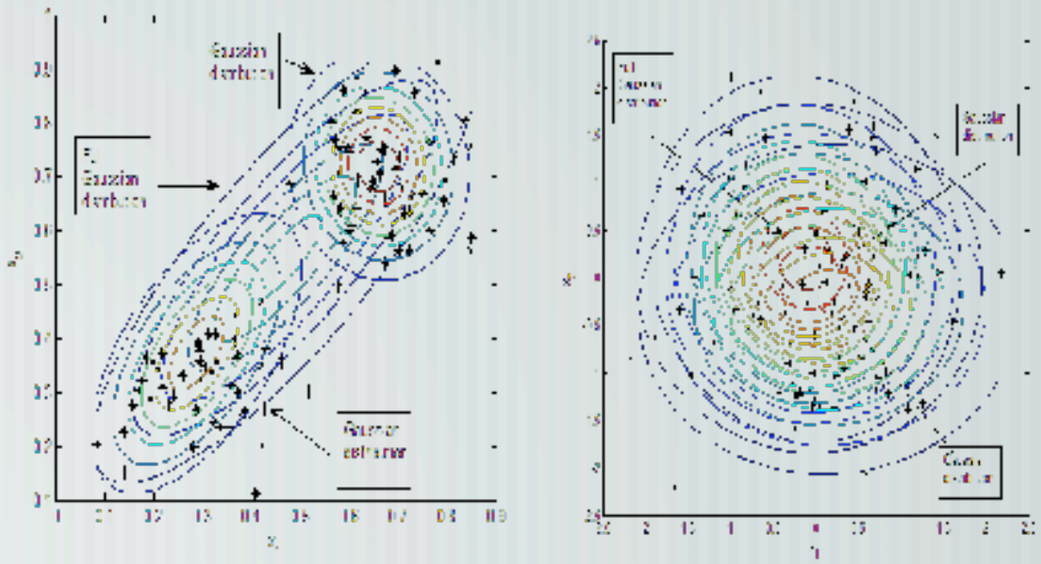
NEC

Clustering method based on “negative entropy”, a measure of the distance from gaussianity of a distribution. For each couple of contiguous (linear Fisher’s discriminant) clusters **A** and **B** in the sample, relations:

$$\mathrm{NegE}(\mathbf{A} \cup \mathbf{B}) < \mathrm{NegE}(\mathbf{A}) + \mathrm{NegE}(\mathbf{B})$$

$$\mathrm{NegE}(\mathbf{A} \cup \mathbf{B}) < \mathbf{D} \qquad (\mathbf{D} \text{ constant})$$

are checked. Whether at least one is true, **A** and **B** are replaced by **C = A** ∪ **B**.



The overall algorithm

1. PPS determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped into near knots on the surface of the sphere.
2. NEC aggregates clusters from PPS to a (a-priori unknown) number of final bundles.
3. These clusters are examined and “interesting” ones are selected.

Two free parameters to be set are the number of latent variables for PPS (“resolution” of the initial clustering) and the critical value(s) of dissimilarity threshold D for NEC.

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (provided that no initial cluster is empty); critical values for **D** are classically determined by two similar methods both embodying a **stability criterion** :

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{st} are those for which a plateau is visible.
2. **Dendrogram analysis**: the stability threshold(s) D_{st} can be determined observing the number of branches at different levels of the graph.

The overall algorithm

1. PPS determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped into near knots on the surface of the sphere.
2. NEC aggregates clusters from PPS to a (a-priori unknown) number of final bundles.
3. These clusters are examined and “interesting” ones are selected.

Two free parameters to be set are the number of latent variables for PPS (“resolution” of the initial clustering) and the critical value(s) of dissimilarity threshold D for NEC.

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (provided that no initial cluster is empty); critical values for **D** are classically determined by two similar methods both embodying a **stability criterion** :

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{st} are those for which a plateau is visible.
2. **Dendrogram analysis**: the stability threshold(s) D_{st} can be determined observing the number of branches at different levels of the graph.

The overall algorithm

1. PPS determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped into near knots on the surface of the sphere.
2. NEC aggregates clusters from PPS to a (a-priori unknown) number of final bundles.
3. These clusters are examined and “interesting” ones are selected.

Two free parameters to be set are the number of latent variables for PPS (“resolution” of the initial clustering) and the critical value(s) of dissimilarity threshold D for NEC.

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (provided that no initial cluster is empty); critical values for **D** are classically determined by two similar methods both embodying a **stability criterion** :

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{st} are those for which a plateau is visible.
2. **Dendrogram analysis**: the stability threshold(s) D_{st} can be determined observing the number of branches at different levels of the graph.

“Training” of the algorithm

Unsupervised clustering in colours space is performed and clusters mainly populated by quasars (according the knowledge-base at our disposal) **are selected; then these clusters are exploited for the candidate quasars selection.**

To determine the critical dissimilarity threshold we rely not only on a stability requirement. Given the following definition:

[cluster is “successful”] Def ⟺ **[its fraction of confirmed quasars is higher then a fixed value]**

