Multivariate Evolutionary Analyses in Astrophysics

Didier Fraix-Burnet

MULTIVARIATE EVOLUTIONARY ANALYSES IN ASTROPHYSICS

Didier Fraix-Burnet
Université Joseph Fourier - Grenoble 1 / CNRS
Laboratoire d'Astrophysique de Grenoble (LAOG) UMR 5571
BP 53, F-38041 GRENOBLE Cedex 09, France
Email: fraix@obs.ujf-grenoble.fr

September 2010

Abstract

The large amount of data on galaxies, up to higher and higher redshifts, asks for sophisticated statistical approaches to build adequate classifications. Multivariate cluster analyses, that compare objects for their global similarities, are still confidential in astrophysics, probably because their results are somewhat difficult to interpret. We believe that the missing key is the unavoidable characteristics in our Universe: evolution. Our approach, known as Astrocladistics, is based on the evolutionary nature of both galaxies and their properties. It gathers objects according to their "histories" and establishes an evolutionary scenario among groups of objects. In this presentation, I show two recent results on globular clusters and earlytype galaxies to illustrate how the evolutionary concepts of Astrocladistics can also be useful for multivariate analyses such as K-means Cluster Analysis.

1 Introduction

We are now able to study galaxies in great detail, identifying individual stars, gas and dust clouds, as well as different stellar populations. Imagery brings very fine structural details, and spectroscopy provides the kinematical, physical and chemical conditions of the observed entities at different locations within a galaxy. For more distant objects, information is scarcer, but deep systematic sky surveys gather spectra for millions of galaxies at various redshifts. The amount of data on galaxies, their number, their diversity, their complexity and that of their evolution, suggest that they should be envisaged as a population or an ensemble of populations. This implies the use of the appropriate statistical tools.

Like paleontologists, we observe objects from the distant past (galaxies at high redshift), and like evolutionary biologists, we want to understand their

relationships with nearby galaxies, like our own Milky Way. Consequently, a "galactogenesis" can be advantageously approached by considering phylogenetics methods.

2 Why be multivariate?

The description of a given galaxy requires many observables, most of them derived from a spectrum. Usual classifications, often inspired by the Hubble tuning fork, use only a very few properties. Even if bivariate plots or correlations are clear and useful, they are incomplete. Worse, they are merely the projection onto a 2-D diagram of a multivariate parameter space. This projection is generally expected to increase the dispersion of the plot. Anyhow, it is difficult to represent many data with only bivariate plots, and any classification necessarily requires an arbitrary binning of one or several parameters.

Multivariate analyses are still not much used in astrophysics. One basic tool, the Principal Component Analysis, is relatively well-known (e.g. Cabanac et al., 2002; Recio-Blanco et al., 2006), but this is not a clustering tool in itself. A very few attempts to apply multivariate clustering methods have been made very recently (Chattopadhyay and Chattopadhyay, 2006, 2007; Chattopadhyay et al., 2007, 2008, 2009a,b; Fraix-Burnet et al., 2009, 2010). Sophisticated statistical tools are used in some areas of astrophysics and are developing steadily, but multivariate analysis and clustering techniques have not much penetrated the community. It is true that the interpretation of the results are not always easy.

3 Why be evolutive?

Evolution, an unavoidable fact, is also not correctly taken into account in most classification methods. By mixing together objects at different stages of evolution, most of the physical significance and usefulness of a classification is lost. In practice, the evolution of galaxies is often limited to the evolution of the properties of the entire population as a function of redshift (Bell, 2005). Since environment (the expanding Universe) and galaxy properties are so much intricate, this kind of study is relevant to a first approximation. However, recent observations have revealed that galaxies of all kinds do not evolve perfectly in parallel, as illustrated for instance by the so-called downsizing effect which shows that large galaxies formed their stars earlier than small ones (e.g. Neistein et al., 2006). New observational instruments now bring multivariate information at different stages of evolution, and in various evolutive environments. In this multivariate context, we believe that the notion of "evolution", easy to understand for a single parameter, is advantageously replaced by "diversification".

The transformation of galaxies is a complex process (Fraix-Burnet et al., 2006b,c) that cannot be disentangled with only a very few observables. For instance, the elliptical shape of galaxy can be obtained through the monolithic collapse of a big cloud of gas, or by big mergers. To find which process has shaped a given galaxy, many observables are required. Only a multivariate and evolutive analysis can distinguish different histories.

