An Overview of Spectral Synthesis of Stellar Populations

Robert W. O'Connell

Astronomy Department, University of Virginia P.O. Box 3818, Charlottesville, VA 22903-0818, USA

Abstract. The key step in the spectral synthesis of stellar populations is comparison of the models to the data. I discuss some general aspects of this comparison, including the information content of spectra, answerability and nonuniqueness, the importance of observational precision, and the limitations of using small data subsets such as broad band colors.

1. Introduction

The prospect of deducing the evolutionary history of galaxies from their integrated spectra has tantalized astronomers since the pioneering work of Whipple (1935), 60 years ago. At that time the most serious difficulty for integrated light methods was the lack of a basic understanding of stellar structure and evolution. We now have far more than a basic understanding in these areas, as the many beautiful models presented at this conference demonstrate. The techniques for integrated light analysis of stellar populations now occupy somewhat the same place in extragalactic astronomy as does radioactive decay dating in archaeology or geology. The entire field relies on the basic chronologies developed through integrated light analysis even if the specialists are slow to reach consensus on the details. Since we (the specialists) tend to focus on those details, it is easy to forget how far we have come since Whipple. Today we argue vigorously over factors of 2 in age or metal abundance; in the 1930's, ages of galaxies were uncertain within a factor of perhaps 100, and the concept of metal enrichment did not even exist.

The recent literature, not to mention attendance at this conference, shows burgeoning interest in integrated light techniques for stellar population analysis. I think there are two main stimuli for this: one technical, the other scientific. The technical stimulus is the proliferation of powerful workstations and interactive software. Spectral synthesis has always depended on bringing together very large and disparate data sets (both observational & theoretical). This may be as complicated a data management problem—in a conceptual sense—as any in astronomy, and it was frankly intimidating in the mainframe computing environment of only 15 years ago. That discouraged people from learning these techniques and certainly constrained the variety of modeling attempted.

The major scientific stimulus is the wealth of emerging information on the high redshift universe. We have found direct evidence for galaxy evolution, and there is a compelling fascination in following its progress. High redshift systems are obviously much too distant to observe individual stars, so we must use their integrated light properties to probe their stellar populations and evolution.

The central issue on which this conference is intended to focus is: how well can we do this? **How well can integrated light analysis of a stellar population determine the following as a function of time: the star formation rate, chemical abundance ratios, and the initial mass function?** I take these to be the generally agreed-upon goals of current population studies. The key step in the process is neither the acquisition of data nor the generation of synthetic models for populations, but rather the *comparison* between the two. It is an oddity of the field that this confrontation is not often discussed. One basic limitation in the process is widely recognized, namely the "nonuniqueness" problem; but, with a few exceptions, it has not been examined quantitatively. Here, I would like to explore the general problem of comparing integrated light observations and theoretical spectral synthesis models and review some rules of thumb derived from the experience of synthesis practitioners.

2. Integrated Light versus Color Magnitude Diagrams

First, a brief comment on the relative merits of integrated light and color-magnitude diagram techniques. There is a school of thought which claims that integrated light is seriously inferior to star-by-star analysis using CMD's. This would be an unhappy conclusion, since star-by-star techniques are useless at large distances. The fair assessment is that the two techniques are complementary. Their relative utility in a given situation depends on some generalized comparison of their signal-to-noise ratios. The classic counterexample to the superiority of CMD techniques was Morgan's demonstration (e.g. Morgan & Mayall 1957) from its integrated spectrum that the bulge of M31 was metal rich and could not be a Pop II system, contrary to Baade's (1944) famous CMD analysis. At the time it was not possible to obtain the same information from CMD data; even today, because of image crowding, the mean metallicity of the bulge population is hard to extract from star-by-star data. Furthermore, it is the physics of stellar interiors and atmospheres which is largely responsible for the well-publicized ambiguities (e.g. age vs. metallicity) which appear in integrated light studies. CMD studies cannot avoid similar difficulties, as the literature on controversies over interpreting star clusters voluminously attests. One must be careful whichever technique is used.

3. Answerability and Nonuniqueness

Answerability is an inelegant term to remind us that we need to formulate questions about stellar populations which are answerable given the data and models in hand. While this may seem obvious, doing so can be the hardest part of the problem. There do exist unanswerable questions in this business!

