Multivariate Statistical Analysis Software Technologies for Astrophysical Research Involving Large Data Bases

A project supported by the NASA AISRP program, 1991 — 1994,
through the NASA Grant # NAS5-31348 and USRA contract # 5555-32
(USRA contract covering the period 1993 — 1994)

P.I.: S.G. Djorgovski, California Institute of Technology

The Final Technical Report

Table of Contents:

Summary

This report describes the results obtained in the course of the entire project, which was initially funded through a NASA AISRP grant NAS5-31348, and through USRA contract # 5555–32.

In the early phases of the project, we developed a user-friendly package for multivariate statistical analysis of small and moderate-size data sets, called STATPROG. The package was tested extensively on a number of real scientific applications, and has produced real, published results.

Subsequently, the bulk of the effort was in the development and testing of a major package used to process and analyse the data from the digital version of the Second Palomar Sky Survey (some 3 Terabytes of raw pixel information). This system, called SKICAT, incorporates the latest in machine learning and expert systems software technology, in order to classify the detected objects objectively and uniformly, and facilitate handling of the enormous data sets from digital sky surveys, and other sources. The system was developed as a major collaborative effort between our group, and the JPL Artificial Intelligence group.

The SKICAT system provides a powerful, integrated environment for the manipulation and scientific investigation of catalogs from virtually any source. The system serves three principal functions: image catalog construction, catalog management, and catalog analysis. Through use of the GID3* Decision Tree artificial induction software, SKICAT automates the process of classifying objects within CCD and digitized plate images. To exploit these catalogs, the system also provides tools to merge them into a large, complex database which may be easily queried and modified when new data, or better methods of calibrating or classifying the old, become available. The most innovative feature of SKICAT is the facility it provides to experiment with and apply the latest in machine learning technology to the tasks of catalog construction and analysis. The very same classification learning software used to create the classifiers in SKICAT's automated image cataloging tools are available for use on any SKICAT data set, or even data from external sources. SKICAT provides a unique environment for implementing these tools for any number of future scientific purposes.

Initial scientific verification and performance tests have been made using galaxy counts and measurements of galaxy clustering from small subsets of the survey data, and a search for very high redshift quasars. These tests helped uncover and fix several minor problems, and excercized the software in a real-life situation. All of the tests were successful, and produced new and interesting scientific results.

Attachments to this report give detailed accounts of the technical aspects of the SKICAT system, and of some of the scientific results achieved to date.

1. Introduction. The goals of the project

This report describes the technical and scientific results obtained in the course of the project which was initially funded through a NASA AISRP program grant, and then, in the last year of the total period of performance, through an USRA contract. Since there was no distinct boundary within the total period of performance, this report covers the entire project, and not just the work done under the USRA contract.

A substantial portion of the work was done as a collaborative effort between our group at Caltech, funded through this program and other sources, and the Artificial Intelligence group at the JPL (funded separately). While the Caltech contributions were substantial, our funding alone could not have accomplished the work and the results described here. Further cost sharing was through the P.I.'s NSF Presidential Young Investigator award, which covered many of the scientific verification tests and applications, and from Palomar Observatory, which paid one half of the salary of the postdoctoral fellow (Dr. de Carvalho) who worked on this project for the past year or so.

The motivation and the goals behind our work were to confront the problem of extracting interesting scientific results from vast amounts of data, with a minimum of loss and waste, in our fields of interest, viz., astronomy and space science. Raw data, no matter how expensively obtained, are no good without an effective ability to process them quickly and thoroughly, and to refine the essence of scientific knowledge from them. We approached the problem with a belief that many of the advanced tools needed for this task already exist in the various fields of computer science and statistics. Our practical goal was to bridge the gap between the disciplines, and introduce the modern data management and analysis software technologies into astronomy and astrophysics.

Our philosophy throughout has been to seek existing, applicable software and algorithms from the public domain wherever possible, and minimize independent programming effort (except for the interfaces, etc.). We did not want to reinvent any wheels, but to identify the most promising tools from the vast amount of scattered software, available in the open scientific literature or commercially, and to assemble some particularly useful pieces into working scientific packages. We verified their effectiveness and improved their design by attacking some real-life scientific problems. In the end, we provided packaged tools for working astronomers, who have to deal with large amounts of data, and extract a maximum science from it.

Our work proceeded in two stages: First, we developed a simple, but very effective and scientifically productive multivariate statistical analysis package (STATPROG). We utilized as much as possible existing or published routines and algorithms, with some programming and development of our own, and after an extensive comparisons, testing, and evaluation, put together a user-friendly, science-ready package. The package was scientifically validated through some real advances and discoveries, published in major astronomy journals, as documented in a number of references listed in the Bibliography. The work on STATPROG was done entirely during the NASA/AISRP funding stage of this project, before the USRA contract.

We then embarked on a larger and more ambitious effort, which constituted the bulk of our work, and which was the subject of N. Weir's Ph.D. thesis at Caltech. We were very fortunate in this endeavor to start an extremely productive and mutually beneficial collaboration with the JPL AI group, and in particular Drs. R. Doyle and U. Fayyad.

Our initial motivation there was to facilitate the scientific exploitation of the digital scans of the nearly 3000 photographic plates comprising the Second Palomar Observatory Sky Survey (POSS-II). The scans will ultimately add up to about 3 Terabytes of pixel data, an unprecedented amount of image information in the optical/IR astronomy. These scans will be the highest quality set of images covering the entire northern sky produced to date, and will almost certainly not be surpassed for at least a decade. Their potential scientific value is enormous, if only the relevant information can be extracted quickly and efficiently. We estimate that ultimately $> 5 \times 10^7$ galaxies and $\gtrsim 2 \times 10^9$ stars should be detected on the POSS-II plates, reaching down to the 22nd B magnitude. As an illustration, this exceeds the *entire* IRAS survey by three orders of magnitude in the number of objects alone, and with much more information per object!

To provide for the construction, classification, and analysis of object catalogs from this three Terabyte imagery data set, we developed a software system we call the Sky Image Cataloging and Analysis Tool (SKICAT). The system incorporates the latest techniques from the fields of machine learning and artificial intelligence, and is probably one of the first major applications of such modern software technology to astronomy. The system consists of roughly three layers of information processing and analysis. The first one, which generates catalogs of automatically classified objects from the raw plate scans and CCD calibration images, and the second one, where image catalogs are matched and manipulated, have been covered in this project. The third layer, in which a powerful toolbox of modern data analysis algorithms is to be applied for scientific exploitation of the catalogs was only started, and the work was temporarily suspended by the termination of our funding. We are now pursuing alternative funding sources to continue this work.

SKICAT is a collection of new and borrowed, commercial and public domain, software products which have been integrated for a common purpose. The current version of SKICAT uses the Sybase commercial database package for catalog storage and management. The system is thereby designed to manage a data base constantly growing and improving with time. With consistent command line and X-windows interfaces, the programs collectively meet the following three demands of standard astronomical surveys: catalog construction, management, and analysis. We have already demonstrated SKICAT's successful application to the digitized POSS-II. The system is already beginning to produce real science, and we believe that it will become scientifically useful to the astronomical community at large. Within it, data from a variety of different wavelengths could be retrieved and cross-analyzed within the same powerful environment.

The lead part of this report gives a synthetic summary of the principal achievements. Technical details and specific results are described extensively in the Attachments, which represent papers submitted to refereed journals (two of them have been already practically accepted for publication, as of this writing). The Attachments thus constitute the real

technical description of the results of this work. Additional results and interim reports can be found in the papers listed in the Bibliography.

2. The STATPROG package for multivariate data analysis

The STATPROG package consists of a number of standalone programs, originally developed under the VMS operating system. The software is written entirely in the standard Fortran 77, and has been ported to Unix Sparcstation platforms. We have systematically explored the available software resources, combined them in a homogeneous system, and tested them on real-life astronomical research problems. We have sampled some software from widely available, public-domain sources: the *Numerical Recipes* library and its companion volume (Press *et al.*), the monograph *Multivariate Data Analysis* by Murtagh and Heck, the *Gaussfit* package, available from Dr. Jeffries at the Astronomy Department, University of Texas at Austin, the MDRACE package, available from the Statistics Department, University of California at Berkeley, and several routines published in various astronomical journal papers. We also did some of our own coding of simple statistical diagnostics and fitting routines. The sources of the codes and their evaluation will be prepared in the later stages of this project.

The prototype package assembles a number of algorithms and routines, providing simple statistics, data handling, covariance analysis, Principal Component Analysis (PCA), bivariate optimization, and several versions of least squares fitting routines has been developed, running under the VMS operating system. The data input is through simple, standard ASCII files, combining any number of the leading header records, followed by data records (one per data vector) listed in a free-format column-by-column format. The package is very easy to use.

We performed the initial tests of the package on synthetic data, and then tackled some real astrophysical problems: systematics of properties of elliptical galaxies and their globular cluster systems. This excercise was both scientifically successful (with several papers published or in press in major journals so far, plus a large number of conference papers; see the Bibliography), and it also provided the valuable feedback, leading to a number of small design modifications and improvements. Such tests "under the fire" are the only way of providing a scientifically credible and useful software package.

The initial version of the package has been exported to several sites, both within the U.S., and abroad (Europe, and Brazil), for an independent evaluation by other astronomers. Their reactions were both useful and positive. We believe that STATPROG will become a valuable tool for the astronomical and space science research community.

3. The SKICAT system: Background and motivation

The initial motivation for the Sky Image Cataloging and Analysis Tool (SKICAT) was to facilitate the scientific exploitation of the Palomar - STScI Digitized Sky Survey, based on the scans of the nearly 3000 J, F, and N photographic plates comprising the Second Palomar Observatory Sky Survey (POSS-II). The scans will ultimately add up to about 3 terabytes of pixel data. These scans will be the highest quality set of images covering the

entire northern sky produced to date, and will almost certainly not be surpassed for at least a decade. Their potential scientific value is enormous, if only the relevant information can be extracted quickly and efficiently. We estimate that ultimately $> 5 \times 10^7$ galaxies and $> 5 \times 10^8$ stars should be detected on the POSS-II plates, reaching down to the 22nd B magnitude.

To provide for the construction, classification, and analysis of object catalogs from this three Terabyte imagery data set, the JPL Artificial Intelligence Group and Caltech Astronomy developed a software system we call SKICAT. The system incorporates the latest techniques from the fields of machine learning and artificial intelligence, and is probably one of the first major applications of such modern software technology to astronomy.

The system is described in some detail in the Attachment A to this report. Only a brief description will be given here. The SKICAT system is envisioned to consist of three layers of information processing and analysis. The first one, which generates catalogs of automatically classified objects from the raw plate scans and CCD calibration images is now complete. The second one, where image catalogs are matched and manipulated is now practically complete, with further refinements and capabilities being added to it on a continuous basis. The third layer, in which a powerful toolbox of modern data analysis algorithms is to be applied for scientific exploitation of the catalogs, was only partly completed, due to the termination of our funding. We are now in the process of seeking resources to complete this stage, and will do so as the future funding allows.

Put briefly, SKICAT is a collection of new and borrowed, commercial and public domain, software products which have been integrated for a common purpose. With consistent command line and X-windows interfaces, the programs collectively meet the following three demands of standard astronomical surveys: catalog construction, management, and analysis.

We first wrote and integrated the tools necessary for constructing object catalogs from the plate and CCD sequence images. Next, we applied state-of-the-art machine learning technology to develop an object classification method which is accurate at levels a full magnitude fainter than in previous automated Schmidt-based photographic sky surveys. As a result, we obtained more than twice the density of classified galaxies in our catalogs relative to previous ones. We next developed the machinery for matching multiple plate and CCD catalogs into a single "matched catalog", as well as a mechanism for performing sophisticated queries thereof.

No existing software, such as FOCAS or DAOPHOT (two commonly used astronomical software packages), was able to meet the complex demands of cataloging the Gigabyte images comprising a single plate scan, much less manage and match the few thousand plate catalogs that will comprise the whole POSS-II. Given that we had to design these management tools from scratch, we chose to generalize SKICAT to eventually accommodate astronomical catalogs from sources other than plates or CCDs (e.g., IRAS or ROSAT) with a modest amount of programming effort. Thereby data from a variety of different wavelengths could be retrieved and cross-analyzed within the same powerful environment.

The current version of SKICAT uses the Sybase commercial database package for catalog storage and management. To use a catalog within SKICAT, it must be registered

in the SKICAT system tables, where a complete description and history of every catalog loaded to date is maintained. Catalog revisions, that might result from deriving new and improved plate astrometric solutions or photometric corrections, are also logged. The system is thereby designed to manage a data base constantly growing and improving with time.

To maintain a reliable inventory, catalogs must be read from and written to external storage using a SKICAT interface. Catalogs may be matched, object by object, with other catalogs to form a matched catalog. This catalog contains independent entries for every measurement of every object detected in the constituent catalogs. The matched catalog may be queried using a sophisticated filtering and output mechanism to generate a so-called object catalog, containing just a single entry per matched object. For example, a user may requests all objects from a large sky region covered by multiple plates of the same or different passbands, specifying exactly which object attributes to report and from which source. Such queries may generate either additional Sybase objects tables or ASCII files, thus maintaining a considerable flexibility for different applications.

One of the most novel aspects of SKICAT is the facility to query overlap regions in the matched catalog and to dynamically update the constituent catalogs (their photometry, astrometry, classifications, etc.) in light of these results. The query tool may in turn be used to create a static, distributable data product from the current set of matched plate catalogs. However, the essential feature of SKICAT is that it maintains a "living," growing data set, instead of a data archive fixed for all time.

The third major component of SKICAT as envisioned, is a set of programs for survey data exploration and analysis. This includes the STATPROG library of multivariate statistical analysis routines we have developed earlier as part of this project, and much more. For example, we started to incorporate the neural networks and the GID3*/OBtree decision tree induction software used to produce the plate object classifier implemented in the AutoPlate script of SKICAT. We also started the work introducing unsupervised Bayesian classification algorithms, such as AUTOCLASS, for a more sophisticated and model-independent exploration of large data spaces. These programs might later be used to train and produce classifiers for scientific uses of the digitized POSS-II, or any other catalogs, that we had never anticipated.

4. The star-galaxy classification problem

A key technical and scientific problem in this kind of work is the objective star-galaxy classification. The accuracy and reliability of object classifications really determines the scientifically useful depth of a sky survey, regardless of the flux detection limits achieved. A paper describing our results and the work on the star-galaxy classification problem is given in the Attachment B to this report, and it gives all the details.

Briefly, in our early work on this problem, we experimented with the template-fitting approach as used in the FOCAS package, and Neural Net (NN) classifiers. We have applied this software to the star-selection classification problem with great success, achieving a better than 95% success rate on test data using a set of nine input attributes after training on

only a few hundred objects. We developed code which implements a multi-layer perceptron artificial NN model for non-linear regression and classification. The software provides for an arbitrary number of layers and nodes at the input, output, and hidden level, as well as a broad choice of linear and non-linear activation functions. A variety of optimization methods are available, including gradient descent-based standard back-propagation and highly efficient conjugate gradient and variable metric methods. The latter reduce network training time by more than an order of magnitude over the traditional method. We also did some research into the possibility of incorporating formal error estimates in the form of a covariance matrix associated with the network outputs for any given input, which is a novelty in the field of Neural Nets.

We then tried a different approach to the problem of automatic objective classification, using Decision Tree algorithms (ID3, GID3*). We applied the GID3* decision tree algorithm developed by Fayyad and a neural network to the task of selecting a set of stars from a relatively bright sample of objects. These are subsequently used to generate the point spread function, which is used in a template matching procedure for constructing more accurate classification attributes. Our tests indicated that both approaches worked comparably well, achieving $> 95\%$ success rates. We, therefore, chose to stick with the GID3* method, as it produces a readily comprehensible set of classification rules, unlike the neural network. Our tests on the actual PDS data indicated that we can perform star selection with $< 1\%$ error rate. When the final set of attributes produced by template matching are included, we are able to perform star/galaxy/artifact classification with $> 95\%$ accuracy down to $\sim 20^m$ in the B_J band, and $> 90\%$ accuracy down to $B_J = 21^m$. Thus, Decision Tree algorithms have been used as the principal object classification tools within SKICAT. More details are given in the Attachment B.

Finally, we started explorations of unsupervised learning algorithms such as AUTOCLASS, to the analysis of object catalogs derived from the digitized POSS-II. Our goal was to explore the of the power of unsupervised learning techniques to classify objects meaningfully and perhaps to discover previously unrecognized object categories in digital sky surveys. Our primary finding is that AUTOCLASS was able to form several sensible categories from a few simple attributes of the object images, separating the data into four recognizable and astronomically meaningful classes: stars, galaxies with bright central cores, galaxies without bright cores, and stars with a visible "fuzz" around them. In an independent experiment we found out that the two types of galaxies have distinct color distributions (the more concentrated class being redder, as indeed expected if they are predominantly early Hubble types), although no color information was given to the program! This illustrates the power of unsupervised classification techniques to discriminate between astronomically distinct types of objects on the basis of data alone. We believe that the application of such algorithms to large-scale astronomical sky surveys can aid in cataloguing the detected objects, and may even have the potential to discover new categories of objects. Thus, we believe that this remains a very interesting and promising area for the future work.

5. The SKICAT system: Description and current status

The single-plate reduction is accomplished by a parent unix script which calls subordinate routines for reading in and processing the plate image. The plate is broken into a set of 13 by 13 overlapping footprint images, which are analysed separately, and then combined in the master plate catalog (the full plate scans are over 23,000 by 23,000 pixels, or about 1 Gigabyte, which is too large to handle efficiently).

One of the most novel aspects of SKICAT is the facility to query overlap regions in the matched catalog and to dynamically update the constituent catalogs (their photometry, astrometry, classifications, etc.) in light of these results. The query tool may in turn be used to create a static, distributable data product from the current set of matched plate catalogs. However, the essential feature of SKICAT is that it maintains a "living," growing data set, instead of a data archive fixed for all time. This is one of the real novelties in our work, never before attempted in the astronomical data processing at large, especially in sky surveys.

We started exploring the unsupervised clustering and objective automatic classification techniques. For example, we investigated AUTOCLASS unsupervised classification software developed at NASA Ames, and explored other Bayesian inference and cluster analysis tools. These software tools may be capable of *independent or cooperative discoveries*, and their application may greatly enhance the productivity of practicing scientists.

Effectively, by crossing the wavelength boundaries and creating a synergy of space-based and ground-based data from surveys covering large fractions of the entire sky, we are approaching a new level of complexity in astronomical source catalogs. Furthermore, the catalogs we generate will be constantly changing, growing in size and scope, and improving in time, as new and better data come in. This is *an entirely new concept of an astronomical data catalog:* a downloadable, growing data base with which one interacts using semi-intelligent software robots (knowbots); no more dusty, immutable printed volumes! The tools we developed are generic to this concept of hypercatalogs. There is a fusion of the data and the information tools, and it is that new ground, at least within astronomy.

6. The initial scientific verification tests

While the principal thrust of this work was technical and software technology oriented, the validity of the data products and the software systems which generate them, as well as the power of the sophisticated data analysis tools (such as many functions of SKICAT) can be really verified only through an application to a real scientific problem. This testing on the fire line is an indispensable part of the system shakedown. We thus attacked, in a limited way, several important scientific problems using some of the preliminary catalogs generated by SKICAT. These are only initial, but still scientifically substantial experiments; they pave the way for the future pipeline processing and scientific exploration of digitized POSS-II, which should be funded separately elsewhere. Here we used them as test cases to excercize the system. Indeed, they helped us uncover and fix numerous "features" in the system.

One basic test of our galaxy photometry, parameter definition and measurement procedures, and star-galaxy classification, are galaxy counts as a function of magnitude. This is one of the traditional tests of cosmology (pioneered by Hubble), and it provides us with a sensitive test of internal consistency and accuracy. A detailed paper dealing with these tests is presented in the Attachment C to this report. Briefly, we have demonstrated an unprecedented level of accuracy and internal consistency relative to all previous studies using a comparable sky survey material. Since the raw data quality has not changed from the previous studies, our improvements are clearly due to the superior software technology now implemented within the SKICAT system.

A related test are studies of the large-scale structure using two-point correlation functions for the galaxies. The preliminary results here are equally encouraging. We presented them as a conference paper, and we will turn them into a journal paper shortly.

Another project which provides a stringent test of our star-galaxy classification and catalog matching procedures is the search for high-redshift ($z > 4$) quasars, using peculiar colors. The trick here is to select on the average one $z > 4$ quasar per approximately 10^5 foreground stellar images. The first results of this work are starting to come in, and the first luminous $z > 4$ quasar selected using this AI-based software technology from the sky survey scans has just been discovered at Palomar about two weeks ago! It is the first one of many more to come. This work is a part of Julia Smith's Ph.D. thesis at Caltech.

The accuracy of our star-galaxy classifications is also being tested through spectroscopy, a completely independent technique. This has been done by ourselves during our quasar search (virtually all objects classified as being stellar indeed turned out to be stellar), and by our colleagues who are conducting a massive redshift survey at Palomar: they find that the accuracy of our star-galaxy classifications is at least a factor of five higher than in the previous surveys using a comparable plate material. This work has been funded separately by the NSF, but it provides a valuable verification of our efforts.

Finally, we have started an exploration of the huge data bases resulting from the sky survey to discover and define objective catalogs of groups and clusters of galaxies. This work is also being funded separately, and it will provide valuable feedback to further refine and enhance our algorithms in the third layer of SKICAT.

We thus conclude that our software and algorithms have passed the initial scientific verification tests with flying colors. They are now starting to produce real science, and are being used by several independent groups for different projects.

7. Prospects for the future work

On the scale of a couple of years from now, the storage technologies may be good enough to revisit the cataloguing and classification problem in a whole new light: iterative or feedback catalog generation. The current practice (including SKICAT) is to measure the images once in a predetermined way, and derive the object parameters and classifications from these measurements (e.g., moments of the light distribution, etc.). Once measured, pixels are not revisited, since the image data volume is too bulky to keep on line. If history is any guide, this technical limitation may change very quickly. It would

then be possible to have intelligent object-finding and classifier algorithms automatically *redefine the measurement process*, i.e., go back to the pixels and measure some new object parameters if deemed necessary. This may be naturally accomplished using the so-called genetic algorithms, which are capable of evolving and self-improvement. We are not aware of any application of such tools in astronomy so far, yet this has a natural appeal. It would represent a truly novel approach to astronomical catalog generation.

The basic mode of our work has been to search for existing tools and software technologies on the cutting edge of applied statistics, machine intelligence and related fields, and apply them to specific and very pressing problems of astronomical data analysis. In this, we have already developed a successful set of tools, first STATPROG, and then, in collaboration with the JPL AI group, SKICAT. We thus hope to continue our role as a conduit between the communities of observational astronomers on the one side, and the applied software technology and computer science experts on the other. We are well positioned to do so, and we have a considerable and an ever growing credibility in the astronomical community. For example, astronomers involved with the planned Sloan Digital Sky Survey, astronomers at STScI involved with the HST Guide Star Catalogs, astronomers at IPAC and JPL involved in planning of the Two Micron All-Sky Survey, and some U.S. astronomers involved with the Rosat sky survey, expressed an enthusiastic interest in our work so far, and are keen to import our software and methodology. We welcome that as an additional source of an external scientific evaluation of our products. With the anticipated scientific results we hope to achieve based on these enabling information technologies, astronomers and space scientists will pay a serious attention to this interface of astronomy and computer science, and we hope to stimulate other groups to start similar efforts and collaborations.

Whereas we have approached this work with a specific application in mind, viz., the 3 Terabytes of digitized POSS-II burning holes in our pockets, we have understood from the start the universality of the problem, and of the proposed technical solutions we are trying to develop. These techniques are clearly and directly applicable to a wide variety of astronomical imaging applications, especially sky surveys of any sort: IRAS, Rosat, and those from the anticipated future missions. There are also potential ground-based applications of interest to NASA, e.g., the searches for Earth-crossing asteroids, where a substantial portion of the sky would be covered a few times per night, every night; our software can be almost directly ported to that problem. In addition to the efficient analysis of vast amounts of new data, these techniques can be also used to explore the existing data archives, and have a potential of revolutionizing the archival research (e.g., the HST archive, reanalysis of IRAS or HEAO-B data, etc.). This great universality should attract a very broad constituency of science users, probably with a multitude of applications which have never occurred to us...

It has thus been our long-term ambition from the onset of this effort to develop modern software tools for astronomy and space science of the turn of the century, and lay down the information processing infrastructure for the imminent data flood which is upon us. We thing we choose the exactly right path, in establishing an excellent and productive working relations with experts from the NASA-sponsored Artificial Intelligence community, and we hope to broaden this synergy on both sides. We see this as a first stage

of a larger technology transfer process, in our case from the applied computer science to a basic science of astronomy. Perhaps there is an even more fundamental undercurrent here: Information is the steam of the second industrial revolution, and here we are trying to make some good engines. If we are successful, others might be stimulated to try, and this may be a model of the growth process for the postindustrial economy.

8. Software distribution and and dissemination of results

We have received a very substantial interest from the astronomical community, upon the presentations of our work at various professional meetings. In particular, groups working on the Sloan Digital Sky Survey, the Two Micron All-Sky Survey, the HST Guide Star Catalog, a Center for Astrophysics group doing a deep redshift survey, a University of California consortium planning an ultra-deep survey with the Keck telescope, a JPL group planning a survey for the Earth-crossing asteroids, and numerous others. There is also an international component, from the COSMOS plate scanning machine group in Scotland, the ESO/ESTEC group in Germany, two groups in France, and a group in Brazil. There is thus a considerable and substantial interest in the astronomy and space science communities, both for the SKICAT system itself, and for the data products it is now generating. We have also seen a lot of interest from the astronomical software specialist community, at the various ADASS and AAS conferences, and other gatherings.

Catalog management aspects of SKICAT could be used directly for many data archive systems, a subject which is of a considerable and growing interest in the astronomy community.

A modified version of SKICAT, with a special data interface, has been used successfully by our collaborators at JPL and a group of planetary scientists, to search for and catalog millions of small volcanos on Venus, from the Magellan radar synthetic images. This illustrates very directly the broad applicability of our software and methodology.

All of the code is adequately documented internally. All of it is the standard C and Fortran, and in unix shell script language.

While so far we have been communicating with the interested groups on a case by case basis, we will establish a more systematic and orderly distribution of the software and object catalogs. The software itself (except, of course, for the commercial parts for which a license has to be purchased, such as Sybase) will be deposited in at least two NASA software distribution sites, along with the proper documentation. Several useful documents exist or are being completed now, and will be deposited in the form of LATEX and PostScript files:

- *SKICAT Users Manual*
- *SKICAT Installation Guide*
- *SKICAT Plate and CCD Processing Reference*
- *SKICAT Plate and CCD Processing Cookbook*
- *SKICAT Database Reference*

The papers describing the system and the relevant parts of the methodology are now submitted to the professional journals (see the Attachments); other papers in conference proceedings also cover some specific aspects of the work. They constitute an extended, and obviously fully public, form of documentation. We plan to publish further results as they are produced.