4 Classification, complexity, evolution

Multivariate clustering methods compare objects with a given measure and then gather them according to a proximity criterion. There are two main classes. Firstly, distance analyses are based on the overall similarity derived from the values of the parameters describing the objects. The choice of the most adequate distance measure for the data under study is not unique and remains difficult to justify a priori. The way objects are subsequently grouped together is also not uniquely defined. Secondly, methods based on characters (a trait, a descriptor, an observable, or a property, that can be given at least two states characterizing the evolutionary stages of the object for that character) compare objects in their evolutionary relationships (Wiley et al., 1991). Here, the "distance" is an evolutionary cost simply measured by the number of changes of the parameter values (or character states). Groupings are then made on the basis of shared or inherited characteristics, and are most conveniently represented on an evolutionary tree.

Character-based methods like cladistics are better suited to the study of complex objects in evolution, even though the relative evolutionary costs of the different characters is not easy to assess. Distance-based methods are generally faster and often produce comparable results, but the overall similarity is not always adequate to compare evolving objects. In any case, one has to choose a multivariate method, and the results are generally somewhat different depending on this choice (Buchanan and Collard, 2008). However, the main goal is to reveal a hidden structure in the data sample, and the relevance of the method is mainly provided by the interpretation and usefulness of the result.

We must note that taking all available parameters blindly can kill the multivariate and evolutive analysis. One dangerous component is a hidden correlation, such as a size effect, that creates a redundancy. A less known caveat is due to spurious correlations, due to independant variables that vary as function of a non-necessarily obvious parameter. This is especially the case with the time or the stage of evolution. Two quantities can be totally unrelated but if they vary both with time in a more or less monotonic way, then they appear to be correlated. For instance, all photometric quantities for galaxies are affected by the stellar evolution. In such a case, a cladistic

analysis yields a regular tree showing the regular stellar evolution (e.g. Fraix-Burnet, 2006).

Multivariate evolutionary classification in astrophysics has been pioneered by the author (Fraix-Burnet et al., 2006a,c, 2009, 2010; Fraix-Burnet, 2009). Called astrocladistics, it is based on cladistics that is heavily developed in evolutionary biology. Astrocladistics has been first applied to galaxies (Fraix-Burnet et al., 2006b, 2010) because they can be shown to follow a transmission with modification process when they are transformed through assembling, internal evolution, interaction, merger or stripping. For each transformation event, stars, gas and dust are transmitted to the new object with some modification of their properties. Cladistics has also been applied to globular clusters (Fraix-Burnet et al., 2009), where interactions and mergers are probably rare. These are thus simpler stellar systems, even though we have firm evidence that internal evolution can create another generation of stars and that globular clusters can lose mass. Basically, the properties of a globular cluster strongly depend on the environment in which it formed (chemical composition and dynamics), and also on the internal evolution which includes at least the aging of its stellar populations. Since galaxies and globular clusters form in a very evolving environment (Universe, dark matter haloes, galaxy clusters, chemical and dynamical environment), the basic properties of different objects are related to each other by some evolutionary pattern.

5 A more pertinent physical interpretation

An obvious difficulty for a physicist in general is to intepret the results of multivariate analyses using his models that mostly result from a set of equations and are more conveniently presented by curves on bivariate plots. Interestingly enough, these models are multivariate, especially in astrophysics, and the resolution of the set of equations yields a "population" of possible results often called a grid of models. As a result, some parameters are set to sensible values, and the corresponding models are then compared to some observables. These observables can also have been truncated by setting some other observables in order to simplify the information.

It appears that we must here compare two populations, one of real objects and one of models, in a multivariate space. We show here two examples of multivariate (and evolutive) analyses of astrophysical objects showing that such approaches are both more direct, objective and physically pertinent.

Figure 1 shows the cladogram obtained for globular clusters of our Galaxy and the projection of the partitioning on pair plots for the four parameters used for the analysis: logTe, that measures the temperature of stars that are at a specific point in their evolution, Fe/H, MV that is the total visible intensity (magnitude) and roughly indicates the mass of the globular cluster,

