Supervised and Unsupervised Classification – The Case of IRAS Point Sources

H.-M. Adorf
Space Telescope – European Coordinating Facility
European Southern Observatory
Garching b. München, F.R. Germany

E.J.A. Meurs
European Southern Observatory
Garching b. München, F.R. Germany

Abstract

Progress is reported on a project which aims at mapping the extragalactic sky in order to derive the large scale distribution of luminous matter. Our approach consists in selecting from the IRAS Point Source Catalog a set of galaxies which is as clean and as complete as possible. The decision and discrimination problems involved lend themselves to a treatment using methods from multivariate statistics, in particular statistical pattern recognition. Two different approaches – one based on supervised Bayesian classification, the other on unsupervised data-driven classification – are presented and some preliminary results are reported.

1 Introduction

The Infrared Astronomical Satellite (IRAS) was launched in January 1983 and successfully operated for a period of about 300 days, during which more than 96 % of the sky was surveyed at an angular resolution between $\sim 0.5'$ and $\sim 2'$ depending on wavelength (Beichman *et al.* 1985). The Point Source Catalog (PSC) resulting from the IRAS mission constitutes an attractive database for classification pursuits. With a high level of homogeneity and almost complete sky coverage, the PSC provides positions and infrared fluxes at four wavelengths for a total of $\sim 250\,000$ sources. IRAS looked relatively unhampered through much of the Galaxy, but nearer the galactic centre the high source density causes noticeable source confusion along the galactic plane.

On the basis of their infrared colours (flux ratios), sources contained in the PSC can to a large degree be separated into four main categories, as has been demonstrated in a number of studies (Chester 1986, Lawrence *et al.* 1986, Wolstencroft *et al.* 1986, Habing 1987, Soifer *et al.* 1987). Exploiting this property, Meurs and Harmon (1988) have produced sky maps for these source categories, while aiming specifically at a homogeneous map of almost the entire extragalactic sky. The other source categories that could be distinguished are stars, a very thin galactic component (which may largely consist of H II regions) and a broader and more diffusely distributed galactic

component that is related to star forming regions (where also the Gould Belt can be recognised).

A successful and reliable selection of extragalactic objects from the IRAS PSC will be highly interesting for various research projects. The importance of a homogeneous and complete all-sky sample of galaxies for a dipole anisotropy determination has already been demonstrated (Harmon *et al.* 1987). Although cosmological inhomogeneities may be less prominent in the sky distribution of spiral galaxies (mainly recorded in the PSC) when compared to ellipticals, the all-sky homogeneity of the IRAS data render them an important sample for studying sky distribution features. Besides this, the forthcoming Co-Added Catalog may be expected to contain ellipticals as well (*cf.* Knapp 1987). Other areas where such an extragalactic IRAS sample may be beneficial include studies of luminosity functions and of statistical relations involving IRAS data. Furthermore, regions much nearer the galactic plane become accessible for all kinds of research (see *e.g.* Dow *et al.* 1988).

2 A sidestep into data analysis methodology

Data analysis may be subdivided into two categories, *exploratory* and *confirmatory*. With confirmatory data analysis one tries to corroborate – or falsify – a specific preconceived hypothesis. Exploratory data analysis, on the other hand, aims at discovering regularity or structure inherent to a given data set "with no preconceived notions or precise questions in mind" (Friedman 1986). In exploratory mode, the data analyst must be open to several equally legitimate structures in the data. The exploration phase logically precedes the confirmation phase and is a prerequisite for forming any hypothesis to be tested.

Exploratory data analysis is commonly performed by constructing a classification scheme over the set of data points (objects), where the abstract clustering task can be defined as follows (Fisher and Langley 1986): "Given: A set of objects, O. Goal: Distinguish clusters (*i.e.* subsets of O) $s_1, \ldots, s_n$, such that the intra-cluster object similarity of each s_i tends to be maximized, and the inter-cluster object similarity over all s_j's tends to be minimized." If successful, such a classification procedure results in a data description, which is more condensed and therefore easier to communicate than the original set itself. Creating a classification is also a typical first step in developing a theory about a collection of observations (Stepp and Michalski 1986b).

