nisms (magnetic fields and energy sources), gleaned from evidence of spectral line changes during the rotational modulation. We want to understand how the chromospheric structure and magnetic heterogeneities behave according to the major stellar parameters, viz. mass, age, composition and rotation rate, and the present observations will provide us with a key to understanding both chromospheric heating mechanisms, and the dynamo mechanism in late-type stars.

The Increasing Importance of Statistical Methods in Astronomy

A. Heck, Observatoire Astronomique, Strasbourg, France
F. Murtagh, The Space Telescope European Coordinating Facility, European Southern Observatory*
D. Ponz, European Southern Observatory

> You may ask:
> "What can a hard headed statistician offer to a starry eyed astronomer?"
> The answer is:
> "Plenty."
>
> Narlikar (1982)

Generalities

In the past, astronomers did everything individually, from the conception of a project to the collection of data and their analysis. As the instrumentation became more complex, teams had to be set up and they progressively included people (astronomers or otherwise) specialized in technology. Today it is practically impossible to run a project at the forefront of astronomical research without the help of these technologists.

In a similar way, one can already see that, at the other end of the chain, teams will have to include also people specialized in methodology to work on the collected data. And we are not thinking here only of image processing (which is a natural consequence of sophisticated technology), but mainly of a methodology applicable to already well-reduced data. This is actually the only way to face the challenge put to us by the accumulation of data.

Compared to the past, we are indeed collecting now a huge amount of data (see e.g. Jaschek, 1978), and the rate will speed up in the next decades. Just think that the Space Telescope will send down, over an estimated lifetime of 15 years, the equivalent of 14×10^{12} bytes of information, which means a daily average of 4×10^{9} bytes! But even if we exclude this special case of ST, we have now at our disposal more and more instruments which are collecting observations faster and faster. And these data are more and more diversified. The rate of data accumulation is higher than the rate of increase of the people able to work on them.

Thus, we will have to work on bigger samples if we want to take advantage and fully use the information contained in all these data, globally and individually. We might well live at the end of the period when a significant number of astronomers are spending their lives investigating a couple of pet objects. If not, what would be the use of collecting so many data?

One way to work efficiently on large samples is to apply, and if necessary to develop, an adequate statistical methodology. If Nature Is consistent, the results obtained by applying the tools developed by the mathematicians and the statisticians should not be in contradiction with those obtained by physical analyses.

However, do not let us say what we did not say: the statistical methodology is not intended to replace the physical analysis. It is complementary and it can be efficiently used to run a rough preliminary investigation, to sort out ideas, to put a new ("objective" or "independent") light on a problem or to point out sides or aspects which would not come out in a classical approach. A physical analysis will have anyway to refine and interpret the results and take care of all the details.

Probably the most important statistical methods, for astronomical problems, are the multivariate methods such as Principal Components Analysis (PCA) and Cluster Analysis. The former allows the fundamental properties to be chosen for a possibly large number of observational parameters. This is clearly an important task, since the apparent complexity of a problem will necessarily grow with improvement in observational techniques.

The problem of clustering is that of the automatic classification of data. Clustering methods can also be employed to pick out anomalous or peculiar objects. These techniques all work at will in a multidimensional parametric space, while graphically, and also classically in statistics, it is difficult to get results from more than two dimensions. These statistical methods are often considered as descriptive rather than inferential and, since astronomy is fundamentally a descriptive science, they would appear to be ideally suited for problems in this field.

In the same way that instrumentation should not be employed without respecting its conditions of use, algorithms should not be applied as black boxes by non-specialists without paying attention to their applicability constraints and their result limitations. Forgetting this golden rule is the best way to contribute to the bad reputation of statistics while ruining from the start any attempt at elaborating relevant conclusions.

Maybe somewhat paradoxically, astronomers have not been among the quickest to realize the potentialities of the "modern" statistical methodology. One of us (AH) became interested in 1974–75 and produced among the first papers in the field. But the idea was in the air and applications started to multiply. He therefore suggested the holding of a first meeting on "Statistical methods in astronomy". It took place in September 1983 at Strasbourg Observatory with the European Space Agency as co-sponsor (the proceedings were published as ESA SP-201).

This was the first opportunity to bring together astronomers using various statistical techniques on different astronomical objects and to review the status of the methodology, not only among astronomers, but also with invited statisticians. The

* Affiliated to the Astrophysics Division, Space Science Department, European Space Agency.

colloquium was a real success and another one is planned, most likely in 1987. In addition, a working group, concerned with the application of statistical methodology to astronomical data, is being set up. A newsletter should keep interested persons in touch and informed of the various activities in the field.