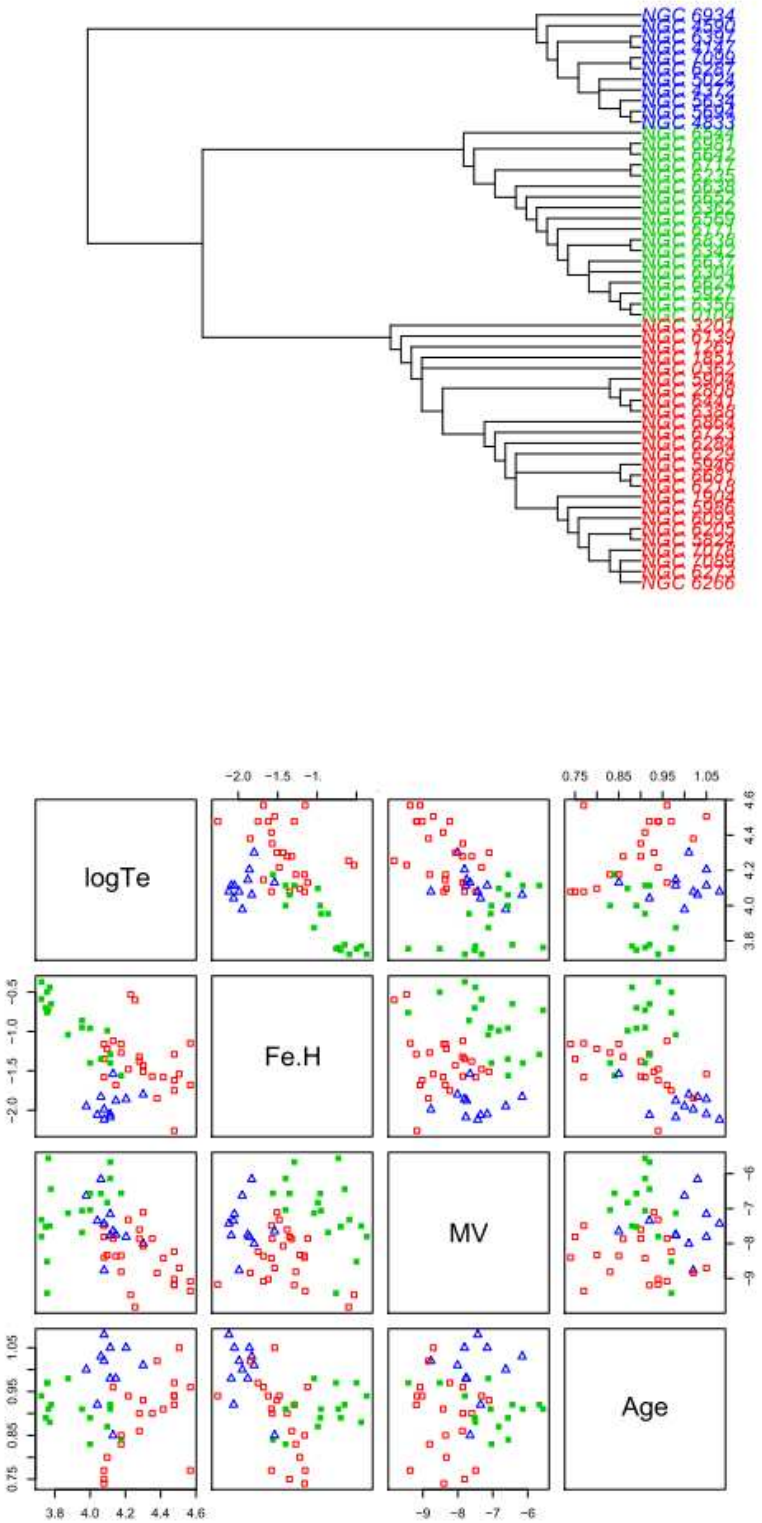

Figure 1: *Top*: cladogram obtained for the globular clusters of our Galaxy. *Bottom*: projection of the partitioning on pair plots with the four parameters used for the cladistic analysis. From Fraix-Burnet et al. (2009).