There is no simple test for answerability. The only way to proceed is to make trials using simulated data sets (or real ones, where the answers are known from some other method). There are two intertwined issues: the *sensitivity* of the observables to the population parameters of interest and the *uniqueness* of the outcome. An answerable question does not necessarily imply a unique interpretation. Rather, it is one where the ambiguity in the outcome is small compared to the range of astrophysical interest. In terms of making fits to data, a result is ambiguous, or "nonunique", if *(i)* another model can be found which yields an equivalent statistical fit; *(ii)* the other model is astrophysically plausible; and *(iii)* the astrophysical implications of the other are distinct. All of the terms used here are to some extent a matter of judgement. The point, however, is that it is important to explore answerability from the standpoint of all three criteria and to be ready to reformulate the questions asked if necessary.

As a benchmark, we might recall the simplicity of the original stellar populations concept proposed by Baade (1944). He stipulated only *two* population types: today we would characterize those as very young and metal rich (Pop I) and very old and metal poor (Pop II). These two types have such different integrated spectra that it would be trivial to distinguish between them in integrated light. If all real populations were this simple, there would be no need for this conference! Today we understand that a pure population is specified not by a binary choice but by 3 or more parameters, each of which range continuously, most over orders of magnitude. Equally important, we recognize that most real populations are *mixtures* of pure populations.

For a concrete but simple example of a mixture problem (in age), consider the spectra shown in Figs. 1 and 2. Fig. 1 shows the optical spectra of the nuclei of three galaxies at $\sim$8 Å resolution with $S/N \sim 30$. Anyone familiar with stellar spectra and elementary stellar evolution would quickly be able to assign rough ages of <50, $\sim$500, and $\gtrsim$5000 Myr to these three spectra. Distinguishing the three on the basis of the presented observables is as easy as would be the case with pure Pop I and Pop II spectra. (Relative changes in spectral signatures are roughly proportional to $\Delta \log t$; the differences present here for $\Delta \log t \sim 1$ are much larger than the $\delta m \sim 0.03$ mag precision of the spectra)

In Fig. 2 I have combined the three spectra with equal weights at 4500Å. The resulting composite resembles that of a distant spiral galaxy, where disk and bulge light are inseparable. The special signatures of each constituent can still be recognized because of the good signal-to-noise ratio, but this is only a qualitative remark. The relevant answerability issue here is how well we can *invert* the mixed spectrum into its constituents and derive the fraction of the light contributed by each. This is equivalent to determining its star formation history.

4. The Population Information Content of Spectra

Answerability obviously depends on the *information content* of the data. The experience of synthesis practitioners shows that, as an abstract function,

$$\text{Information} = \mathbf{f}\,(NID,\, S/N,\, \lambda/\delta\lambda,\, \lambda_{\max}/\lambda_{\min})$$

where NID is the number of independent data points, S/N is the signal to noise ratio, $\delta\lambda$ is the minimum resolvable wavelength, and the ratio of maximum to minimum wavelength measures the wavelength baseline of the data. The first partial derivative of $\mathbf{f}$ with respect to any of these variables is positive (at least when the variables are small). By "information" here, of course, we mean information on the populations, not on the spectra. The appearance