In addition to the specific software distribution, we are now looking into the distribution of catalogs via Internet and WWW. In order to make these vast amounts of data easily accessible, we will have to make the software available through the same venue. The network fashions change rapidly, and the exact mechanism by which we will accomplish this is still under consideration. This may well have a substantial educational component. Presumably the production and distribution of the catalogs will be funded separately, and it does not come under the scope of the present contract.

We emphasize that we have a substantial vested interest in seeing that our work is used by the community. We will thus make every effort to make it easily accessible.

Bibliography of papers supported in part or whole by this grant/contract

"Towards a Digitized Second Palomar Sky Survey: Initial Reduction and Star/Galaxy Classification", Weir, N., Djorgovski, S., Fayyad, U., and Doyle, R. 1991, *Bull. Am. Astron. Soc.* **23**, 1434.

"Systematic Differences Between the Field and Cluster Ellipticals", de Carvalho, R., and Djorgovski, S. 1992, *Astrophys. J. Letters* **389**, L49.

"On the Formation of Globular Clusters in Elliptical Galaxies", Djorgovski, S., and Santiago, B.X. 1992, *Astrophys. J. Letters* **391**, L85.

"Dynamical Evolution Effects on the Hot Stellar Populations in Globular Clusters", Djorgovski, S., and Piotto, G. 1992, *Astron. J.* **104**, 2112.

"Galaxy Manifolds and Galaxy Formation", Djorgovski, S. 1992, in G. Longo, M. Capaccioli, and G. Busarello (eds.), *Morphological and Physical Classification of Galaxies*, p. 337. Dordrecht: Kluwer.

"Systematics of Galaxy Properties: Clues About Their Formation", Djorgovski, S. 1992, in R. de Carvalho (ed.), *Cosmology and Large-Scale Structure in the Universe*, *A.S.P. Conf. Ser.* **24**, 19.

"Properties of Dwarf Spheroidals", Djorgovski, S., and de Carvalho, R. 1992, in G. Longo, M. Capaccioli, and G. Busarello (eds.), *Morphological and Physical Classification of Galaxies*, p. 379. Dordrecht: Kluwer.

"Properties of Brightest Cluster Members", Djorgovski, S., de Carvalho, R., Shlosman, I., and Schombert, J. 1992, in G. Longo, M. Capaccioli, and G. Busarello (eds.), *Morphological and Physical Classification of Galaxies*, p. 427. Dordrecht: Kluwer.

"Systematic Differences Between the Field and Cluster Ellipticals", de Carvalho, R., and Djorgovski, S. 1992, in B. Barbuy and A. Renzini (eds.), *The Stellar Populations of Galaxies*, Proceedings of the IAU Symp. #149, p. 400. Dordrecht: Kluwer.

"Surface Photometry of Southern Ellipticals and Fundamental Plane Solutions for the Fornax Cluster Galaxies", Penereiro, J., Djorgovski, S., de Carvalho, R., Gorjian, V., and Thompson, D. 1992, in R. de Carvalho (ed.), *Cosmology and Large-Scale Structure in the Universe*, *A.S.P. Conf. Ser.* **24**, 123.

"Surface Photometry and Fundamental Plane Solutions for Elliptical Galaxies in the Virgo and Coma Clusters", Gorjian, V., Djorgovski, S., de Carvalho, R., Penereiro, J., and Weir, N. 1992, in R. de Carvalho (ed.), *Cosmology and Large-Scale Structure in the Universe*, *A.S.P. Conf. Ser.* **24**, 129.

"The Manifold of Low Surface Brightness Dwarf Galaxies", de Carvalho, R., and Djorgovski, S. 1992, in R. de Carvalho (ed.), *Cosmology and Large-Scale Structure in the Universe*, *A.S.P. Conf. Ser.* **24**, 135.

"Multivariate Analysis of Hickson's Compact Galaxy Groups", Djorgovski, S., Weir, N., and de Carvalho, R. 1992, in R. de Carvalho (ed.), *Cosmology and Large-Scale Structure in the Universe*, *A.S.P. Conf. Ser.* **24**, 141.

"Applying Machine Learning Classification Techniques to Automate Sky Object Cataloguing" Fayyad, U., Doyle, R., Weir, N., and Djorgovski, S. 1992, in *Proceedings of the International Space Year Conference on Earth & Space Science Information Systems*, Pasadena, CA, February 1992.

"Automating Sky Object Classification in Astronomical Survey Images", Fayyad, U., Doyle, R., Weir, N., and Djorgovski, S. 1992, in J. Zytkow (ed.), Proceedings of the ML-92 Workshop on *Machine Discovery (MD-92)*, Ninth International Conference on Machine Learning, Aberdeen, Scotland, p. 117, San Mateo, CA: Morgan Kaufman Publ.

"Multivariate Statistical Analysis Software Technologies for Astrophysical Research Involving Large Data Bases", Djorgovski, S. 1992, in *Proc. of the Applied Information Systems Research Program Workshop II*, pp. D-4 and E-1, Washington: NASA/ISB/OSSA.

"The Palomar Observatory - STScI Digital Sky Survey: I. Program Definition and Status", Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I.N., and Laidler, V. 1992, *Bull. Am. Astron. Soc.* **24**, 750.

"The Palomar Observatory - STScI Digital Sky Survey: II. The Scanning Proces", Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I.N., and Sturch, C. 1992, *Bull. Am. Astron. Soc.* **24**, 741.

"An Analysis of the Palomar Observatory - STScI Digital Sky Survey: Catalog Construction and Initial Results", Weir, N., Djorgovski, S., Fayyad, U., and Doyle, R. 1992, *Bull. Am. Astron. Soc.* **24**, 741.

"Digitized POSS-II: Initial Scientific Tests using Galaxy Number Counts", Weir, N., Djorgovski, S., and Fayyad, U. 1992, *Bull. Am. Astron. Soc.* **24**, 1139.

"Digitization of the Second Palomar Sky Survey: Program Definition and Status", Djorgovski, S., Weir, N., and Lasker, B. 1992, IAU Commission 9 Working Group on Wide-Field Imaging Newsletter #2, 29.

"On the Effects of Cluster Density and Concentration the Horizontal Branch Morphology: The Origin of the Blue Tails", Fusi Pecci, F., Ferraro, F., Bellazzini, M., Djorgovski, S., Piotto, G., and Buonanno, R. 1993, *Astron. J.* **105**, 1145.

"What Determines the Stellar Mass Functions in Globular Clusters?", Djorgovski, S., Piotto, G., and Capaccioli, M. 1993, *Astron. J.* **105**, 2148.

"Multivariate Analysis of Globular Cluster Systems in Early-Type Galaxies", Santiago, B.X., and Djorgovski, S. 1993, *M.N.R.A.S.* **261**, 753.

"The Meaning and the Implications of the Fundamental Plane", Djorgovski, S., and Santiago, B.X. 1993, in J. Danziger *et al.* (eds.), proceedings of the ESO/EIPC Workshop on *Structure, Dynamics, and Chemical Evolution of Early-Type Galaxies*, ESO publication No. 45, 59.

"Families of Early-Type Stellar Systems", Djorgovski, S. 1993, in G. Smith and J. Brodie (eds.), *The Globular Cluster - Galaxy Connection*, *A.S.P. Conf. Ser.* **48**, 496.

"SKICAT: A Machine Learning System for Automated Cataloging of Large Scale Sky Surveys", Fayyad, U.M., Weir, N., and Djorgovski, S. 1993, in *Proceedings of the Tenth International Conference on Machine Learning*, p. 112. San Mateo, CA: Morgan Kaufmann Publ.

"SKICAT: A System for the Scientific Analysis of the Palomar – ST ScI Digital Sky Survey", Weir, N., Djorgovski, S., Fayyad, U., Roden, J., and Rouquette, N. 1993, in A. Heck and F. Murtagh (eds.), *Astronomy from Large Data Bases II*, ESO publication No. 43, 513.

"The Second Palomar Sky Survey", Reid, I.N., and Djorgovski, S. (for the POSS-II photographic and digital survey teams) 1993, in B.T. Soifer (ed.), *Sky Surveys: Protostars to Protogalaxies*, *A.S.P. Conf. Ser.* **43**, 125.

"Towards the Analysis of the Digital POSS-II: Catalog Construction and Classification Results", Weir, N., Djorgovski, S., Fayyad, U., Doyle, R., and Roden, J. 1993, in B.T. Soifer (ed.), *Sky Surveys: Protostars to Protogalaxies*, *A.S.P. Conf. Ser.* **43**, 135.

"Using Machine Learning Techniques to Automate Sky Survey Catalog Generation", Fayyad, U., Weir, N., Roden, J., Djorgovski, S., and Doyle, R. 1993, in K. Krishen (ed.), *Sixth Annual Workshop on Space Operations, Applications, & Research (SOAR-92)*, NASA CP-3187, p. 340.

"SKICAT: A Cataloging and Analysis Tool for Wide Field Imaging Surveys" Weir, N., Fayyad, U., Djorgovski, S., Roden, J., and Rouquette, N. 1993, in R. Hanisch, R. Brissenden, and J. Barnes (eds.), *Astronomical Data Analysis Software and Systems III*, *A.S.P. Conf. Ser.* **52**, 39.

"Automated Analysis of a Large-Scale Sky Survey: The SKICAT System", Fayyad, U.M., Weir, N., and Djorgovski, S. 1993, in G. Piatetsky-Shapiro (ed.), *Proceedings of AAAI-93 Workshop on Knowledge Discovery in Databases*, p. 1, Washington: AAAI Press.

"Multivariate Statistical Analysis Software Technologies for Astrophysical Research Involving Large Data Bases", Djorgovski, S. 1993, in *Proc. of the Applied Information Systems Research Program Workshop III*, pp. C-4 and E-26, Washington: NASA/ISB/OSS.

"A Multivariate Analysis of Galaxy Cluster Properties", Ogle, P.M., and Djorgovski, S. 1993, *Bull. Am. Astron. Soc.* **25**, 839.

"A New Database of Globular Clusters Parameters: Distributions of Cluster Properties and Correlations Between Them", Djorgovski, S., and Meylan, G. 1993, *Bull. Am. Astron. Soc.* **25**, 885.

"Digitized POSS-II: Galaxy Number Counts in Two Colors Over a Multi-Plate Region", Weir, N., Djorgovski, S., and Fayyad, U. 1993, *Bull. Am. Astron. Soc.* **25**, 890.

"Galaxy Number Counts in Two Colors From Digitized POSS-II: Evidence for Galaxy Evolution at Low Redshifts?", Weir, N., Djorgovski, S., and Fayyad, U. 1993, *Bull. Am. Astron. Soc.* **25**, 1398.

"Initial Measurements of the Galaxy Angular Two-Point Correlation Function from the Digitized Palomar Observatory Sky Survey", Brainerd, T.G., Weir, N., Djorgovski, S., and Fayyad, U. 1993, *Bull. Am. Astron. Soc.* **25**, 1400.

"Cataloging of the Northern Sky From Digitized POSS-II: A Progress Report", Djorgovski, S., Weir, N., and Fayyad, U. 1993, *Bull. Am. Astron. Soc.* **25**, 1469.

"Cataloguing of the Northern Sky from the POSS-II using a Next-Generation Software Technology", Djorgovski, S., Weir, N., and Fayyad, U. 1993, IAU Commission 9 Working Group on Wide-Field Imaging Newsletter #4, 14.

"Analysis of the Palomar-STScI Digital Sky Survey: an Overview of the SKICAT System", Weir, N., Djorgovski, S., Fayyad, U., and Roden, J. 1993, IAU Commission 9 Working Group on Wide-Field Imaging Newsletter #4, 15.

"The Initial Results on Galaxy Counts and Searches for $z > 4$ Quasars using the Palomar-STScI Digital Sky Survey", Weir, N., Djorgovski, S., and Fayyad, U. 1993, IAU Commission 9 Working Group on Wide-Field Imaging Newsletter #4, 54.

"Machine Learning for Automated Catalog Generation", Fayyad, U., Atkinson, D., Djorgovski, S., and Weir, N. 1993, *NASA Information Systems Newsletter*, Nov. 1993, p. 3.

"The Galactic Globular Cluster System", Djorgovski, S., and Meylan, G. 1994, *Astron. J.* **108**, 1292.

"Processing and Analysis of the Palomar – STScI Digital Sky Survey Using a Novel Software Technology", Djorgovski, S., Weir, N., and Fayyad, U. 1994, in D. Crabtree, R. Hanisch, and J. Barnes (eds.), *Astronomical Data Analysis Software and Systems III, A.S.P. Conf. Ser.* **61**, 195.

"Cataloging the Northern Sky Using a New Generation of Software Technology", Weir, N., Djorgovski, S., Fayyad, U., Smith, J.D., and Roden, J. 1994, in H. MacGillivray *et al.* (eds.), *Astronomy From Wide-Field Imaging*, Proceedings of the IAU Symp. #161, p. 205. Dordrecht: Kluwer.

"Automated Cataloging and Analysis of Sky Survey Image Databases: the SKICAT System", Fayyad, U., Weir, N., and Djorgovski, S. 1994, in *Proc. of the 2nd International Conference on Information and Knowledge Management (CIKM-93)*, Washington: ISCA/ACM, in press.

"Cataloging and Exploration of the Digitized POSS-II", Djorgovski, S., de Carvalho, R., Schombert, J., Weir, N., Smith, J., Barton, E., Fayyad, U., and Roden, J. 1994, *Bull. Am. Astron. Soc.* **26**, 915.

"Digitizing the Sky", Djorgovski, S. 1994, *Nature* News & Views, **370**, 18.

"The Fundamental Plane Correlations for Globular Clusters", Djorgovski, S. 1995, *Astrophys. J. Letters* in press.

"Automated Star/Galaxy Classification for Digitized POSS-II", Weir, N., Fayyad, U., and Djorgovski, S. 1995, *Astron. J.* in press

"Initial Galaxy Counts From Digitized POSS-II", Weir, N., Djorgovski, S., and Fayyad, U. 1995, *Astron. J.* in press

"Automated Analysis and Exploration of Image Databases: Results, Progress, and Challenges", Fayyad, U., Smyth, P., Weir, N., and Djorgovski, S. 1995, *Journal of Intelligent Information Systems* 4, 1. Dordrecht: Kluwer.

"Clustering Analysis Algorithms and Their Applications to Digital POSS-II Catalogs" de Carvalho, R., Djorgovski, S., Weir, N., Fayyad, U., Cherkauer, K., Roden, J., and Gray, A. 1995, in R. Hanisch, *et al.* (eds.), *Astronomical Data Analysis Software and Systems IV*, *A.S.P. Conf. Ser.* in press.

"Applications of Clustering Analysis and Unsupervised Classification Algorithms to Digitized POSS-II", de Carvalho, R., Djorgovski, S., Weir, N., Fayyad, U., Roden, J., Gray, A., and Cherkauer, K. 1995, *Bull. Am. Astron. Soc.* in press.

Attachment A:

A description of the SKICAT system

From a paper submitted to the *Publ. of the Astron. Soc. of the Pacific*, November 1994

Current status: Being refereed

Expected publication date: mid-1995

The SKICAT System for Processing and Analyzing Imaging Sky Surveys

Nicholas Weir[†]
Usama M. Fayyad[‡]
S. Djorgovski[†]
Joseph Roden[‡]

[†] Palomar Observatory
California Institute of Technology 105-24
Pasadena, CA 91125
weir@fritz.caltech.edu
george@deimos.caltech.edu

[‡] Jet Propulsion Laboratory
California Institute of Technology 525-3660
Pasadena, CA 91109
fayyad@aig.jpl.nasa.gov
roden@aig.jpl.nasa.gov

To be submitted to *Publications of the Astronomical Society of the Pacific*

November 16, 1994

Received:

Accepted:

Running Title: *The SKICAT System*

Abstract

We describe the design and implementation of a software system for producing, managing, and analyzing catalogs from the digital scans of the Second Palomar Observatory Sky Survey. The system (SKICAT) integrates new and existing packages for performing the full sequence of tasks from raw pixel processing, to object classification, to the matching of multiple, overlapping Schmidt plates and CCD calibration frames. We describe the relevant details of constructing SKICAT plate, CCD, matched, and object catalogs. Plate and CCD catalogs are generated from images, while the latter are derived from existing catalogs. A pair of programs complete the majority of plate and CCD processing in an automated, pipeline fashion, with the user required to execute a minimal number of pre- and post-processing procedures. Some of the most critical aspects of the image catalog construction process are the steps required for assuring consistent detection and attribute measurement across different plates, particularly measurements of magnitudes and attributes used for classification. We apply a modified version of FOCAS for the detection and photometry, and new software for matching catalogs on an object by object basis. SKICAT employs modern machine learning techniques, such as decision trees, to perform automatic star-galaxy-artifact classification with a $> 90\%$ accuracy down to $\sim 1^m$ above the plate detection limit. The system also provides a variety of tools for interactively querying and analyzing the resulting object catalogs.

keywords: image processing, database management, sky surveys

1 Introduction

The critical needs of observational astronomers have shifted from the exclusive realm of instrumentation to include that of advanced data analysis. The rate and quality of the data regularly produced by modern instruments frequently overwhelm the tools available to exploit them. Because of this mismatch, astronomers are forced to develop new methods and systems in order to make full use of modern astronomical data sets for producing meaningful scientific results timely and efficiently.

One such data set, large even by modern day standards, is the Second Palomar Observatory Sky Survey (POSS-II, Reid *et al.* 1991). When complete, this photographic northern-sky survey will cover 894 fields spaced 5° apart in three passbands: blue (IIIa-J + GG 395), red (IIIa-F + RG610), and near-infrared (IV-N + RG9). While the photographic survey is still under way, ST ScI and Caltech have begun a collaborative effort to digitize the complete set of plates (Djorgovski *et al.* 1992; Lasker *et al.* 1992; Reid and Djorgovski 1993). Both the photographic survey and the plate scanning are hoped to be $> 90\%$ complete by the end of 1997. The resulting data set, the Palomar-STScI Digital Sky Survey, will consist of ~ 3 TB of pixel data: ~ 1 GB/plate, with 1 arcsec pixels, 2 bytes/pixel, 20340^2 pixels/plate, for all survey fields in all three colors. In conjunction with the plate survey, we are also conducting an intensive program of CCD calibrations using the Palomar 60-inch telescope, using the Gunn-Thuan *gri* bands.

Given the enormous resources devoted to conducting such surveys, it is natural to pay special attention to how, using present day technology, one can make most effective use of the data once they are available. Attention to this detail, with an understanding of its increasingly general applicability, prompted the work described in this paper.

Caltech Astronomy and the JPL Artificial Intelligence Group have been engaged in a collaborative effort to integrate state-of-the-art computing methods for facilitating the scientific exploitation of POSS-II, applying the latest and most effective technology for performing any number of analyses of the data. The traditional means of extracting useful information from imaging surveys is through the construction of object catalogs. Thanks

to developments in the fields of pattern recognition and machine learning, it is now possible to reliably construct such catalogs objectively and automatically with a higher degree of accuracy than ever before.

2 Overall Design

The Sky Image Cataloging and Analysis System (SKICAT) was designed to facilitate the creation and use of catalogs from large, overlapping imaging surveys, and in particular, the scans of the Palomar-STScI Digital Sky Survey (DPOSS). The purposes of the software utilities comprising SKICAT generally fall into three main categories: catalog construction, catalog management, and catalog analysis. The relationship of these processes is illustrated in Figure 1. For reducing scans of POSS-II, the first step in SKICAT processing is catalog construction, which results in individual image catalogs. These, in turn, are registered within the SKICAT database management system and matched, object by object, with other catalogs to create a matched catalog of objects appearing in the survey. A matched catalog, or any individual image catalogs, may subsequently be queried in a variety of sophisticated ways to facilitate maintenance or analysis of the data.

While our interest in DPOSS provided the initial motivation for the development of SKICAT, these tools are quite general and applicable to a broad range of data reduction and analysis problems. For example, the catalog construction software could be rather easily adapted to processing large-scale CCD or infrared imaging surveys. Likewise, the catalog management and analysis tools are useful for integrating and making use of an even wider variety of data sources (*e.g.*, matching radio and x-ray sources with their counterparts from optical surveys).

Currently, SKICAT provides utilities for generating catalogs from two types of images, although it was designed to handle any number of types in the future. One image type consists of a plate scan from the Palomar-ST ScI Digitized POSS-II (DPOSS) survey. The other, a CCD image, is used for photometric calibration and training the star/galaxy classifiers applied to DPOSS catalogs. Step-by-step instructions for processing plates and CCDs from raw pixel into catalog form appear in the *SKICAT Plate and CCD Processing*

Cookbook (Weir *et al.* 1994a) and the *SKICAT User's Manual* (Weir *et al.* 1994b).

In this first section, we provide an overview of the steps involved in catalog construction, as well provide an introduction to the catalog management and analysis tasks supported by SKICAT. In the section which follows, we provide a more detailed discussion of the scientifically relevant details of the plate catalog construction processes. In the final section, we describe how matched and object catalogs are constructed within SKICAT.

2.1 Catalog construction

2.1.1 Processing plates

The heart of SKICAT is a collection of programs for the quasi-automatic processing of DPOSS plates from raw pixel to classified catalog form. Starting with a 1-GB digitized plate exabyte tape from ST ScI, SKICAT provides the tools for transferring the pixel data to SKICAT format, measuring the plate sky level and image boundaries, and determining a photographic density-to-intensity relation. The user then initiates a script, AutoPlate, which automates the process of cataloging the plate as a set of overlapping 2048^2 pixel image 'footprints'.

The three most critical elements of plate processing are detection, photometry, and classification. By using the Faint Object Classification and Analysis System (FOCAS, Jarvis and Tyson 1979; Valdes 1982a) for image detection and measurement, SKICAT is able to reach close to the faintest reliable limits of the plate scans, *i.e.*, down to a typical equivalent limiting B magnitude of $\sim 22^m$ for galaxies. In addition, by measuring quasi-asymptotic rather than isophotal magnitudes, using local sky estimates from annuli surrounding each object, and adapting the measurement thresholds within and across each plate to adjust for differences in sky level, noise, and pixel-to-pixel correlation, we are able to obtain very consistent photometry within and across plate boundaries. Details of our methods for performing photometry and the resulting accuracy appear in Weir, Djorgovski, and Fayyad (1994).

For classification, SKICAT benefits from the application of recent developments in machine learning. In particular, it utilizes the GID3* and O-Btree decision tree induction

software (Fayyad 1991; Fayyad and Irani 1992; Fayyad and Irani 1993), together with the Ruler system (Fayyad, Weir, and Djorgovski 1993) for combining multiple trees into a robust collection of classification rules. These algorithms work by using measurements of a training set of classified objects and inferring an efficient set of rules for accurately classifying each example. The rules are simply conjunctions of multiple "if...then.." clauses, which condition upon any of eight different object parameters to determine an object's classification. The real advancement in using this type of classifier relative to those used in most large-scale surveys to date is twofold: first, we are able to condition upon a larger and more diverse set of attributes; second, we allow the computer to decide what are the optimal number and form of the rules. Additionally, this technique readily generalizes to other, more difficult forms of classification, such as distinguishing galaxies by their morphology.

We have created separate sets of classification rules for objects from J and F band survey plates. We used CCD calibration data, which generally have superior image quality, to construct the training sets used to train the plate object classifiers. Classifications derived from the CCD data, more reliable than "by eye" estimates from the plates themselves, were matched to plate measurements to form the training sets. The measurements used to perform classification are a set of robust, renormalized object parameters that we found to be distributed in a stable fashion within and across plates. By training the algorithms to classify based on these attributes, we were able to nearly completely remove the effect of PSF variation across a given plate, or even between different plates. Average accuracy of star-galaxy classifications as a function of magnitude may be determined from tests using independent CCD-classified plate data. In both the J and F bands, we found the accuracy to drop below $\sim 90\%$ at about the same equivalent magnitude level, $B \sim 21.0^m$. This is $\sim 1^m$ above the plate detection limits, and nearly 1^m better than what was achieved in the past with similar data. This increase in depth effectively doubles the number of galaxies available for scientific analysis, relative to the previous automated Schmidt surveys. The details of our classification methods and results are presented in Weir, Fayyad, and Djorgovski (1994).

Plate X,Y to RA,Dec assignment, like object classification, is automatically performed in the final stages of catalog construction. Currently, the astrometric transformation is performed based on the astrometric solutions provided by ST ScI as part of their plate scanning operation, but improved solutions are easily implemented. As both astrometric assignment and final object classification rely only upon existing catalog measurements, not raw pixel data, they may be easily repeated at later times using a different set of classification rules or improved astrometric solution coefficients. SKICAT provides database manipulation tools that facilitate the continuous refinement of catalogs as better calibration, or even entirely new algorithms, become available.

2.1.2 Processing CCDs

CCD catalogs are constructed using most of the same tools as are applied to plate data. A script called AutoCCD, analogous to AutoPlate, is used to quasi-automatically process an image from pixel into catalog form. The primary differences between plates and CCDs are in the forms of pre- and post-processing that are applied. In particular, a whole host of standard CCD calibration procedures (*e.g.*, de-biasing, flat-fielding, photometric calibration, etc.), far different from those for plates, must be followed before running AutoCCD. In addition, we found FOCAS's built-in classifier to provide very accurate results on the CCDs down to the plate detection limit, which is our magnitude limit of interest. We were, therefore, able to let FOCAS automatically classify each object, with just a quick follow-up check by eye, producing excellent quality data without the need for much human interaction or more sophisticated classification algorithms.

CCD data are used for two purposes in our work with DPOSS. First, they provide "true" object classifications, at very faint levels, for our classifier training sets. Because the CCD images are of higher resolution and signal to noise ratio (SNR) than digitized plates, we are able to assign accurate classifications to objects whose morphology is not reliably distinguishable, even by an expert, when looking at the plate image alone. Through the machine learning process, the aim is to train the computer to consistently classify these faint objects, thereby enabling it not just to mimic a human's performance, but actually

improve upon it.

The second, most important, purpose for the CCD measurements is to provide photometric calibration for the plate catalogs. We use CCD exposures in the Gunn-Thuan (Thuan and Gunn 1976) g, r, and i bands to calibrate the IIIa-J, IIIa-F, and IV-N plate data, respectively. These CCD bandpasses provide a reasonable match to the photographic emulsion plus filter passbands. Details of how we perform our CCD photometry and the level of accuracy we achieve appear in the paper Weir, Djorgovski, and Fayyad (1994).