When no *a priori* information is given for the association of the objects with categories, the classification process is said to be *unsupervised*. For quite some time unsupervised classification or "learning without a teacher" was widely felt to be impossible, and, indeed it is not uniquely possible in general (Cooper 1969a, b) because quite often the same data can be organised in different ways (Fisher and Langley 1986). In particular, difficulties may arise from overlapping or interleaved categories, or when the data stem from a continuous distribution displaying no natural decomposition into classes.

Nevertheless, unsupervised classification is not only possible in a wide range of situations, as has been shown for example by Cooper and Cooper (1964), but also of

importance, because *supervised* classification may be inconvenient, too costly or even impossible, for one of the following reasons:

(i) classes may be unknown, *e.g.* because data are coming from a new instrument *(problem novelty)*;
(ii) the dimensionality of the feature space may be too high for easy visualisation *(problem complexity)*;
(iii) it may be difficult to separate the various populations from each other *(problem difficulty)*;
(iv) the number of objects to be considered may be very large *(problem size)*.

Given the need for *unsupervised* classification and the potential difficulties a human analyst may run into, the intriguing question is, to what extent the class formation task could be carried out by a computer, with its capability of analysing large data sets *automatically* and *objectively*.

Past work on automated generation of classes was performed under the headings of numerical taxonomy and cluster analysis. Numerical taxonomy offers a number of algorithmic clustering techniques: *e.g. optimisation*, which attempts to construct an optimal partition of the data set into mutually exclusive classes; *hierarchical clustering*, which forms a classification tree over the object set; and *clumping*, which allows for overlapping classes. Kurtz (1983), Murtagh (1986, 1987), and Murtagh and Heck (1987a, b) review the application of some of these techniques to astronomical problems.

Recently the automated generation of classification schemes has attracted researchers from the area of artificial intelligence (see the bibliography by Kedar-Cabelli and Mahadevan 1986). From an artificial intelligence perspective, numerical taxonomy can be viewed as a first step towards *conceptual clustering* (Fisher and Langley 1986, Michalski and Stepp 1983, Stepp and Michalski 1986a, b), an artificial intelligence technique which aims at identifying higher level (conceptual) descriptions of object groups. However, the claim that conceptual clustering is superior to numerical taxonomy has been criticised by Dale (1985).

Once a partition of feature space into classes has been established, *supervised* techniques from the well-founded theory of statistical inference, in particular decision theory, can be used to classify additional objects into classes derived from "training sets". (For various aspects of pattern recognition and classification see Watanabe 1969, Grasselli 1969, Duda and Hart 1973, Bock 1974, Batchelor 1978, Melsa and Cohn 1978, Fu 1980, Hand 1981, Sklansky and Wassel 1981, Kulkarni 1986, Jain 1987, Mantas 1987.)

3 Supervised classification of IRAS point sources

A first step towards a proper classification of the IRAS PSC, applying multivariate statistical methods and concepts from decision theory, was made by Meurs *et al.* (1988). The distributions of data points in four-dimensional feature space – three infrared colours and one flux, as in Meurs and Harmon (1988) – were represented by multivariate Gaussian distributions. These were fitted to training sets for each source