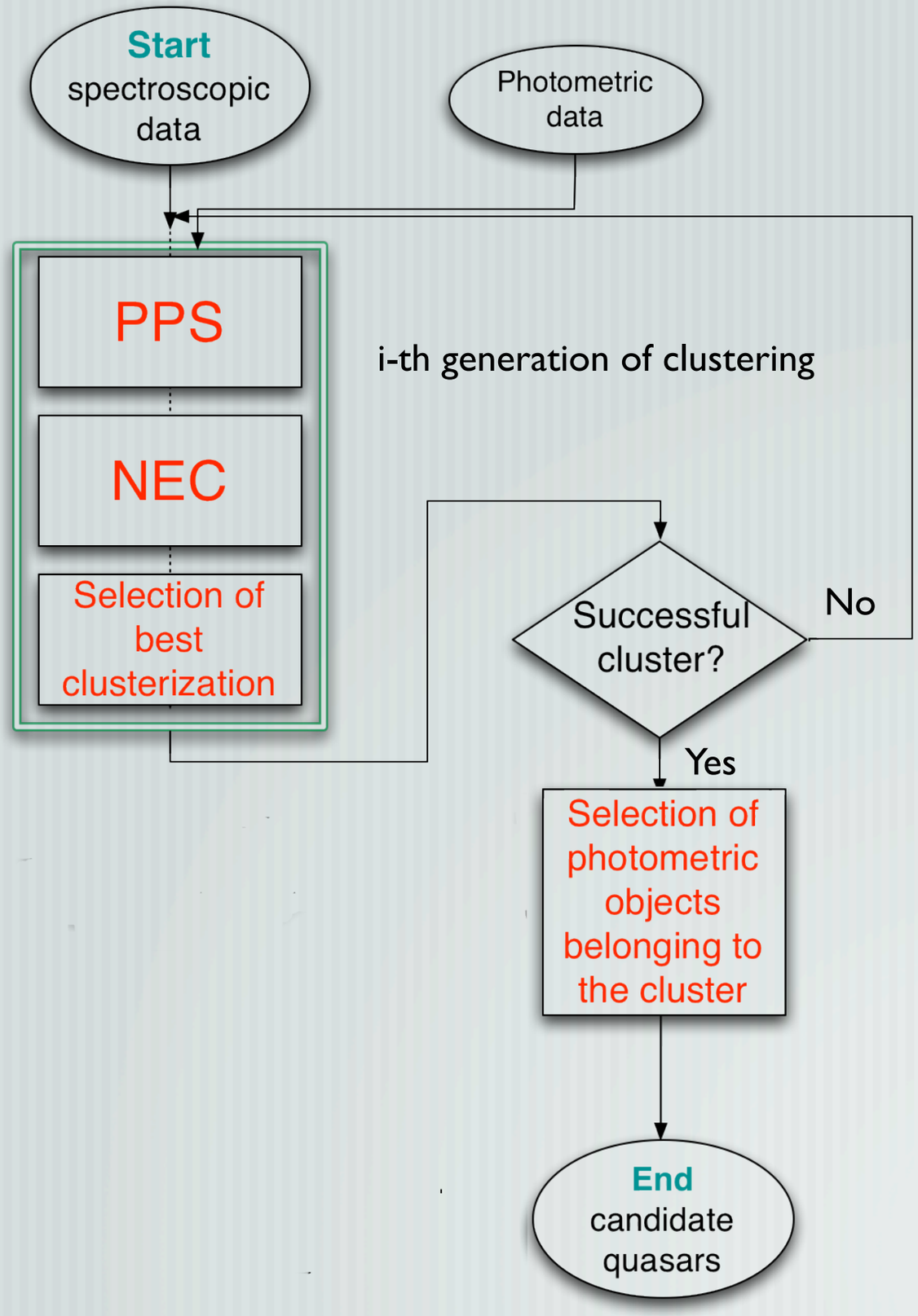
we ask D to maximise the **Normalised Success Ratio** (NSR):

$$NSR = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}$$

The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found. The overall efficiency of the process e_{tot} is the sum of weighed efficiencies e_i for each generation:

$$e_{tot} = \sum_{i=1}^{n} e_i$$

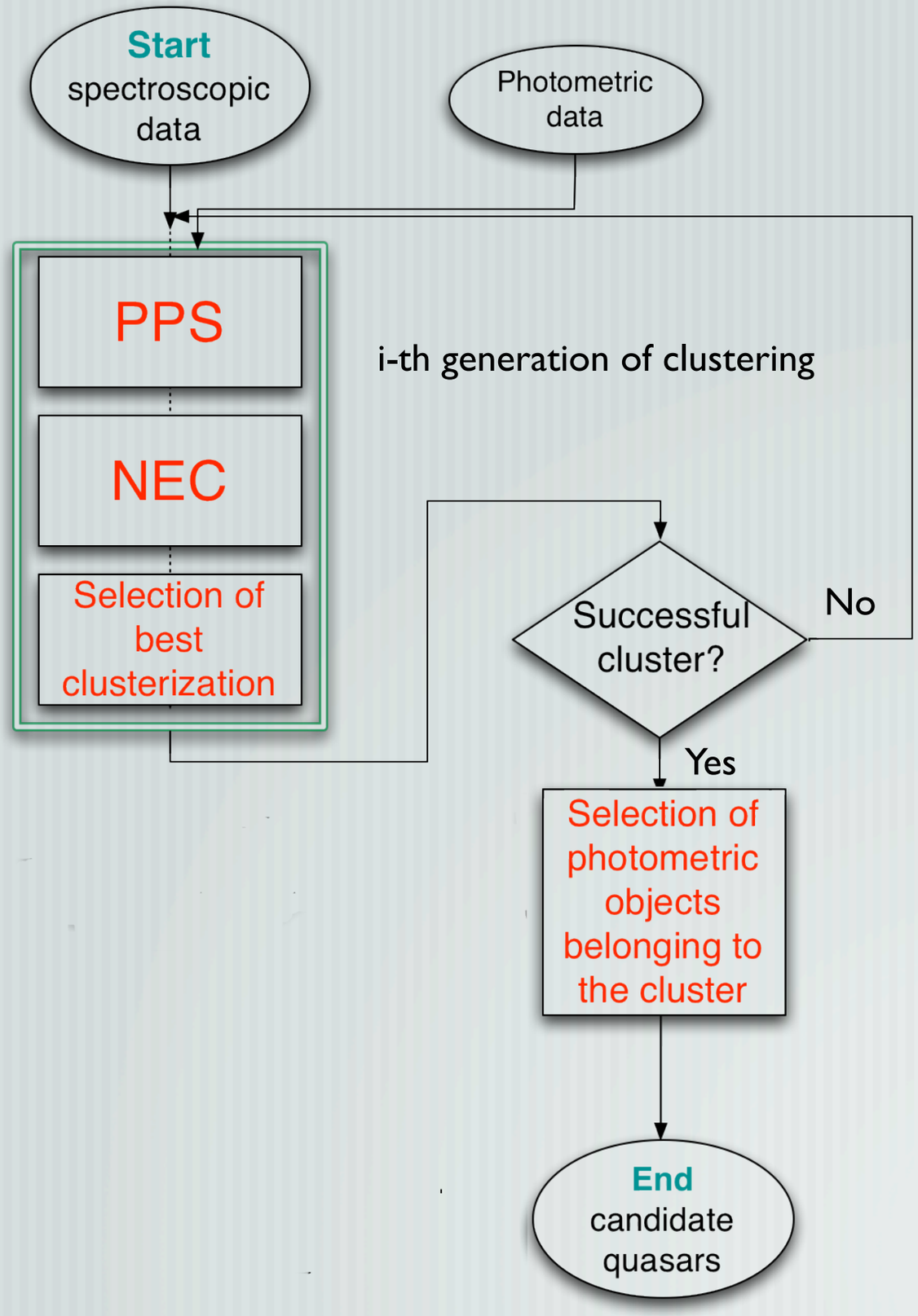
Selection of candidate quasars



1st approach: both spectroscopic and photometric objects put into the same clusterization: selection of candidate quasars as those objects belonging to clusters where spectroscopic confirmed quasars ("tracers") lie.

IT'S SIMPLE AND STRAIGHTFORWARD, BUT...

Selection of candidate quasars



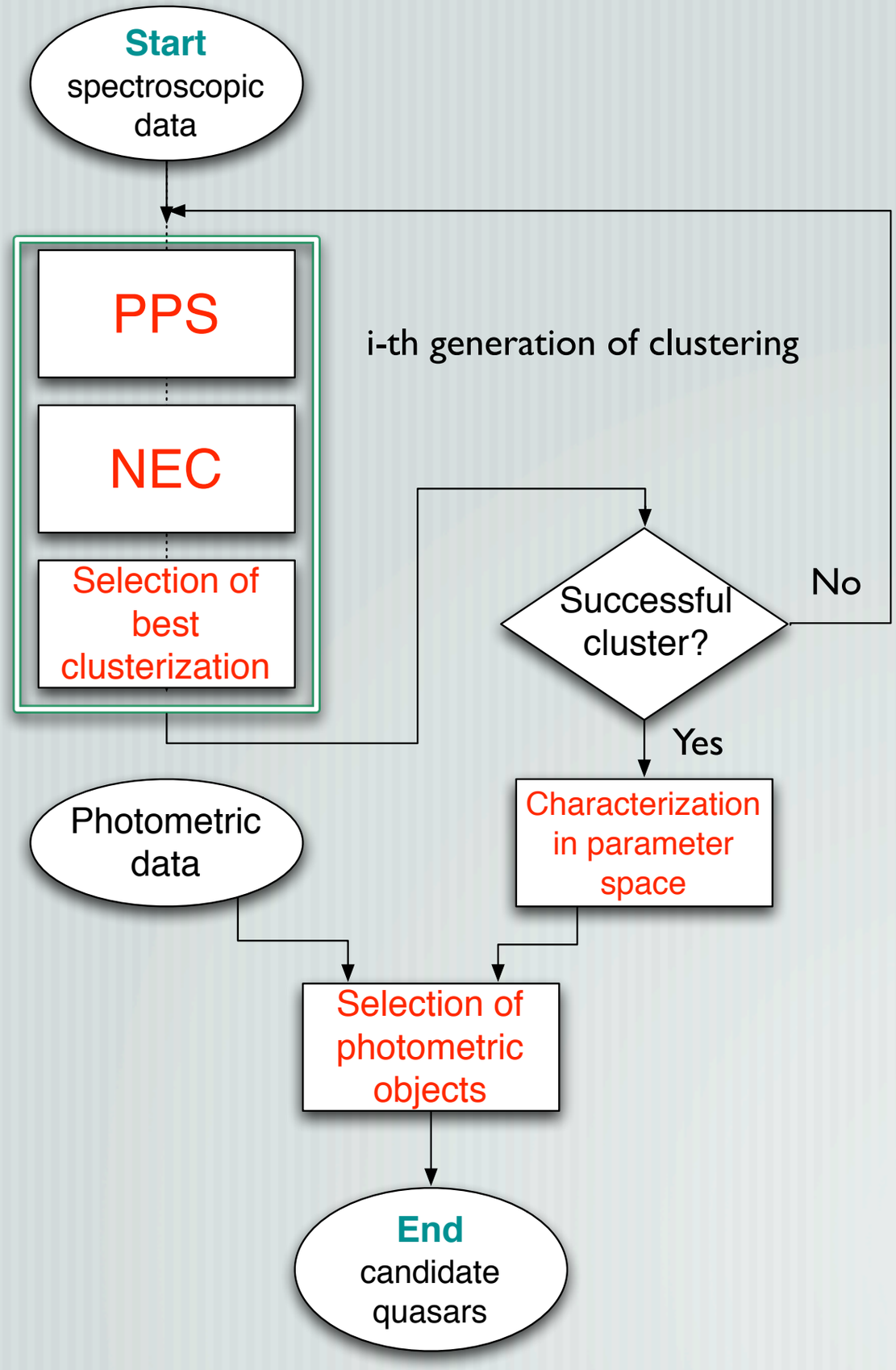
1st approach: both spectroscopic and photometric objects put into the same clusterization: selection of candidate quasars as those objects belonging to clusters where spectroscopic confirmed quasars ("tracers") lie.