We shall present in the following a few examples of applications which do not exhaust all the possibilities and can only give a partial idea of the variety of the problems that can be tackled with this methodology.

Photometry Versus Spectroscopy

Since practically nothing had been done when AH started working in the field, he was first concerned with the applicability of this "modern" statistical methodology to astronomy. Thus he decided to put the algorithms on the trial bench of stellar data because there were a lot of such data available and because the corresponding physics was well established (and could be efficiently used for comparison).

A priori, however, it might be more appropriate to apply this methodology to non-stellar objects because their physics is less developed and because the data are more heterogeneous, incomplete, very diverse (continuous, discrete, binary, qualitative, . . .) and are thus challenging for the methods.

In stellar astronomy, an interesting subject was to study the interface between photometry and spectroscopy, especially in the framework of stellar classification. In the past, photometric and spectroscopic data were thought of as conflicting by some astronomers, but today most consider these techniques to be complementary. With P. C. Keenan (1973), we now "take it as evident that the aims of stellar classification are the same whether we work by direct photoelectric photometry through filters, by spectrophotometry, or by visual classification of spectrograms".

How well could, for instance, photometric indices of a star allow us to predict its spectral classification? Apart form its scholastic interest, this question has direct practical observational implications since photometric colours are much more quickly collected than good spectrograms, and with a smaller instrument for a given object.

The first step (Heck, 1976) has been to point out by multivariate statistical algorithms the most significant indices or group of indices in the *uvby* β photometric system for different ranges of spectral types. The photometric catalog (Lindemann & Hauck, 1973) was considered as a set of numerical data and the only physical hypothesis which intervened was a correlation between the effective temperature and the spectral types. The results were in full agreement with those of Strömgren (1966, 1967) obtained by physical analysis. The correct indices or groups of indices were selected as discriminators of luminosity and effective temperature.

The second step (Heck et al., 1977) was to directly confront photometric and spectroscopic data by applying clustering algorithms in an adequate multivariate space to the photometric indices of the same catalog (considered again purely as a set of numerical data). Then looking at the continuity of the arborescences or the homogeneity of the groups obtained from the point of view of the spectral classification, it transpired that about 8 % of the stars where deviating significantly from the ranges in spectral types and luminosity classes they should naturally belong to. These discrepancies resulted either from a wrong spectral type (as some re-determinations indicated), from poorly determined photometric indices, or simply because photometric indices, even when well selected, give information of a type different from the spectral features, and also from different wavelength coverages.

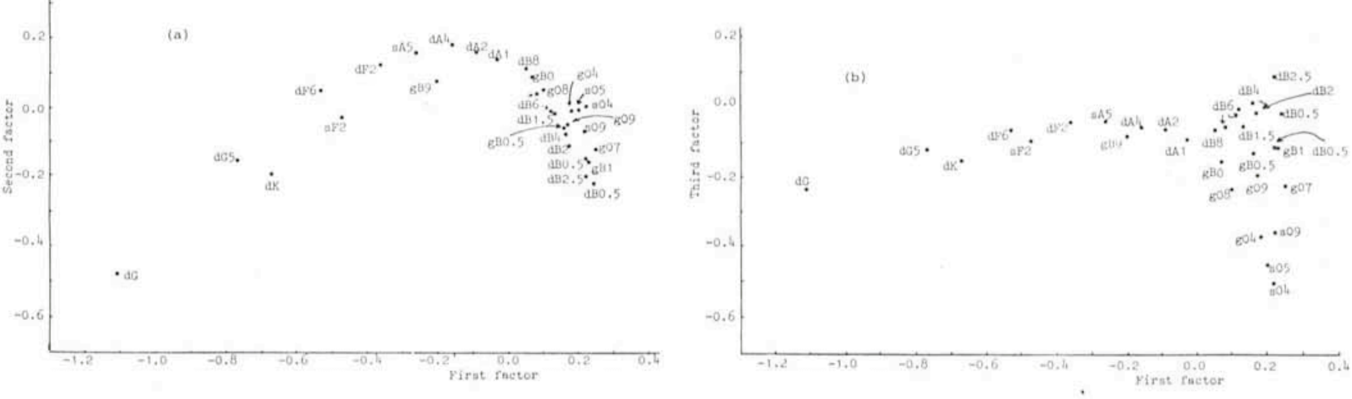
Might this mean that spectral classifications could be predicted from *uvby* β indices with about 90 % chance of being correct? This was investigated in a couple of subsequent papers (Heck and Mersch, 1980; Mersch and Heck, 1980) by elaborating an algorithm involving isotonic regression (working on the ranks of the spectral subtypes and luminosity classes) and stepwise multiple regressions (selecting the most significant indices or combinations of indices).