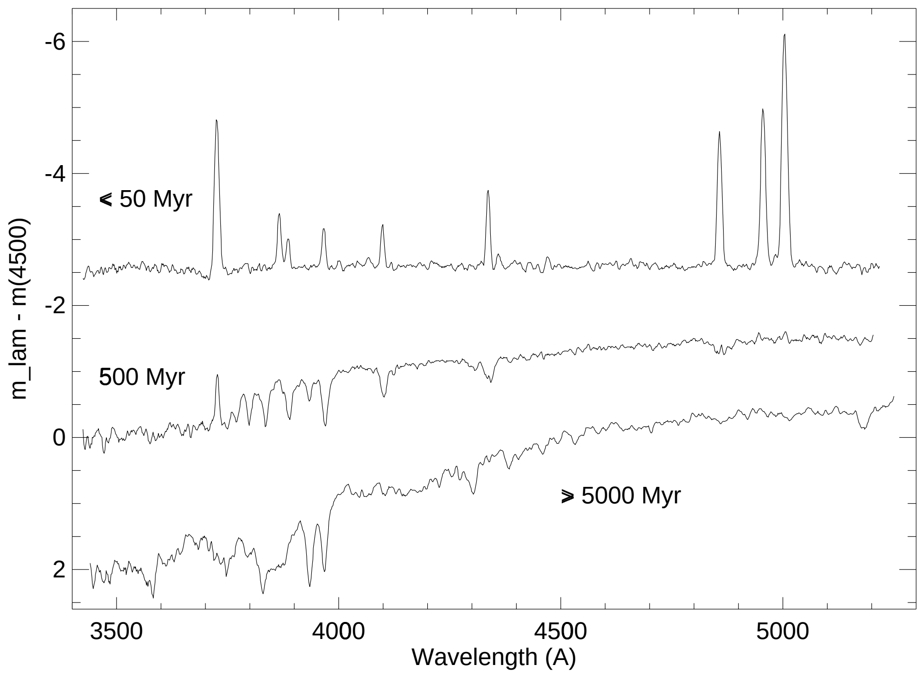

Figure 1. Nuclear spectra of three galaxies, with arbitrary offsets. Approximate light-weighted mean ages are assigned as indicated. The three are not highly composite, although the emission line in the 500 Myr spectrum demonstrates the presence of a minority component which is much younger than the mean population.

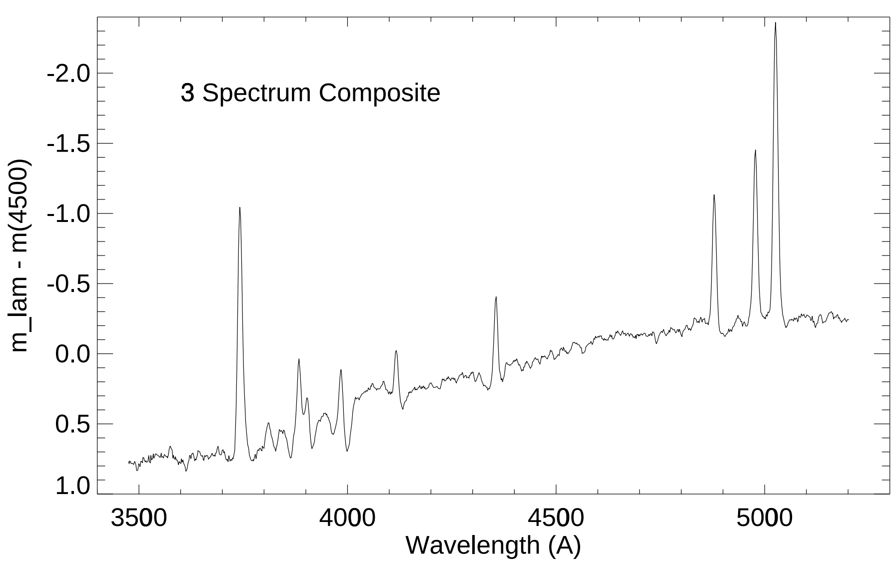

Figure 2. A composite formed by combining the three spectra in Fig. 1 with equal weights at 4500Å. How well can this spectrum be inverted to yield its three components?

of high quality spectra can be deceptive, because the amount of spectroscopic information is usually *much larger* than the amount of population information. There is no simple scaling: population information is not directly proportional to the number of data points.

An important answerability study by Pickles & van der Kruit (1990) emphasizes some of these points. They considered how well spectrophotometry between 3,600 and 10,000 Å with 13.5 Å resolution could invert a test mixture containing three old populations with a range of age and metal abundance. They found that their nonlinear programming algorithm could indeed perform reliable inversions, but only if the photometry had $S/N > 20$. (My own tests of a linear programming algorithm yielded a similar S/N requirement.) In their formulation of the problem, there were 14 components in the input spectral library for which light fractions needed to be determined; since the light fractions must sum to 1.0, there were 13 unknowns. Their dataset of some 470 data points therefore contained 36 times as many observations as unknowns. Despite this surfeit of data, they were not able to invert their test spectrum unambiguously if $S/N < 20$: that is, *additional data points could not compensate for low precision.* On the other hand, Pickles & van der Kruit did not explore the redundancy of their data in the high precision limit: e.g. how few data points would have permitted reliable inversions if all had, say, $S/N = 50$?