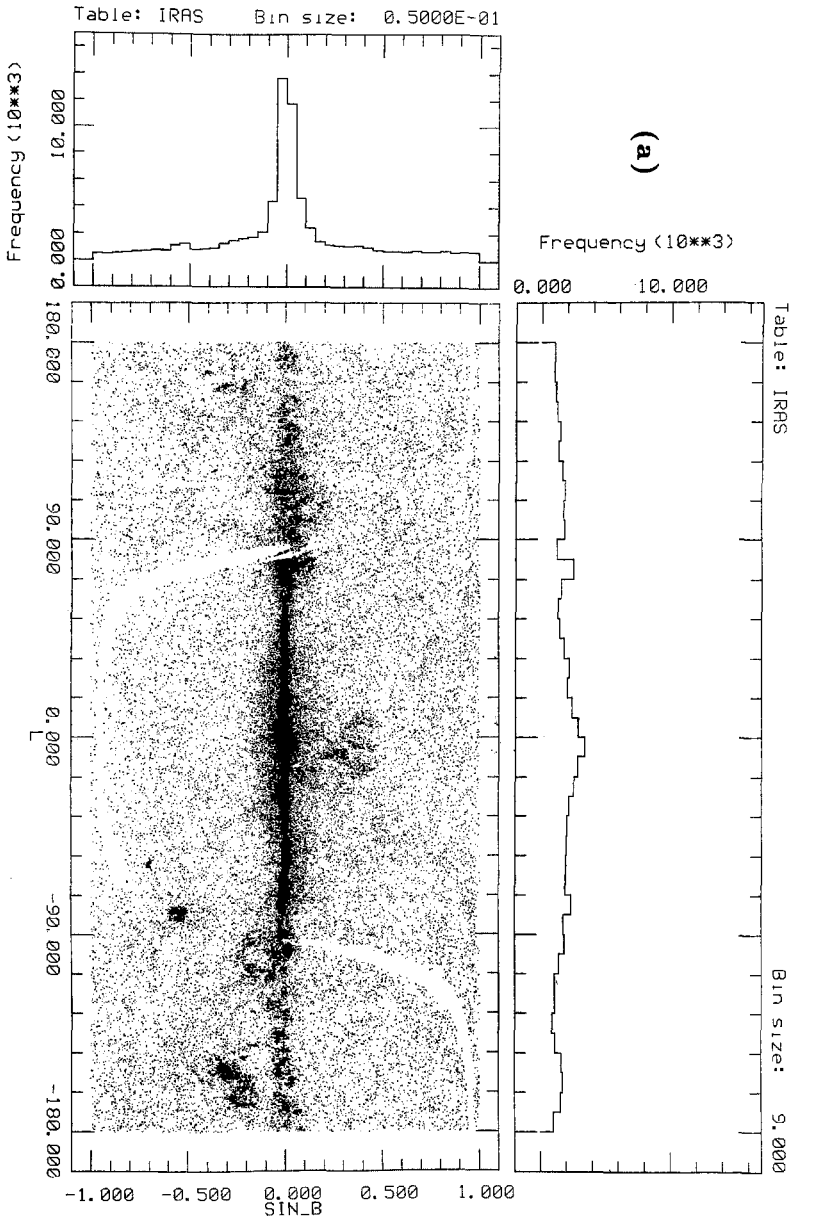
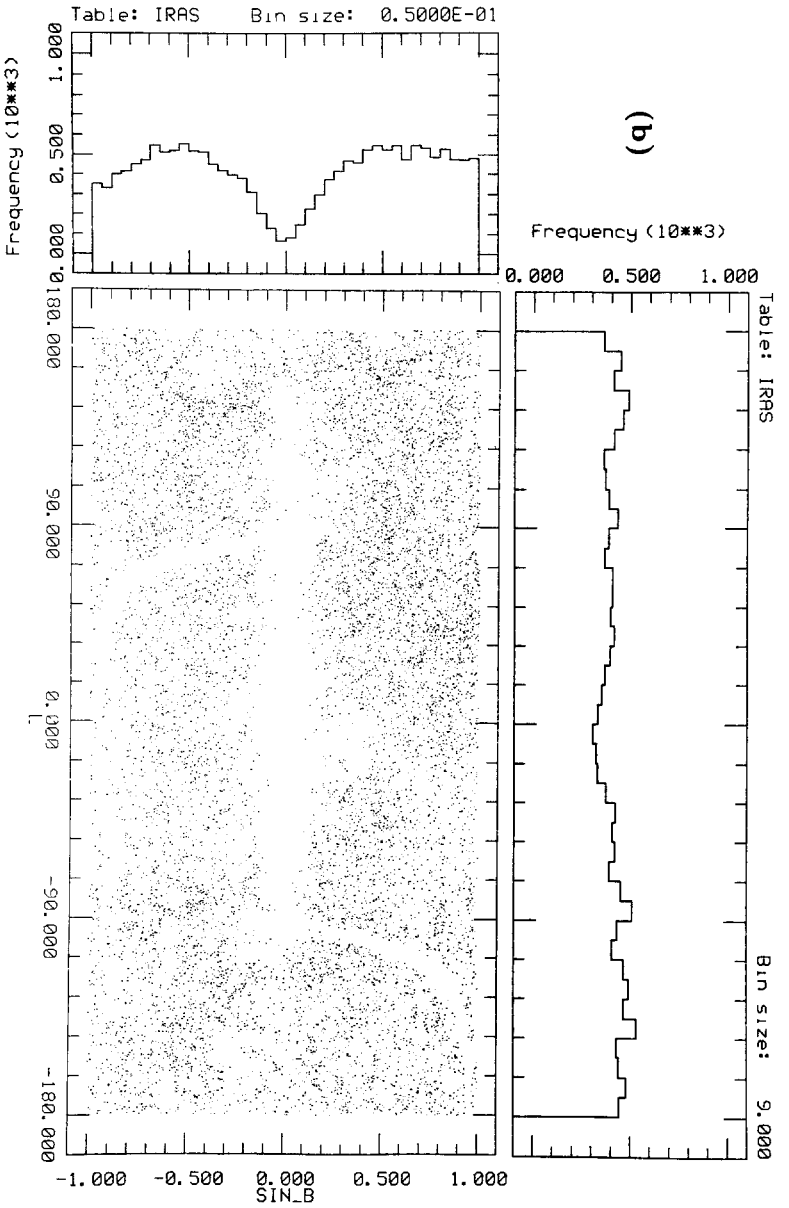