IT'S SIMPLE AND STRAIGHTFORWARD, BUT...

PPS performs the best projection according to an estimated probability distribution function from the parameter space to the latent space, which can be heavily modified injecting new points (photometric objects) into the initial sample.

...SO SOMETIMES DOESN'T WORK!

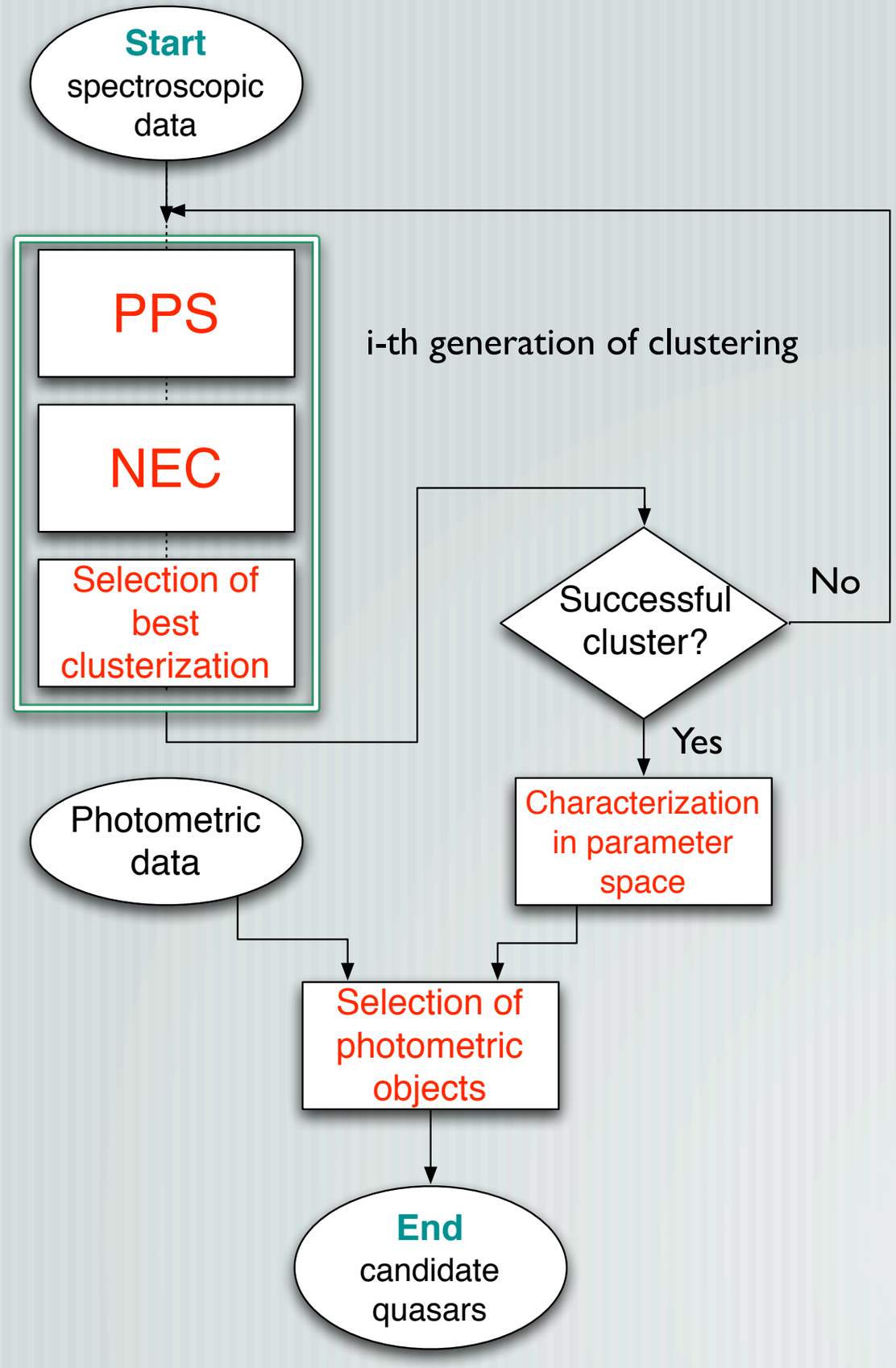
Selection of candidate quasars



2nd approach: characterization of successful clusters obtained using "training" phase objects. Constraints on parameters are applied to photometric sample for candidate quasars determination by AstroGrid "Colour Cutter".

IT WORKS!

Selection of candidate quasars



2nd approach: characterization of successful clusters obtained using "training" phase objects. Constraints on parameters are applied to photometric sample for candidate quasars determination by AstroGrid "Colour Cutter".

IT WORKS!

...BUT IT NEEDS A LITTLE GRAIN OF SALT...