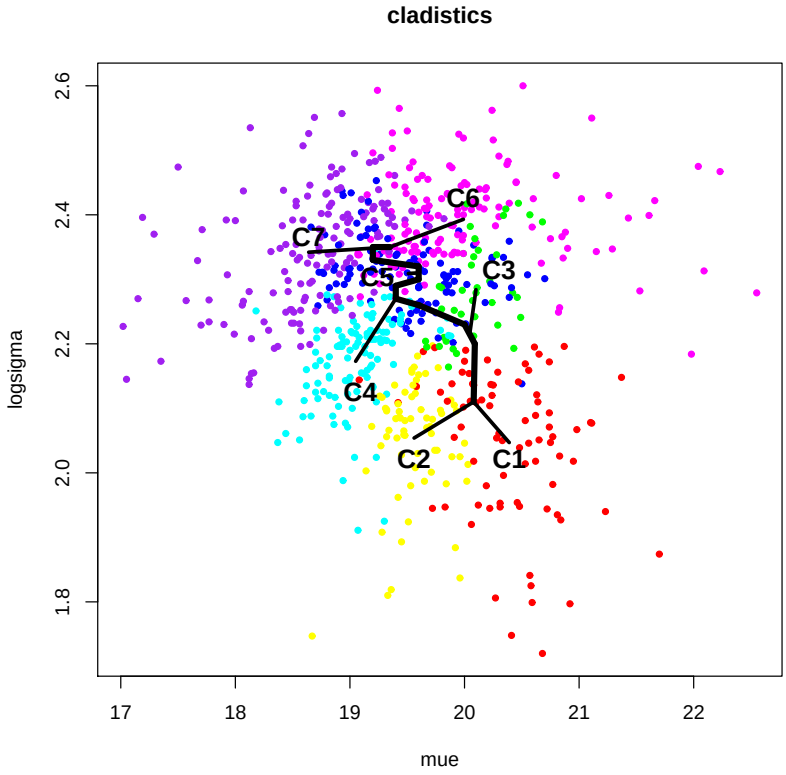

Figure 2: The fundamental plane of 699 galaxies, showing the partitioning and projection of the tree obtained by a cladistic analysis (Fraix-Burnet et al., 2010).

and Age that can be measured quite precisely because all stars of a given globular cluster are formed nearly at the same time. However Age is not an intrinsic property discriminating evolutionary groups since it evolves in the same way for all. But we gave it a half weight to arrange the objects within each group (Fraix-Burnet et al., 2009). Three groups are identified. The first one (in blue) has on average the lower ratio Fe/H that measures the proportion of heavy atomic elements that are processed within stars. This group is consequently considered as more primitive. It is obivous that this partitioning would be impossible to obtain with only bivariate plots.

Looking at other parameters (such as orbital elements, kinematics, more refined chemical abundances...) revealed clear characteristics that allowed us to infer that each group formed during a particular stage of the assembly history of our Galaxy. The blue group is the older one. It formed during the dissipationless collapse of the protogalaxy. They are located mainly in the outer halo. The red group belongs to the inner halo and the corre-