It resulted that there was an 80 % chance of predicting the spectral type within one spectral subtype for luminosity classes I to IV and an 87 % chance for luminosity class V. As far as luminosity was concerned, there were 44 %, 28 % and 56 % chances of predicting it correctly for luminosity classes I, III, and V respectively. This might point out some inability of the *uvby* β photometry *alone* to discriminate properly the luminosity and it would be worthwhile to undertake a similar study with different photometry.

Ultraviolet Spectral Classification

The previous investigations were more centered on studying the applicability of the "modern" statistical methodology to astronomical data than on tackling new fields. An opportunity came recently with the classification of IUE low-dispersion stellar spectra.

The International Ultraviolet Explorer (IUE), launched on 26 January 1978 and still operating in an observatory mode, is the most successful astronomical satellite up to now. Details can be found in Boggess et al. (1978a and b). We shall only recall here that it is collecting low- and high-resolution spectra of all kinds of celestial objects in the ultraviolet wavelength range (UV) covering about 1150 to 3200 Ångströms.



Distribution of the group barycenters in the plane of the first two PCA factors (a) and of the first and third PCA factors (b) derived from the IUE spectral data. Each center has been named according to the spectral mode in the multivariate space of the first twenty PCA factors. The UV spectral symbols have been introduced in Heck et al. (1984).

From earlier work on data collected by the S2/68 experiment on board the TD 1 satellite, it had been shown that stars which are spectrally normal in the visible range do not necessarily behave normally in the ultraviolet range and vice versa (see Cucchiaro et al., 1978, and the references quoted therein). Consequently, MK spectral classifications defined from the visible range cannot simply be extrapolated to the UV.

A UV stellar classification program, supported by a VILSPA workshop on the same subject (proceedings published as ESA SP-182) was then initiated in order to define from IUE low-resolution spectra smooth spectral sequences proper to the UV and describing the stellar behavior in the UV while staying as far as possible in accordance with the MK scheme in the visible.

The first volume of a reference atlas has been produced (Heck et al., 1984), together with reference sequences and standard stars. The considerable underlying classification work has been carried out following a classical morphological approach (Jaschek and Jaschek, 1984) and it essentially confirmed that there is no one-to-one correspondence between the UV and visible ranges.

Stellar spectral classifications are more than taxonomical exercises aiming just at labelling stars and putting them in boxes by comparison with standards. They are used for describing fundamental physical parameters in the outer atmosphere of the stars, to discriminate peculiar objects, and for other subsidiary applications like distance determinations, interstellar extinction and population synthesis studies.

It is important to bear in mind that the classification systems are built independently of stellar physics in the sense that they are defined completely by spectral features in selected standards in a given wavelength range (see e.g. Jaschek, 1979, and Morgan, 1984). If the schemes are based on a sufficiently large number of objects, it appears easily that they are intimately linked with the physics, but not necessarily of the same stellar regions if they refer to different wavelength ranges. Consequently, the discrepancies reported between the MK system and the UV frames are not too surprising.

Moreover, the only way to confirm independently the correctness of the UV classification frame introduced in the atlas was to remain in the same wavelength range. Therefore, statistical algorithms working in a multidimensional parametric space were applied to variables expressing, as objectively as possible, the information contained in the continuum and the spectral features (Heck et al., 1985). This was done through, on the one hand, an asymmetry coefficient describing the continuum shape and empirically corrected for the interstellar reddening, and, on the other hand, the intensities of sixty objectively selected lines (which included all the lines retained as discriminators in the atlas).

These line intensities were weighted in a way we called the "variable Procrustean bed method" because, contrary to a standard weighting where a given variable is weighted in the same way for all the individuals of a sample, the spectral variables were weighted here according to the asymmetry coefficient which varies with the star at hand. The algorithm applied to the set of the variables consisted of a Principal Components Analysis and a Cluster Analysis.