This approach permits "tweaking" candidate selection

Loose contraints →
- efficiency **e**
- completeness **c**
- n° generations

Tight contraints →
- efficiency **e**
- completeness **c**
- n° generations

Experiments: data

Two different samples were used for experiments:

1. **Optical**: sample derived from SDSS database table "Target" queried for quasars candidates, containing ~ $1.11 \cdot 10^5$ records and ~ $5.8 \cdot 10^4$ confirmed quasars ('specClass == 3 OR specClass == 4').

2. **Optical + NIR**: sample derived from positional matching ('best') between SDSS-DR3 database view "Star" queried for all objects with spectroscopic follow-up available and detection in all 5 bands (u,g,r,i,z) with high reliability for redshift estimation and line-fitting classification ('specClass') and high S/N photometry, and UKIDSS-DR1 star-like ('mergedClass == -1') objects fully detected in each of the four lasSurvey bands (Y,J,H,K) and clean photometry. This sample is formed by 2192 objects.

Optical

candidate quasars

4 colours (1)

Optical + NIR

star-like objects

7 colours (2)

Optical

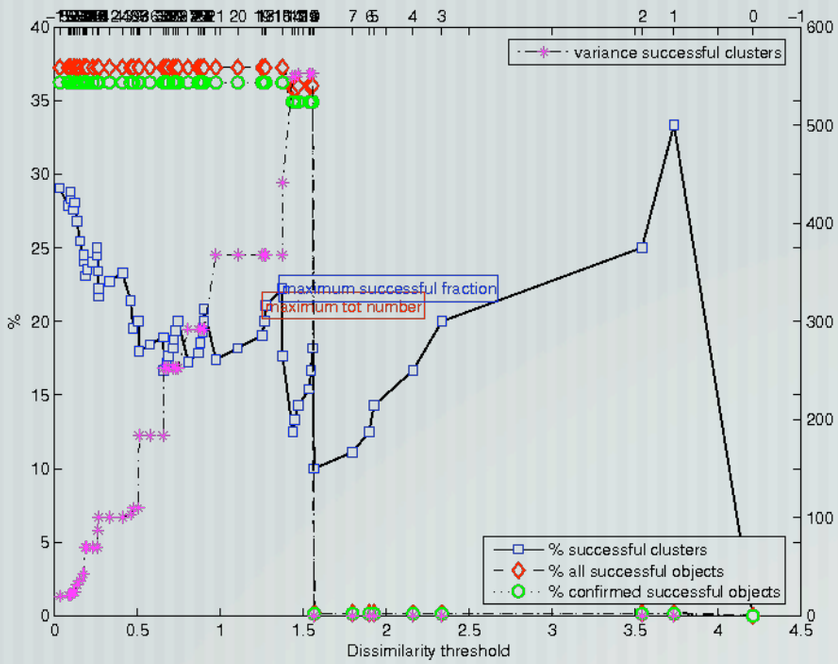
star-like objects

4 colours (3)

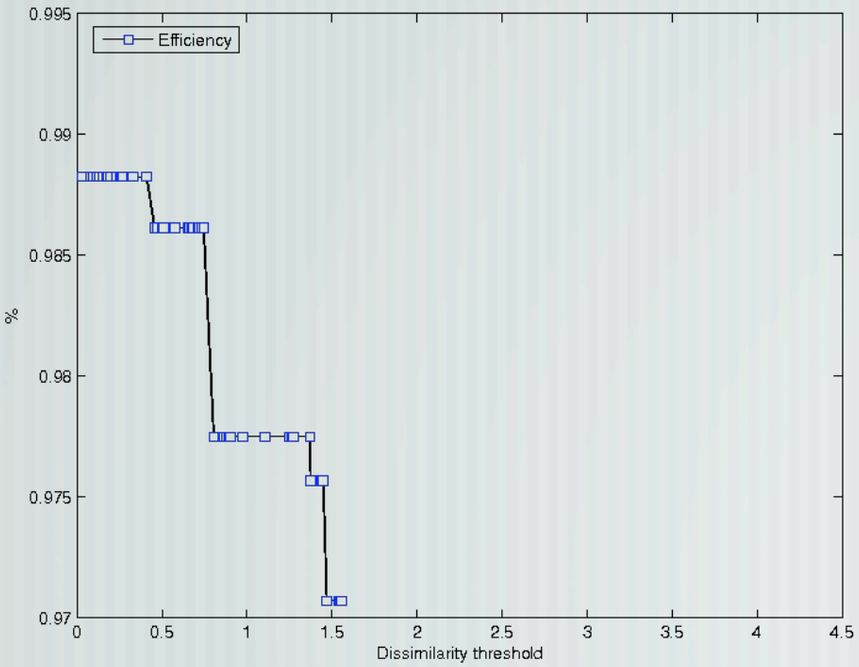
Experiment (2): “training” - candidate selection

”Training”

NSR



Estimated efficiency



Candidate selection

To assess the reliability of the algorithm, the same objects used for the “training” have been re-processed using photometric information only, and results have been checked for consistency.