5. Data Subsets and the 2-Color Dilemma

The latter question is important because for distant galaxies we usually must deal with only a few observables. In fact, most of the literature on spectral analysis of populations uses only a small subset of the kind of spectral information present in Fig. 1. Popular subsets include broad-band colors, emission line strengths, absorption line strengths, and spectral discontinuities. It is evident from Fig. 2 that unless these were chosen with care it would be difficult to solve the inversion problem posed there. The *selection of the optimum data subset* for a particular application is one of the "answerability" issues.

The most successful application of small data subsets to populations is the method by which the number of ionizing stars, and hence the recent star formation rate, can be estimated from a few emission line strengths. This technique is so easy and familiar that the qualification "recent" is often overlooked. The emission lines are sensitive only to the star formation rate during a tiny window—the past $\sim$ 5 Myr or only 0.05% of the lifetime of a galaxy—which is in no way representative of the long-term star formation history of a galaxy. We shouldn't be misled into thinking this success is characteristic of the general problem.

Despite their ubiquitous appearance in population studies, UBV colors offer a prime example of a subset which is *not suitable for determining a galaxy's star forming history unless strong a priori assumptions are made.* This was pointed out some time ago by Larson & Tinsley (1978) and others but is worth illustrating again. Fig. 3 is a two-color diagram in its standard sense but with flux ratios, rather than colors, plotted (cf. Rabin 1981, Frogel 1985). The ratios are given by $F_U/F_V = 10.0^{-0.4(U-V)}$, etc. The solid triangles correspond to single generations of stars with different ages (we ignore abundance differences

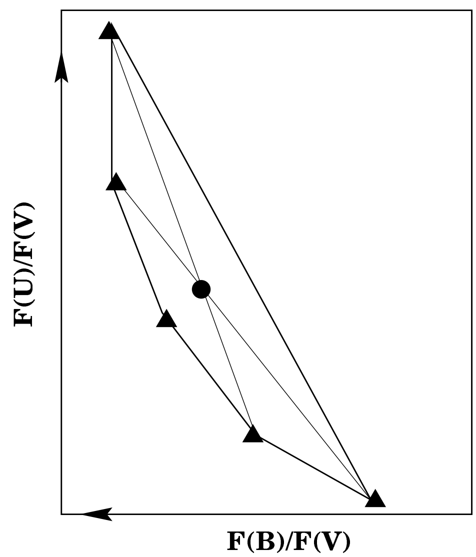

Figure 3. A standard UBV 2-color diagram, in flux ratio form, illustrating the "2-color dilemma". The triangles represent single generation populations, youngest at the top. The region inside the thicker boundary lines contains all possible combinations of these. The data point (solid circle) can be synthesized by many combinations of the populations, two of which are shown.

here). Any real population will be some combination of these single generations. In this kind of figure, any combination of two particular generations lies on a straight line connecting them. All possible composites therefore lie in the polygonal "feasible" region bounded below by the single generation locus as outlined on Fig. 3.

The dilemma here is obvious: most feasible points can be reached by *many* different combinations of the basic generations. I show one data point with the solid circle. This point could be the result of either of the two combination lines shown on the plot or many others (involving 3 or more generations) which are not shown. There could well be a fit corresponding to the familiar exponentially declining star formation rate. All the fits would have zero residuals but distinct astrophysical implications. They are excellent examples of ambiguity in spectral synthesis. Note that this nonuniqueness does not result from the "degeneracy" of the basic generations since, at least as plotted here, those are independent (i.e. cannot be synthesized from one another). The dilemma arises simply from the limited information content of the two colors.

This particular ambiguity has generated its share of prominent controversies. Among the best known is the debate over the ages of high redshift radio galaxies, e.g. 0902+32 (Lilly 1988, Chambers & Charlot 1990, Bithell & Rees 1990, Rocca-Volmerange & Guiderdoni 1990). Lilly interpreted his broadband observations of this redshift 3.4 object to indicate a relatively old (1-2 Gyr) galaxy in which a much more recent burst of star formation occurred. An object at this redshift which is already 1-2 Gyr old would imply a surprisingly high formation redshift. The other authors preferred alternatives with a dominant intermediate age (300 Myr) population and a lower implied formation redshift. It turns out this situation corresponds to the dilemma shown in Fig. 3: either interpretation is possible because of the ambiguity of broad band colors (exacerbated here by low S/N and the fact that the observed bands are not widely separated in the rest frame). The problem was illustrated again in the case of 4C 41.17 by Chambers, Miley, & van Breugel (1990), who found four distinct models which fit the observed colors equally well. Other examples of broad-band color ambiguities were discussed by Schweizer & Seitzer (1992) and Fritze-v. Alvensleben & Gerhard (1993) in the context of merging galaxies.