2.2 Catalog management

Once the image catalogs are constructed, they must be registered within the SKICAT database. Modifications and updated versions of the catalogs are maintained through database management software and tracked by the SKICAT system. The structure of the SKICAT database was specifically designed to facilitate the creation and classification of image catalogs, comparison of object photometry and classifications, revision of object measurements, and the construction of larger, matched catalogs.

For each plate or CCD image, the catalog construction scripts generate a header and features table, together comprising what we term a SKICAT catalog. A detailed description of the most commonly referenced SKICAT database terms appears in *Appendix A*. The header table consists of columns of parameters used to guide the catalog construction process, the name of the image from which the catalog was derived, the location of the image on offline storage, comments, and other information necessary to identify the data source and reconstruct the catalog from scratch if necessary. The features table contains one row for each detected feature in the image. The columns represent the measured attributes of each feature. Approximately 50 parameters per object are measured and saved in the individual plate and CCD catalogs.

After the construction process, catalogs within SKICAT must be registered in the SKICAT system tables, where a complete description and history of every catalog loaded to date is maintained. Catalog revisions, that might result from deriving new and improved plate astrometric solutions or photometric corrections, are also logged. Multiple versions of

each image catalog may exist, each reflecting a different processing history. The SKICAT system tables also keep track of which catalogs are currently loaded on-line, or physically loaded on disk. SKICAT provides tools for quickly and easily saving/loading catalogs off-line/on-line. Only registered catalogs may be moved to/from off-line storage or matched with other catalogs.

Multiple, overlapping catalogs can be matched into a special SKICAT data structure called the matched catalog. The matched catalog consists of a matched features table and a table of those catalogs comprising it. The matched features table contains independent entries for every measurement of every object detected in the constituent catalogs. Because of size and speed considerations, not every attribute may feasibly be saved within the matched catalog, but a sufficiently small subset of parameters is generally more than adequate for most uses of the data. Of course the saved catalogs themselves provide a complete archive of the full list of parameters if they are ever needed. SKICAT allows for multiple matched catalogs to be on-line at once, and they may be saved and loaded to/from off-line storage and a new one created at any time.

The matched catalog may be queried using a sophisticated filtering and output tool to generate a so-called object table, which contains just a single entry per matched object. With this tool, the user may, for example, generate a distributable data product, such as a galaxy list, from the current set of matched plate catalogs. The tool may also be used to perform consistency checks within catalog overlap regions, or to perform specialized scientific analysis over large survey regions. For example, a user may request a listing of all stars within a well-defined section of sky covered by multiple J and F plates, specifying exactly which object attributes to report (*e.g.*, magnitude, RA, Dec, etc.) and from which source (specific J plates, average of all F plates, etc.).

Catalogs may be easily altered using a procedure that allows arbitrary operations on table columns. This user simply specifies the C code which describes the computation for the column value as a function of any other column values, external data files, or constants. The utility automatically generates the necessary code for transforming the table and executes it. This utility is used in a number of contexts in the SKICAT system,

including the computation of right ascension and declination, as well as for applying the classification rules. In the same way, catalogs may be re-calibrated or otherwise adjusted in light of new or improved data. Such updates might include applying a field-effect correction to a plate's list of magnitude or performing new classifications using an improved rules set.

A catalog may also be modified by using a utility that updates selected columns from corresponding columns in the matched catalog. This procedure would be appropriate if, for example, the entries in a matched catalog were calibrated, and the calibrated measurements needed to be passed back to the original catalogs for archival purposes. An updated catalog could subsequently be re-registered as a new version of the existing catalog. Both the original and new header information would now be saved in the system, maintaining a complete history of catalog revisions. Via this mechanism, SKICAT is designed to maintain a "living," growing database, instead of a data archive fixed for all time.

2.3 Catalog analysis

The third layer of SKICAT, which is still under development, will consist of a powerful tool box of modern data analysis algorithms to be applied for survey data space exploration and the scientific analysis of the catalogs. It will facilitate more sophisticated scientific investigations of these expanding survey data sets, including a multivariate statistical analysis package, and a wide variety of Bayesian inference tools, objective classifiers, and other advanced data management and analysis packages and algorithms.

The analysis tools included in the current version of SKICAT are the GID3*/O-Btree decision tree induction software and Ruler program for classification learning, as well as the extremely useful collection of stream processing routines included in the standard FOCAS distribution. The very same classification learning software which was used to create the classifiers in SKICAT's plate cataloging script are available for use on any SKICAT data set, or even data from external sources. SKICAT provides an environment for implementing these tools to train and produce classifiers for scientific uses of the DPOSS, or any other catalogs.

We also intend to explore the potential of *machine-assisted discovery*, where modern,

artificial intelligence-based software tools automatically explore large parameter spaces of data and draw a scientist's attention to unusual or rare types of objects, or nonobvious clusters of objects in parameter space. We have begun applying the Autoclass (Cheeseman *et al.* 1988) unsupervised classification software to DPOSS, with plans to implement this and other Bayesian inference and cluster analysis tools within SKICAT in the future.

2.4 Application environment

The SKICAT system is largely written in C, Unix shell scripts, and FORTRAN, and it is portable across Unix systems. As mentioned before, SKICAT is built around and incorporates a number of preexistent software packages: FOCAS routines for image detection and measurement; the GID3*/O-Btree/Ruler induction software for object classification; and the Sybase commercial relational database management system (DBMS) for maintaining and accessing the data. While SKICAT was developed using these packages, none are irreplaceable. Each package serves its purpose and, because of the modularity of the system, could be substituted for another which performs the same function. In addition, SKICAT provides quick and easy access to most system utilities through a common X-Windows graphical user interface, while users familiar with Unix can access the same utilities directly from the Unix command line.

SKICAT was designed so that all database system operations specific to Sybase would be transparent to the user. The user interfaces and underlying Unix utilities have been designed to allow the user to select and specify subsets of catalogs using a slightly expanded version of the industry standard SQL (Standard Query Language). This extended query language provides additional features of specific interest to users in astronomy. For example, unit conversion capabilities have been provided to allow the user to specify positional values in a variety of astronomical units (*e.g.*, hours, minutes, and seconds in addition to degrees and radians). Most database operations controlled through the SKICAT software are implemented using SQL, so that it would be relatively easy to replace the underlying DBMS if the need arose.

3 Constructing Plate Catalogs

In this section, we provide more detail on the steps involved in constructing a catalog from a DPOSS scan. Additional details may be found in Appendix B. Aside from the initial pre-processing steps, the process of cataloging a CCD image is very analogous to that for a plate. We provide the details of these operations in Appendix C.

3.1 Pre-processing

Once a POSS-II plate has been scanned by ST ScI, only a few manual steps are required before it may be pipeline processed using a Unix command-line-based program called AutoPlate, or the X-windows-based graphical user interface to it. A digitized POSS-II plate scan is provided in the form of pixel data consisting of arbitrarily scaled photographic densities. Each DPOSS plate image is 23,040 $\times$ 23,040 in size. After defining the plate boundaries, and the sky and saturation densities, the first step in processing the plate is to perform the photographic density to arbitrary intensity conversion. A SKICAT program automatically retrieves the portion in the southwest corner of each image that contains the 16 sensitometry spots that appear in each POSS-II plate. This program assists the user in running an IRAF script to measure the 16 spots and compile a list of the densities. It then prompts the user to interactively fit an 'HD' curve to the data points, providing a density to intensity transformation for the plate scan.

The mathematical formula we use to fit the measured plate densities (D) to relative intensities (I) is:

$$\log I = \frac{P(D)}{(D_S - D) \times (D_T - D)} \tag{1}$$

where $P(D)$ is a polynomial function of the density, and the saturation and toe densities, D_S and D_T, are those corresponding to fully exposed and unexposed portions of the plate, respectively. The polynomial coefficients, together with the toe and saturation values, establish the conversion applied to each pixel value whenever image blocks are subsequently loaded and mosaiced to form larger images. As the average sky density is generally far above the toe level, it is usually desirable to avoid fitting the polynomial to the lowest few intensities, thereby improving the fit in the other portions of the curve. Similarly, the most

nearly saturated point or two is also generally ignored. After several iterations adjusting the relevant parameters, we have found it possible to reduce the residual between the fit and all accepted data points to less than 5% in intensity.

There is a long history to efficiently modeling the HD curve. The method employed by ST ScI (Russel *et al.* 1990), for example, involves a more complicated formula and averaging many plates together. By their own admission, however, they find the more complicated expression to be overkill for the linear part of the curve of most interest. In addition, we found considerable variation of the curve among different plates, requiring independent fits. As described in Weir, Djorgovski, and Fayyad (1994), we also find the instrumental magnitudes resulting from these fits to be extremely consistent from plate to plate, in the sense of only requiring a single zero point offset to match them. This provides, in our opinion, the most important test of the validity of our linearization scheme.

3.2 AutoPlate processing

AutoPlate is a C-Shell script which executes a suite of other scripts, C code, and Fortran programs to conduct the pipeline processing of plate scans from their raw pixel form to SKICAT catalogs. The steps involved include everything from loading the pixel data from exabyte tape, to image detection and measurement, to catalog construction and quality control. The majority of image processing functions are accomplished using FOCAS routines, while Sybase is used for database management.

3.2.1 Overlapping footprints

Each plate is analyzed as a set of 13 $\times$ 13 overlapping 'footprint' images. After pre-processing, a plate scan exists on exabyte tape as 23 Vax VMS savesets of 23 images each (see Figure 2). These image blocks are pasted together to form image footprints, which form an overlapping grid covering the entire plate (see Figure 4). Each footprint is 2048^2 in size, with a minimum overlap between adjacent footprints of 272 pixels, or ~ 4.5 arcmin. The large overlap allows all but the largest objects to be reliably measured in this piecemeal fashion, while providing a quality control check and statistics on footprint

dependent measurement errors. In fact, analysis of these errors indicate that the systematic errors induced by processing the scan in this fashion are at least an order of magnitude below random image measurement errors.

A number of distinct levels of processing are applied to each footprint, leading to the construction of individual footprint features tables. Footprints are identified by a row number within the plate and by a column number within that row. They are created and processed a row at a time, from bottom (south) to top. Up to nine image blocks must be mosaiced together to form a single footprint image; up to three rows of image blocks must be loaded on disk to form an entire row of footprint images. As each footprint row is processed, AutoPlate loads the necessary image blocks from tape and deletes unnecessary blocks from disk.

Consecutive footprint images, from left (east) to right, are created just prior to their processing. Up to two rows of footprints are always on disk, facilitating the detection of vertical mismatches between footprint tables. Each row of footprint features tables is saved to the plate features table only after passing a number of quality control checks meant to assure uniformity of catalog construction. This process is described in greater detail in the Quality Control section below.

3.2.2 Image analysis

Footprint images are analyzed in a few ways prior to object detection. First, a quality control check is performed by measuring correlations between alternating pairs of pixel rows in the plate scan. This check was developed in response to problems detected in the first batch of ST ScI scans. These correlations resulted from the scanning machine not taking equal size vertical steps before raster scanning from the right or left side of the plate. The problem seems to have been corrected, and all previously corrupted plates were re-scanned. Nonetheless, we still perform the check as a part of our production system.

Next, AutoPlate creates a re-binned version of the image with one pixel per 8×8 in the original. This scale matches that of the 'sky' image produced by the FOCAS detection algorithm. To provide the FOCAS algorithm with a good first guess of the footprint sky, the

value is initially estimated by binning the image into blocks of 64^2 pixels each, accumulating the median and quartile sigma[1] for each block, then accumulating the median and quartile sigma for all of the block measurements. Images of the sky median and sky sigma are saved at this reduced (one pixel per 64×64) scale. This robust estimation procedure provides relatively accurate initial sky and sky sigma values, even when relatively large and bright sources exist in the image. Seeded with these values, the FOCAS detection and background estimation procedures have been found to work well. We were able to test the accuracy of this approach by applying it to the simulated plate images we describe in Weir, Djorgovski, and Fayyad (1994), which were also used to help optimize the choice of detection and measurement parameters.

AutoPlate also estimates the pixel-to-pixel correlation (horizontal and vertical combined) within each footprint. For this measurement, in addition to applying the same binning and median filtering procedure as above, AutoPlate excludes all pixels two and a half sigma above the sky level. This technique was found to provide an extremely robust and accurate measurement for all levels of pixel blurring, even when large saturated objects appear in the image.

3.2.3 Object detection

The basic processes of object detection and measurement are accomplished using only slightly modified versions of the standard FOCAS routines (Jarvis and Tyson 1979; Valdes 1982a). Algorithmic details of these programs may be found in the FOCAS documentation (Valdes 1982b). Here we describe how we apply these functions and what are the relevant parameter settings.

Just prior to object detection, a FOCAS catalog is automatically initialized for the current footprint. The appropriate header values are determined in AutoPlate based upon the current footprint row and column numbers, and from information derived from the

[1] We define a quartile sigma as 0.7415 times the difference between the 75th and 25th percentile values, a robust estimate of the sample standard deviation that is insensitive to outliers. For a Gaussian distribution, this is virtually identical (in the limit of large sample sizes) to the standard deviation defined in the normal way.

plate image header. The FOCAS 'detect' command then uses the header parameters for driving its object detection and sky estimation procedure. Details of the detection process appear below. The result of this command is a catalog of features, or contiguous pixels a certain threshold above the background, and meeting a minimum area and signal to noise ratio (SNR) requirement. The FOCAS detect command also produces an estimate of the sky with a one pixel per 8×8 resolution. If this estimate significantly differs from the median sky image computed previously, an error is reported and processing ceases.

For optimal sensitivity, the FOCAS detection algorithm applies a threshold equal to some number of estimated standard deviations (sky sigma) above the locally estimated sky. The assumed sky sigma is the robust value computed for the footprint, as described in the Image Analysis section above. However, because of spatially varying pixel-to-pixel correlation within each plate scan, using the same multiple of sky sigma as the threshold for all footprints would not result in the same detection sensitivity.

To compensate for this effect and approach a common level of sensitivity between and within plates, we sought to derive a factor by which to scale the measured sky sigma so as to make it correspond to approximately one standard deviation in an *unblurred* version of each footprint. To establish this scaling factor as a function of measured blur, we created a simulated footprint image matching the average noise[2] and object number statistics of real footprints, then we convolved it with a series of Gaussians of different width. Given the convolution kernel, the appropriate scale factor is simply the square root of the inverse of the sum of squares of the normalized kernel elements. By measuring the pixel-to-pixel R^2 for each image, we are able to empirically derive a mapping from measured (square) correlation to scale factor. We found a sixth order polynomial to provide a good fit to the relation (see Figure 5). We also established the relation using a blank simulated sky image and derived virtually identical results, lending confidence in the robustness and accuracy of our correlation estimation procedure.

[2]The appropriate level of uncorrelated, Gaussian random noise was determined in an iterative fashion. First, we found a Gaussian kernel which, when convolved with the image, produced a degree of blur, as measured by the pixel-to-pixel correlation, closely approximating that of an average footprint. We then found that noise amplitude which, after convolution, resulted in a measured sky sigma closely matching that of an average footprint.

We then used 2.5 times this scale factor times the estimated sky sigma as our detection threshold. The additional detection parameters required by FOCAS include a minimum object area, "significance limit" for object detection, and pre-detection blurring kernel. We require every object to comprise six contiguous pixels. We set the significance limit to -100, which is equivalent to turning off this SNR requirement (see the FOCAS manual for details). We used the built-in FOCAS blurring function, which is simply:

1	2	3	2	1
2	3	4	3	2
3	4	5	4	3
2	3	4	3	2
1	2	3	2	1

The FOCAS detection algorithm works by convolving the image with this kernel, then searching for contiguous pixels with values greater than the locally estimated sky by the specified detection threshold. To adjust for the convolution, which is meant to improve the sensitivity of the detection algorithm, the detection threshold is scaled by the square root of the inverse of the sum of squares of the normalized kernel elements. Note this is the same blurring correction we applied earlier to account for the correlation induced by the scanning process.

Our choice of detection parameters, in particular our scaling correction for pixel-to-pixel correlations, results in relatively consistent sensitivity as a function of plate quality, as evidenced by the relative uniformity of object density we detect from footprint to footprint and plate to plate. Our choice of threshold, minimum area, significance limit, and pre-detection blurring were chosen after extensive tests on both real and simulated images, establishing some feel for the trade-off between completion (percentage of real objects detected) and contamination (percent of detected objects which are not real). On simulated images, this combination of parameters resulted in an average FOCAS detection isophote corresponding to roughly 2.0 *uncorrelated* sky sigma, which is sufficiently far into the noise as to pick up every object readily detectable by eye. It also resulted in what we considered a manageable number of detections per footprint and plate, in excess of the density saved in previous Schmidt plate surveys. Typical galaxy detection limits for the J and F DPOSS

plates are found to be 21.0^m to 21.5^m in g and 20.1^m to 20.6^m in r, respectively. For point sources, the limit can extend up to half a magnitude fainter.

3.2.4 Object measurement

The local sky brightness for each feature is measured using the FOCAS 'sky' command. It measures the median pixel value in an annular region surrounding each feature, avoiding pixels that are within the detection isophote of another feature. The accuracy and systematic effects of this sky measuring algorithm are addressed in Weir, Djorgovski, and Fayyad (1994), where we discuss details of our photometry.

After obtaining the sky estimate, additional attributes for each feature are measured using the FOCAS 'evaluate' routine. The total number of measurements number more than 30, including those in Table D: The indicated magnitudes are instrumental and are computed according to:

$$m = 30.0 - 2.5 \log L$$

where L is the luminosity, or sky-subtracted integrated intensity. The offset of 30.0 is arbitrary and was chosen to make the instrumental magnitudes approximate the final calibrated values within a magnitude or two. The aperture magnitudes are computed using a five arcsec radius. The 'total' magnitude and area are computed by 'growing' the detection isophote out a pixel at a time in all directions until the total area is at least twice the original. This magnitude is meant to provide a flux measurement less biased with respect to surface brightness profile, approximating something like an asymptotic or true total magnitude. The cost for decreased systematic error is greater sensitivity to sky subtraction, integration over more noisy pixels, and hence, increased random error (relative to isophotal or aperture magnitudes). A substantial portion of the paper Weir, Djorgovski, and Fayyad (1994) is dedicated to an analysis of the photometry obtained from DPOSS using SKICAT, including the results of detailed simulation studies.

FOCAS also sets a number of flags for each feature, each of which is saved as an attribute. These flags indicate such things as whether the object touches the edge of the footprint, the object is below the sky level in integrated intensity, the object's size exceeds

current FOCAS limits, there are saturated pixels in the object, or the object was not split at any level by the FOCAS deblending routine. Additional useful attributes are obtained by taking non-linear combinations of some of those listed in Table D. For example, using the intensity weighted second moments, we can calculate the ellipticity and position angle of each feature. Additional attributes, the so-called 'revised' ones described below, are defined by the position of a feature within the statistical distribution of that footprint's features within some measured parameter space (*e.g.*, within the plane defined by the first radial moment and the total magnitude).

3.2.5 Object deblending

After each feature in a footprint has been evaluated, SKICAT next applies the FOCAS 'splits' command. Effectively, this routine runs the detection algorithm on every existing feature, but using successively higher thresholds. 'Islands' detected at a given threshold are entered into the catalog as distinct features, and all attributes are remeasured for them. The 'parent's' flux is divided between the 'children' according to the ratio of isophotal fluxes obtained using the higher threshold. This process continues recursively until no more islands are detected.

All parents and intermediate children (*i.e.*, a feature's full family tree) are saved within the FOCAS catalog and likewise within SKICAT. Each feature is referenced by an entry and subentry number. A parent and all of its children share the same entry number. Children are distinguished by the hierarchically constructed subentry number: subsequent generations append additional digits to the end. The leaf or leaves in a feature's family tree correspond to indivisible objects and are marked as such by a flag attribute.

We note that improvements can certainly be made to the deblending process. For example, other methods could be used to improve the quality of the photometry of the deblended objects, better take deblending into account when matching overlapping images, handle the extreme crowding conditions to be found in lower Galactic latitude POSS-II plates, etc. Nonetheless, we find the present implementation to be more than sufficient even for detailed analyses of higher latitude plates, and that it at least represents a step

above reduction without the use of deblending at all, as in the case of some previous surveys (*e.g.*, APM, Maddox *et al.* 1990).

3.2.6 Classification related measurements

An additional set of attributes are measured solely for the purpose of facilitating feature classification. Four revised attributes are determined by automatically estimating and subtracting the 'stellar locus' from the parameters M_{core}, the magnitude of the brightest 3 $\times$ 3 pixel region, of total intensity L_{core}; the log of the isophotal area, log A; the intensity weighted first moment radius, r_1; and S, where

$$S = \frac{A}{\log[L_{core}/(9 \times I)]}.$$

and I is the average intensity of the detection isophote. The stellar locus is the attribute value as a function of magnitude around which point sources are fairly narrowly distributed, at least at brighter magnitudes. As described in Weir, Fayyad, and Djorgovski (1994), we have found that the resulting revised attributes are relatively insensitive to footprint-to-footprint, and even plate-to-plate, variations, and are thus robust parameters for use in feature classification.

In order to derive even more powerful classification attributes, we form an empirical estimate of the PSF for each footprint. Along with magnitude and ellipticity, the four revised attributes are fed as input to a decision tree classifier, which culls out a list of 'sure-thing' stars. This represents a significant application of machine learning technology to the classification task. A FOCAS routine then adds images of these stars to form a two-dimensional PSF template.

Using the PSF template, the FOCAS 'resolution' routine determines the best-fitting 'scale' (α) and 'fraction' (β) values, which parameterize the fit of a blurred (or sharpened) version of the PSF to each feature (Valdes 1982a). The template used to model each feature is of the form:

$$t(r_i) = \beta s(r_i/\alpha) + (1 - \beta)s(r_i)$$

where r_i is the position of pixel i, α is the broadening (sharpening) parameter, and β

is the fraction of broadened PSF. This template-based approach is the core of FOCAS's Bayesian classification method. Objects are classified as stars, galaxies, artifacts, etc., according to their maximum likelihood (best-fitting) location within two-dimensional scale and fraction space. Extensive tests performed by Valdes (1982a) indicate that one can achieve significantly higher accuracy in star/galaxy separation with this template-fitting approach versus simpler approaches employed previously. Weir and Picard (1991) explicitly tested the use of these two techniques on digitized Schmidt plate data and confirmed this result.

In the present version of SKICAT, we combine these resolution parameters along with total magnitude, ellipticity, and the four revised attributes described above to form an even higher dimensional space in which to perform feature classification. Actual classification is run as a post-processing procedure, using the measured attributes within the plate catalog. One can thereby alter the existing, or create an entirely new, classifier and apply it to a catalog at any future date. The classifier currently applied to plate features within SKICAT was generated using the GID3*/O-Btree and Ruler decision tree induction programs. A full description of how it was created and the results we have achieved on actual plate data appears in Weir, Fayyad, and Djorgovski (1994). The net effect is that by employing this new technology, we are able to go about a magnitude deeper in achieving accurate object classifications, resulting in approximately three times larger classified object catalogs than in previous surveys using comparable data.

3.2.7 Quality control tests

Each individual footprint FOCAS catalog, and its corresponding revised attribute list, is joined into a Sybase table for subsequent processing. As a quality control check, the current footprint features table is matched with the tables of the footprints to its left and bottom, if they exist. If any major discrepancies are detected in the mean or standard deviation of measurements in the overlap, processing is halted and an error reported. Otherwise, AutoPlate appends these results to a summary file characterizing the footprint row.

After a row is complete, Autoplate searches the footprint summary file for outliers

and trends, halting the program if it encounters any problems. If none are found, the previous row of footprints is added to the Sybase plate catalog and any auxiliary files are saved. First, the row's footprint summary file is appended to the corresponding file for the plate. Next, each footprint's compressed original, sky, median sky, and sky sigma images are pasted into corresponding composite images for the entire plate. Footprint specific parameters are appended to a footprints file. All features with central coordinates in a nonredundant portion of the plate image are added to the plate features table, while features whose outer isophotes extend beyond any single footprint's boundaries are saved to a border objects list. Generally these are features which appear at the edge of the plate. In addition, AutoPlate appends to a list of footprint overlap statistics, and summary thereof. Data for the previous row are deleted after each of these operations is complete.

After all rows have been processed, the system checks the footprint summary file for outliers and trends among footprint statistics in the vertical direction. Provided none are found, catalog generation is complete, a plate catalog header is created (if it was not already) and all remaining footprints and image blocks are deleted.

3.2.8 Data products

The final products of an AutoPlate run are a SKICAT catalog, consisting of a Sybase format features table and header table, and several auxiliary files. The plate catalog resides on the Sybase disk partition while the auxiliary files are saved within a Unix directory hierarchy created specifically for that plate. The auxiliary files include the following images: a re-binned version of the plate scan containing the average of every 8×8 pixels in the original; the 'sky' image produced by the FOCAS detection algorithm at the same scale; images of the median and quartile sigma of the plate scan at a one pixel per 64×64 scale. Besides providing an overall reality check of the AutoPlate process, these images may be valuable for future scientific programs, such as searches for low surface brightness galaxies.

In addition, SKICAT saves each of the FOCAS 'areas' files produced for each footprint. These files contain a run-length encoding of all the pixels comprising every feature in each image. This information may prove useful in the future for locating the precise extent of

a feature when all of the imagery data, in addition to catalogs, are readily available online for querying and analysis.

The other auxiliary files produced by AutoPlate are those produced and used for quality control purposes. They include a footprint statistics file, containing lists of statistics measured for each footprint (*e.g.*, number of features detected, average sky level, etc.) which are used to detect trends and outliers among the footprints along any given plate row or column. The other quality control file contains lists of all of the overlap statistics measured between adjoining footprints.

3.3 Post-processing

After a plate catalog has been created by AutoPlate, there are still a few operations which must be performed as a part of the plate's standard pipeline processing. These include the assignment of Right Ascension and Declination (RA,Dec) to each object, as well as classification. As neither of these operations require access to the pixel data themselves, one is able to re-run either of these multiple times in the future using new and better coefficients or algorithms.

3.3.1 Astrometric transformation

The J2000 RA and Dec of the central pixel (specified in plate standard coordinates by the **XC** and **YC** attributes) of each feature is calculated using coefficients in the plate catalog header. These coefficients are initially provided by ST ScI and are supposed to be good to $\sim$ 0.5 arcsec RMS accuracy over scales less than about a square degree. When in the future better plate solution coefficients are available, it is simply a matter of entering them in the catalog header, then re-executing a catalog modifying procedure to assign a new RA and Dec to each feature.