Confusion matrix

algorithm \ labels	quasars	stars
quasars	794 (94.3 %)	22 (5.7 %)
stars	48 (3.5 %)	1327 (96.5 %)

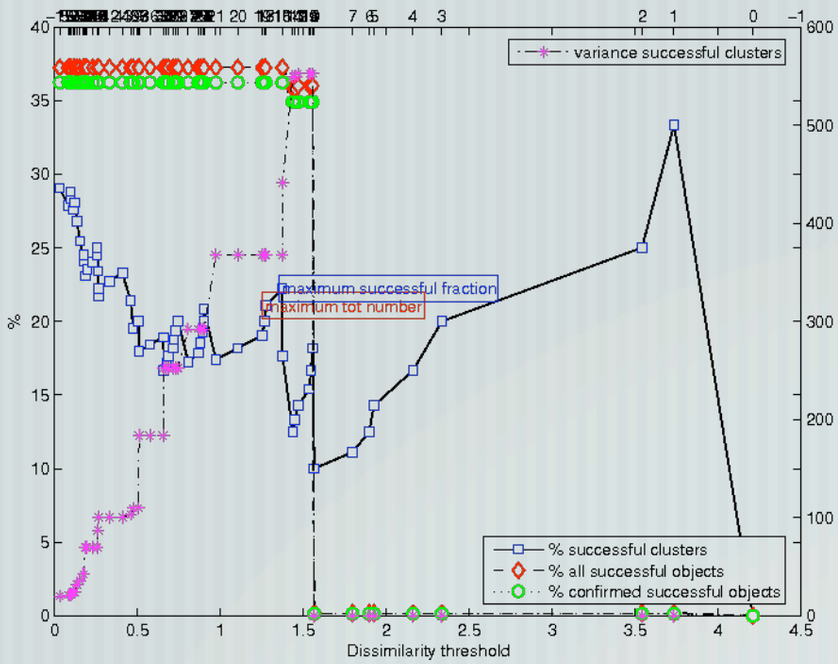
c = 94.3 %

e = 97.3 %

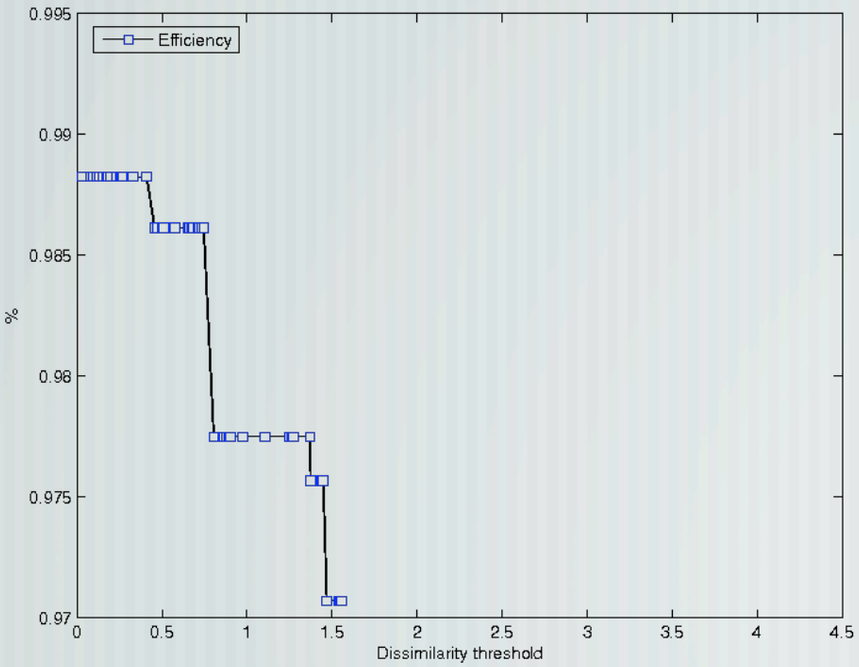
Experiment (2): “training” - candidate selection

”Training”

NSR



Estimated efficiency



Candidate selection

To assess the reliability of the algorithm, the same objects used for the “training” have been re-processed using photometric information only, and results have been checked for consistency.

Confusion matrix

algorithm \ labels	quasars	stars
quasars	794 (94.3 %)	22 (5.7 %)
stars	48 (3.5 %)	1327 (96.5 %)