6. Solving the Combination Problem

Since limited information content is inevitable, especially for high redshift galaxies, we must learn to deal with the consequent implications for interpretation. The ambiguity illustrated above is a general problem, present at some level in all integrated light analyses regardless of the number of data points. I think the only sensible way to address this problem is to employ mathematical algorithms which can determine the *full range of arbitrary combinations of basic populations which fit the data within the errors.* To put this another way, we should impose as few *a priori* assumptions about the history of galaxies as possible. The simplest approach would use pure, single-generation models as the basic populations. Once the range of possible combinations is established, one can begin winnowing them on the basis of other information (e.g. see the papers on the radio sources discussed above).

There are two reasons for allowing *arbitrary* combinations of basic populations. *(i)* Despite 40 years of intensive effort, we have developed little secure theoretical intuition regarding the evolution of galaxies and therefore have little reason to impose *a priori* constraints on star forming histories. *(ii)* Real star formation histories are probably *strongly discontinuous.* Many galaxies which have been examined closely seem to have undergone discrete star forming episodes which are not well represented by any smooth function (such as an exponential). The evidence ranges from the Local Group (e.g. the beautiful example of the Carina Dwarf discussed by Smecker-Hane et al. at this conference) to distant systems in clusters and the field (see reviews by O'Connell 1994, Freedman 1995).

Fitting algorithms of the kind I have in mind are not widely used. The general method of searching for best fitting single-generation combinations has been applied by at least five different groups since 1980, most systematically by Bica, Alloin and their collaborators (e.g. Bica 1988), who use observed spectra of star clusters to represent single generations. Most of the optimizing algorithms now employed do not readily explore the full space of solutions near the optimum fit. However, Peck (1980) developed a linear programming algorithm which systematically determines all the basic solutions near the optimum; this was first applied to broad-band data (for M31) by Wu et al. (1980). Techniques of this type can deal with the combination problem.

The number of viable alternative combinations can be *greatly reduced* by increasing the information content of the data being analyzed. Next to increasing the number of data points, increasing their *precision* is the best way to reduce ambiguity. My impression is that for the kinds of questions often being asked about stellar populations, $S/N = 20$ is the threshold below which ambiguity becomes very serious. I would like to suggest, as a goal, that we strive to acquire *data on spectral energy distributions with a precision of 1% (i.e. $S/N = 100$) in flux ratios over a long wavelength baseline.* The 1% goal is challenging but not ridiculous. Most data employed in studies of nearby galaxies now has $S/N \gtrsim 30$. The widely-used Lick Observatory spectral index dataset often meets or exceeds the 1% criterion (e.g. Gorgas et al. 1993), though only over short wavelength baselines. Ironically, it was actually easier to achieve 1% photometry with older photomultiplier-tube technology than with today's CCD's, which, as multielement detectors, are harder to calibrate. The widespread practice of using small entrance apertures, which are sensitive to differential atmospheric refraction and seeing/guiding effects, also reduces precision. Small apertures are often used for better spectral resolution; however, since this simultaneously affects S/N in flux ratios there is an answerability issue here which needs to be considered in a trade-off study.

We will often fall short of the 1% criterion, and presumably will almost always do so at high redshift, but we should recognize that the farther we are from that kind of standard, the less reliable will be our population analyses.

7. Fitting Criteria

An underlying assumption of this whole discussion is that we will ask our models to *fit the data to observational precision.* Surprisingly, this is a standard which is often ignored in population analysis. There are many published examples of ostensibly "good" comparisons between data and models which involve residuals of many times the observational error (I will refrain from citations here). Although it is the observer's responsibility to provide realistic and detailed estimates of error from all sources, it is the modeler's responsibility to examine the goodness-of-fit quantitatively. It is not healthy for modelers to attribute large residuals to observational effects which may not exist (cf. Kepler 1609).