3.3.2 Classification

The plate features classifier provided with SKICAT was generated using the GID3*/O-Btree and Ruler programs, and is implemented as a procedure executed by a more general

utility for modifying columns within a database table. By applying a set of rules that condition upon a subset of the parameters in a plate features table, the procedure provides a classification to each object. An entry within a plate's header table specifies the classifier rules file to use. Therefore, it is simply a matter of editing this field and re-running the appropriate column modifying procedure to apply a new and improved rules-based classifier to the catalog. Similarly, an entirely different plate classification algorithm could be designed in the future and implemented as an alternative column modifying procedure.

3.3.3 Bright object editing

Currently, the SKICAT user is required to hand create a list of the 'bad regions' within the plate, such as areas corrupted by bright stars. The *SKICAT Plate and CCD Processing Cookbook* provides a description of how to create such a list using the SAOImage display program. One detects the bad regions by analyzing the 8×8 binned average of the full scan image produced by AutoPlate. By displaying this image, the user may easily pick out and mark the 100 or so brightest objects in the scan which will have been poorly processed by AutoPlate. It is particularly important to mark the regions surrounding bright stars, as their halos and spikes are split into sometimes hundreds of small artifacts which may be mistaken for real objects in the catalog (*e.g.*, see Figure 7).

At this time, the bad regions list is not used to filter or flag entries in the SKICAT plate catalog itself, but rather for subsequent filtering of ASCII data files generated by queries of the plate or matched catalog. Details of how this filtering is performed are in the Queries section of the *SKICAT Plate and CCD Processing Cookbook*. We also note that the entire process of bright object detection will also be automated in the near future.

3.3.4 Catalog registration

Once all of the aforementioned processes are complete, the plate catalog is ready for registration into the SKICAT catalog management system. This loads the catalog header information into the SKICAT System Tables, allowing it to be matched with other catalogs or saved to/loaded from tape. At this time, the plate catalog, along with the auxiliary

files, are generally saved on an archive tape, and plate processing is complete.

4 Constructing Matched and Object Catalogs

4.1 Matched catalogs

SKICAT provides the ability to match features from multiple plate and CCD catalogs based on the similarity of their measured positions in celestial (RA,Dec) coordinates. This procedure is essential for analyzing objects measured in multiple bandpasses, such as finding optical IDs of non-optical sources; constructing object lists spanning multiple overlapping images; and for performing consistency checks of object measurements and classifications. Details of the data structures pertaining to the matched catalog appear in Appendix D.

The process of adding a catalog to the matched catalog involves matching each feature in the catalog to the nearest object meeting certain criteria within the matched catalog, after solving for a small systematic X,Y offset between the two. To perform this matching, the filtered source catalog is broken down into a user-specified number of solid angle 'segments'. A best fit transformation in X and Y is solved for using a robust fitting algorithm and applied to each segment when it is matched. To optimize this process, the catalog should be split into as many segments as necessary to allow for systematic deviations in its astrometric accuracy.

For each segment, the matcher attempts to minimize the overall match error (defined as the average matched feature difference) separately in X and Y by repeating the matching process until the errors meet specified criteria. For each feature in a segment, the matcher attempts to find the closest feature within some search radius within the matched catalog, offsetting by the previous iteration's match error in X and Y. These errors are accumulated over each iteration to form a mean offset. The initial search radius is given by the user; subsequently it is determined as some multiple of the measure standard deviation in the previous iteration's offsets. These average offsets and the standard deviations are computed only for a quartile-sigma clipped fraction of the matches from the previous iteration, in order to exclude outliers from the estimate. This matching and estimation procedure repeats until the iteration's match error in both X and Y is less than some multiple of the estimated error in the mean offset. The matcher then performs a final pruning of the

matched object list, passing only those matches with a residual Chi-squared error less than some threshold.

The matcher then assigns each feature an identification number according to the match results. Features with no corresponding object in the matched catalog are assigned the default next ID, which is then incremented. For each feature from the segment, a row including a user-specified subset of attribute columns is appended to the matched catalog's features table. The match and converge process is repeated for each segment of the catalog. After each segment has been matched, information about the input catalog is added to system files detailing the contents of the matched catalog.

4.2 Object catalogs

While the matched catalog is the most comprehensive form of database produced by SKICAT, it is generally too unwieldy for direct use in large scale survey analysis. By allowing a virtually unlimited number of independent feature entries per object, very little data reduction actually takes place in the matching process. Although in practice, one generally limits the number of attributes saved in the matched catalog, this still leaves unsolved the problem of combining the multiple measurements that are usually present for any given attribute and feature.

To provide the user with power and flexibility in accessing the matched catalog for scientific analysis and calibration, we developed a sophisticated database querying mechanism. This program summarizes data from the matched catalog to form an object catalog, which by our definition contains just one entry per object. The query program has two primary inputs: a filter and an output specification file. The filter basically defines the conditions that an object, or its constituent features, must meet in order to be passed on for output. A full description of the filter language appears in the *SKICAT Users Manual* and specific useful examples appear in the Query section of the *SKICAT Plate and CCD Processing Cookbook*. These filter conditions might include a requirement on the number of features measured per object, that an object be measured in a particular catalog, that an object not be measured in a particular passband, that an object's magnitude falls within a

certain range, etc. The most important filter specification is of an allowable RA and Dec range, as the matched catalog is sorted on those fields. All queries to the matched catalog should specify the most restrictive RA and Dec limits possible, for most efficient retrieval of the data.

The output specification file defines which attribute columns to pass on from the query and how to combine multiple measurements into one. For example, the following output specification would produce a table containing the following five columns: the object ID, RA, and Dec from plate J442, and calibrated J and g magnitudes derived from a combination of all feature measurements for each object:

```
ObjectId/j442 %d
RA/j442 %d
Dec/j442 %d
Mag/C/J %d
Mag/A/g %d
```

To the right of the column/source specifiers are format codes, indicating how to print the column value if the output is directed to a text file. For this output specification to result in a valid query, the filter must have restricted its output to those objects detected in plate J442 for which there is at least one g (CCD) measurement, since we are requesting output from both these sources. The specification `Mag/A/g` refers to the average (`A`) of all calibrated g magnitudes measured for that object. The preceding specification asks for the object's J magnitude not necessarily from plate J442, but from that particular feature that was measured closest to the center (`C`) of its source catalog (and, therefore, presumably the least susceptible to field effects).

Using the query program, the user can combine the data in the matched catalog in most ways needed for subsequent scientific use. To facilitate the construction of the filter and output specification files, we created an X-windows interface to the program (see Figure 10). Using either program, the user has the option of producing another Sybase table or an ASCII text file. The former is of use if the user might wish to perform subsequent

queries of the resulting table using any of the available Sybase database management tools. A Sybase table is also the most appropriate form for a catalog one might wish to make available on-line, through the Astrophysical Data System, for example. An ASCII file, on the other hand, though inefficient, is an almost universally accepted format for general purpose or homemade analysis programs.

We also developed a similar query mechanism and graphical user interface for filtering and outputting portions of any Sybase table, such as a plate or CCD features table, or even an object table produced by the query mechanism. Using these programs, one can perform all of the same basic filtering and output operations, but without the functionality related to handling multiple entries per object. Again, the resulting tables may be produced in either Sybase or ASCII format.

After the successive application of the tools described in this chapter, from creating individual plate and CCD catalogs, to matched catalog construction, to the generation of user-specified object catalogs, the user will have reduced the raw pixel data into a form suitable for systematic study. Following the next chapter, in which we describe our classification methods in more detail, we will present results derived from the application of these SKICAT programs to actual DPOSS data.

This work was supported at Caltech in part by NASA AISRP contract NAS5-31348, the Caltech President's fund, and NSF PYI award AST-9157412, and at JPL under a contract with NASA. The POSS-II is partially funded by grants to Caltech from the Eastman Kodak Co., the National Geographic Society, the Samuel Oschin Foundation, NSF grants AST 84-08225 and AST 87-19465, and NASA grants NGL 05002140 and NAGW 1710. We acknowledge the efforts of the POSS-II team at Palomar, the scanning team at STScI, and the rest of the SKICAT team at JPL.

A Appendix - Database Definitions

Below is a description of the most commonly used database terms within the SKICAT system:

A **feature** is the set of measurements (magnitude, surface brightness, position angle, etc.) of a unique object contained in a catalog. For example, a star may be a feature within a catalog, as might be a galaxy or a satellite trail detected on a plate.

A **table** is a collection of data organized by row and column, where each row has a value (or space for a value) for every column in the table. For example, a list of galaxies may be organized in the form of a table, with one row per galaxy (feature) and one column per galaxy measurement. SKICAT tables are stored and manipulated using Sybase. Therefore, all references to tables refer specifically to the Sybase data structures of the same name.

A **catalog** consists of a features table and a header table. These are data sets produced by Autoplate and AutoCCD. A features table contains one row for each feature appearing in the catalog. The header table contains information relevant to the entire catalog (image source, date of creation, etc.) and is generally used for reference purposes.

An **object** is a unique image artifact or physical sky object (*i.e.*, star, galaxy, etc.) to which there may correspond multiple features within distinct catalogs. For example, the *object* M87, which lies in the overlap of two plates, would appear as a *feature* within both plates' catalogs.

A **matched features table** contains features from multiple, matched catalogs. Features at the same RA and Dec position (within astrometric uncertainty) are considered to be different measurements or features of the same object. They are assigned a common object ID during the matching process.

A **matched catalog** consists of a matched features table and a table listing those catalogs comprising it. New catalogs are added to it by matching each new feature with existing matched features (objects). The user controls which subset of measurements to include in the matched features table and also specifies parameters affecting the matching algorithm. In a reverse operation, selected columns within catalog features tables may be

updated from their corresponding entries in the matched features table.

Objects tables are produced by filtering and outputting selected columns of object entries from any individual catalog or the matched catalog. They might be generated for catalog calibration, specialized scientific analysis, or as distributed data products (such as the PNSC). These tables may also be queried and manipulated using the SKICAT table manipulation tools.

B Appendix - Plate Processing Details

B.1 Digitized POSS-II Scan Data Format

The plate pixel data, consisting of arbitrarily scaled photographic densities, are provided by ST ScI as a single file, two bytes per pixel, on a single VMS backup saveset on exabyte tape. For processing by SKICAT, the single pixel file is transferred to another exabyte as 23 VMS backup savesets, each containing 23 image 'blocks'. The scanned image is broken into these more manageable image blocks, of at most 1024×1024 pixels, to facilitate retrieval and processing.

The following additional files produced by ST ScI are also necessary for processing a plate:

scan_name.gsh	- Plate scan header file
snap_name.hhh	- Snapshot image header file
snap_name.hhd	- Snapshot image pixels

The scan header contains parameters, such as the plate name, band, and astrometric solution coefficients, which are eventually loaded into the plate catalog header. The 'snapshot' image is a sparsely sampled (one pixel per $\sim 33 \times 33$) version of the plate scan, useful not only as a reality check, but for determining the usable portion of the scan image. Figure 3 depicts such a snapshot.

One must analyze the snapshot image to determine the plate sky and saturation densities and the image boundaries. These parameters are listed in Table 1 The pixel positions in the snap image should be multiplied by 32.914 to match the plate dimensions. The pixel values must be multiplied by 1.5259×10^{-4} to convert to properly scaled densities.

B.2 Running AutoPlate

AutoPlate is designed to automatically perform all levels of processing for the footprints in all columns of all rows of a plate. However, if it becomes necessary to restart the script at a particular stage of plate processing (due to, for example, a prior system failure), control parameters supplied at run time can force it to begin at a specified level of the processing of the footprint at a specified column of a specified row. Any subsection of a plate may be processed or reprocessed with the same facility.

The AutoPlate script may be invoked either directly from the C shell prompt or from within the **xautoplate** graphical user interface described in the *SKICAT Users Manual* (see Figure 6). The parameters that control AutoPlate are specified in a file, the name of which must be supplied as the sole command-line parameter when AutoPlate is initiated. The parameter specification file details the data to be processed, the initial processing level, and the footprint row and column at which to begin and end processing. A detailed description of the parameters in this file is described in an appendix within the *SKICAT Users Manual.* It is automatically produced by a separate initialization program that is run prior to plate processing.

In addition to the parameters file, the only additional inputs required by AutoPlate (and referenced in the parameters file) are the file containing the plate density-to-intensity transform coefficients and the plate header file provided by ST ScI. Assuming all the necessary image blocks do not already reside on disk, the exabyte containing the raw pixel data must also be loaded on the appropriate tape device.

C Appendix - Processing CCD Images

C.1 Pre-processing

The construction of CCD catalogs is similar to the process of constructing plate catalogs, although simpler. As with plate data, there are a number of preliminary steps before an image is ready for processing. In particular, the CCD image should be reduced (*i.e.*, debiased, flat-fielded, calibrated, etc.) according to standard astronomical procedures. Methods and specific software for performing these tasks on DPOSS calibration sequences obtained using the Palomar 60 inch telescope are described in the *SKICAT Plate and CCD Processing Cookbook*.

After these standard CCD reduction tasks are performed, the image is nearly ready to be run through the catalog processing script. The user must first run an initialization script in order to create and load a parameters file containing header and control information for subsequent processing. To the extent it is possible, this program loads the necessary values from the image header itself. Otherwise, the user must enter the values, such as image center RA/Dec, descriptive name, date of observation, and photometric calibration coefficients manually.

C.2 AutoCCD processing

Like its sister AutoPlate, the AutoCCD script takes a parameter specification file as its sole argument and, in turn, calls a collection of programs, primarily from FOCAS, to construct a SKICAT catalog from the indicated CCD image. All of the same sky and object attributes measured for plate images are measured for CCDs, using the same routines. Unlike AutoPlate, there is not a corresponding X-Windows interface.

After initial object detection, measurement, and splitting, the script attempts to automatically generate a list of stars with which to form the empirical PSF estimate. It tries to do so by first looking for the stellar locus in a plot of intensity weighted first moment radius (the FOCAS **IR1** parameter) versus magnitude. After estimating the stellar **IR1** parameter, AutoCCD uses it to create a filtered catalog of candidate stars. It then feeds this catalog to a FOCAS script which iteratively prunes the list until some maximum level

of dispersion in **IR1** is achieved. The script then allows the user to view and prune the candidate stars before actually forming the template.

Next, the script runs the FOCAS 'resolution' routine, which measures the same scale and fraction attributes as described in the AutoPlate section above, and based upon these values, applies a simple set of default rules for classifying objects as stars, fuzzy stars, galaxies, or artifacts. The script then allows the user to review the image in order to facilitate changes to the FOCAS-provided object classifications. If a good PSF template was formed and the data are of sufficiently high resolution and quality, FOCAS will do an excellent job of classifying the objects, generally beyond the detection limits of DPOSS. Even better classifications are no doubt achievable with the CCD data by applying machine learning to derive more complex rules, and SKICAT was designed to facilitate just that. However, we found the quality of the standard FOCAS classifications more than sufficient for our present purposes: to facilitate photometric calibration and construction of training sets for plate object classification.

Once the construction of the FOCAS catalog is complete, meaning all attributes have been measured and classifications assigned, a final routine transforms the FOCAS format catalog into a SKICAT catalog. The latter is comprised of the CCD header, which contains information from both the FOCAS catalog header and the AutoCCD parameters file, and a features table of the exact same format as that of a plate catalog.

C.3 Post-processing

C.3.1 Astrometric transformation

One has three options for setting the RA,Dec coordinates of the objects in a CCD catalog, depending on what, if any, other catalogs covering the same field currently exist in the SKICAT database. Ideally a plate catalog covering the CCD field has already been created, in which case a SKICAT tool performs the following operations. Using the approximate position of the CCD frame saved in the CCD's header file, the program automatically searches the relevant portion of the plate catalog and tries to match the two. The program automatically restricts the search to objects classified as stars within an intermediate

magnitude range, as one expects these objects to provide the most consistent and precise astrometry. It then allows the user to interactively view and correct the matches it finds (see Figure 8). One can accept or reject any of the suggested matches before allowing the program to solve for the astrometric solution. First, the program finds the transformation matrix from CCD to plate standard coordinates. Then, the program applies the plate's standard to celestial coordinate transformation polynomials to compute RA and Decs. This same interface would be useful for scientific projects involving the association of astrometric coordinates with deep CCD images not even used for calibration.

If an overlapping plate catalog is not available, but a CCD catalog is, the user may execute an analogous script which determines the astrometric solution using the other CCD's celestial coordinates. In this case, it derives a single matrix expressing shift, shear, scale, and rotation for converting directly from X,Y to RA,Dec coordinates.

As a final resort, the RA and Dec values of a CCD catalog may be derived by assuming the image is rotated counterclockwise relative to nominal (*i.e.,* north to the top and east to the left) an amount indicated by the header's position angle column, and centered on the approximate position saved in the header. This procedure should only be used in the event no other SKICAT catalog exists covering the same field of view. Ultimately, all CCD catalogs should be astrometrically calibrated using the plate catalog to which they most directly apply. This will minimize matching error when the catalogs are eventually matched.

C.3.2 CCD registration

At this point, the catalog is ready for registration into the SKICAT catalog management system. As with plate catalogs, a catalog must be registered in order to be matched with other catalogs or saved off-line in a manner such that it can be reloaded by SKICAT.

C.3.3 Photometric calibration

As with the astrometric assignment of CCD catalogs, the user has a choice of photometric calibration methods, depending on what catalogs are already loaded in the system. One

method performs the calibration assuming a default color term, meaning a user-specified color is applied when performing the instrumental to calibrated magnitude transformation given by 2. Independent default colors are assumed for stellar and non-stellar objects and are specified within the CCD header file. This routine also uses header columns containing the magnitude zero point offset term (A), extinction term (B), color term (C), exposure time (t), and airmass ($sec(z)$) parameters to derive the calibrated magnitude (m) saved in the CCD features table. These parameters are used in the relation:

$$m = m_{inst} + 2.5\log(t) + A + Bsec(z) + C(g - r), \tag{2}$$

where m_{inst} is the measured instrumental magnitude and $(g - r)$ is the default color term applied. Any of the four instrumental magnitudes available in the CCD features table may be substituted for m_{inst}

Once red and blue catalogs of the same CCD field have been created and matched together within SKICAT, one may calibrate the magnitudes of each using actual color information. One has three options, depending on whether one wants to update either the red or blue catalog, or both. One command takes the names of corresponding blue and red CCD catalogs and updates the magnitudes of both by simultaneously solving the relation 2 for objects measured in both the blue and the red. Unmatched objects are not affected. Alternative programs exist in the event that one only wants to update one of the two catalogs using this method.

D Appendix - Matched Catalog Data Structure

The primary data structure comprising the matched catalog is the MatchedFeatures table, which contains one row for each feature added from each constituent catalog. The MatchedFeatures table contains a user-defined subset of the columns from the catalog features tables. Features are linked together by an ObjectId column which indicates which object each feature is associated with (see Figure 9). A MatchedCatalogs table indicates those catalogs which have been added to the matched catalog. A third table, named MatchedCount, maintains a running count of the number of features associated with each ObjectId and is maintained simply for improved query performance.

The user can modify the parameters which control the matching process by setting parameters in the MatchProc table. The list of columns from the catalog features table which are included in the matched catalog is maintained in the MatchColumns table. The parameters which control the process of adding a catalog to the matched catalog (located in the MatchProc table) are:

NextObjectId: the next unused ObjectId used to uniquely identify objects.

MaxObjectDistance: the maximum allowable distance between two matched features in arcsec.

XSeg: the number of segments (in X dimension) to break the catalog into for matching.

YSeg: the number of segments (in Y dimension) to break the catalog into for matching.

QSigmaClip: The quartile-sigma clipping threshold for computing offset means and standard deviations.

SearchNumSig: The search radius applied for second and subsequent iterations, in terms of measured offset standard deviations.

ErrMax: the maximum Chi-squared positioning error in X and Y for a match to be accepted.

ConvergeMode: 0 for automatic convergence, 1 for manual convergence.

ConvergeScale: the maximum allowable average match difference in X or Y, in terms of estimated error in the mean offset, for convergence.

MaxNumPasses: the maximum number of matching passes for auto convergence, exact number of passes for manual convergence.

References

Cheeseman, P. *et al.*1988, in *Proc. Fifth Machine Learning Workshop, San Mateo*, Morgan Kaufmann, 54.

Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I., and Laidler, V. 1992, *BAAS*, **24**, 750.

Ellis, R. 1987, in *Observational Cosmology, IAU Symp. 124*, ed. A. Hewitt, G. Burbidge, and L. Z. Fang, Dordrecht: Reidel, 367.

Fayyad, U. 1991. Ph.D. thesis, EECS Dept. The University of Michigan.

Fayyad, U. and Irani, K. 1992, in *Proceedings of the Tenth National Conference on Artificial Intelligence AAAI-92, San Jose, CA.*

Fayyad, U. and Irani, K. 1993, in *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, France*, Morgan Kauffman, in press.

Fayyad, U., Weir, N., and Djorgovski, S. 1993, in *Proceedings of AAAI-93 Workshop on Knowledge Discovery in Databases, Washington D.C.*, ed. G. Piatetsky-Shapiro, AAAI/MIT Press, 1.

Jarvis, J. and Tyson, J. 1979, *SPIE Proc. on Instrumentation in Astronomy*, **172**, 422.

Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I., and Sturch, C. 1992, *BAAS*, **24**, 741.

Maddox, S., Sutherland, W., Efstathiou, G., and Loveday, J. 1990, *MNRAS*, **243**, 692.

Reid, I. *et al.*1991, *Publ. Astron. Soc. Pac.*, **331**, 465.

Reid, N. and Djorgovski, S. 1993, in *Sky Surveys: Protostars to Protogalaxies*, ed. B. T. Soifer, A.S.P. Conf. Ser. #43, 125.

Russel, J. L., Lasker, B. M., McLean, B. J., Sturch, C. R., and Jenker, H. 1990, *AJ*, **99**, 2059.

Thuan, T. X. and Gunn, J. 1976, *Pub. Astron. Soc. Pac.*, **88**, 543.

Valdes, F. 1982a, *SPIE Proc. on Instrumentation in Astronomy IV*, **331**, 465.

Valdes, F. 1982b, *FOCAS User's Manual*, (Tucson: NOAO).

Weir, N., Djorgovski, S., and Fayyad, U. 1994, *AJ*, submitted.

Weir, N., Djorgovski, S., Fayyad, U., and Roden, J. 1994a, *SKICAT Plate and CCD Processing Cookbook*, (Pasadena: JPL/Caltech).

Weir, N., Fayyad, U., and Djorgovski, S. 1994, *AJ*, submitted.

Weir, N., Fayyad, U., Roden, J., and Djorgovski, S. 1994b, *SKICAT User's Manual*, (Pasadena: JPL/Caltech).

Weir, N. and Picard, A. 1991, in *Digitised Optical Sky Surveys*, ed. H.T. MacGillivray and E.B. Thomson, Dordrecht: Kluwer Academic Publisher, 225.

Figure 1: An overview of the SKICAT system.

Figure 2: A plate scan is saved as 23 Vax VMS savesets (rows) of 23 image 'blocks' each. Each image block consists of 1024×1024 pixels, except at the right and top edges, where one dimension is only 512.

Figure 3: ST ScI produces a 'snapshot' image for every plate scan. It contains one sample pixel per every $\sim 33 \times 33$) in the full scan. The snapshot may be used to quickly and easily check general qualities of the scan.

Figure 4: A plate scan is analyzed as a set of 13×13 overlapping footprint images of 2048^2 pixels each. Not only is this approach computationally convenient, but it provides greater sensitivity to position-dependent plate effects. It also facilitates quality control via the systematic comparison of the overlap regions.

Figure 5: Given the measured image blur (R^2), we establish the appropriate factor by which to scale the measured sky sigma to approximate that of an unblurred version of the same image.

Figure 6: The X-Windows catalog construction interface within SKICAT.

Figure 7: The regions surrounding bright stars must be avoided when analyzing the plate catalogs generated by SKICAT, as it typically splits these objects into dozens, or even hundreds, of spurious artifacts.

Figure 8: SKICAT automatically searches a plate catalog for the region overlapping a CCD frame. The program returns with a list of suggested matches and displays the overlapping portions of the two catalogs in graphical form, as shown above (plate to the left, CCD to the right). The displayed coordinates are those of the plate scan. On a workstation, the matched objects are color coded as well as numbered, allowing the user to easily identify and remove spurious matches from the list.

Figure 9: An overview of the SKICAT object matching process.

Figure 10: The X-Windows catalog query interface within SKICAT.

Label		Description
xmin	-	Minimum useable X coordinate in plate image
xmax	-	Maximum useable X coordinate
ymin	-	Minimum useable Y coordinate
ymax	-	Maximum useable Y coordinate
spotxmin	-	Beginning of spots boundary in X
spotymax	-	End of spots boundary in Y
sky	-	Density of the sky at plate center
saturation	-	Saturation density of the plate

Table 1:

Label		Description
XC	-	x position (center of maximum 3 × 3 pixel integrated intensity)
YC	-	y position
MCore	-	core magnitude (from maximum 3 × 3 pixel integrated intensity)
MAper	-	aperture magnitude (from integrated intensity within aperture)
MIso	-	isophotal magnitude (from integrated intensity within detection isophote)
MTot	-	total magnitude (from integrated intensity within 'grown' isophote)
SLi	-	sigma of sky subtracted integrated intensity (luminosity) within detection isophote
SSBr	-	local sky sigma
Ispht	-	isophote brightness (average intensity along detection isophote)
Area	-	isophotal area (area within detection isophote)
TArea	-	total area (area within 'grown' isophote)
XAvg	-	average x width
YAvg	-	average y width
ICX	-	x intensity weighted centroid
ICY	-	y intensity weighted centroid
IXX	-	xx intensity weighted second moment
IXY	-	xy intensity weighted second moment
IYY	-	yy intensity weighted second moment
IR1	-	intensity weighted first moment radius
IR3	-	intensity weighted third moment radius
IR4	-	intensity weighted fourth moment radius
CX	-	x unweighted centroid
CY	-	y unweighted centroid
XX	-	xx unweighted second moment
XY	-	xy unweighted second moment
YY	-	yy unweighted second moment
R1	-	unweighted first moment radius

Table 2:

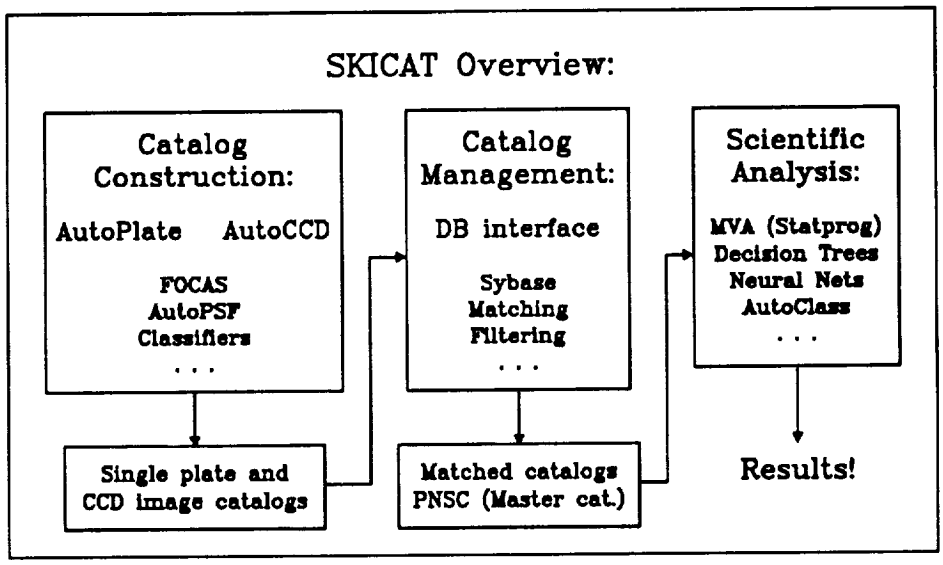
SKICAT Overview:
Catalog Construction:
AutoPlate
AutoCCD
FOCAS
AutoPSF
Classifiers
. . .
Catalog Management:
DB interface
Sybase
Matching
Filtering
. . .
Scientific Analysis:
MVA (Statprog)
Decision Trees
Neural Nets
AutoClass
. . .
Single plate and CCD image catalogs
Matched catalogs PNSC (Master cat.)
Results!