Fig. 1a–b. Sky distribution (a) of all IRAS Point Sources except stars (a sample similar to that in Meurs *et al.* 1988), and (b) of the set of "galaxies" as found by the maximum likelihood classifier. Histograms at the top and left show how source density varies with galactic longitude and (sine of) galactic latitude, respectively. (Note the different scales of the histograms!)



The latitude distribution in (b) clearly shows the effect of missing sources along the galactic plane in the galactic centre direction. There is a decrease in source density towards the southern galactic pole consistent with the north-south anisotropy found for IRAS galaxies (see Clowes *et al.* 1987); an additional slight decrease at high galactic latitudes, north and south, may be attributed to those parts of the empty strips (representing the 4 % of the sky not covered in the PSC) which run parallel to the longitude axis.

category obtained from sky regions where one category at a time could be expected to dominate. (From the beginning stars were essentially excluded from consideration, by applying an infrared colour cut and requiring good fluxes at longer wavelengths). The Gaussian approximation appears appropriate for the galaxy category; the distributions of the other two categories show some non-Gaussian structure, suggesting a different distribution type or a division into two subcomponents. A maximum-likelihood decision strategy established a category separation very comparable to, though probably better than, the more intuitive "handcrafted" approach of Meurs and Harmon (1988). The separation was only slightly modified when a maximum-a-posteriori decision strategy was used which takes into account the relative population of the categories considered.

From the sky map (*Fig. 1b*) displaying the resulting set of "galaxies" – the source distribution of the IRAS PSC (stars excluded) is shown for comparison in *Fig. 1a* – it is obvious that the apparent all-sky galaxy distribution is far from being homogeneous, reflecting the confusion problem in the area surrounding the galactic centre. Indeed it is remarkable that, near the galactic anti-centre, galaxies can be found practically in the middle of the galactic plane.

4 Unsupervised classification of IRAS point sources

Our steps into the area of unsupervised classification are mainly motivated by our interest to see whether the non-Gaussian distributions mentioned above would naturally split into two or more subcomponents. Also we are curious to see whether an unsupervised classification procedure, having access only to information intrinsic to the data, would find the same three categories used in supervised classification.