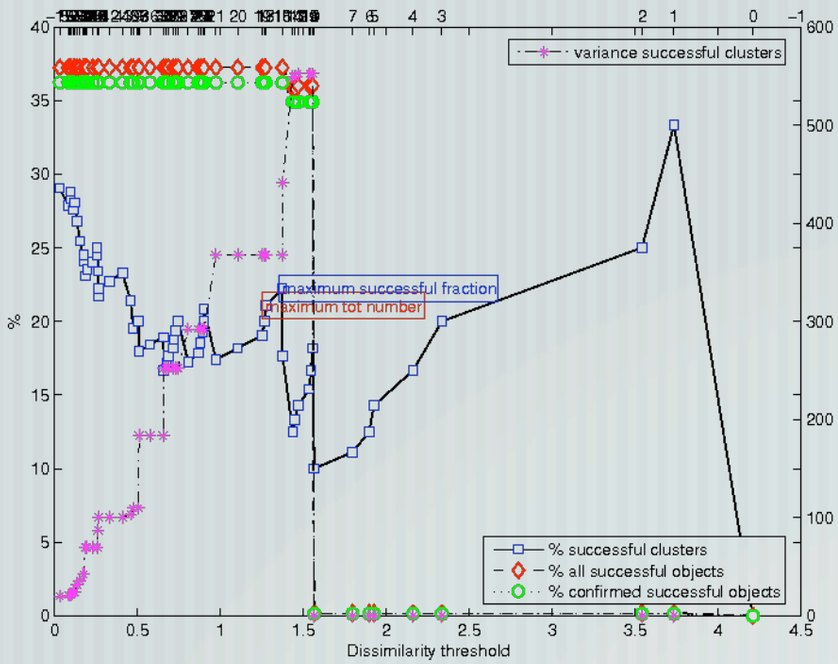
c = 94.3 %

e = 97.3 %

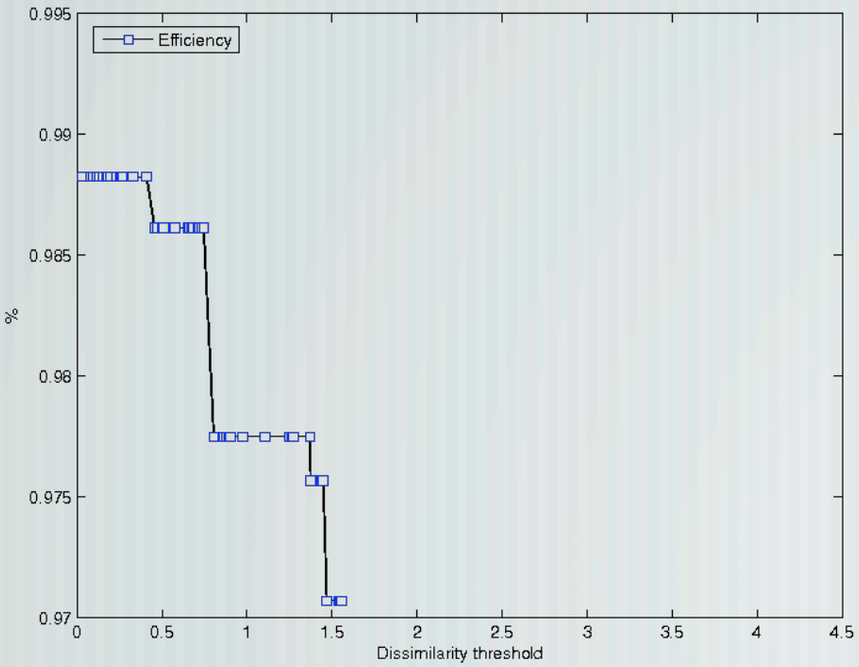
Experiment (2): “training” - candidate selection

”Training”

NSR



Estimated efficiency



Candidate selection

To assess the reliability of the algorithm, the same objects used for the “training” have been re-processed using photometric information only, and results have been checked for consistency.

Confusion matrix

algorithm \ labels	quasars	stars
quasars	794 (94.3 %)	22 (5.7 %)
stars	48 (3.5 %)	1327 (96.5 %)

