THE ASTROPHYSICAL JOURNAL, 426:340–344, 1994 May 1


# STELLAR SPECTRAL CLASSIFICATION USING AUTOMATED SCHEMES

R. K. GULATI AND RANJAN GUPTA
IUCAA, Post Bag 4, Ganeshkhind, Pune 411007, India
gulati@iucaa.ernet.in;ranjan@iucaa.ernet.in

AND

PRADEEP GOTHOSKAR AND SHYAM KHOBRAGADE
NCRA, TIFR Centre, P.O. Box 3, Pune 411007, India
pradeep@gmrt.ernet.in;shyam@gmrt.ernet.in
*Received 1993 September 14; accepted 1993 November 4*

## ABSTRACT

Classification of stellar spectra by human experts, in the past, has been subjective, leading to many non-unique databases. However, with the availability of large spectral databases, automated classification schemes offer an alternative to visual classification. Here, we present two schemes for automated classification of stellar spectra, namely, $\chi^2$-minimization and Artificial Neural Network. These techniques have been applied to classify a complete set of 158 test spectra into 55 spectral types of a reference library. Using these methods, we have successfully classified the test library on the basis of reference library to an accuracy of two spectral subclasses. Such automated schemes would in the future provide fast, uniform, and almost on-line classification of stellar spectra.

*Subject headings:* methods: data analysis — stars: fundamental parameters — techniques: spectroscopic

## 1. INTRODUCTION

Recent improvements in observational instruments have made it possible to resolve spectroscopically a large number of stars in the solar neighborhood, as well as in nearby stellar systems (Gorgas et al. 1993 and references therein). This improvement has generated an increasing number of digitized spectral libraries in the optical region (e.g., Gunn & Stryker 1983; Jacoby, Hunter, & Christian 1984, hereafter JHC; Pickles 1985; Silva & Cornell 1992, hereafter SC) which can be used for the purpose of stellar classification. Stellar spectral classification is not only a tool for labeling individual stars in accordance with a standard library but is also useful in studies like estimating stellar atmospheric parameters and stellar population synthesis. A comprehensive review on the subject of stellar classification has been made by Jaschek & Jaschek (1990).

Classification of a large number of stellar spectra into various closely related classes, which requires extensive human efforts and often subjective judgement, poses an ideal problem for automated classification techniques. Conventionally, spectral classification used to be done by visual inspection of individual stellar spectra. But, with the advent of modern computational techniques, it has become possible to classify large databases uniformly and in considerably short time. Previous attempts of automated spectral classification were based on schemes like cross-correlation and height of correlation function applied to the databases available at that time (see, e.g., Kurtz 1984; Adorf 1986). Taking advantage of a new digital stellar optical library (SC), we present two schemes for automated spectral classification of stars ranging from O to M types.

The first scheme adopts the classical $\chi^2$-minimization which is based on using a reference library of stellar spectra with known classification. Then one compares a star with unknown classification with the reference library for a "best" fit decided by statistical constraints. The second scheme employs the technique of Artificial Neural Network (hereafter ANN). Since the formulation of the ANN algorithm, such as Multilayer Back Propagation Network (MBPN) (Rumelhart, Hinton, & Williams 1986), the technique has been extensively used to classify large databases. In particular, ANN has been used in astronomy to classify galaxies into various Hubble types (Storrie-Lombardi et al. 1992), to identify point sources in the IRAS survey (Adorf & Meurs 1988) and to separate stellar objects from galaxies (Odewahn et al. 1992). ANN has also been used to classify stellar spectral types (von Hippel et al. 1992, 1993).

In this paper, we describe classification of stellar spectra based on the above two schemes. The input database and preprocessing of data are described in § 2. Section 3 describes the $\chi^2$-minimization and ANN schemes. The performance and results of these schemes are discussed in § 4, and the conclusions are presented in § 5.