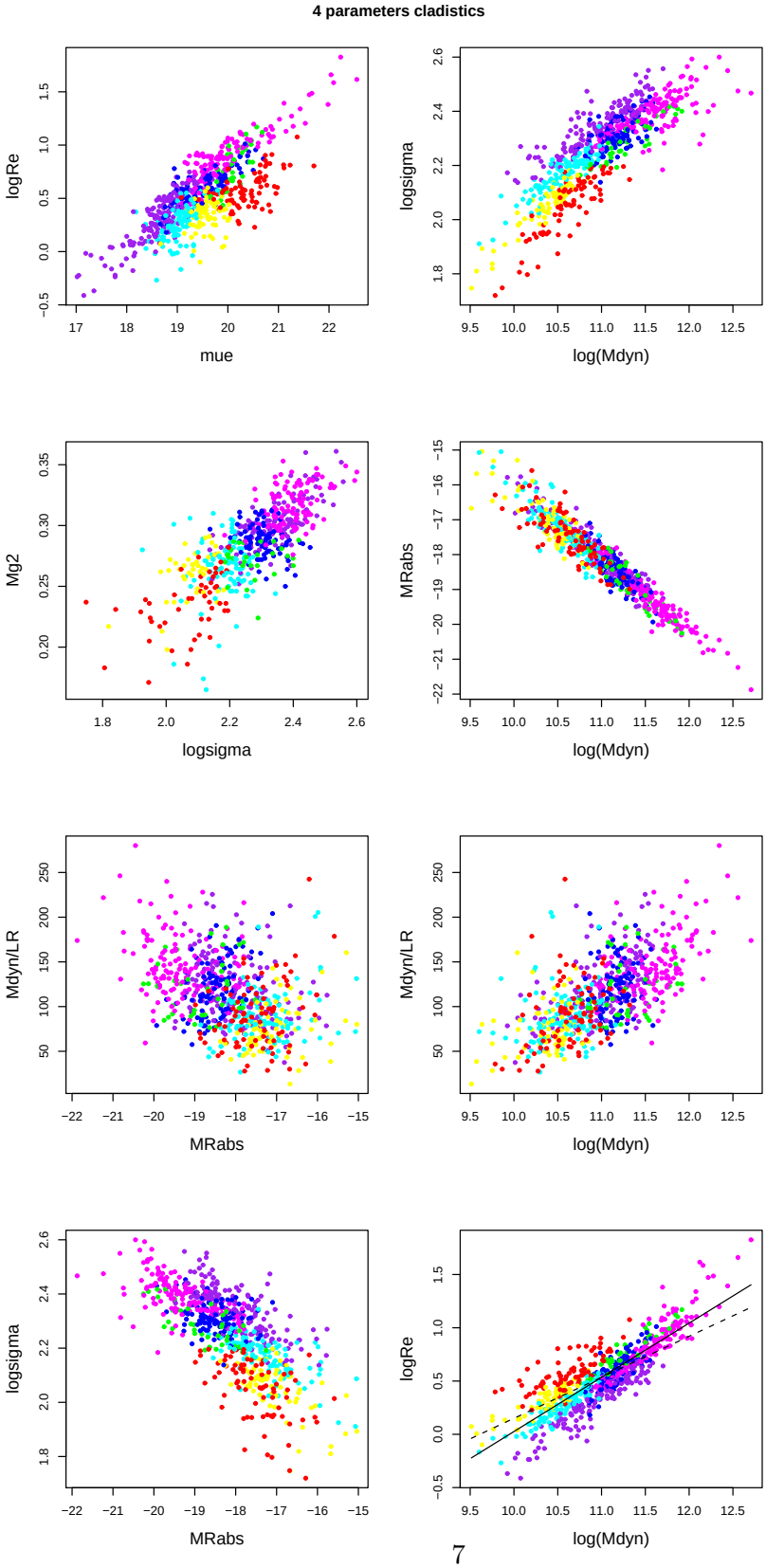

Figure 3: Cluster and cladistic analysis of the fundamental plane of early-type galaxies: bivariate plots showing how correlations differ for each group and for the whole sample. Note in particular the Mg_2 vs $\log \sigma$ plot revealing a spurious correlation between these two parameters (Fraix-Burnet et al., 2010).

sponding clusters formed at a later stage during the dissipational phase of Galactic collapse, which continued in the halo after the formation of the thick disc and its globular clusters. These clusters were very massive before "star evaporation" took place. The latter group (green) formed during an intermediate and relatively short period and comprises clusters of the disk of our Galaxy (all details in Fraix-Burnet et al., 2009).

Another example is given with the fundamental plane of early-type galaxies which is a lonog-known correlation between the central velocity dispersion σ, the surface brightness μ_e and the effective radius r_e. In addition, the metallicity, as measured with the Mg_2 index, plays a role and seems to be correlated with $\log \sigma$. We performed a K-means cluster analysis and a cladistic analysis in parallel (Fraix-Burnet et al., 2010). The partionings are remarkably in agreement. We believe the reason is due to the careful choice of the parameters. For cladistics, they must be informative with respect to diversification, and should not be redundant or incompatible. This requirement is logically pertinent also for any cluster analysis.

Cladistics provides in addition the evolutionary relationships between the groups. On Figure 2, the tree is projected onto the $\log \sigma$ vs μ_e diagram on which the fundamental plane is seen essentialy face-on. Since galaxies are more complicated than globular clusters, the interpretation of the results and all relations between all possible parameters and within each group takes great advantage of numerical simulations. Here again, we are able to derive the probable history of each group of galaxies as well as their relative level of diversification, giving possible sequences of past transforming events such as mergers, accretions or sweeping (for details, see Fraix-Burnet et al., 2010).

A quite interesting finding is that most known correlations are different or even absent when we consider groups individually (Figure 3). This proves that they have different evolution histories. Another noticeable fact is that the well-known correlation between Mg2 and $\log \sigma$ is indeed spurious, or historical. It is due to the fact that each parameter changes with the level of diversification as clearly shown by the placement of the groups (see Figure 3).

6 Conclusion

Undoubtly, the study of galaxies now requires multivariate statistical treatments. Evolution must also be taken into account and the concept of populations seems appropriate and points to the use of methodologies developed elsewhere. Complexity, evolution and classification suggest similar studies as in phylogenetics. Astrocladistics has opened the pathway.

References

Bell, E., 2005. Galaxy Assembly. Planets to Cosmology: Essential Science in Hubble's Final Years. Cambridge: CUP. astro-ph/0408023.