Figure 1:

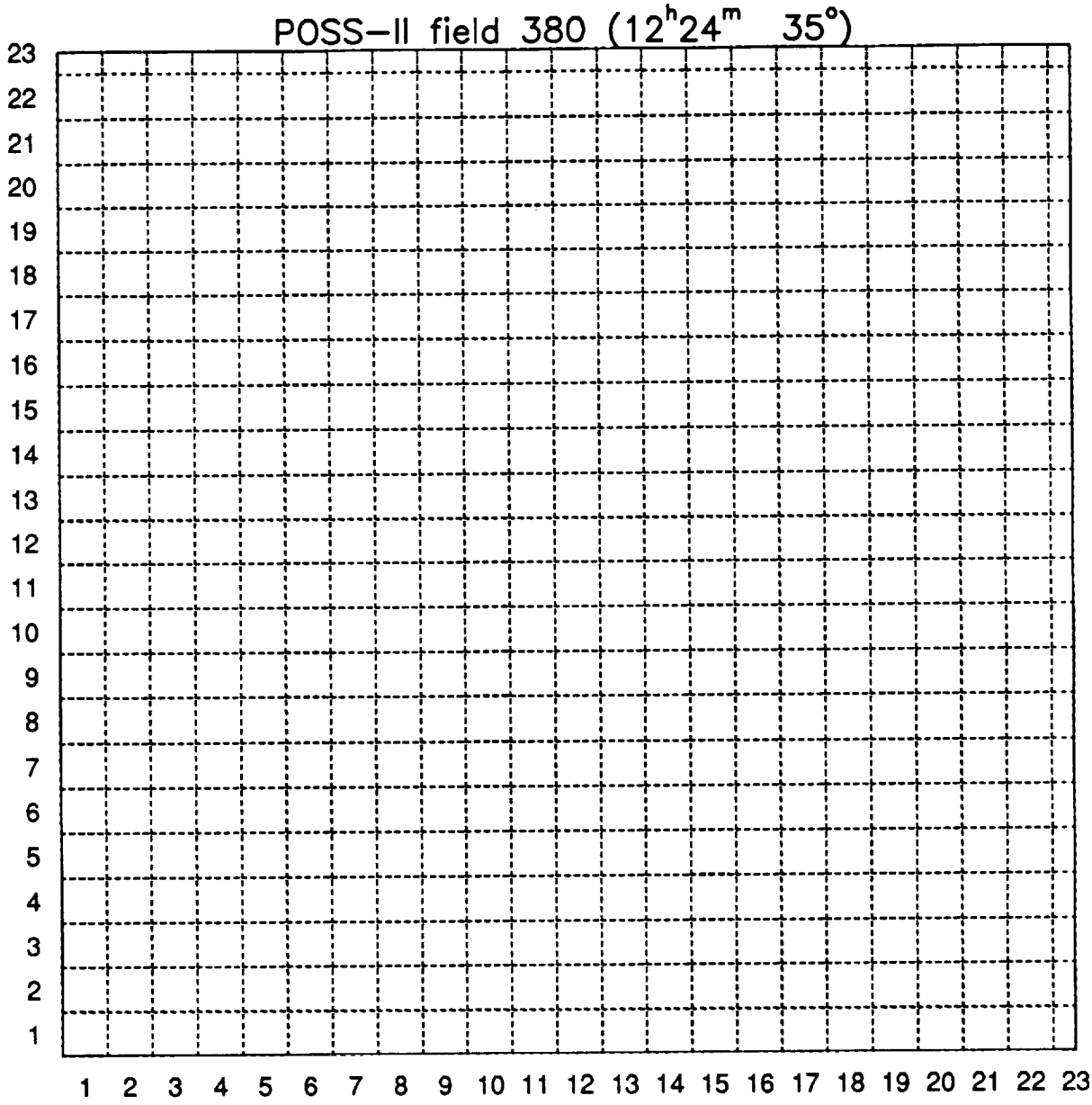
POSS–II field 380 (12h24m 35°)
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

Figure 2:

POSS-II Plate J380 (12^h24^m $+35^o$)

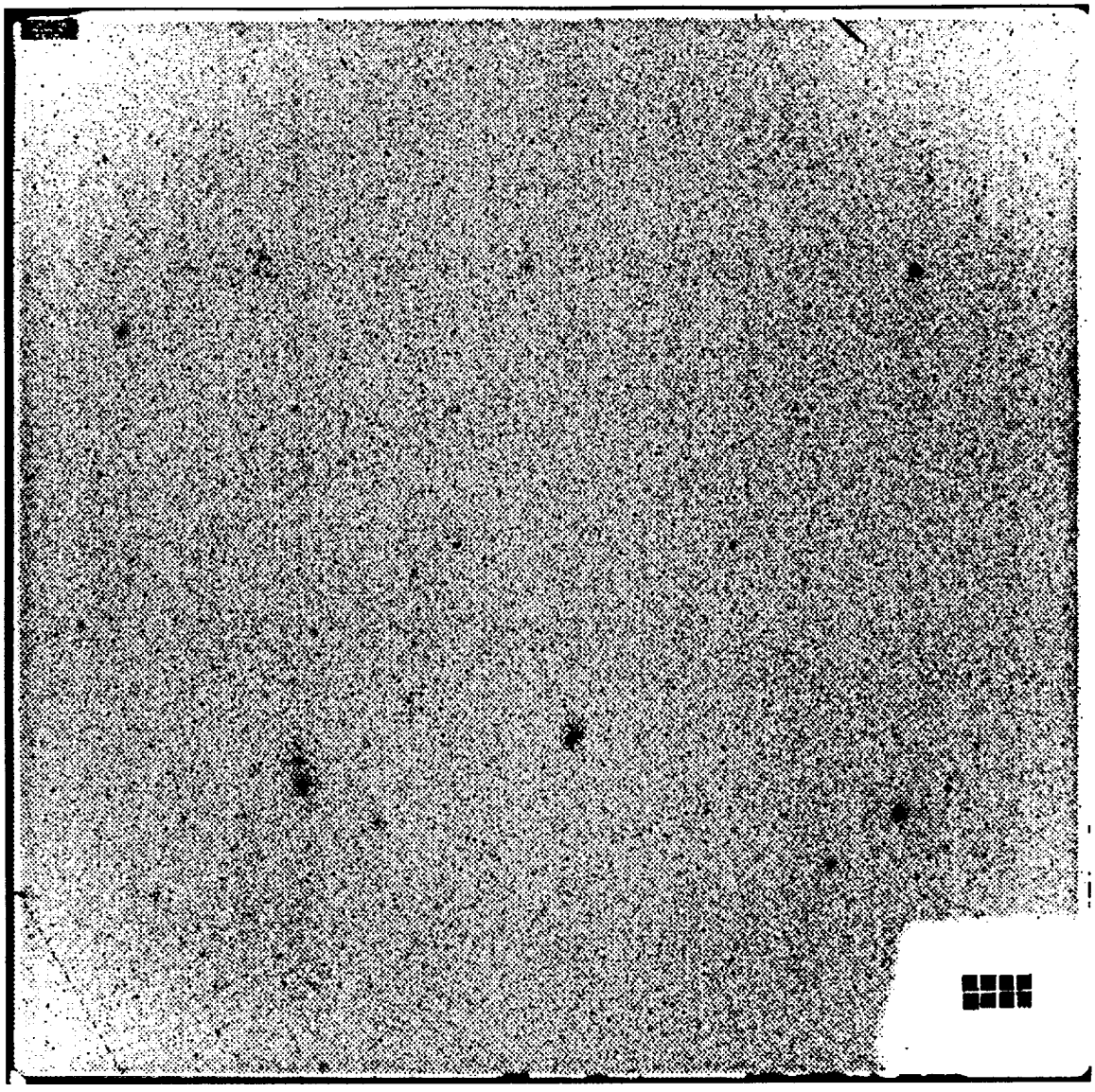

Figure 3:

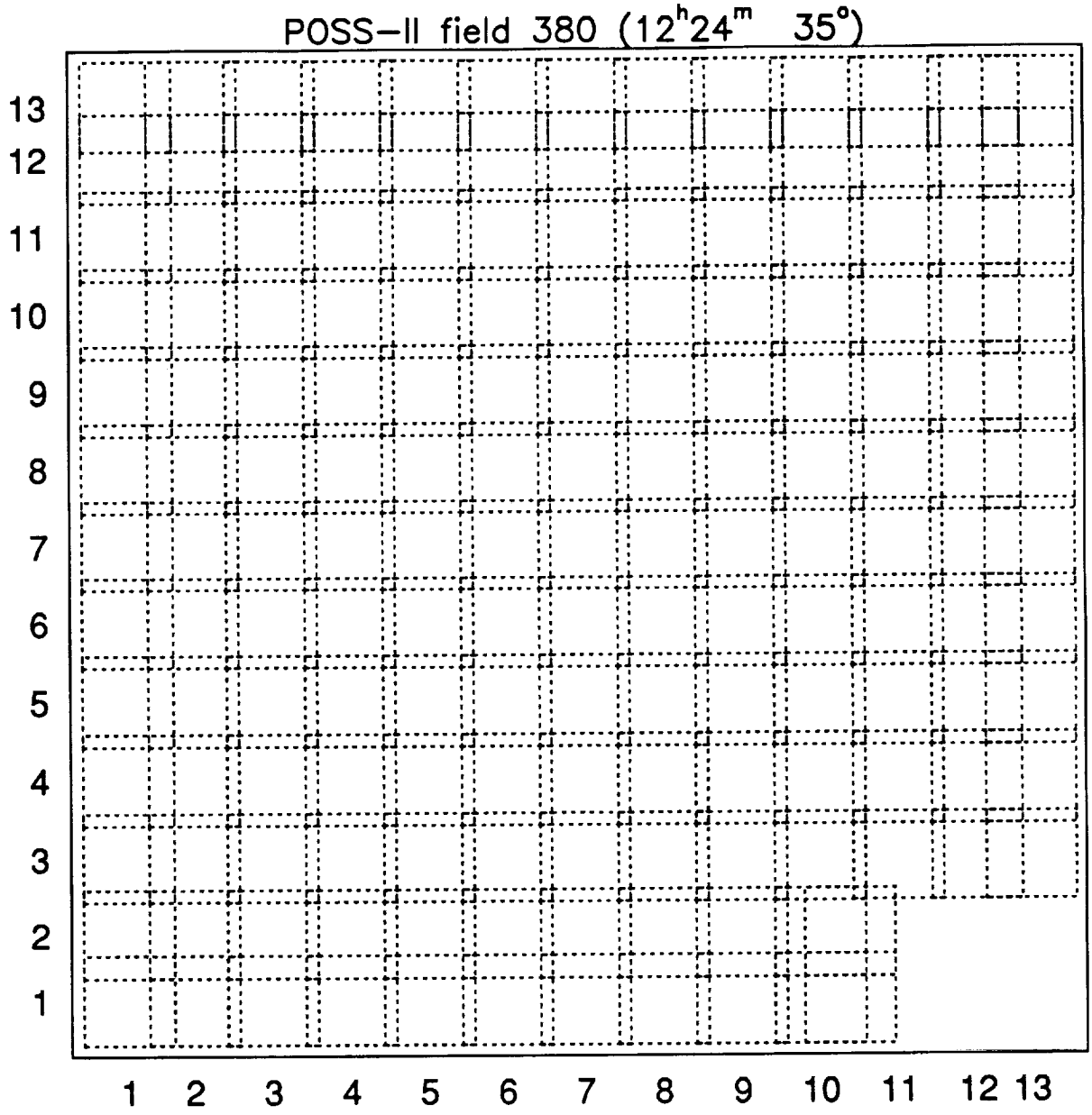
POSS–II field 380 (12h24m 35°)
13
12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13

Figure 4:

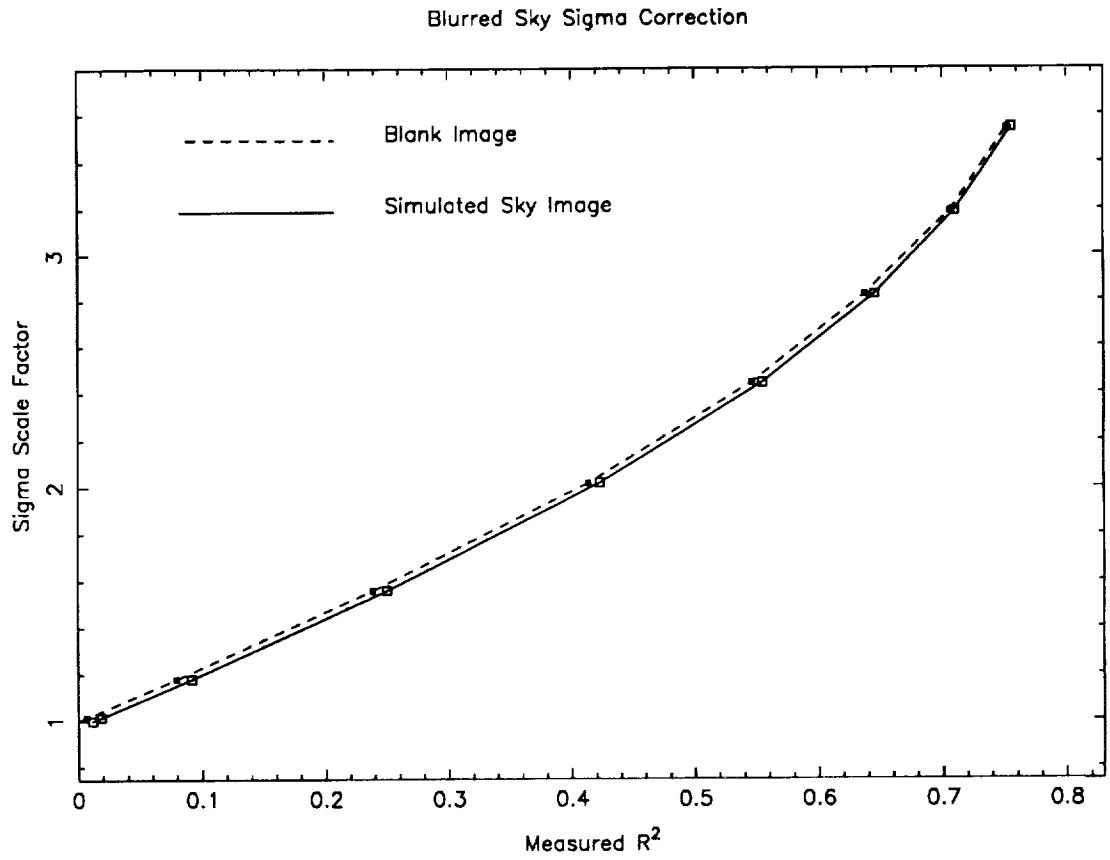
Blurred Sky Sigma Correction
Blank Image
Simulated Sky Image
Sigma Scale Factor
1
2
3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Measured R2

Figure 5:

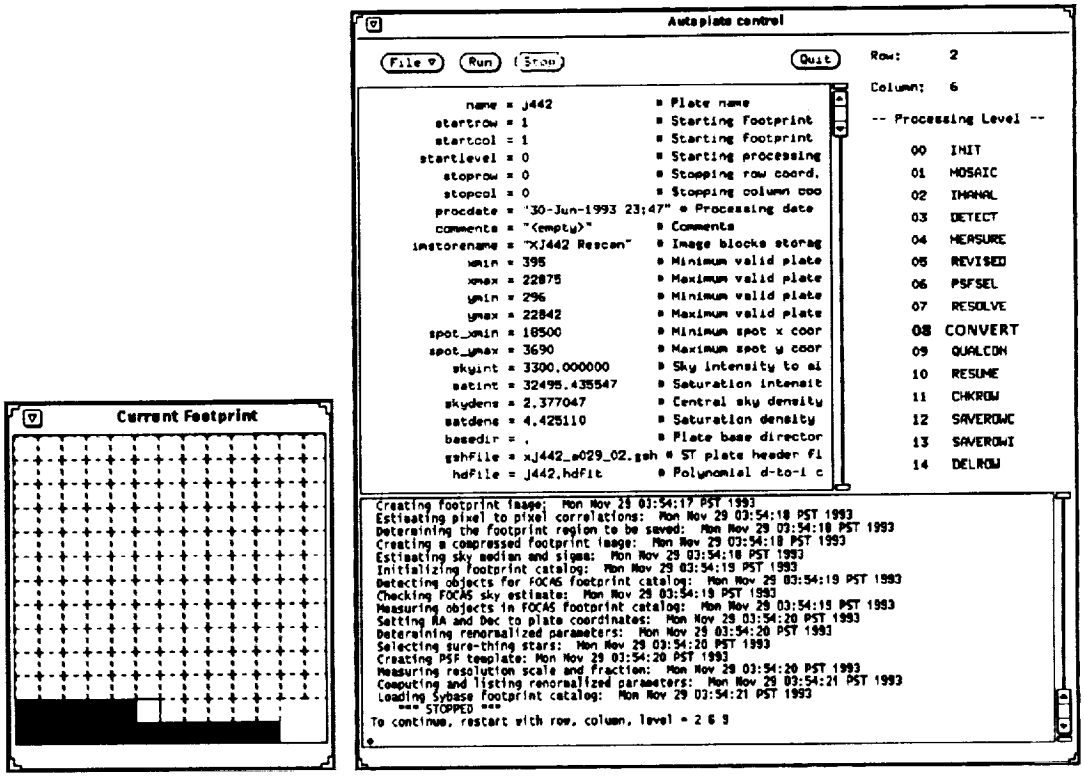

Figure 6:

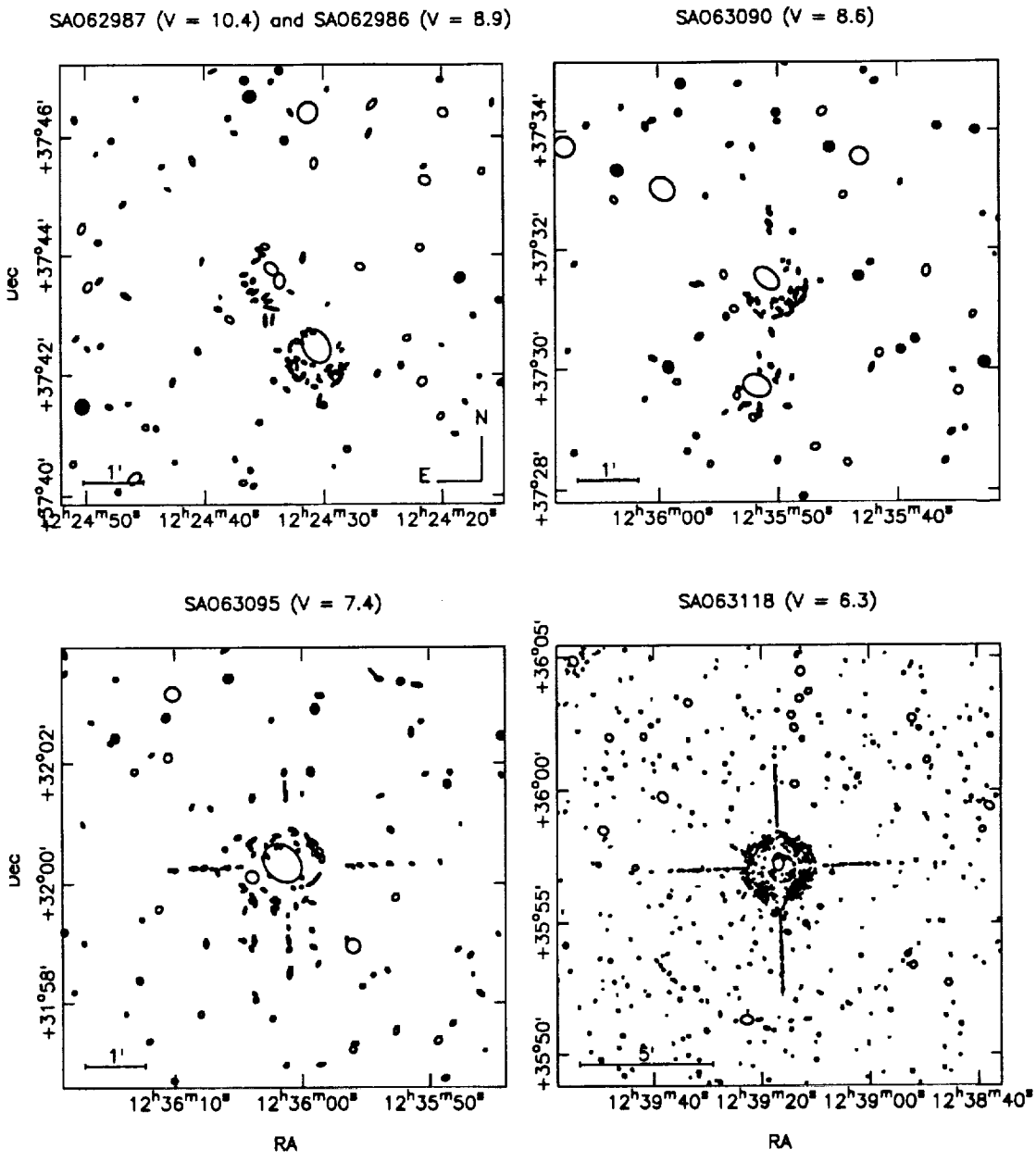
SA062987 (V = 10.4) and SA062986 (V = 8.9)
SA063090 (V = 8.6)
SA063095 (V = 7.4)
SA063118 (V = 6.3)
Dec
RA
N
E
1'
5'
+37°46'
+37°44'
+37°42'
+37°40'
+37°34'
+37°32'
+37°30'
+37°28'
+32°02'
+32°00'
+31°58'
+36°05'
+36°00'
+35°55'
+35°50'
12h24m50s
12h24m40s
12h24m30s
12h24m20s
12h36m00s
12h35m50s
12h35m40s
12h36m10s
12h36m00s
12h35m50s
12h39m40s
12h39m20s
12h39m00s
12h38m40s

Figure 7:

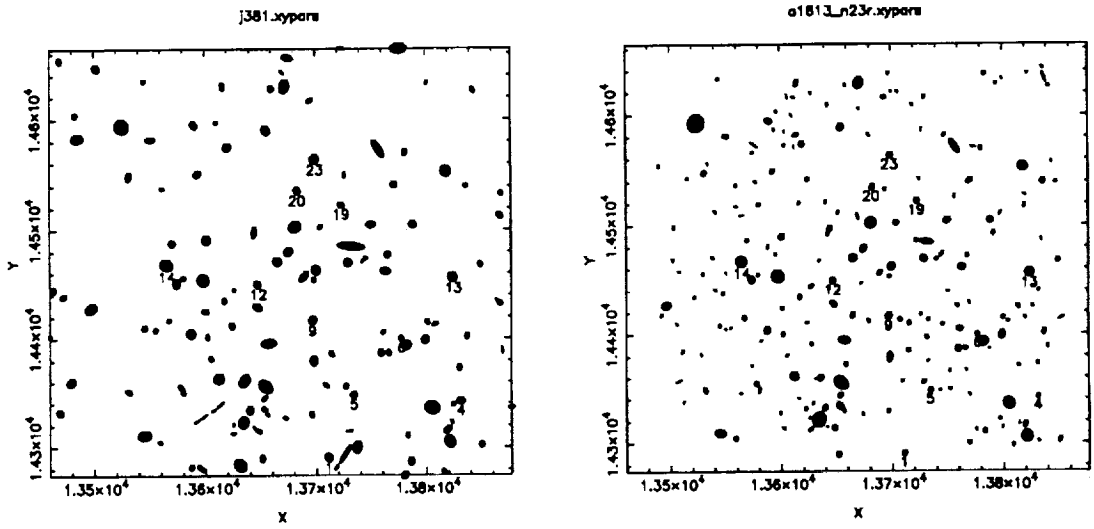
j381.xypars
a1813_n23r.xypars
Y
X
1.43x10^4
1.44x10^4
1.45x10^4
1.46x10^4
1.35x10^4
1.36x10^4
1.37x10^4
1.38x10^4

Figure 8:

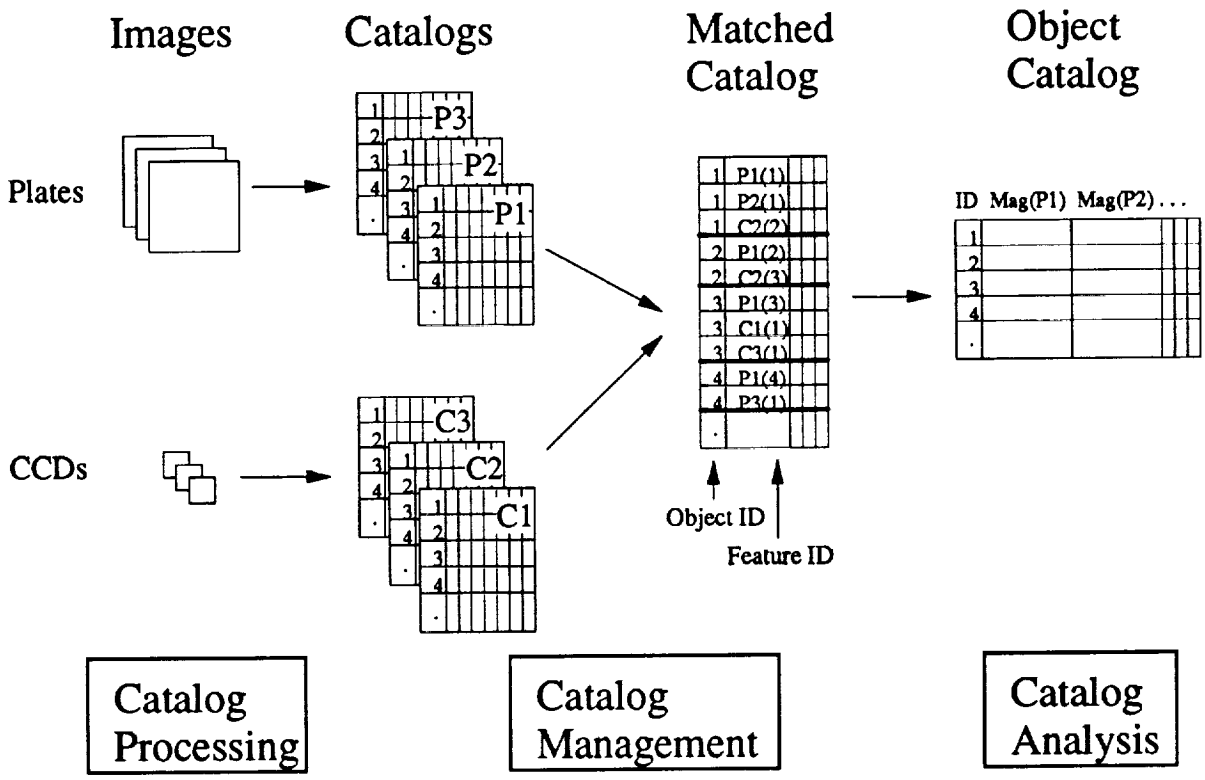
Images
Catalogs
Matched Catalog
Object Catalog
Plates
CCDs
P3
P2
P1
C3
C2
C1
P1(1)
P2(1)
C2(2)
P1(2)
C2(3)
P1(3)
C1(1)
C3(1)
P1(4)
P3(1)
ID Mag(P1) Mag(P2)...
Object ID
Feature ID
Catalog Processing
Catalog Management
Catalog Analysis

Figure 9:

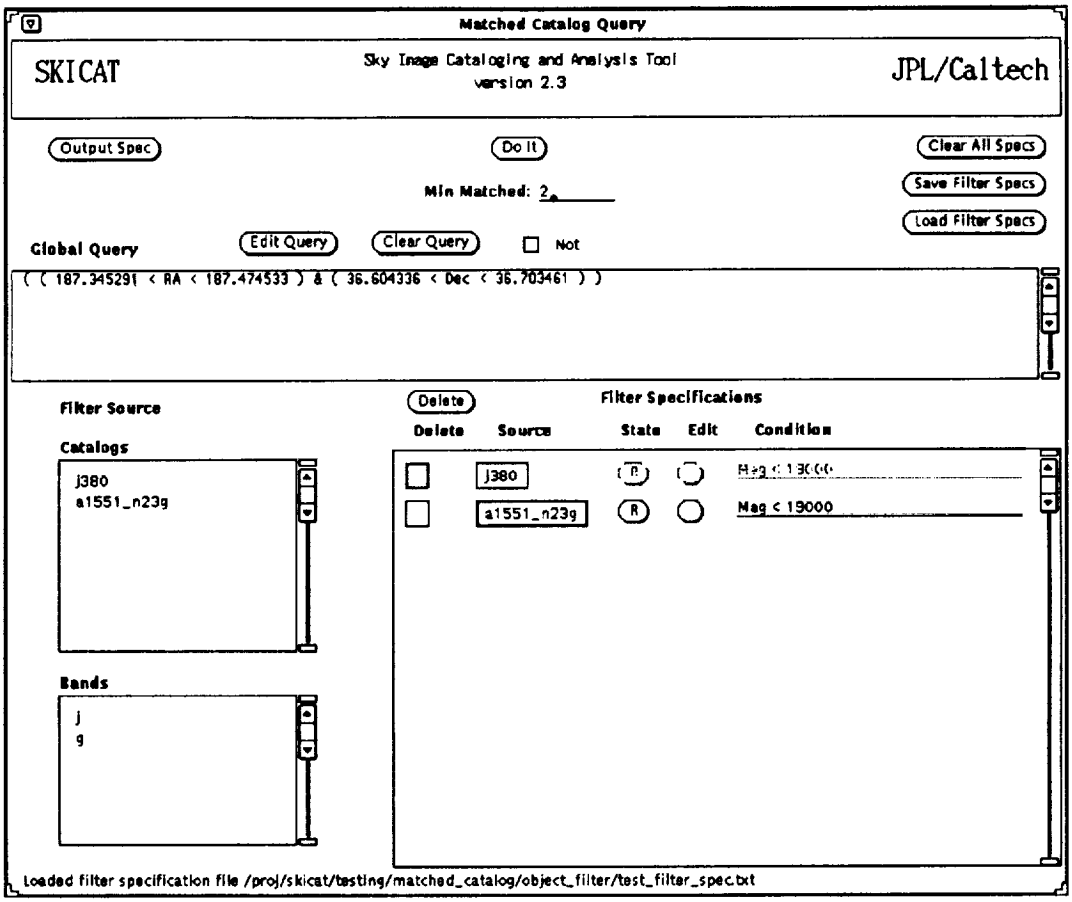
Matched Catalog Query
SKICAT
Sky Image Cataloging and Analysis Tool
version 2.3
JPL/Caltech
Output Spec
Do It
Clear All Specs
Save Filter Specs
Load Filter Specs
Min Matched: 2
Global Query
Edit Query
Clear Query
Not
((187.345291 < RA < 187.474533) & (36.604336 < Dec < 36.703461))
Filter Source
Catalogs
j380
a1551_n23g
Bands
J
g
Delete
Filter Specifications
Delete
Source
State
Edit
Condition
j380
a1551_n23g
R
Mag < 19000
Loaded filter specification file /proj/skicat/testing/matched_catalog/object_filter/test_filter_spec.txt

Figure 10:

Attachment B:

Star-galaxy classification within the SKICAT system

From a paper submitted to the *Astronomical Journal*, September 1994

Current status: Accepted for publication with very minor modifications

Expected publication date: mid-1995

Automated Star/Galaxy Classification for Digitized POSS-II

Nicholas Weir[†]
Usama M. Fayyad[‡]
S. Djorgovski[†]

[†] Palomar Observatory
California Institute of Technology 105-24
Pasadena, CA 91125
weir@fritz.caltech.edu
george@deimos.caltech.edu

[‡] Jet Propulsion Laboratory
California Institute of Technology 525-3660
Pasadena, CA 91109
fayyad@aig.jpl.nasa.gov

Submitted to *The Astronomical Journal*:

Received:

Accepted:

Running Title: *Automated Classification for POSS-II*

Abstract

We describe the automated object classification method implemented in the Sky Image Cataloging and Analysis Tool (SKICAT) and applied to the Digitized Second Palomar Observatory Sky Survey (DPOSS). This classification technique was designed with two purposes in mind: first, to classify objects in DPOSS to the faintest limits of the data; second, to fully generalize to future classification efforts, including anything from classifying galaxies by morphology, to improving the existing DPOSS star/galaxy classifiers once a larger volume of data are in hand. To optimize the identification of stars and galaxies in J and F band DPOSS scans, we determined a set of eight highly informative object attributes. In the eight-dimensional space defined by these attributes, we found like objects to be distributed relatively uniformly within and between plates. To infer the rules for distinguishing objects in this, but possibly any other, high-dimensional parameter space, we utilize a machine learning technique known as decision tree induction. Such induction algorithms are able to determine near-optimal classification rules simply by training on a set of example objects. We used high quality CCD images to determine accurate classifications for those examples in the training set too faint for reliable classification by examining the plate scans by eye. Our initial results obtained from a set of four DPOSS fields indicate that we achieve 90% completeness and 10% contamination in our galaxy catalogs down to a magnitude limit of $\sim 19.6^m$ in r and 20.5^m in g, within F and J plates respectively, or an equivalent B_J of nearly 21.0^m. This represents a 0.5^m – 1.0^m improvement over results from previous digitized Schmidt plate surveys using comparable plate material. We have also begun applying methods of unsupervised classification to the DPOSS catalogs, allowing the data, rather than the scientist, to suggest the relevant and distinct classes within the sample. Our initial results from these experiments suggest the scientific promise of such machine discovery methods in astronomy.

keywords: classification, sky surveys

1 Introduction

The first step in analyzing any imaging sky survey is to identify, measure, and catalog all of the detected objects into their respective classes. Once the objects have been measured and classified, further scientific analysis may proceed.

The accuracy of star/galaxy separation generally determines the effective limiting magnitude, in terms of scientific usefulness, of imaging surveys. This limit is, in very many respects, more important than the object detection limit in terms of its impact on the variety of programs for which the data may be used. For example, in order to effectively use the data to compare against models of star or galaxy counts or colors, measure the angular correlation function of galaxies, or search for high redshift quasars, accurate star/galaxy classification is required at the level of approximately 90%. At the faint end, every additional magnitude to which one can extend this accuracy limit buys one on order of two to three times more classified objects in the catalog. Given the enormous resources put into obtaining the survey data in the first place, it makes sense to fully investigate the very latest technology when approaching the task of object classification, in the hope of squeezing every last bit of scientifically useful information from the survey. This was our motivation when designing and implementing the classification methods described in this paper, which are currently being applied to the digitized scans of the Second Palomar Observatory Sky Survey (POSS-II).

POSS-II (Reid *et al.* 1991) is more than 60% complete as of August, 1994, and will eventually cover 894 fields spaced 5° apart in three passbands: blue (IIIa-J + GG 395), red (IIIa-F + RG610), and near-infrared (IV-N + RG9). The typical limiting magnitudes for point sources in the corresponding J, F, and N bands are 22.5^m, 21.5^m, and 19.5^m, respectively. While the photographic survey is still under way, ST ScI and Caltech have begun a collaborative effort to digitize the complete set of plates (Djorgovski *et al.* 1992; Lasker *et al.* 1992; Reid and Djorgovski 1993). So far, only a subset of the J, F, and N plates have been scanned and processed. Both the photographic survey and the plate scanning are estimated to be $>$ 90% complete circa 1997. The resulting data set, the

Palomar-STScI Digital Sky Survey (DPOSS), will consist of $\sim$ 3 TB of pixel data: $\sim$ 1 GB/plate, with 1 arcsec pixels, 2 bytes/pixel, 20340^2 pixels/plate, for all survey fields in all three colors. In conjunction with the plate survey, we are also conducting an intensive program of CCD calibrations using the Palomar 60-inch telescope, using the Gunn-Thuan *gri* bands. These CCD images serve both for magnitude zero-point calibration and object classification purposes. The plate scans, when complete, will be the highest quality set of digital images covering the entire northern sky produced to date.

The first scientific results obtained using DPOSS, and making use of the classification methods described herein, are measures of blue and red galaxy counts in four POSS-II fields near the North Galactic Pole (Weir, Djorgovski, and Fayyad 1994). Several additional programs, including a high-redshift quasar search and measures of galaxy-galaxy angular correlations, are in progress (Weir *et al.* 1994a).

In order to make most efficient use of DPOSS, and to generally facilitate its scientific exploitation, Caltech Astronomy and the JPL Artificial Intelligence Group have been engaged in a collaborative effort to integrate state-of-the-art computing methods for application to DPOSS. The result of our joint effort is the Sky Image Cataloging and Analysis Tool (SKICAT), a suite of programs designed to facilitate the maintenance and analysis of astronomical surveys comprised of multiple, overlapping images. The classification technology described in this paper was developed as a part of this effort and is implemented within SKICAT (Weir *et al.* 1994b).

Historical methods for classifying image features would preclude the identification of the majority of objects in a DPOSS image, since these objects are too faint for traditional recognition algorithms, or even object-by-object classification by eye. A principal goal of SKICAT was to provide an effective, objective, repeatable, and examinable basis for classifying sky objects at levels beyond the limits of previously existing technology. Of course, due to statistical fluctuations of the data, one may never construct a classifier that will be 100% accurate. One may, nonetheless, aim for the highest statistical accuracy achievable to the greatest possible depth.

A particular difficulty in classifying DPOSS objects is that the scan images vary sig-

nificantly in terms of image quality (*e.g.*, background noise, point spread function shape, etc.) both within and across plate boundaries. This created an important demand on the classification method to be able to cope with this variation and produce consistent results throughout the survey.

The two essential steps in performing automated object classification are to define the space of discriminating attributes characterizing each object, then determine a means of distinguishing objects within that space. The first step is key, as it determines upon what information any classification will be based. We concentrated a significant amount of effort in deriving a set of object attributes which effectively remove the intra- and inter-plate variations described above. The second step is likewise very important, as there are any number of ways, some much more powerful than others, of designing rules that divide the parameters space into regions of like objects.

The approach we chose for this second step was one developed in the field of machine learning, namely using decision tree induction algorithms. These methods are able to automatically induce classification rules based simply upon user-supplied examples. This approach not only provided us with the very effective star/galaxy classifiers that already are being used to produce high-quality DPOSS catalogs, but it will easily allow future users to re-train specialized classifiers (*e.g.*, to identify galaxy morphology), or redo existing star/galaxy classifications as more data become available and/or attribute measurement technology improves.

1.1 Historical approaches

The problem of automatic object classification has been addressed for at least two decades, with a variety of proposed solutions. The most basic approach is to plot one measured attribute versus another and draw a line within that space best separating stars from galaxies. Typically the chosen attributes are magnitude and some measure of object 'peakedness', such as peak intensity, isophotal area, or intensity weighted first moment radius. Because in that space point sources are generally distributed along a fairly well-defined stellar locus, or ridge (see, *e.g.*, Figure 3), such a discriminant function tends to be reasonably accurate

down to moderately faint magnitudes. The shortfalls of this approach are that defining the classifier is very labor-intensive as well as subjective, and at faint levels, stars and galaxies quickly blur together around the locus.

The next level of sophistication is to perform star/galaxy separation in a space defined by some non-linear combination of parameters, rather than raw measurements. For example, simply by plotting the logarithm of isophotal area [**log(Area)**] vs. magnitude, instead of just object area, the stellar locus becomes more linear, making a separator much easier to define and generally more accurate. For classifying objects from COSMOS digitized plate scans, Heydon-Dumbleton, Collins, and MacGillivray (1989) found it optimal to discriminate using one of three different pairwise plots depending on an object's magnitude. The three parameters they plotted versus magnitude were: **G**, a measure of how effectively an image fills the ellipse fitted to its major and minor axes, for bright objects; **log(Area)**, for intermediate objects; and a derived parameter, **S**, which effectively measures the scale of a best fit Gaussian to an object's light distribution, for the faintest objects.

Heydon-Dumbleton, Collins, and MacGillivray (1989) also improved upon the standard method by making the choice of discriminant line more objective. They measured the statistical distribution of objects around the stellar locus as a function of magnitude, setting the star/galaxy separation line some number of standard deviations above the locus mean.

Picard (1991), in his analysis of COSMOS scans of POSS-II F plates, similarly measured the mean and width of the stellar locus in **S** vs. magnitude space, defining a new parameter, ϕ, corresponding to an object's distance from the locus, normalized by the width of the locus at that magnitude. He binned all the measurements for a given plate and computed a value, ϕ_{cut}, corresponding to three times the estimated width of the normalized stellar locus. He would then classify all objects with ϕ less than ϕ_{cut} as stars, the rest as galaxies. Using this approach, he estimated that he was able to achieve on average 90% completion (fraction of all galaxies classified as such) and 10% contamination (fraction of non-galaxy objects classified as galaxies) in his galaxy catalog down to a magnitude of 19.0^m in r.

The APM group (Maddox *et al.* 1990) took a slightly different approach to classifying objects from their scans of J plates from the Southern Schmidt survey. Rather than measuring the distance from the stellar locus in the space of one parameter vs. magnitude, they used a metric involving ten different parameters: peak density, radius of gyration, and image area above each of eight surface brightness levels. Two additional parameters were used to help them distinguish blended objects from galaxies, as no deblending algorithm was applied by the APM real-time software in the course of processing. Using this approach, APM reported a classification accuracy comparable to Picard's at a B_J magnitude of 20.0^m.

A far different method for classifying objects from plate scans was pioneered by Sebok (1979) in his Ph.D. thesis at Caltech. He introduced the concept of Bayesian classification to the problem, estimating the most probable classification of each object based upon its fit to a set of models. While this approach was effective, it was never widely applied to Schmidt plate surveys subsequently.

Sebok's classification method preceded the similar approach devised by Valdes and implemented in modern versions of FOCAS (Valdes 1982). Valdes also applied a technique premised on Bayesian probability theory, but more significantly, he introduced a measurement procedure that results in extremely discriminating object attributes. By selecting a number of objects in an image that are 'sure-thing' stars, FOCAS adds the rasters of the central pixels of these objects to form an empirical estimate of the point spread function (PSF) for that image. Using the 'resolution' routine, FOCAS then fits a model to each object consisting of a pure PSF component and a blurred version of the same. The best-fitting fraction of blurred component and its scale are the two attributes resolution measures and uses for performing object classification. These attributes have never been used in large scale digitized plate surveys to date because computing technology prevented the repeated access to the pixel data, which this technique requires.

FOCAS provides a default set of rules specifying to which class different portions of fraction vs. scale space correspond. Because the distribution of objects in the space of these attributes tends to be relatively invariant from image to image (PSF variations are

effectively taken into account by the fitting process), the default rules are found to provide excellent classification accuracy down to fairly faint levels for a wide variety of images. The user has the option of changing these classification rules, but FOCAS does not provide a way of allowing for more attributes in the rules, or a systematic way for determining a new, optimal set of rules for a particular type of image.

1.2 The machine learning approach

Drawing upon these previous efforts, we chose to measure and calculate those object attributes found to provide the best star/galaxy discrimination. However, unlike most previous approaches, we chose to apply modern methods from the field of machine learning to determine the optimal discriminant functions, or set of classification rules, within the multi-dimensional space of these measurements. The goal when applying these methods is to provide enough examples of accurate classifications to the algorithm to allow it to infer the rules for distinguishing objects in that space. An important advantage of this approach is that one can typically feed a relatively large number of input parameters to the algorithm, allowing it to determine classification rules more complex than those typically devised by humans, generally as a result of examining pairwise plots of attributes. The extra degrees of freedom provided by learning in multi-dimensional parameter space often lead to substantially more accurate classifications. In addition, the rules are formed in an objective, repeatable fashion.

Others have also begun exploring the use of new machine learning methods for the purpose of object classification, perhaps most notably the APS group in Minnesota, who have digitized the plates of the original POSS (Odewahn *et al.* 1992). They applied artificial neural networks to the task of automatically inducing a set of classification rules for objects in their catalog. We, too, experimented with neural nets; however, for reasons discussed below, we chose to use a method involving decision trees, based on the work of Fayyad (1991), for creating the production-line classifier implemented within SKICAT and used on DPOSS.

2 Classifier Induction

For a detailed discussion of decision trees and associated methods of machine learning, we refer the reader to Fayyad (1991) and Fayyad and Irani (1992). Below we include a brief discussion and history of these methods, in particular those we utilize within SKICAT, in addition to a comparison of this approach with neural networks.

2.1 Decision trees

A particularly efficient method for extracting rules from data is to generate a decision tree (Quinlan 1986). A decision tree consists of nodes that represent tests on attribute values. The outgoing branches of a node correspond to all the possible outcomes of the test at the node, thus partitioning the examples at a node along the branches. For example, as illustrated in Figure 1, at the top-most (root) node, the tree may branch left or right depending on whether the object has **log(Area)** less than or greater than $\mathbf{A}_o$. In turn, either of these branches may lead to a node that conditions on the same attribute, a different one, or any combination of the same [*e.g.*, "branch left if $(mag < m_o)$ and $(\phi > \phi_o)$"]. The final nodes in the tree, the leaves, would correspond to an actual classification: star, galaxy, artifact, etc.

In Figure 2 we illustrate a portion of a much larger actual decision tree generated by the O-Btree algorithm (described below) for performing star/galaxy classification. The interval appearing above each node indicates the range in value of the attribute specified in the node above that an object must meet for it to pass along that branch. The dark branches lead to actual classifications. A full path from the root to any particular leaf corresponds to a single classification rule. The number in parentheses within each leaf indicates the number of training examples classified correctly by that rule.

A well-known algorithm for generating decision trees is Quinlan's ID3 (Quinlan 1986) with extended versions called C4 (Quinlan 1990). ID3 starts with all the training examples at the root node of the tree. An attribute is selected to partition the data. For each value of the attribute, a branch is created and the corresponding subset of examples that have the attribute value specified by the branch are moved to the newly created child node. The

algorithm is applied recursively to each child node until either all examples at a node are of one class, or all the examples at that node have the same values for all the attributes. Every leaf in the decision tree represents a classification rule. Note that the critical decision in such a top-down decision tree generation algorithm is the choice of attribute at a node. Attribute selection in ID3 and C4 is based on minimizing an information entropy measure applied to the examples at a node. The measure favors attributes that result in partitioning the data into subsets that have low class entropy. A subset of data has low class entropy when the majority of examples in it belong to a single class. For a detailed discussion of the information entropy selection criterion see Quinlan (1986), Fayyad (1991), and Fayyad and Irani (1992).

2.1.1 The GID3* and O-Btree algorithms

The attribute selection criterion clearly determines whether a "good" or "bad" tree is generated by a greedy algorithm (see Fayyad and Irani 1990 and Fayyad 1991 for the details of what we formally mean by one decision tree being better than another). Since making the *optimal* attribute choice is computationally infeasible, ID3 utilizes a heuristic criterion which favors the attribute that results in the partition having the least information entropy with respect to the classes. There are weaknesses inherent in algorithms like ID3/C4 due to the fact that, for discrete attributes, a branch is created for each value of the attribute chosen for branching. This overbranching is problematic since in general it may be the case that only a subset of values of an attribute are of relevance to the classification task while the rest of the values may not have any special predictive value for the classes. The GID3* algorithm was designed mainly to overcome this problem, generalizing the ID3 algorithm so that it does not necessarily branch on each value of the chosen attribute. GID3* can branch on arbitrary individual values of an attribute and "lump" the rest of the values in a single default branch. Unlike the other branches of the tree which represent a single value, the default branch represents a subset of values of an attribute. Unnecessary subdivision of the data may thus be reduced. See Fayyad (1991) for more details and for empirical evidence of improvement.

The O-Btree algorithm (Fayyad and Irani 1992) was designed to overcome problems with the information entropy selection measure itself. O-Btree creates strictly binary trees and utilizes a measure from a different family of measures that detect class separation rather than class impurity. Information entropy is a member of the class of impurity measures. O-Btree employs an orthogonality measure rather than entropy for branching. For details on problems with entropy measures and empirical evaluation of O-Btree, the reader is referred to Fayyad (1991) and Fayyad and Irani (1992). Both O-Btree and GID3* differ from ID3 and C4 in one additional aspect: the discretization algorithm used at each node to discretize continuous-valued attributes. Whereas ID3 and C4 utilize a binary interval discretization algorithm, we utilize a generalized version of that algorithm which derives multiple intervals rather than strictly two. For details and empirical tests showing that this algorithm does indeed produce better trees, see Fayyad (1991) and Fayyad and Irani (1993). We have found that this capability improves performance considerably in several domains.

2.2 The RULER system

There are limitations to decision tree generation algorithms that derive from the inherent fact that the classification rules they produce originate from a single tree. This fact was recognized by practitioners early on (Quinlan 1986). The basic problem is that in even a good tree, there are always leaves that are overspecialized or predict the wrong class. For example, if there are any measurement errors in the attributes, the decision tree will tend to fit to the noise and, hence, not generalize well to data that are out of sample. The very reason that makes decision tree generation efficient (the fact that data is quickly partitioned into ever smaller subsets) is also the reason why overspecialization or incorrect classification occurs. It is our philosophy that once we have good, efficient decision tree generators, they can be used to generate multiple trees, and from these, only the best rules in each are kept. To implement this strategy, the algorithm RULER was developed (Fayyad *et al.* 1992).

In multiple passes, RULER partitions a training set randomly into a training subset

and test subset. A decision tree is generated from the training set and its rules are tested on the corresponding test set. Using Fisher's exact test (Finney *et al.* 1963), the exact hypergeometric distribution, RULER evaluates each condition in a given rule's preconditions for relevance to the class predicted by the rule. It computes the probability that the condition is correlated with the class by chance[1]. If this probability is higher than a small threshold (say 0.01), the condition is deemed irrelevant and is pruned. In addition, RULER also measures the merit of the entire rule by applying the test to the entire precondition as a unit. This process serves as a filter which passes only robust, general, and correct rules.

By gathering a large number of rules through iterating on randomly subsampled training sets, RULER builds a large rule base of robust rules that collectively cover the entire original data set of examples (*i.e.*, every example is classified by a rule). A greedy covering algorithm is then employed to select a minimal subset of rules that covers the examples. The set is minimal in the sense that no rule could be removed without losing complete coverage of the original training set. Using RULER, we can typically produce a robust set of rules that has fewer rules than any of the original decision trees used to create it, and that generalizes better to out-of-sample data. The fact that decision tree algorithms constitute a fast and efficient method for generating a set of rules allows us to generate many trees without requiring extensive amounts of time and computation.

We implemented the RULER algorithm, in conjunction with GID3* and O-Btree, within SKICAT for the purpose of inducing classification rules by example, and it was used to produce the particular star/galaxy classifiers described subsequently. Throughout this paper, we generally refer to our technique as decision tree induction and the rules as decision trees. We simply note that in practice we are actually referring to the use of decision trees in conjunction with the RULER tree pruning and combining algorithm.

2.3 Decision trees vs. neural nets

In order to compare against other learning algorithms, and to preclude the possibility that a decision tree based approach is imposing *a priori* limitations on the achievable classification

[1]The Chi-square test is actually an approximation to Fisher's exact test when the number of test examples is large. We use Fisher's exact test because it is robust for both small and large data sets.

levels, we tested several neural network algorithms for comparison. The results indicate that neural nets achieve similar performance as decision trees. The learning algorithms we tested were traditional backpropagation, conjugate gradient optimization, and variable metric optimization of a two-layer perceptron (see Hertz, Krogh, and Palmer 1991 for an excellent introduction to perceptrons and neural methods of computation). The latter two are training algorithms that work in batch mode and use standard numerical optimization techniques in changing the network weights. Their main advantage over backpropagation is the significant speed-up in training time.