## 2. THE INPUT DATA

The SC library was used as a database of template stellar spectra. This library has good photometric measurement, wide spectral coverage, and adequate spectral resolution for the purpose of classification. The database has 74 stellar spectra of stars from O to M types spanning the wavelength range of 3510–8930 Å. From these, 55 spectra assigned to solar metallicity O to M–type stars were selected. It should be noted that most of these template spectra were obtained by averaging several stars of similar spectrum-luminosity class (see SC library).

A set of test spectra was obtained from the library compiled by JHC. This library covers the wavelength range 3510–7427 Å for various O to M stellar types. Unlike the SC library, where the majority comprises averaged stellar spectra, the JHC library contains 161 spectra of individual stars; we selected 158

of these of approximate solar composition. Thus, the present study was based on 55 reference spectra and 158 test spectra derived from two libraries.

### 2.1. *Data Preprocessing*

Application of either scheme for stellar classification requires the input database to be preprocessed. We used the wavelength range 3510–6800 Å, which was common to both the libraries. The SC library has a resolution of 11 Å with one sample per 5 Å in this wavelength range. The JHC library has a resolution of 4.5 Å with one sample per 1.4 Å. Thus, it was necessary to reduce the JHC library to the resolution and sampling of the SC library for meaningful comparison.

Each spectrum of the JHC library was convolved with a Gaussian function of appropriate FWHM to reduce the resolution from 4.5 to 11 Å. Further, to match the wavelength points of the JHC library with those of SC, each spectrum was sampled at 5 Å intervals to generate 659 data points between 3510–6800 Å. The spectra from both databases were normalized to a value of 100 at the wavelength of 5450 Å. This preprocessing provided a common database for both the classification schemes.

## 3. CLASSIFICATION SCHEMES

### 3.1. $\chi^2$-*Minimization Scheme*

Let us assume that the reference library fluxes for a particular stellar spectrum are given by $S_i(\lambda_i)$, where $i = 1, n$ for $n = 659$ wavelength points. Similarly, the test library fluxes are defined as $T_i(\lambda_i)$. The reduced $\tilde{\chi}^2$ is defined as

$$\tilde{\chi}^2 = \frac{\sum_{i=1}^{n} (S_i - T_i)^2}{p} \quad (1)$$

(for no statistical weighting, see Bevington 1969, chapters 1–4), where $p$ is equal to the degrees of freedom. Each test spectrum was compared with the 55 reference spectra to compute 55 $\tilde{\chi}^2$ values. The spectral type of the reference spectrum corresponding to the minimum $\tilde{\chi}^2$ was assigned to the test spectrum. This procedure was repeated for the remaining 157 spectra in the test library. The minimization scheme also had a wavelength shift algorithm to shift the test spectrum by $\pm$ a few pixels to the reference frame of the reference spectrum during the $\tilde{\chi}^2$ computation. Although, according to Silva & Cornell (1992), the SC library takes into account the effects of such wavelength shifts, a few JHC spectra required a wavelength shift of $\pm 1$ pixel. The $\chi^2$-minimization scheme required about 90 minutes of processing on a SUN SPARC-10 machine for classifying 158 spectra.

### 3.2. *ANN Scheme*

#### 3.2.1. *Parameterization*

Each spectral class is characterized by absorption features at a few selected wavelengths which are diagnostics of spectral classes (Jaschek & Jaschek 1990). These absorption features change in strength, position, and width with the change in spectral class. In order to improve the efficiency of the network, each spectrum was parameterized in two stages. First, we used fluxes at the selected 67 wavelengths (and at neighboring $\pm 5$ Å) as 160 input parameters. Next, the flux values were normalized to the maximum flux of the corresponding spectrum, and this value of maximum flux was used as an additional parameter. Thus, both the training and the testing set of spectra were characterized by 161 input parameters per spectrum. Figure 1 shows typical stellar spectra for B-, F-, K- and M-type stars, along with the wavelengths at which the fluxes were monitored. In order to display the individual spectra without overlapping, each spectrum has been offset on an arbitrary scale. It is evident from this figure that, as one goes from early to later types, the number of spectral features increases, and they concentrate in the longer wavelength region.