Buchanan, B., Collard, M., 2008. Phenetics, cladistics, and the search for the alaskan ancestors of the paleoindians: a reassessment of relationships among the clovis, nenana, and denali archaeological complexes. *Journal of Archaeological Science* 35, 1683–1694.

Cabanac, R.A., de Lapparent, A., Hickson, P., 2002. *Astronomy & Astrophysics* 389, 1090–1116.

Chattopadhyay, A., Chattopadhyay, T., Davoust, E., Mondal, S., Sharina, M., 2009a. Study of ngc 5128 globular clusters under multivariate statistical paradigm. *Astrophysical Journal* 705, 1533. arXiv:0909.4161.

Chattopadhyay, T., Babu, J., Chattopadhyay, A., Mondal, S., 2009b. Horizontal branch morphology of globular clusters: A multivariate statistical analysis. *Astrophysical Journal* 700, 1768.

Chattopadhyay, T., Chattopadhyay, A., 2006. Objective classification of spiral galaxies having extended rotation curves beyond the optical radius. *The Astronomical Journal* 131, 2452–2468.

Chattopadhyay, T., Chattopadhyay, A., 2007. Globular clusters of local group – statistical analysis. *Astronomy & Astrophysics* 472, 131–140.

Chattopadhyay, T., Misra, R., Naskar, M., Chattopadhyay, A., 2007. Statistical evidences of three classes of gamma ray bursts. *Astrophysical Journal* 667, 1017. arXiv:0705.4020.

Chattopadhyay, T., Mondal, S., Chattopadhyay, A., 2008. Globular clusters in the milky way and dwarf galaxies - a distribution-free statistical comparison. *Astrophysical Journal* 683, 172.

Fraix-Burnet, D., 2006. Determining the evolutionary history of galaxies by astrocladistics: some results on close galaxies, in: D. Barret, F. Casoli, S.C.F.C.T.C., Pagani, L. (Eds.), Journées de la SF2A, Paris (France), Société Française d'Astronomie et d'Astrophysique (SF2A). http://hal.archives-ouvertes.fr/ccsd-00104352.

Fraix-Burnet, D., 2009. Evolutionary Biology Concept, Modeling, and Application. Springer Berlin Heidelberg. chapter Galaxies and Cladistics. Biomedical and Life Sciences, pp. 363–378. arXiv:0909.4164.

Fraix-Burnet, D., Choler, P., Douzery, E., 2006a. Towards a Phylogenetic Analysis of Galaxy Evolution : a Case Study with the Dwarf Galaxies of the Local Group. *Astronomy and Astrophysics* 455, 845–851. astro-ph/0605221.

Fraix-Burnet, D., Choler, P., Douzery, E., Verhamme, A., 2006b. Astrocladistics: a phylogenetic analysis of galaxy evolution I. Character evolutions and galaxy histories. *Journal of Classification* 23, 31–56.

Fraix-Burnet, D., Davoust, E., Charbonnel, C., 2009. The environment of formation as a second parameter for globular cluster classification. *MNRAS* 398, 1706–1714. arXiv:0906.3458.

Fraix-Burnet, D., Douzery, E., Choler, P., Verhamme, A., 2006c. Astrocladistics: a phylogenetic analysis of galaxy evolution II. Formation and diversification of galaxies. *Journal of Classification* 23, 57–78.

Fraix-Burnet, D., Dugué, M., Chattopadhyay, A., Chattopadhyay, T., Davoust, E., 2010. Structures in the fundamental plane of early-type galaxies. *Monthly Notices of the Royal Astronomical Society* accepted for publication. http://fr.arxiv.org/abs/1005.5645.

Neistein, E., van den Bosch, F., Dekel, A., 2006. Natural downsizing in hierarchical galaxy formation. *Monthly Notices of the Royal Astronomical Society* 372, 933–948. astro-ph/0605045.

Recio-Blanco, A., Aparicio, A., Piotto, G., De Angeli, F., Djorgovski, S., 2006. Multivariate analysis of globular cluster horizontal branch morphology: searching for the second parameter. *Astronomy & Astrophysics* 452. http://arxiv.org/abs/astro-ph/0511704.

Wiley, E., Siegel-Causey, D., Brooks, D., Funk, V., 1991. The Compleat Cladist: A Primer of Phylogenetic Procedures. The University of Kansas, Museum of Natural History, Special Publication No. 19.