The results of our comparison between these approaches and decision trees can be summarized as follows. The performance of the neural networks was a fairly unstable function of the random initial network weights chosen prior to training and produced accuracy levels on a sample test set of data varying between 30% (no convergence) and 95%, compared with a 94% accuracy level for a decision tree classifier. The most common range of accuracy averaged between 76% and 84%. To achieve these levels of accuracy, we had to perform multiple trials, each time varying the number of internal nodes in the hidden layer, the initial network weight settings, and the learning rate constant for backpropagation.

Upon examining the results of this empirical study, we concluded that the neural net approach did not offer any clear advantages over the decision tree based learning algorithms. Although neural networks, with extensive training and several training restarts with different initial weights to avoid local minima, could match the performance of the decision tree classifier, the decision tree approach still holds several major advantages. For one, the tree is more easily interpreted than the weights in a neural network (although, admittedly, a list of 20 rules that condition on up to eight parameters is not entirely transparent either). More importantly, the learning algorithms we employ do not require the specification of parameters such as the size of the neural net or the number of hidden layers, nor do they call for random trials with different initial weight settings. There are, in fact, very few free parameters. This makes the decision tree algorithm much easier to implement as a generic tool within SKICAT. Also, the required training time is orders

of magnitude faster than the training time required for a neural network program (*i.e.*, seconds rather than dozens of minutes in some cases).

3 Classification Attributes

In classification learning, the choice of attributes used to define examples is by far the single most important factor determining the success or failure of the learning algorithm. The attributes we use for classification are computed through a combination of image processing and statistical measurement techniques. While they are not expected to be the final advancement in this area, we did find them to provide the most discriminating and uniform characterization of objects detected in DPOSS of any other set of attributes we have encountered. This section provides a detailed description of these attributes and how they are computed.

3.1 Base-level attributes

The eight attributes we use in object classification include a compendium of measures found to be most useful and discriminating in previous surveys. They include:

MTot - the FOCAS total instrumental magnitude;

MCore - the core magnitude, measured from the brightest 3×3 pixel region in the object;

log(Area) - the log of the isophotal area of the object;

Ellip - the ellipticity;

IR1 - the intensity weighted first moment radius:

$$\mathrm{IR1} = \frac{\sum_k i_k r_k}{\sum_k i_k},$$

where i_k is the intensity of pixel k and r_k is its distance from the object's centroid;

S - the parameter defined by Heydon-Dumbleton, Collins, and MacGillivray (1989) and used by Picard (1991), which is a function of object area (a), core intensity (I_{core}, the sum of the central 3×3 pixels), and the average intensity along the detection isophote (p):

$$\mathrm{S} = \frac{a}{\log[I_{core}/(9 \times p)]}.$$

We chose FOCAS total magnitudes for our standard brightness measure for its decreased sensitivity to the surface brightness threshold relative to aperture or isophotal magnitudes (see Weir, Djorgovski, and Fayyad 1994). The other attributes measure the object's symmetry or compactness in one way or another. FOCAS measures the two listed magnitudes and **IR1** directly, while the other three are easily computed from actual measurements. We tested the use of a few additional object parameters, such as additional image moments, but found that they contributed little additional discriminatory power due to their high correlation with one or more of these parameters. There is always the possibility that future researchers will find that some unconsidered parameter helps result in significantly improved classifications, and the machine learning software is fully capable of incorporating additional new parameters as they are discovered. For now, however, we found that this list is sufficient.

Like previous researchers (*e.g.*, Valdes 1982; Heydon-Dumbleton, Collins, and MacGillivray 1989; Picard 1991), we quickly determined that the distribution of these base-level attributes does not exhibit the required invariance between different regions of a single plate, much less across plates. This was exhibited by the low out-of-sample accuracy of the classifiers we produced by training on these attributes alone. Their variability is also clearly evident when one looks at the distribution of these parameters across or within plates. For example, in Figure 3, we plot the distribution of **log(Area)** vs. **MTot** for two 2048^2 pixel sections of plates J380 and J442. We analyze each plate in image sections of this size (which we call footprints) to help account for variations in image quality across the plate (see Weir *et al.* 1994b for a full discussion of our plate reduction procedure). Note that the stellar loci for these two footprints are nonlinear and do not overlay one another. The implication is that a classifier optimized for one of the images would not only be difficult to construct due to the nonlinearity of the stellar locus, but it would certainly be less than optimal for the other image.

Raw measurements of object shape are inherently sensitive to the local background sky level, seeing, and the pixel blurring induced by the scanning process. We therefore expect these measurements to vary from plate to plate and even footprint to footprint. For any

learning algorithm to be able to produce robust classifiers consistent across a large survey area, different attributes are clearly required.

3.2 Derived attributes

As we discussed in Section 1, the resolution routine of Valdes (1982) provides two extremely powerful classification parameters that, by construction, are very uniformly distributed from image to image. In fact, a preliminary study by Weir and Picard (1991) indicated the significant benefits of using the FOCAS approach to object classification on digitized Schmidt plates. They found that using the PSF-fitting algorithm, one could extend the limiting magnitude of classified Schmidt plate catalogs nearly a full magnitude beyond previous limits achieved using historical approaches.

An essential task in employing the resolution technique, however, is to establish an accurate estimate of the PSF for a given image. Only after this is obtained can the resolution scale and fraction parameters be measured. The problem, therefore, naturally breaks up into two separate steps: (1) star selection, the process of automatically deriving a list of candidate stars for generating an empirical PSF template; and (2) final classification, in which the resolution parameters, possibly along with others, are used for assigning all objects to a particular class.

As previous surveys indicate, certain rather simplistic methods are perfectly adequate for performing accurate star/galaxy separation at bright to moderately faint magnitudes: a method involving PSF-fitting is necessary only when approaching a magnitude or so within the detection limit. One need not approach this limit just to produce lists of stars for empirically estimating the PSF template. Using a straightforward approach similar to ones used for final classification in previous surveys, we were able to develop a technique for robustly selecting candidate PSF stars, up to some limiting magnitude, uniformly within and among plates.

The solution we employ is to fit, on a footprint by footprint basis, the stellar locus within four separate parameter vs. magnitude projections, measuring four new attributes in the form of the distance of each object from the stellar ridge in each dimension. We compute

these so-called 'revised' attributes for the M_{core}, **log(Area)**, **IR1**, and **S** parameters described above. We find that in these new parameter spaces, the line distinguishing stars from galaxies is roughly linear and does not vary much from image to image.

Measuring the distance of an object from the stellar locus first requires the ability to delineate the location of the locus. The method we use for automatically tracking the locus in an attribute vs. magnitude parameter space works by computing a histogram of the attribute value in a set of 0.5^m bins spanning the instrumental magnitude ranges $15.5^m - 21.5^m$ in J_{inst} and $15.5^m - 20.5^m$ in F_{inst}. Objects brighter than the lower magnitude limit are typically saturated and must be classified separately; and one has little hope of forming accurate star lists using this type of method at magnitudes fainter than the upper limit.

Our locus tracking algorithm next computes robust estimates of the mode and width of the histogram for each magnitude bin. These mode values and their error estimates (specified by the widths) are then fit by a fourth or fifth order polynomial as a function of magnitude (see Figure 4). The fit is subtracted from each object, effectively bringing the stellar ridge close to the abscissa on an attribute vs. magnitude plot. To assure an optimal fit to the stellar ridge, the algorithm applies the same fitting and subtraction procedure a second time, this time using a third or higher order polynomial. The optimal orders used to perform the fit in the first and second iterations were found to be very consistent across all DPOSS images and were determined separately for each of the four parameters. These fitting parameters were ultimately hard-coded into the measurement process. Other researchers found it useful to renormalize the new attribute values by the width of the stellar locus. Our tests did not indicate significant variations in the widths of the revised attribute distributions from footprint to footprint, so we eliminated this step.

The distribution of the revised parameters derived for the objects shown in Figure 3 appear in Figure 5. As demonstrated in this example, we find that the distribution of objects in revised attribute space differs little between plates. The same holds true for the other revised attributes we compute, as well.

Along with magnitude and ellipticity, the four revised attributes now form a six-

dimensional parameter space in which we perform star/galaxy separation. To produce our star selector classifier, we trained the decision tree induction software on a set of over a thousand objects which one of us (NW) classified by eye from the digitized scans of plates J380 and J442. Subsequent comparison with several hundred much more reliable classifications obtained from CCD images indicated an error rate of less than 5% in the training list constructed by eye.

The star selector we produced had an error rate of less than 3% percent on an out-of-sample list of objects from the same two plates in the instrumental magnitude range 16.5^m to 19.0^m. Subsequent application of the classifier on independent *J and F* data resulted in lists of candidate stars in this magnitude range which we found to be more than accurate enough for use in constructing the PSF template required by FOCAS resolution. Whereas the typical footprint contains between 3500 to 4500 objects, the star selector returns between 500 and 600 objects in the magnitude range listed above. This list of candidate stars is provided to a FOCAS routine which averages the central nine by nine pixels of each object to form the PSF template.

Armed with the template, one is then able to run the FOCAS resolution routine on each object. As described previously, this routine determines the best-fitting scale (α) and fraction (β) values, which parameterize the fit of a blurred (or sharpened) version of the PSF to each object. The template used to model each object is of the form:

$$t(r_i) = \beta s(r_i/\alpha) + (1-\beta)s(r_i)$$

where r_i is the position of pixel i, α is the broadening (sharpening) parameter, and β is the fraction of broadened PSF. In turn, the resolution parameters are combined with the previous six used for star selection in order to perform final object classification.

4 Classification Results

In the course of processing each plate, the attribute measurement tasks described in the previous section, including revised attribute measurement and star selection, are performed fully automatically, as is the task of final object classification. However, in order to produce the classifiers implemented within the DPOSS reduction programs, we were required at some point to manually produce large samples of classified objects for training and testing purposes. We describe how we produced these training samples below. The same steps would be required of any user who might wish to construct their own, specialized classifier, or to improve upon or monitor the quality of the existing classifiers on future data. We follow this discussion with an examination of the results of applying these classifiers to actual DPOSS data.

4.1 Classifier training

In order to obtain training data for classifying faint objects in DPOSS, especially those too faint for recognition by human inspection of the plates alone, we made use of higher resolution (and narrower field of view) CCD imagery obtained from the Palomar 60" telescope. CCD images are being collected systematically in order to photometrically calibrate the Survey (see Weir, Djorgovski, and Fayyad 1994); however, they serve this very important role in the object classification process as well.

For classification purposes, the obvious advantage of a CCD image relative to a plate is higher resolution and signal-to-noise ratio at fainter levels. By matching a CCD image with the corresponding (small) portion of the plate that it covers, one can determine the classes of objects too faint to classify by eye on the plate. By training learning algorithms to classify these faint objects correctly using the attributes derived from the plate image, SKICAT can conceivably classify objects from the survey that even humans would have difficulty classifying.

The training and test data consisted of objects collected from four different plate fields from regions for which we had CCD image coverage, as well as the by-eye classifications used to construct the star selector described in the previous section. To adequately test

the reliability of the classifier, we divided the data into independent training and test sets from different plates. The F plate training sample totaled 1239 objects from plates F381 and F442, while the J sample consisted of 2563 objects from plates J380 and J342.

We trained the decision tree induction and combining algorithms, O-Btree and RULER, separately on the J and F data in order to produce independent classifiers. As a matter of future research, one might attempt to train a classifier which combines the information available for objects matched in multiple images, particularly in two colors. The results of our training were a list of 84 rules for the F plate classifier and 96 for the J's. Each rule is effectively an "if...then..." statement assigning a class to any object meeting its conditions. For both classifiers, each rule conditions upon anywhere from three to six different parameters. By construction, as described in Section 3.2.2, the rules will generate a unique classification for any object within the training set's multidimensional parameter space.

4.2 Comparisons with training and test data

We tested the classifiers on a sample of 1539 objects from plates F380 and F382 and 589 objects from plates J381 and J382. Testing consisted simply of keeping track of the fraction of objects classified correctly or incorrectly as a function of magnitude. It is noteworthy that for a large fraction of these objects, an astronomer would have difficulty reliably determining their classes by examining the corresponding digitized plate images. As an example, see Figure 6, which depicts a star and galaxy as it appears on a plate and on a CCD. These objects are representative of those with a magnitude at the limit of which we would like to perform accurate star/galaxy separation. We have begun spectroscopic follow-up observations of a sample of the small, faint objects, providing another independent check on our faint classifications.

The accuracy we achieved from applying the classifiers on the training and test DPOSS data sets appears in Tables 1 and 2. We estimate the accuracy by measuring the completeness and contamination of a galaxy catalog formed from the sample data. The training results reflect the in-sample accuracy of the classifier, which is largely irrelevant and included

only for completeness. The test set results are indicative of the accuracy of the classifier on independent data and, therefore, reflect the true quality of the classifier. These results are plotted in Figure 7.

Note that on our test data, we achieve approximately 90% completeness and 10% contamination down to $r \sim 19.6^m$ and $g \sim 20.5^m$, or an equivalent B_J of approximately 21.0^m. This reflects an accuracy rate comparable to what previous surveys attained, but at magnitude levels 0.5^m to 1.0^m fainter. Our limited spectroscopic follow-up observations to date are fully consistent with these results.

Though not listed here, we also computed the results of the J classifier on a test set of data from the same plates on which the classifier was trained. The completeness and contamination closely matched that of the test set from independent plates. Therefore, we can expect the performance of the classifiers to be virtually the same for large catalogs of objects from either the training or test sets of plates. We can help confirm this expectation by comparing the consistency of classifications from plate to plate, as we do below.

We also confirmed the relative importance of the resolution attributes for object classification. When the same experiments were conducted using only the six attributes used in star selection, the results were significantly worse. The error rates jumped above 20% for O-BTree, above 25% for GID3*, and above 30% for ID3 at a magnitude of approximately 20.0^m in g. The respective sizes of the trees grew significantly as well. This clearly demonstrates that although learning algorithms improve matters considerably by allowing one to optimally and objectively make use of multiple parameters in the classification process, the choice of parameters is still of first order importance.

4.3 Comparisons in plate overlaps

The tests described above indicate an overall classification accuracy of approximately 90% at a magnitude of approximately 19.6^m in r and 20.5^m in g. If we assume that the probability of an object being correctly classified is independent from plate to plate, this would imply a consistency of classification of approximately 82%. This is the sum of the probabilities of both classifications being correct (0.9^2) or incorrect (0.1^2). Measuring

the consistency of classifications from plate to plate across many different plates provides some measure of the uniformity of plate classification accuracies, if not their actual levels of accuracy. In Tables 3, 4, and 5 we list the consistency of object classification for the large number of objects measured in each pair of overlapping plates of the same color and overlapping plates of the same field but different color. Note that at each magnitude level, the consistencies are in line with the accuracies listed in the previous section assuming independent classifications.

Also notice that the consistency of the classifications between the pairs of plates on which the classifiers were trained (F381/F442 and J380/J442) does not significantly differ from the consistency of other measured pairs. This corroborates the notion that the classification accuracy for these plates as a whole is no better or worse than that for the test plates, despite the fact that the classifiers were trained exclusively on objects from those plates. In this sense, the classifiers are truly robust.

5 Initial Experiments with Unsupervised Classification

We have also begun exploring the application and implementation of unsupervised classification techniques like Autoclass (Cheeseman *et al.* 1988) for the purpose of automated machine discovery. Unlike the so-called supervised methods of classification that we have described so far, where the computer learns how to distinguish user-specified classes within the data, unsupervised classification consists of the computer identifying the statistically significant classes within the data itself. For example, one could employ this type of method to try to systematically detect new classes of objects within astronomical catalogs.

Our own initial experiments in applying Autoclass to DPOSS appear to confirm the validity and usefulness of this approach. After supplying Autoclass with the eight-dimensional feature vectors from a sample of several hundred objects from our four fields, it analyzed the distribution of the objects in this parameter space and suggested four distinct classes within the data. Representative objects from these four classes are presented in Figure 8. Visually, the classes seem to divide into stellar objects, stellar-like objects with a low surface brightness halo, and diffuse or irregular objects with and without a central core. Its success at distinguishing these apparently physically relevant classes based just upon eight image parameters suggests that far richer and innovative results may be in store when ones matches multiple catalogs together, increasing the informational dimensionality of the data set manifold.

6 Concluding Remarks

Through the careful selection and construction of object attributes, and the application of machine learning to derive sets of rules based upon them, we have been able to achieve high rates of classification accuracy at levels up to a magnitude fainter than in previous Schmidt surveys. By examining a set of four fields in two colors, we have verified that galaxy catalogs produced from DPOSS using this technique appear to be consistently complete and contaminated across multiple plates. In fact, in testing our classifiers on completely independent plate data, we found them to produce 90% complete galaxy catalogs down to an equivalent B_J magnitude of approximately 21.0^m. There is no *a priori* reason why, without any further work, these very same classifiers should not result in exactly the same accuracy rates for all future high Galactic latitude DPOSS plates. However, we note that by accumulating more and better overlapping CCD and plate data, one may be able to train classifiers that are able to generalize even better.

A significant additional benefit of the classification approach we describe is that it easily generalizes to the construction of any number of object classifiers for any purpose in the future. Provided the astronomer is able to construct a suitably large enough sample of objects for both testing and training, the same technology may be applied for a wide variety of scientific purposes. To facilitate the construction of such sets, we have implemented a tool within SKICAT that allows the user to display individual objects from a DPOSS plate scan and assign a classification to each. One may also, as we have done, use the extensive object matching technology within SKICAT to retrieve attributes from one set of catalogs (*e.g.*, plates) and classifications from their matched counterparts in others (*e.g.*, CCDs). It is our hope that with the availability of tools such as SKICAT and Autoclass, and the demonstrated scientific value they add, such advanced data analytic techniques may see more widespread use in the future.

This work was supported at Caltech in part by NASA AISRP contract NAS5-31348, the Caltech President's fund, and NSF PYI award AST-9157412, and at JPL under a contract with NASA. The POSS-II is partially funded by grants to Caltech from the Eastman Kodak Co., the National Geographic Society, the Samuel Oschin Foundation, NSF grants AST 84-08225 and AST 87-19465, and NASA grants NGL 05002140 and NAGW 1710. We acknowledge the efforts of the POSS-II team at Palomar, the scanning team at STScI, and the SKICAT team at JPL, most especially Joe Roden.

References

Cheeseman, P. *et al.*1988, in *Proc. Fifth Machine Learning Workshop, San Mateo*, Morgan Kaufmann, 54.

Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I., and Laidler, V. 1992, *BAAS*, **24**, 750.

Ellis, R. 1987, in *Observational Cosmology, IAU Symp. 124*, ed. A. Hewitt, G. Burbidge, and L. Z. Fang, (Dordrecht: Reidel), 367.

Fayyad, U. 1991. Ph.D. thesis, EECS Dept. The University of Michigan.

Fayyad, U., Doyle, R., Weir, N., and Djorgovski, S. 1992, in *Proceedings of the ML-92 Workshop on Machine Discovery (MD-92), Aberdeen, Scotland*, ed. J. Zytkow, Morgan Kaufmann, 117.

Fayyad, U. and Irani, K. 1990, in *Proceedings of the Eighth National Conference on Artificial Intelligence AAAI-90, Boston, MA.*

Fayyad, U. and Irani, K. 1992, in *Proceedings of the Tenth National Conference on Artificial Intelligence AAAI-92, San Jose, CA.*

Fayyad, U. and Irani, K. 1993, in *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, France*, Morgan Kauffman, in press.

Finney, D., Latscha, R., Bennett, B., and Hsu, P. 1963, *Tables for Testing Significance in a 2x2 Contingency Table*, (Cambridge: Cambridge University Press).

Hertz, J., Krogh, A., and Palmer, R. 1991, *Introduction to the Theory of Neural Computation*, (Redwood City, CA: Addison-Wesley).

Heydon-Dumbleton, N. H., Collins, C. A., and MacGillivray, H. T. 1989, *MNRAS*, **238**, 379.

Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I., and Sturch, C. 1992, *BAAS*, **24**, 741.

Maddox, S., Sutherland, W., Efstathiou, G., and Loveday, J. 1990, *MNRAS*, **243**, 692.

Odewahn, S., Stockwell, E., Pennington, R., Humphreys, R., and Zumach, W. 1992, *AJ*, **103**, 318.

Picard, A. 1991. Ph.D. thesis, California Institute of Technology.

Quinlan, J. 1986, in *Machine Learning, Vol. 1, No. 1.*

Quinlan, J. 1990, in *Machine Learning: An Artificial Intelligence Approach Vol. III, San Mateo, CA*, ed. Y. Kodratoff and R. Michalski, Morgan Kaufmann.

Reid, I. *et al.*1991, *Publ. Astron. Soc. Pac.*, **331**, 465.

Reid, N. and Djorgovski, S. 1993, in *Sky Surveys: Protostars to Protogalaxies*, ed. B. T. Soifer, A.S.P. Conf. Ser. #43, 125.

Sebok, W. 1979, *AJ*, **84**, 1526.

Valdes, F. 1982, *SPIE Proc. on Instrumentation in Astronomy IV*, **331**, 465.

Weir, N., Djorgovski, S., and Fayyad, U. 1994, *AJ*, submitted.

Weir, N., Djorgovski, S., Fayyad, U., Smith, J., and Roden, J. 1994a, in *Astronomy From Wide-Field Imaging, IAU Symp. #161*, ed. H. MacGillivray *et al.*, Dordrecht: Kluwer, 205.

Weir, N., Fayyad, U., Djorgovski, S., and Roden, J. 1994b, in prep.

Weir, N. and Picard, A. 1991, in *Digitised Optical Sky Surveys*, ed. H.T. MacGillivray and E.B. Thomson, Dordrecht: Kluwer Academic Publisher, 225.

□

Figure 1: In this sample decision tree, one starts at the top node (root), following the appropriate path to a final leaf (class) based upon the truth of the assertion at each node.

□

Figure 2: A portion of a much larger actual decision tree generated by the O-Btree algorithm for performing star/galaxy classification. The interval appearing above each node indicates the range in value of the attribute specified in the node above that an object must meet for it to pass along that branch. The dark branches lead to actual classifications. The number in parentheses within each leaf indicates the number of training examples classified correctly at that node.

Figure 3: The distribution of **log(Area)** vs. **MTot** in sections of plates J380 and J442. Note that the stellar locus is nonlinear and different for each plate. The locus shows similar variance even within plates.

Figure 4: The **log(Area)** attribute and the locus fit to its distribution before each iteration of the locus subtraction algorithm.

Figure 5: The distribution of the revised **log(Area)** vs. instrumental magnitudes in plates J380 and J442 after the two-step locus fitting and subtracting process.

Figure 6: The top two images are from the scan of plate J442. Each object has a g magnitude of approximately 20.0. The bottom two images are of the same objects but from CCD frames. Our classifier correctly classified the left object as a star and the right as a galaxy, despite their almost indistinguishable appearance on the plate. The higher quality CCD images allowed us to provide reliable classifications to these objects which we would otherwise be unable to use in classifier training or testing.

Figure 7: The accuracy of our star/galaxy separation technique is depicted by the completeness (fraction of galaxies classified as such) and contamination (fraction of non-galaxies classified as galaxies) measured within our test set of data.

Figure 8: Each row consists of representative objects from one of the four classes discovered in the DPOSS data by Autoclass. It appears one can relate each type to physically, not just statistically, distinct classes of objects.

	Training Set		Testing Set	
r mag	completeness	contamination	completeness	contamination
16.56	1.000	0.000	0.857	0.000
16.96	1.000	0.000	0.833	0.062
17.43	0.938	0.032	0.966	0.034
17.95	0.979	0.000	0.885	0.042
18.50	0.966	0.012	0.878	0.133
19.07	0.969	0.054	0.929	0.103
19.64	0.985	0.043	0.895	0.094
20.21	0.948	0.081	0.906	0.247
20.75	0.950	0.102	0.902	0.260

Table 1: The completeness (fraction of galaxies classified as such) and contamination (fraction of non-galaxies classified as galaxies) for the samples of F plate objects used for classification training and testing. The training samples are from plates F381 and F442. The testing samples are from plates F380 and F382.

	Training Set		Testing Set	
g mag	completeness	contamination	completeness	contamination
16.68	1.000	0.000	***	***
17.17	0.857	0.077	***	***
17.67	0.935	0.033	1.000	0.000
18.18	0.956	0.030	1.000	0.091
18.69	0.989	0.021	1.000	0.050
19.21	0.963	0.037	0.966	0.097
19.73	0.954	0.019	0.925	0.098
20.25	0.964	0.024	0.892	0.065
20.77	0.891	0.039	0.861	0.151
21.30	0.806	0.167	0.796	0.204
21.81	0.848	0.200	0.774	0.250

Table 2: The completeness and contamination for the samples of J plate objects used for classification training and testing. The training samples are from plates J380 and J442. The testing samples are from plates J381 and J382. Too few objects of bright magnitude were available to provide a statistically significant test below $g = 17.5^{m}$.

r mag	F380/F381 (8682)	F380/F442 (1357)	F381/F382 (9246)	F381/F442 (3865)	Average
16.23	0.933	0.947	0.880	0.886	0.912
16.56	0.952	0.967	0.957	0.964	0.960
16.96	0.952	0.839	0.968	0.971	0.932
17.43	0.972	0.870	0.937	0.925	0.926
17.95	0.964	0.983	0.957	0.984	0.972
18.50	0.941	0.919	0.961	0.969	0.948
19.07	0.893	0.899	0.921	0.875	0.897
19.64	0.825	0.855	0.826	0.852	0.840
20.21	0.746	0.773	0.761	0.749	0.757
20.75	0.750	0.738	0.681	0.743	0.728
21.25	0.746	0.753	0.718	0.775	0.748

Table 3: The fraction of objects classified consistently as a function of magnitude in the overlap of the listed plates. These rates are consistent with the accuracies listed in Table 1. The number of objects tested in each overlap is listed below the field names.

g mag	J380/J381 (8553)	J380/J442 (1418)	J381/J382 (9659)	J381/J442 (3850)	Average
15.73	0.548	0.533	0.623	0.538	0.561
16.20	0.913	0.846	0.899	0.860	0.880
16.68	0.977	1.000	0.954	1.000	0.983
17.17	0.976	0.964	0.975	0.962	0.969
17.67	0.962	0.972	0.979	0.972	0.971
18.18	0.975	1.000	0.985	0.964	0.981
18.69	0.958	0.984	0.970	0.962	0.968
19.21	0.915	0.927	0.942	0.890	0.918
19.73	0.857	0.911	0.881	0.874	0.881
20.25	0.755	0.812	0.780	0.820	0.792
20.77	0.688	0.759	0.717	0.690	0.713
21.30	0.673	0.671	0.717	0.665	0.681
21.81	0.736	0.701	0.706	0.661	0.701

Table 4: Same as Table 3, but for J plates.