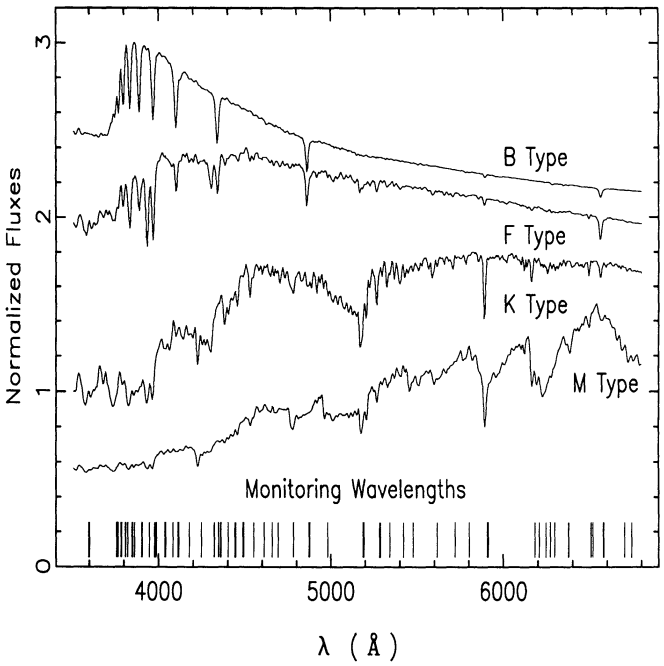


FIG. 1.—Four examples of the template spectra used to train the network. Vertical lines indicate the monitoring wavelengths (and $\pm 5$ Å around them) which were used to derive the 161 fluxes as input to the network.

#### 3.2.2. *The MBPN*

The network consisted of three layers, namely, an input layer with 161 nodes, one or more hidden layers, and an output layer having 55 nodes, each representing a spectral class. The algorithm consists of two stages of operation, namely, the training session and the testing session. The weights $w_{ji}$ over the interlayer connections were generated ramdomly at the start of the training session. Input $x_j$, to each node, $j$, was computed by the weighted sum of the output from all nodes from the previous layer, while the output, $y_j$, of the node $j$, was computed using a sigmoid transfer function of the form

$$y_j = \frac{1}{(1 + e^{-x_j})} . \quad (2)$$

The output $y_j$ at the output layer was compared with the desired output $d_j$ using an error function of the form

$$\delta_j = y_j(1 - y_j)(d_j - y_j) . \quad (3)$$

The error was then propagated backward from output to input layer to update the weight of each connection as follows:

$$w_{ji}^{n+1} = w_{ji}^{n} + \eta\delta_j y_j + \alpha(w_{ji}^{n} - w_{ji}^{n-1}) , \quad (4)$$

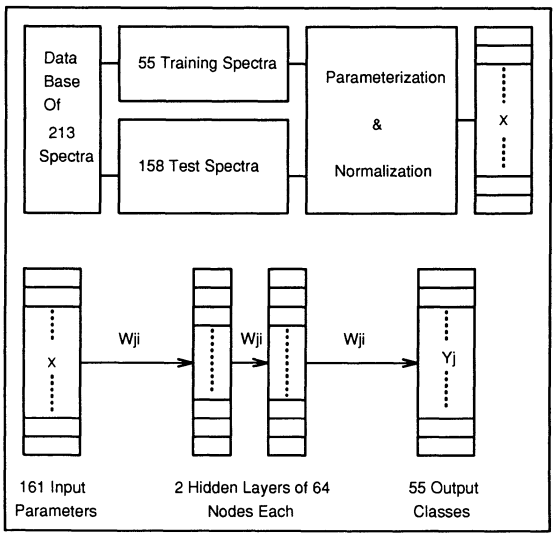


FIG. 2.—Schematic for the preprocessing of data and the configuration of ANN. The data was divided into a training set of 55 spectra and a testing set of 158 spectra, which were parameterized and normalized. ANN consists of 161 input parameters, two hidden layers each with 64 nodes and 55 output nodes, each corresponding to one spectrum-luminosity class.