Various approaches are currently being pursued: 'Conventional' cluster analysis has been performed on a randomly selected subset of the IRAS PSC (stars excluded) with encouraging results (see Murtagh 1988). In another approach we are using the AutoClass program developed at the NASA/Ames Research Center (Cheeseman *et al.* 1987). AutoClass was designed for studying problems of machine learning and is implemented in Common Lisp. Similar to the Bayes classifier described above, AutoClass rests on the assumption that pattern classes can be described by conditional probability density functions of known (multivariate Gaussian) form. AutoClass can be used in two modes, supervised and unsupervised. In unsupervised mode AutoClass does not require the number of classes to be specified in advance. Instead, it uses an information-theoretic criterion that for a given data set simultaneously determines an optimal partition of feature space and an optimal number of classes. Thus AutoClass appears well-suited to the task of finding additional object categories beyond those already known, or to optimally splitting categories that show non-Gaussian distributions in feature space.

Preliminary investigations with AutoClass of a sample from the IRAS PSC are suggestive but inconclusive. Areas requiring further investigation include:
(i) *convergence properties:* it appears that the search for a globally optimal set of classes sometimes ends trapped in a local optimum;

(ii) *performance:* AutoClass in its present form appears to be significantly slower than, *e.g.* more conventional cluster analysis methods, and so may be inappropriate for very large datasets. On the other hand, the attractiveness of the method makes it important to assess whether this is intrinsic to the method or can be circumvented by a more efficient implementation.

5 Summary

Using supervised and unsupervised classification methods we have attempted an 'intrinsic classification' of a relevant subset of the IRAS Point Source Catalog (essentially excluding stars) with the immediate goal of selecting a maximally *clean*, *complete* and *unbiased* set of galaxy candidates. It appears that the extragalactic sources to a large degree can be separated from other, galactic sources contained in the IRAS PSC. This was achieved by constructing a Bayesian classifier from suitably chosen training sets for each source category, with sources being selected on the basis of three infrared colours and one infrared flux. A representative subset of the PSC sources considered was also subjected to unsupervised classification by feeding it to the AutoClass program. The results initially obtained with this program suggest its potential as a powerful tool for exploring unknown data sets, but reaching this state will without doubt require further development efforts.

Acknowledgements

Part of this work relied on the availability of the AutoClass program, which was kindly provided by Peter Cheeseman and his group at NASA/Ames and installed at ESO by Mark D. Johnston, STScI.

References

Batchelor, B.G., 1978. In *Pattern Recognition, Ideas in Practice*, ed. Batchelor, B.G., Plenum Press, New York and London, p. 65.

Beichman, C.A., Neugebauer, G., Habing, H.J., Clegg, P.E., Chester, T.J. (eds.), 1985. *Explanatory Supplement to the IRAS Catalogs and Atlases*, US Govmt. Print. Off.

Bock, H.H., 1974, *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Göttingen.

Cheeseman, P., Stutz, J., Freeman, D., Self, M., 1987. In *Proc. 1987 GSFC Conf. on Space Applic. of AI* (forthcoming).

Chester, T.J., 1986. In: *Light on Dark Matter*, ed. Israel, F.P., Reidel, Dordrecht, p. 3.

Clowes, R.G., Savage, A., Wang, D., Leggett, S.K., MacGillivray, H.T., Wolstencroft, R.D., 1987. *Mon. Not. R. astr. Soc.*, **229**, 27p.

Cooper, D.B., Cooper, P.W., 1964. *Information and Control*, **7**, 416.

Cooper, P.W., 1969a. In *Automatic Interpretation and Classification of Images*, ed. Grasselli, A., p. 97.

Cooper, P.W., 1969b. In *Methodologies of Pattern Recognition*, ed. Watanabe, S., p. 97.

Dale, M.B., 1985. *IEEE Trans. Pattern Analysis Machine Intelligence*, **PAMI-7**, 241.

Dow, M.W., Lu, N.Y., Houck, J.R., Salpeter, E.E., Lewis, B.M., 1988. *Astrophys. J.*, **324**, L51.

Duda, R.O., Hart, P.E., 1973. *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.

Fisher, D., Langley, P., 1986. In *Artificial Intelligence & Statistics*, ed. Gale, W.A., Addison-Wesley, Reading, Mass., p. 77.

Friedman, J.H., 1986. In: *Data Analysis in Astronomy II*, eds. Di Gesù, V., Scarsi, L., Crane, P., Friedman, J.H., Levialdi, S. Plenum Press, New York, p. 3.