The individual classifications resulting from the morphological approach used for the atlas were fully confirmed, and ipso facto the discrepancies with the MK classifications in the visible range. The groups resulting from the Cluster Analysis displayed good homogeneity and an excellent discrimination for spectral types and luminosity classes, especially in the early spectral types which were well represented in the sample used for this study. The standard stars are located in the neighborhood of the barycenters of the groups (see figure).

Currently the contributions of the successive principal axes resulting from the Principal Components Analysis are being investigated in greater detail, and we are looking forward to including more data from the IUE archive in order to refine the conclusions.

Star and Galaxy Separation

Survey work on many plates rapidly encounters problems of processing very large numbers of objects. One current theme of research is to simplify the carrying out of, and make use of the results of, such surveys as the ESO/Uppsala survey of southern galaxies (see Lauberts and Valentijn, 1983). Firstly, the use of multivariate methods in classifying data derived from images is being studied; and secondly, novel approaches are being looked at for the classification of galaxies. In this section, we will look at each of these in turn.

In discriminating between objects on survey plates, the first question which arises is the choice of parameters to extract. At present the object searching algorithm in MIDAS outputs information regarding 20 variables for each object found. Using Prinicpal Components Analysis easily allows it to be seen if all of these variables are necessary – in fact, we have usually found that about 2 or 3 variables (e.g. isophotal magnitude, relative gradient) are sufficient. These provide approximately as much "information" as the original set of variables. For classifying the objects into the major classes (i.e. stars, galaxies, plate defects), the usefulness of the 20-odd variables produced at present is being investigated. We are considering other shape parameters, such as the moments, and hope to be shortly in a position to suggest to the user a sequence for carrying out an analysis such as the following: choose a particular set of variables to characterize the objects studied; run this through a Principal Components Analysis in order to arrive at a best-fitting pair of variables (a linear combination of those chosen) which can be plotted and studied; then use these as input to a clustering program in order to determine the major groups of objects present. Such an approach will never replace the expert (consider for example the range of variables which are candidates for star/galaxy discrimination, and some of which are reviewed by Kurtz, 1983); however, in providing useful analytic tools, it can increase the performance of the expert and indicate to him/her further interesting aspects which would not have been appreciated if overshadowed by the sheer quantity of data to be analyzed.

The large quantity of data, of course, in itself demands the provision of increasingly automated means of analysis. Progress in an expert system to determine galaxy types (Thonnat, 1985) will probably always be hampered by large computational time requirements if sophisticated pattern matching algorithms are not at the core of such systems. Therefore, it is being attempted to assess the potential for classifying galaxies – at least into the major types – by using for each galaxy its magnitude versus surface brightness curve (see Lauberts and Valentijn, 1983). A novel curve matching technique has been developed, a measure of similarity thereby determined, and a clustering carried out on the basis of such similarities. Results obtained so far (Murtagh and Lauberts, 1985) show consistency with a human expert's classification into ellipticals and spirals.

Statistical Algorithms in MIDAS

In the current version of MIDAS we have included commands for some of the basic methods of multivariate statistical analysis. The data matrix is structured as a table where the different objects are associated with rows and the variables are associated with columns. The methods currently available are:

– Principal Components Analysis, to produce the projection of the data matrix onto the principal axes.

– Cluster Analysis, using hierarchical clustering with several agglomerative criteria (single link, complete link, minimum variance, etc.).

– Fast iterative non-hierarchical clustering methods.

In this context, the tables in MIDAS provide a bridge between the raw data and the algorithms for analysis. Data originally in the form of images or catalogs can be put into the analysis program by structuring the extracted information as tables, a natural way of representing the objects in the parameter space.

These commands are in an experimental state. Work is ongoing in making more statistical methods available within the interactive framework of MIDAS. Special attention will be given to the friendliness of usage by means of display facilities and easy interaction. Unlike many statistical packages commercially available, MIDAS offers the advantage of integrating image processing algorithms with extensive graphics capabilities and, of course, the statistical methods.

The linking-up of data collection and of statistical data analysis – of database creation and of an important use to which a database is put – is also of singular importance. The future existence of an ESO and of a Space Telescope archive creates exciting possibilities for the possible use of multivariate statistical procedures on a large scale. A step of far-reaching implications was taken a few years ago when the large-scale archiving of data was linked to the down-stream analyzing (by multivariate statistical methods) of such data: this was when Malinvaud, head of the French statistical service (INSEE), strongly linked the two together (Malinvaud and Deville, 1983). Multivariate statistical analysis of data requires that the data collection be competently carried out; and, in return, it offers the only feasible possibility for condensing data for interpretation if the data is present in very large quantities.