where $w_{ji}$ is the weight at iteration $n$; $\eta$ and $\alpha$ are the gain and momentum factors, respectively, which control the rate of convergence. The process was repeated for each input pattern from all the classes until the error in the network output was reduced to a preassigned threshold value. The final frozen weights were used during the testing session to classify the large database. Figure 2 shows the schematic of parameterization of the data and architecture of the ANN used.

#### 3.2.3. *Implementation*

We explored a variety of network configurations for the present application. Table 1 shows the network performance estimated from the linear correlation coefficients, $r$, and the standard deviations, $\sigma$, of the network and the catalog classification. The set of 55 reference spectra was used to train the network with $\eta$ and $\alpha$ values set to 0.1. The weights were optimized until the error on the output was reduced to 0.0002. Each network, typically, took about 12 hr to converge on a SUN SPARC-10 machine during the training session. The network was then applied to a set of 158 test spectra, previously unexposed to the network. The classification of 158 test spectra was, typically, done in less than 1 minute on a SPARC-10 machine. Network classification was compared to the catalog classification according to the JHC library, by computing the linear correlation coefficients and the standard deviations. These parameters are expressed in terms of code units defined in § 4.1. The correlation coefficients and the standard deviations, as seen from Table 1, indicate that the network with two hidden layers, each having 64 nodes, is optimal, as it uses the minimum number of nodes to give the best performance. The performance is only marginally better than that of a single hidden layer with a 128-node configuration, which requires considerably more training time. We have thus restricted further discussion to the configuration of two hidden layers with 64 nodes in each.

TABLE 1

NETWORK CONFIGURATIONS

| Number | Input | Nodes/Hidden Layer | Hidden Layers | Output | $r$ | $\sigma$ |
|---|---|---|---|---|---|---|
| 1..... | 161 | 128 | 1 | 55 | 0.9924 | 217.6 |
| 2..... | 161 | 64 | 1 | 55 | 0.9908 | 240.5 |
| 3..... | 161 | 64 | 2 | 55 | 0.9922 | 219.8 |
| 4..... | 161 | 32 | 2 | 55 | 0.9879 | 263.3 |
| 5..... | 161 | 16 | 2 | 55 | 0.9830 | 330.1 |

## 4. DISCUSSION

### 4.1. *Spectral Class Coding*

The performance of two schemes was judged using a quantitative correlation analysis. Since the spectral classes are conventionally named in an alphanumeric fashion, we devised a coding method to assign a unique number to each spectral type as follows:

$$\text{Code Number} = 1000.0 \times \text{A1} + 100.0 \times \text{A2} + (1.5 + 2 \times \text{A3}), \quad (5)$$

where A1 is the main spectral type of the star (i.e., O to M types coded as 1 to 7), A2 is the subspectral type of the star (coded from 0.0 to 9.5), and A3 is the luminosity class of the star (i.e., I to V classes coded as 0 to 4). For example, in the present scheme of numbering, stars B2I and G9.5 V would be coded as 2201.5 and 5959.5, respectively.

### 4.2. *Performance*

The performance of two classification schemes was evaluated by the correlation analysis. The list of 158 spectra classified by each scheme (in terms of the spectral classes given in Table 3 of SC) was correlated with respect to the catalog classification given in JHC (see Table 1 of JHC). Figures 3*a* and 3*b* show the correlation plots of catalog classification relative to the ANN and $\chi^2$ classification, respectively. We also plotted the correlation of ANN with that of $\chi^2$ classification in Figure 3*c* to see the consistency of the two schemes. These plots show remarkable agreement between the $\chi^2$, ANN, and the catalog classifications. Further analysis of these plots by fitting straight lines to the data shows that the standard deviations of the data around the best-fit lines are 200 to 220 units in terms of our coding scheme. The slope of the best-fit line was 0.97 to 1.02, close to the ideal value of 1.00, while the intercepts of the lines were found to be within the standard deviations. The correlation coefficients for these plots are close to 0.993. Table 2 shows the list of these parameters; standard deviations, $\sigma$, slopes and intercepts of the best-fitted lines with their errors, $m \pm \Delta m$ and $c \pm \Delta c$, and the correlation coefficients, $r$, for two schemes.