Fig. 4 illustrates some of these points. It is taken from our "homework" exercise, which is discussed by Arimoto elsewhere in these proceedings. The figures shows two fits to the standard "E1" spectral energy distribution made by my optimizing linear programming algorithm. The input library of observed stellar SED's (at discrete wavelengths) had solar abundance, except for a super-solar giant branch. Assumptions and methods were similar to those in O'Connell (1976). The good news is that it is possible to get fits with only $\sim 4\%$ mean residuals. The less good news is *(i)* the ambiguity: the two models, which differ in mean assigned age by 6 Gyr, fit almost equally well; and *(ii)* systematic residuals which are similar in both fits and which evidently cannot be cured by adjusting the age of the population.

Both models give similar fits because the constraints in them, which represent our knowledge of stellar physics, are slightly adjustable. For instance, the slope of the main sequence IMF is allowed to vary with mass, and the ratio of the number of giants to MS turnoff stars is permitted to change by up to $\pm 30\%$. It may be that the latter relaxation is too large, given our present understand-

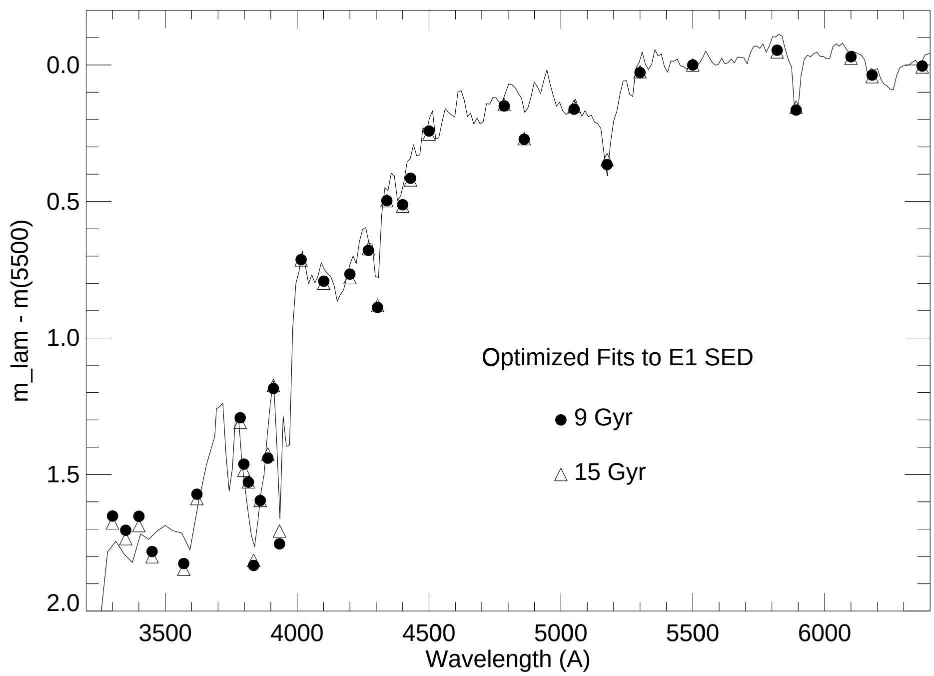

Figure 4. Fits to the standard "E1" spectrum for two single-generation models using a linear programming synthesis technique. The full region fitted extended to 11,000 Å.

ing of evolution, and that tighter constraints would permit better discrimination between models. Nonetheless, it is important that modeling techniques realistically account for the *uncertainty in our theoretical understanding* as well as in the observations.

One type of systematic residual is found at some (though not all) strong absorption features, where the observed line strengths are weaker than the predicted ones. This is probably caused by the finite velocity dispersion in the E1 SED, although aperture smoothing effects may also be involved. Obviously, these are yet other parameters which should be incorporated in the fits. An abundance mismatch between the library and galaxy could also be a factor. Note that the Mg I, Na I, and TiO spectral features in the 5100–6300 Å region are fit within 1–2% despite residuals at shorter wavelengths which may indicate that these models are inappropriate. This illustrates the desirability of including a large wavelength baseline.

It is not clear what causes the systematic residuals below 3600Å. All of the data below 4000Å is important for estimating the turnoff temperature in the population. The discrepancy may represent an abundance mismatch, an error in the foreground reddening, or calibration problems in the near UV. Its significance cannot be evaluated without knowing the observational error of each data point. The overall goodness of fit also cannot be evaluated without such information; if the E1 data actually met our 1% criterion, then neither model shown (having 4σ residuals) would be acceptable! The message is that a proper assessment of the models, and their astrophysical implications, must include a detailed, quantitative analysis of the goodness-of-fit.