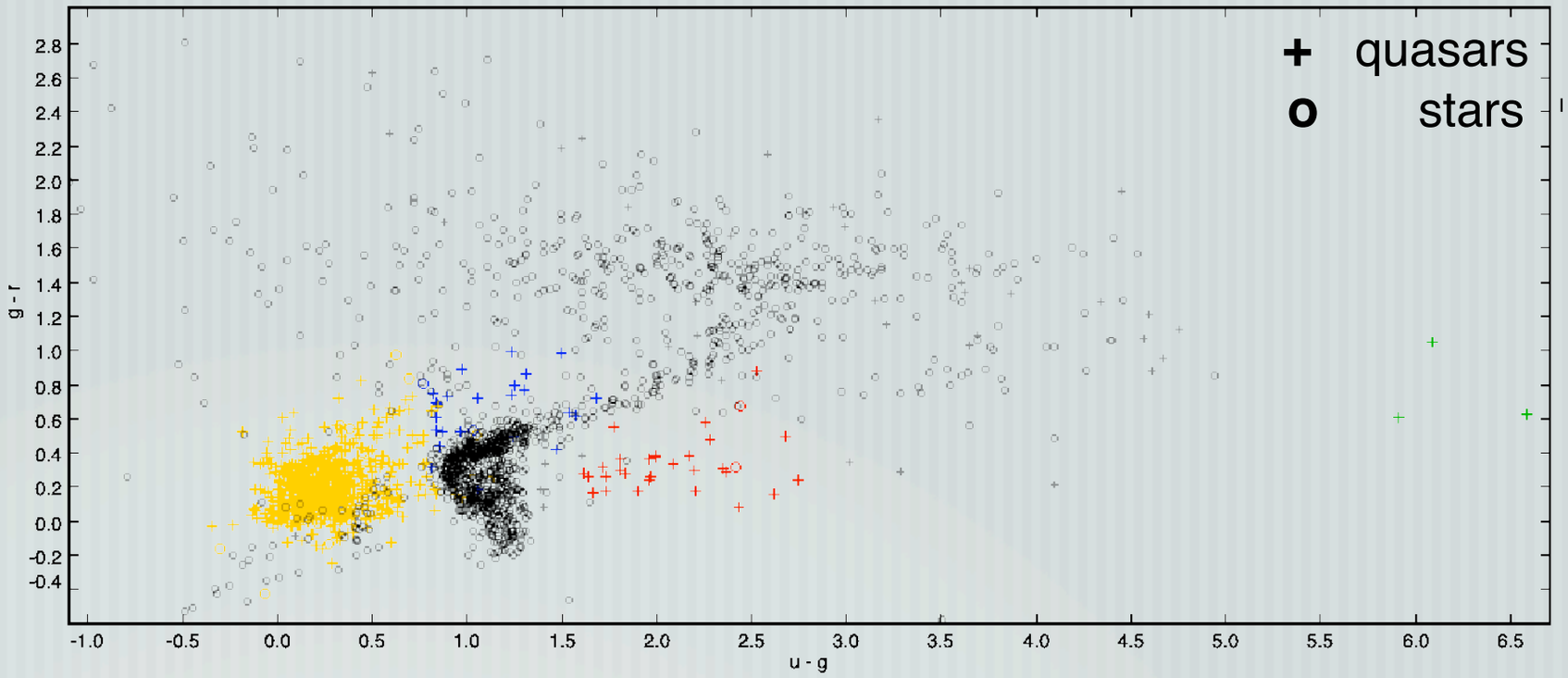
c = 94.3 %

e = 97.3 %

Experiment (2): comparison with SDSS

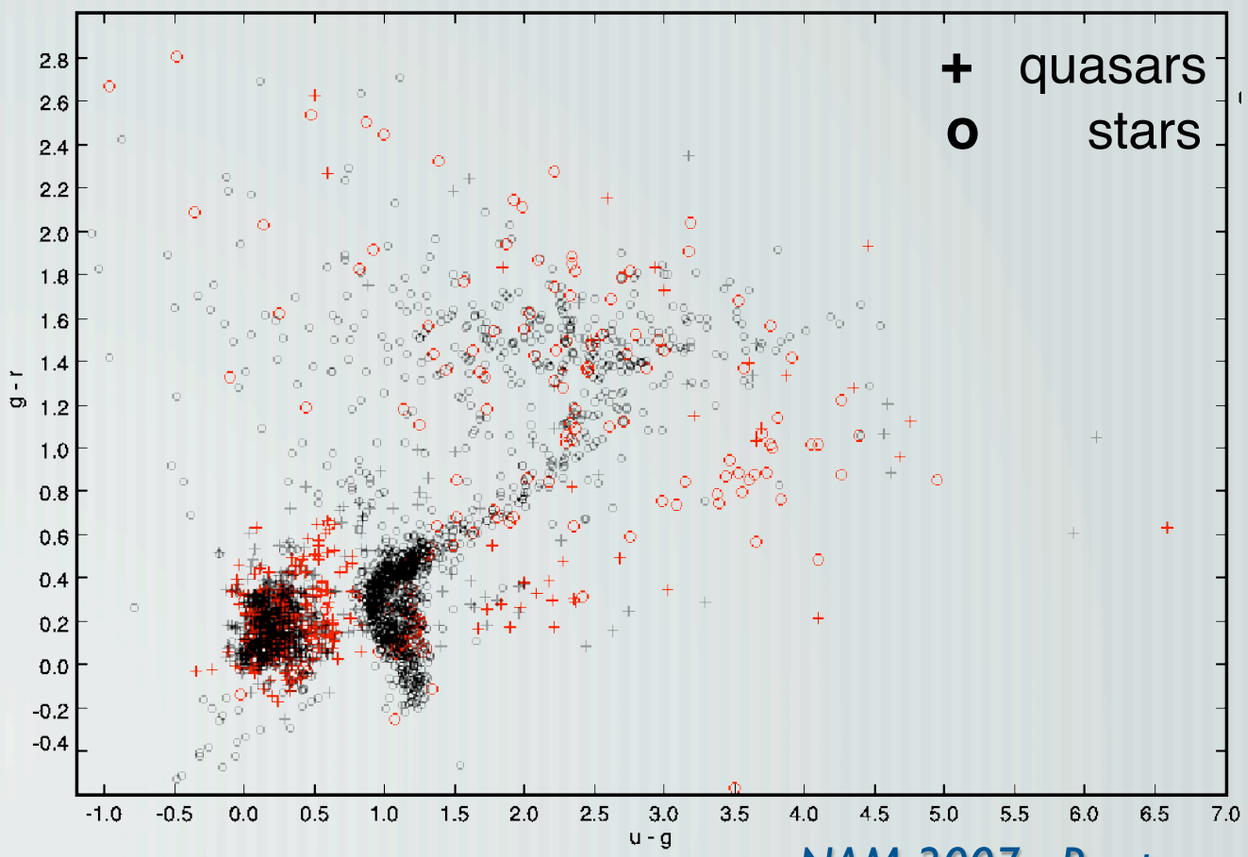
u - g vs **g - r**: **PPS + NEC**

Differently coloured symbols indicate members of different successful clusters. Black symbols are members of stellar clusters (not-successful).



u - g vs **g - r**: **SDSS**

Only a fraction (43%) of these objects have been selected as candidate quasars by SDSS targeting algorithm in first instance: their classification has been achieved thanks to other spectroscopic follow-up programs (star and unknown objects).



Experiments: first results*

Sample	Parameters	Labels	e_{tot}	c_{tot}	n_{gen}	n_{suc_clus}
Optical quasars candidates (1)	SDSS colours	'specClass'	85 % (± 0.5 %)	92.4 % (± 0.5 %)	2	(5,4)
Optical + NIR star-like objects (2)	SDSS colours + UKIDSS colours	'specClass'	94.3 % (± 0.5 %)	97.4 % (± 0.5 %)	3	(4,8,6)
Optical + NIR star-like objects (3)	SDSS colours	'specClass'	87 % (± 0.7 %)	92.7 % (± 0.6 %)	3	(3,6,7)

* D'A., Walton, Longo 2007, in preparation

Conclusions

1. **Unsupervised clustering algorithms** are increasingly promising tools for classification and targeting tasks as new mixed surveys are starting and wealth of archival data gets available.
2. **Ex-novo candidate quasars selection**: comparison with SDSS candidate quasar selection algorithm shows, in general, the usefulness of a more sophisticated approach to the study of the distribution of quasars in colour space, and in particular, Near Infrared luminosities are necessary to partly remove the degeneracy between stars and quasars.

Future work

1. **Cluster members identification**: quasars belonging to different successful clusters have different photometric properties and astrophysical underlying common features. Understanding these similarities can improve observational knowledge of quasars.
2. **Selection of quasars candidates**: determine whether and under which condition 1st approach to selection of quasars candidates is feasible.
3. **Virtual Observatory**: these and others data mining tools (PPS, NEC, MLP) are to be implemented as web services in the VO environment AstroGrid to provide the astronomical community versatile and customizable tools for clustering and classification.