Yet another way of evaluating the performance of two classification schemes is to study the coincidence of $\chi^2$ and ANN

TABLE 2

COMPARATIVE PERFORMANCE OF TWO SCHEMES

| Parameters | Catalog vs. $\chi^2$ | Catalog vs. ANN | ANN vs. $\chi^2$ |
|---|---|---|---|
| $\sigma$ ..... | 202.9 | 219.8 | 200.3 |
| $m \pm \Delta m$ ..... | $0.999 \pm 0.009$ | $1.019 \pm 0.010$ | $0.974 \pm 0.009$ |
| $c \pm \Delta c$ ..... | $58.44 \pm 41.90$ | $-60.70 \pm 45.40$ | $148.61 \pm 40.59$ |
| $r$ ..... | 0.9931 | 0.9922 | 0.9932 |

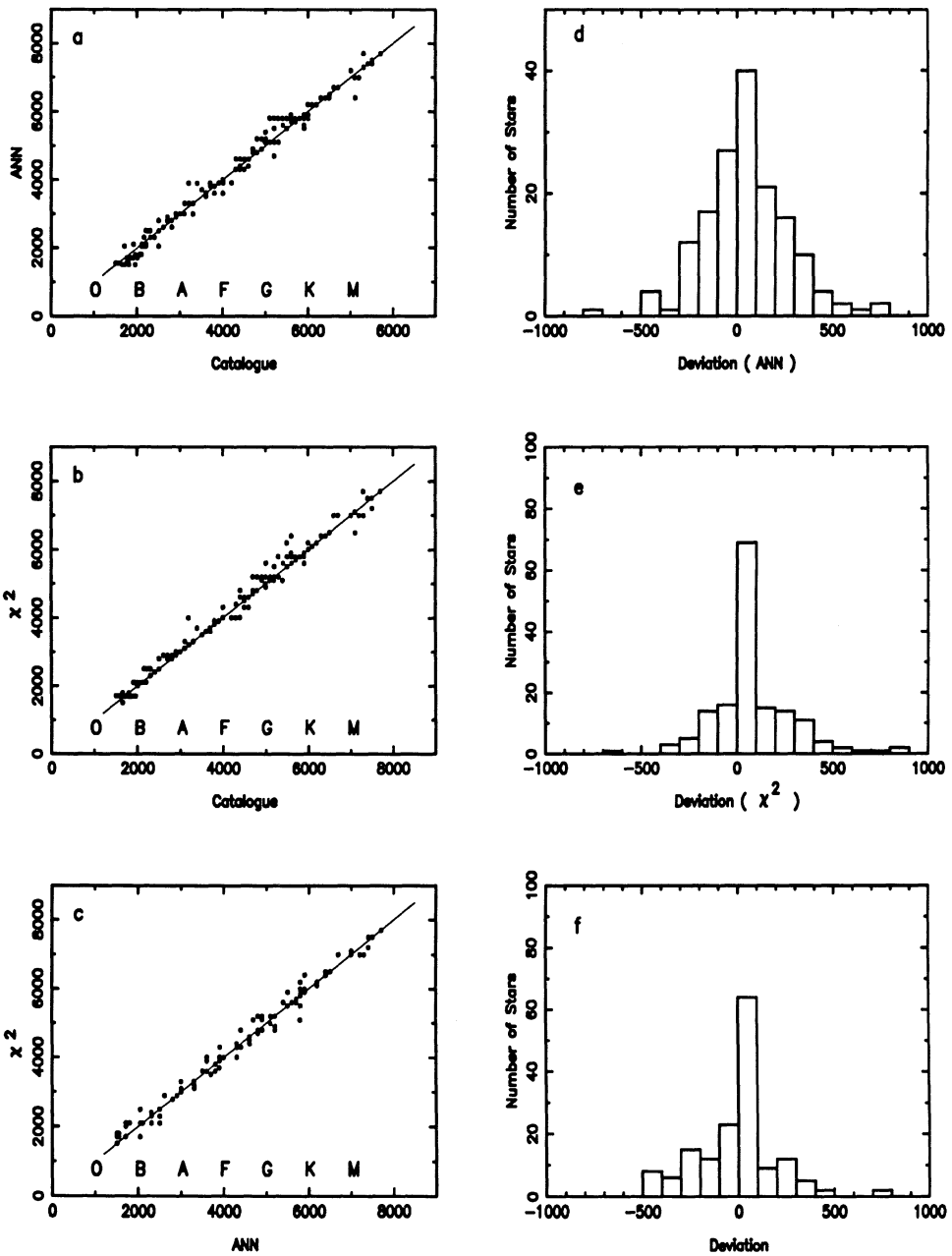


FIG. 3.—Correlation plots for catalog classification compared to ANN (*a*) and $\chi^2$ (*b*) classifications. ANN classification is compared to $\chi^2$ in (*c*). The frequency distribution of stars in each case is shown in (*d*), (*e*), and (*f*), respectively. The data was binned around the best-fit line, with the width of each bin corresponding to 100 units of spectral classification code.

classifications with that of the catalog. Figures 3*d* and 3*e*, respectively, show the frequency distribution of stars in bins which are the deviations of ANN and $\chi^2$ classifications from their respective values on the best-fitted lines. Similarly, Figure 3*f* shows the frequency distribution of stars in bins for ANN classification with respect to $\chi^2$ classification. Each bin was 100 spectral class code units (one subclass) wide and was centered around the best-fitted line. The coincidence study indicates that, on average, 76% of the total 158 stars were within 1 $\sigma$ in the case of catalog versus $\chi^2$ classification, while this statistic was about 68% in the case of catalog versus ANN classification.