8. Conclusion

Let me close by restating some of the main points made above and adding a few new ones. These are mostly generalities, since more detailed quantitative recommendations would depend on the specific application.

- Spectral synthesis modeling and data sets are now good enough to justify fully quantitative comparisons based on sophisticated mathematical optimizing techniques. The linear and nonlinear "programming" methods are most promising.
- The population information content of spectral energy distributions is much more limited than the spectral information content. It is not directly proportional to the number of data points, and it is strongly influenced by observational precision, spectral resolution, and wavelength coverage.
- The questions we ask about populations must be answerable given the data and models in hand; only deliberate quantitative testing can decide answerability and evaluate ambiguities. Testing is especially important where only small data subsets, such as broad band colors, are available. A given subset has a limited range of applicability.
- We need to establish a set of benchmark models and data sets (synthetic test cases or nearby galaxies and star clusters where independent information on stellar populations is available) with which to validate modeling and fitting techniques. We should also explore the variety of optimizing techniques available and select a few as standards.
- Fitting algorithms should routinely test for non-stellar components: extinction or scattering from dust, emission lines, nonthermal continuum, etc. Optimizing synthesis techniques can usually handle these easily, given sufficient spectral resolution and long wavelength baselines.
- We should strive for 1% observational precision in flux ratios.
- We should insist that modelers critically evaluate their fits to the data quantitatively.
- Because the histories of real galaxies are often strongly discontinuous, we should employ fitting techniques which can determine the full range of arbitrary combinations of basic populations which fit the data. Fitting methods should explicitly allow for uncertainties in our theoretical understanding.

The outcome of comparisons between models and data is not one-sided: synthesis studies of galaxies will soon begin to act as constraints on stellar evolution theory, rather than vice versa. As the data and our fitting methods improve, we may often be faced with situations where no model provides a satisfactory fit. The "supermetallicity" effect (Spinrad & Taylor 1971) was an early example. Although they mean agonizing reappraisals of both the data and models, such discrepancies should be welcomed because they point us to new phenomena and ultimately to a better understanding of the physics of both stars and galaxies.

Acknowledgments. This work has been supported in part by NASA grants NAGW-2596, NAGW-4106, and NAG5-700.

9. References

References

Baade, W. 1944, ApJ, 100, 137

Bica, E. 1988, A&A, 195, 76

Bithell, M., & Rees, M. J. 1990, MNRAS, 242, 570

Chambers, K. C., & Charlot, S. 1990, ApJ, 348, L1

Chambers, K. C., Miley, G. K., & van Breugel, W. J. M. 1990, ApJ, 363, 22

Freedman, W. L. 1995, in Stellar Populations (IAU Symposium No. 164), eds. P. C. van der Kruit & G. Gilmore (Dordrecht: Kluwer), p. 165

Fritze-v. Alvensleben, U., & Gerhard, O.E. 1993, A&A, 285, 751

Frogel, J. A. 1985, ApJ, 298, 528

Gorgas, J., Faber, S. M., Burstein, D., Gonzales, J. J., Courteau, S., & Prosser, C., ApJS, 86, 153

Larson, R. B., & Tinsley, B. M. 1978, ApJ, 219, 46

Lilly, S. 1988, ApJ, 333, 161

Morgan, W. W., & Mayall, N. U. 1957, PASP, 69, 291

O'Connell, R. W. 1976, ApJ, 206, 370

O'Connell, R. W. 1994, in Nuclei of Normal Galaxies: Lessons from the Galactic Center, ed. R. Genzel (Dordrecht: Kluwer), p. 255

Peck, M. L. 1980, ApJ, 238, 79

Pickles, A. J., & van der Kruit, P. C., A&AS, 84, 421

Rabin, D. M. 1981, PhD Thesis, California Institute of Technology

Rocca-Volmerange, B. & Guiderdoni, B. 1990, MNRAS, 247, 166

Schweizer, F., & Seitzer, P. 1992, AJ, 104, 1039

Spinrad, H., & Taylor, B. J. 1971, ApJS, 22, 445

Wu, C.-C., Gallagher, J. S., Peck, M., Faber, S. M., & Tinsley, B. M. 1980, ApJ, 237, 290