Fu, K.-S., 1980. *IEEE Trans. Comput.*, **C-29**, 845.

Grasselli, A., (ed.), 1969. *Automatic Interpretation and Classification of Images*, Academic Press, New York.

Habing, H.J., 1987. In *IAU Symp. 122, Circumstellar Matter*, ed. Appenzeller, I., Reidel, Dordrecht, p. 197.

Hand, D.J., 1981. *Discrimination and Classification*, John Wiley & Sons, Chichester.

Harmon, R.T., Lahav, O., Meurs, E.J.A., 1987. *Mon. Not. R. astr. Soc.*, **228**, 5p.

Jain, A.K., 1987. In *Pattern Recognition – Theory and Applications*, eds. Devijver, P.A., Kittler, J., Springer-Verlag, Berlin - Heidelberg, p. 1.

Kedar-Cabelli, S.T., Mahadevan, S., 1986. In *Machine Learning, An Artificial Intelligence Approach* **II**, eds. Michalski, R.S., Carbonell, J.G., Mitchell, T.M., Morgan Kaufmann, Los Altos, Cal., p. 671.

Knapp, G.R., 1987. In *Proc. Conf. Cooling Flows in Clusters and Galaxies* (in preparation).

Kulkarni, A.D., 1986. In *Electronic and Electron Physics*, **66**, ed. Harkes, P.W., Academic Press, Orlando, Fla., p. 309.

Kurtz, M.J., 1983. In *Statistical Methods in Astronomy*, **ESA SP-201**, p. 47.

Lawrence, A., Walker, D., Rowan-Robinson, M., Leech, K.J., Penston, M.V., 1986. *Mon. Not. R. astr. Soc.*, **219**, 687.

Mantas, J., 1987. *Pattern Recognition*, **20**, 1.

Melsa, J.L., Cohn, D.L., 1978. *Decision and Estimation Theory*, McGraw-Hill, New York.

Meurs, E.J.A., Adorf, H.-M., Harmon, R.T., 1988. In *Astronomy from Large Databases — Scientific Objectives and Methodological Approaches*, eds. Heck, A., Murtagh, F., ESO Conf. Proc., p. 49.

Meurs, E.J.A., Harmon, R.T., 1988. *Astr. Astrophys.* (submitted).

Michalski, R.S., Stepp, R., 1983. *IEEE Trans. Pattern Analysis Machine Intelligence*, **5**, 396.

Murtagh, F., 1986. In *Data Analysis in Astronomy II*, eds. Di Gesù, V., Scarsi, L., Crane, P., Friedman, J.H., Levialdi, S., Plenum Press, New York, p. 31.

Murtagh, F., 1987. In *Conf. Image Analysis and Processing*, Cefalù, (forthcoming).

Murtagh, F., 1988. *These proceedings*, p. 308.

Murtagh, F., Heck, A., 1987a. *Astr. Astrophys. Suppl.*, **68**, 113.

Murtagh, F., Heck, A., 1987b. *Multivariate Data Analysis*, Reidel, Dordrecht.

Sklansky, J., Wassel, G.N., 1981. *Pattern Classifiers and Trainable Machines*, Springer, New York.

Soifer, B.T., Houck, J.R., Neugebauer, G., 1987. *Ann. Rev. Astr. Astrophys.*, **25**, 187.

Stepp, R.E., Michalski, R.S., 1986a. *Artif. Intell.* **28**, 43.

Stepp, R.E., Michalski, R.S., 1986b. In *Machine Learning, An Artificial Intelligence Approach*, **II**, eds. Michalski, R.S., Carbonell, J.G., Mitchell, T.M., Morgan Kaufmann, Los Altos, Cal., p. 471.

Watanabe, S. (ed.), 1969. *Methodologies of Pattern Recognition*, Academic Press, New York.

Wolstencroft, R.D., Savage, A., Clowes, R.G., MacGillivray, H.T., Leggett, S.K., Kalafi, M., 1986, *Mon. Not. R. Astr. Soc.*, **223**, 279.