It may be noted that the values for A3 in equation 5 (the luminosity class) get comparatively lower weight when we plotted the four-digit code number in Figures 3*a*, 3*b*, and 3*c*. However, we separately investigated the consistency of only the luminosity class between the catalog classification and that by both schemes. It was found that 64% of the total sample was classified correctly. This can be illustrated by taking an example of a particular star which has been classified by either technique, say as 3755.5 in terms of code units which refer to A7.5 III spectrum-luminosity class. Considering the two-subclass error by our schemes (corresponding to 200 code units), the extremes of classification could be either A5.5 or

A9.5, and its luminosity classification is correct to 1 $\sigma$ confidence level. Hence, these schemes can classify spectra in two dimensions, namely, spectral and luminosity classes.

Some general comments can be made on these correlation plots and histograms.

1. The 55 spectral types from the SC library do not cover all the possible spectral types between O to M–type stars, and this has led to an incomplete set of reference spectral classes. The classification of 158 stars, thus, has discrete gaps, as seen in the correlation plots.

2. The correlation plots also show horizontal grouping of stars at specific spectra types. This is due to the template spectra of the SC library, which were derived from average spectra of many stars of similar spectral type.

3. From the scatter plots, it is clear that most of the stars fall close to the line with 45° slope. This suggests that both the schemes reproduce the catalog classification quite satisfactorily. But the scatter of correlation of $\chi^2$ and ANN is less, indicating that there is better agreement for classification between two schemes as compared to the catalog classification. Further, it is observed that both the methods show a positive deviation from the best-fit line which can be attributed to the inherent fuzziness and nonuniform averaging done to generate the reference spectra.