A Collaborative Future

Current trends in astronomical research not only create prospects for statistical methods to be used, but for reasons mentioned in this article they require them. The flow will not be just one-way however: statisticians will also learn from the problems of astronomy. Computational problems related to the large amounts of data which must be handled, the best ways to treat missing values and mixed qualitative-quantitative data, and even the most appropriate statistical methods to apply – all these and many more currently unforeseen issues will lead to a very fruitful and productive interaction between methodologist and astronomer over the coming years.

References

Boggess, A. et al. 1978a, *Nature* **275,** 377.

Boggess, A. et al. 178b, *Nature* **275,** 389.

Cucchiaro, A., Jaschek, M., Jaschek, C. 1978, *An atlas of ultraviolet stellar spectra,* Liège and Strasbourg.

ESO 1985, MIDAS Operating Manual No. 1.

Heck, A. 1976, *Astron. Astrophys.* **47,** 129.

Heck, A., Albert, A., Defays, D., Mersch, G. 1977, *Astron. Astrophys.* **61,** 563.

Heck, A., Egret, D., Jaschek, M., Jaschek, C. 1984, IUE low-dispersion spectra reference atlas. Part 1. Normal stars. ESA SP-1052.

Heck, A., Egret, D., Nobelis, Ph., Turlot, J.C. 1985, Statistical classification of IUE low-dispersion stellar spectra, in preparation.

Heck, A. and Mersch, G. 1980, *Astron. Astrophys.* **83,** 287.

Jaschek, C. 1978, *Q.J. Roy. Astron. Soc.* **19,** 269.

Jaschek, C. 1979, in *Classification Spectrale,* Ecole de Goutelas, ed. D. Ballereau, Obs. Meudon.

Jaschek, M. and Jaschek, C., 1984, in *The MK Process and Stellar Classification,* ed. R.F. Garrison, David Dunlap Obs., p. 290.

Keenan, P.C. 1973, in *Spectral Classification and Multicolour Photometry,* ed. Ch. Fehrenbach and B.E. Westerlund, D. Reidel Publ. Co., Dordrecht, p. 3.

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

Lauberts, A. and Valentijn, E.A. 1983, *The Messenger* **34,** 10.

Lindemann, E. and Hauck, B. 1973, *Astron. Astrophys. Suppl.* **11,** 119.

Malinvaud, E. and Deville, J.C. 1983, *J. Roy. Statist. Soc.* A 146, 335–361.

Mersch, G. and Heck, A. 1980 *Astron. Astrophys.* **85,** 93.

Morgan, W.W. 1984, in *The MK Process and Stellar Classification,* ed. R.F. Garrison, David Dunlap Obs., p. 18.

Murtagh, F. and Lauberts, A. 1985, Comm. to Fourth Meeting of Classification Societies, Cambridge.

Narlikar, J.V. 1982, Indian J. Statist. 44, 125.

Strömgren, B. 1966, *Ann. Rev. Astron. Astrophys.* **4,** 433.

Strömgren, B. 1967, in *The Magnetic and Related Stars,* ed. R. Cameron, Mono Book Corp., Baltimore, p. 461.

Thonnat, M. 1985, INRIA (Centre Sophia Antipolis) Report No. 387.

NEWS ON ESO INSTRUMENTATION

The following information on instrumentation has been provided by the Optical Instrumentation Group.

The ESO Multiple Object Spectroscopic Facility "OPTOPUS"

OPTOPUS is a fiber-optics instrument intended for multiple-object spectroscopy with the Boller & Chivens spectrograph and a CCD detector at the 3.6 m telescope. Using the Optopus system, the spectra from up to 47 independent objects located within a 33 arcmin field can be simultaneously recorded.

Overall View of the System

For multi-object observations, the B & C spectrograph is mounted on a separate frame within the Cassegrain cage of the 3.6 m telescope and a special fiber optics adaptor is fixed to the Cassegrain flange. The adaptor serves as a support for metal templates (starplates) containing precisely drilled holes (corresponding to the objects of interest for a given observed field) into which the individual fibers are connected. The fibers, serving the purpose of a flexible light transport from focal plane to spectrograph, are terminated together at their output ends in a closely packed row which replaces the conventional B & C entrance slit.

For guiding and alignment purposes, each starplate must also contain bundle connector holes for two guidestars, which