4. Close examination of the scatter plots indicates that there are only a few stars ($\sim$5%) which fall outside the 2 $\sigma$ level. We looked into the spectral classification of these cases using the SIMBAD database operated at CDS, Strasbourg, France and found that the classification of these stars are based on photometry and/or spectra with poor spectral resolution. The classification of these stars needs further investigation, as evidenced by the case of HD 112872, for which the JHC classification is G6 III, while our schemes classify it as G9V, and one of the classifications of this star in SIMBAD is G8V. The literature classifications for this star are based on objective-prism plates of low resolution. The star cannot possibly be a giant, since it has a large measured proper motion, which suggests that it is a nearby star where high-luminosity stars rarely exist. Also, the star lies toward north Galactic pole, where the reddening effect is too small to affect its observed spectrum. Hence, HD 112872 is most likely a dwarf star.

## 5. CONCLUSION

We have demonstrated that a uniform classification of a large number of stellar spectra can be done in considerably short time using the techniques of $\chi^2$-minimization and ANN. The correlation of classification by these schemes with the catalog classification is about 0.993, and we are able to classify stars within an accuracy of two subspectral types. We were also able to classify them into various luminosity classes. Further improvement in accuracy and resolution of the classification is possible if a more exhaustive library of spectrum-luminosity class for a given metallicity is made available.

We were motivated for the present work because of the availability and easy access to a large number of digital stellar libraries. It is expected that larger databases will be accessible, which will require fast classification, and this is where automated schemes will be highly desirable.

The future scope of this technique envisages determination of fundamental parameters for stellar atmospheres, namely, effective temperature, gravity, and metallicity from stellar spectra. The globular clusters are being studied by multiobject, fiber-fed spectrographs, and a large number of spectra in a single frame are available. In order to classify these spectra on-line, we require such automated techniques.

We thank the IUCAA Astronomical Data Center for providing the spectral catalogs used in the current work. We also thank an anonymous referee for helpful suggestions and comments. This research has made use of the SIMBAD database operated at CDS, Strasbourg, France.

## REFERENCES

Adorf, H. M. 1986, in Data Analysis in Astronomy II, ed. V. Di Gesú, L. Saasi, P. Craine, J. H. Friedman, & S. Levialdi (New York: Plenum), 61
Adorf, H. M., & Meurs, E. J. A. 1988, in Large-Scale Structure Of The Universe: Observational and Analytical Methods, ed. W. C. Seitter, H. W. Duerbeck, & M. Tacke (Heidelberg: Springer-Verlag), 315
Bevington, P. R. 1969, Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill)
Gorgas, J., Faber, S. M., Burstein, D., Gonzalez, J. J., Courteau, S., & Prosser, C. 1993, ApJS, 86, 153
Gunn, J. E., & Stryker, L. L. 1983, ApJS, 52, 121
Jacoby, G. H., Hunter, D. A., & Christian, C. A. 1984, ApJS, 56, 257 (JHC)
Jaschek, C., & Jaschek, M. 1990, The Classification of Stars (Cambridge: Cambridge Univ. Press)
Kurtz, M. J. 1984, in The MK Process and Stellar Classification, ed. R. F. Garrison (Toronto: Dunlap Obs.), 136
Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992, AJ, 103, 318
Pickles, A. J. 1985, ApJS, 59, 33
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533
Silva, D. R., & Cornell, M. E. 1992, ApJS, 81, 865 (SC)
Storrie-Lombardi, M. C., Lahav, O., Sodre, L., & Storrie-Lombardi, L. J. 1992, MNRAS, 259, 8
von Hippel, T., Irwin, M., Storrie-Lombardi, M., & Storrie-Lombardi, L. 1992, in Proc. 3d DAEC Workshop Feedback of Chemical Evolution on Stellar Content of Galaxies, ed. D. Alloin & G. Stasinska (Meudon: Observatoire de Paris), 73
———. 1993, MNRAS, in press.