$\overline{r+g}$ mag	Field 380 (7096)	381 (8456)	382 (7660)	442 (7900)	Average
15.95	0.656	0.795	0.777	0.986	0.803
16.45	0.983	0.953	0.992	0.953	0.970
16.95	0.948	0.982	0.962	0.970	0.966
17.45	0.977	0.964	0.966	0.951	0.964
17.95	0.981	0.964	0.986	0.948	0.970
18.45	0.940	0.952	0.967	0.950	0.952
18.95	0.926	0.926	0.928	0.943	0.931
19.45	0.866	0.859	0.901	0.885	0.878
19.95	0.804	0.768	0.818	0.787	0.794
20.45	0.729	0.682	0.763	0.713	0.722
20.95	0.718	0.681	0.684	0.690	0.693
21.45	0.733	0.736	0.672	0.678	0.705

Table 5: The fraction of objects classified consistently as a function of average g and r magnitude in the overlap of the J and F plates covering the indicated fields. The number of objects tested in each overlap is listed below the field names.

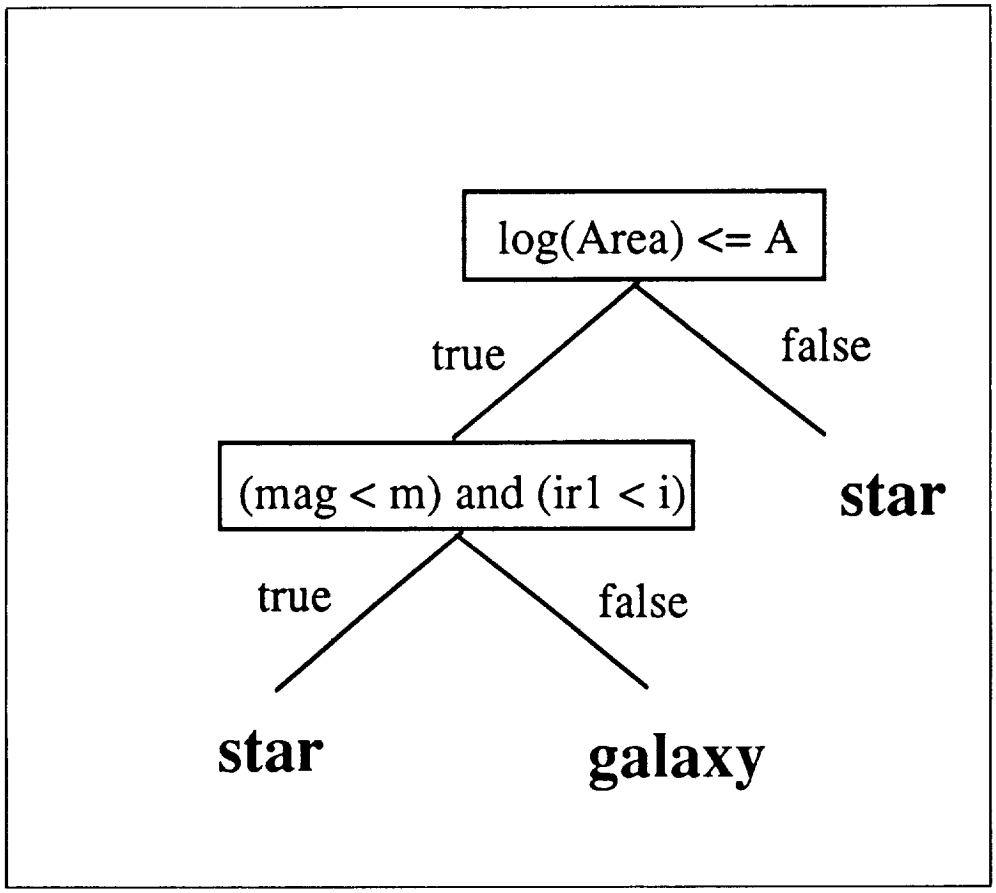
log(Area) <= A
true
false
(mag < m) and (ir1 < i)
star
true
false
star
galaxy

Figure 1:

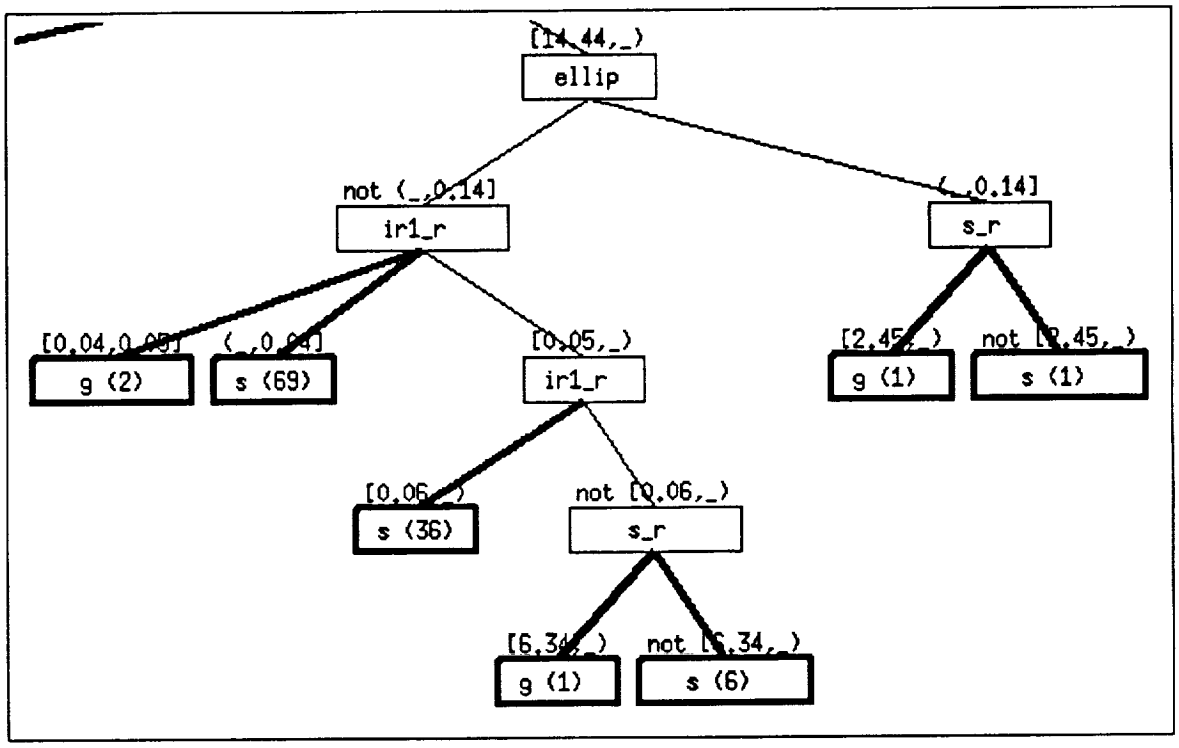
[14.44,_)
ellip
not (_,0.14]
ir1_r
(_,0.14]
s_r
[0.04,0.05]
g (2)
(_,0.04]
s (69)
[0.05,_)
ir1_r
[2.45,_)
g (1)
not [2.45,_)
s (1)
[0.06,_)
s (36)
not [0.06,_)
s_r
[6.34,_)
g (1)
not [6.34,_)
s (6)

Figure 2:

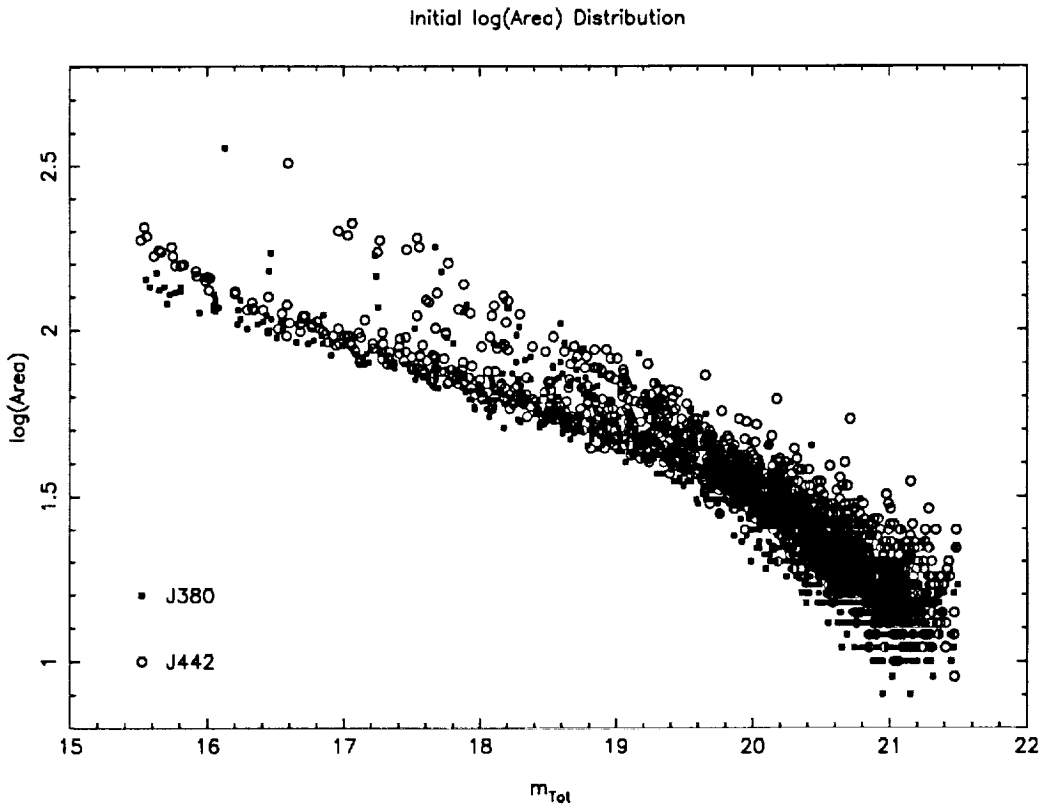
Initial log(Area) Distribution
log(Area)
2.5
2
1.5
1
J380
J442
15
16
17
18
19
20
21
22
m_{Tot}

Figure 3:

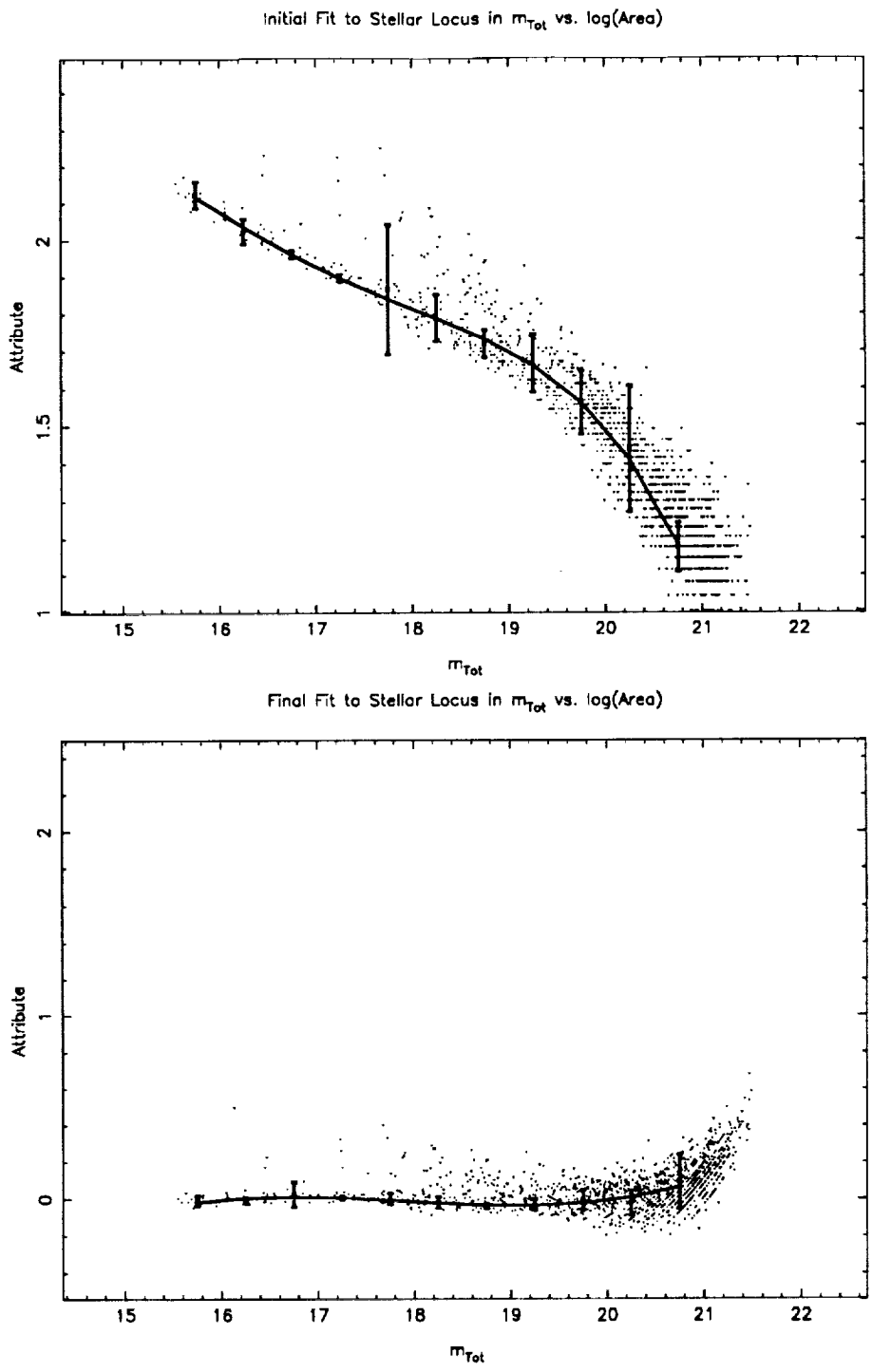
Initial Fit to Stellar Locus in m_{Tot} vs. log(Area)
Attribute
m_{Tot}
Final Fit to Stellar Locus in m_{Tot} vs. log(Area)
Attribute
m_{Tot}

Figure 4:

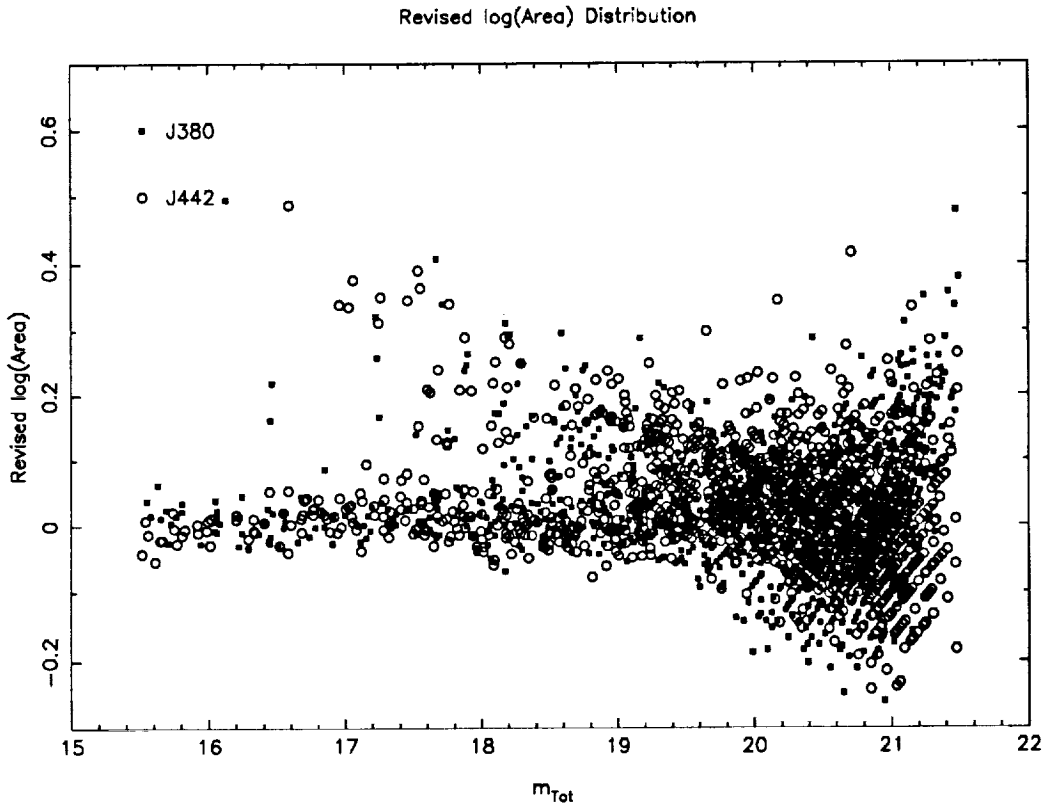
Revised log(Area) Distribution
J380
J442
Revised log(Area)
0.6
0.4
0.2
0
−0.2
15
16
17
18
19
20
21
22
m_{Tot}

Figure 5:

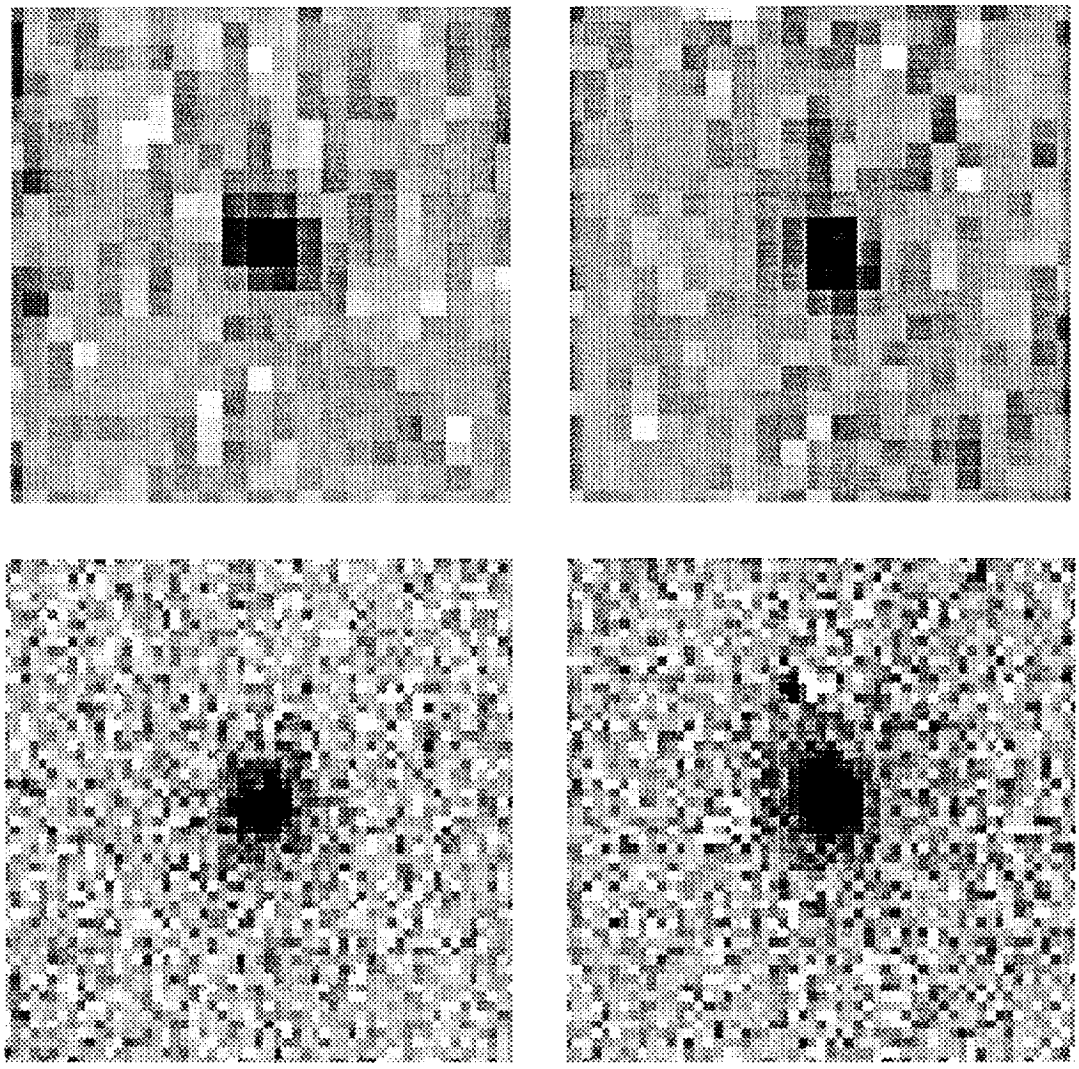

Figure 6:

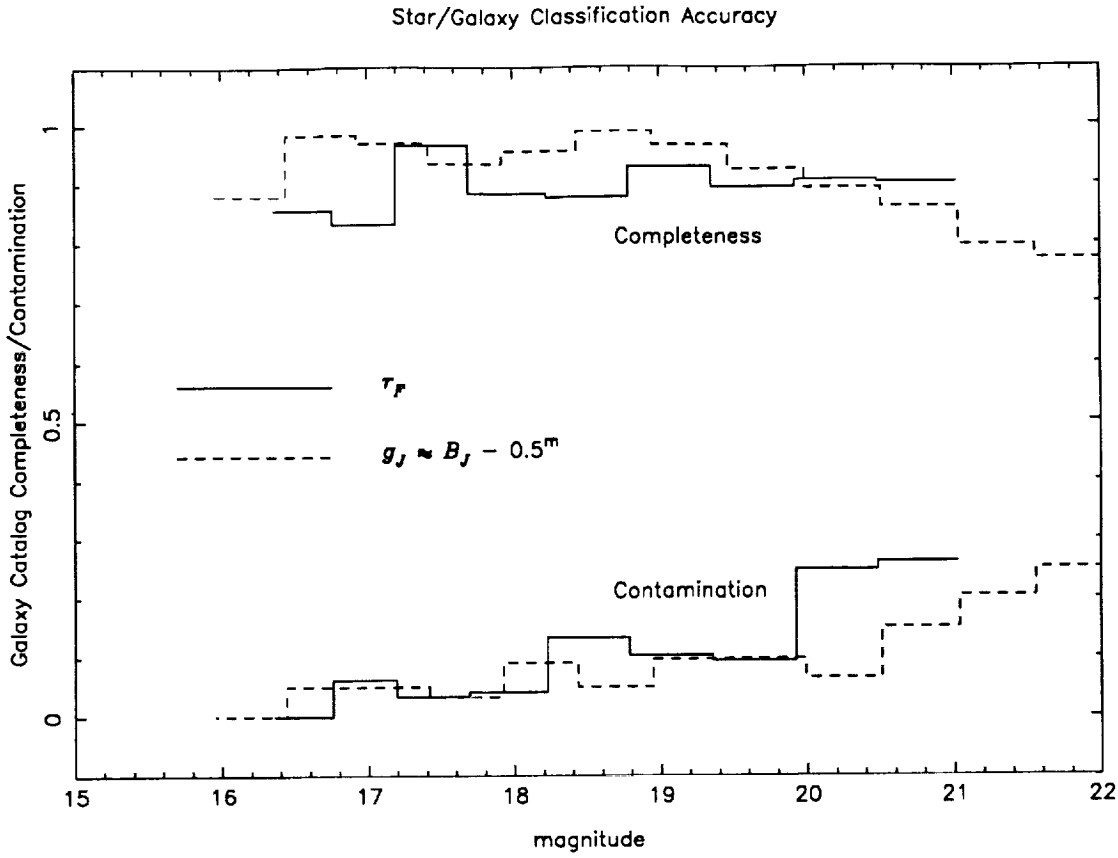
Star/Galaxy Classification Accuracy
Galaxy Catalog Completeness/Contamination
0
0.5
1
Completeness
r_F
$g_J \approx B_J - 0.5^m$
Contamination
15
16
17
18
19
20
21
22
magnitude

Figure 7:

Attachment C:

The initial galaxy counts and photometry tests

From a paper submitted to the *Astronomical Journal*, September 1994

Current status: Accepted for publication with very minor modifications

Expected publication date: mid-1995

Initial Galaxy Counts from Digitized POSS-II

Nicholas Weir[†]
S. Djorgovski[†]
Usama M. Fayyad[‡]

[†] Palomar Observatory
California Institute of Technology 105-24
Pasadena, CA 91125
weir@fritz.caltech.edu
george@deimos.caltech.edu

[‡] Jet Propulsion Laboratory
California Institute of Technology 525-3660
Pasadena, CA 91109
fayyad@aig.jpl.nasa.gov

Submitted to *The Astronomical Journal*:

Received:

Accepted:

Running Title: *Galaxy Counts from POSS-II*

Abstract

In our first analysis of the Digitized Second Palomar Observatory Sky Survey (DPOSS), we examine galaxy counts on an overlapping set of four survey fields near the North Galactic Pole, in both the J and F passbands. Through detailed simulations of a subset of these data, we were able to analyze systematic aspects of our detection and photometric procedures, as well as optimize them. We discuss how we calibrate the plate magnitudes to the Gunn-Thuan g and r photometric system using CCD sequences obtained in a program devoted expressly to calibrating DPOSS. Our technique results in an estimated plate-to-plate zero point standard error of under 0.10^m in g and below 0.05^m in r, for J and F plates, respectively. Using the catalogs derived from these fields, we compare our differential galaxy counts in g and r with those from recent Schmidt plate surveys as well as predictions from evolutionary and non-evolutionary (NE) galaxy models. While we find some significant differences between our measurements and others, particularly at the bright end, we find generally good agreement between our counts and recent NE and mild evolutionary models calibrated to consistently fit bright and faint galaxy counts, colors, and redshift distributions. The consistency of our results with these predictions provides additional support to the view that very recent ($z < 0.1$) and exotic galaxy evolution, or non-standard cosmology, may not be necessary to reconcile these diverse observations with theory.

keywords: galaxy counts, galaxy evolution, large scale structure, photometry, sky surveys

1 Introduction

1.1 Background

Counts of galaxies per magnitude per square degree, $A(m)$, are among the most time-honored observations in extra-galactic astronomy. At face value, however, they are also among the most uninterpretable, as they represent the projection and convolution of so many different physical effects. Galaxy counts are traditionally modeled assuming one or more galaxy luminosity functions (LFs) for a few different galaxy types. Each type is characterized by a model spectral energy distribution (SED). One evolves these galaxy distributions and spectra with time over a grid of different luminosity evolution models and cosmologies, in turn, 'observing' and counting the galaxies in a given bandpass after applying the appropriate K-correction. In comparing actual measurements to predictions, the standard null hypothesis has been a model with no evolution (NE) and a cosmology of, *e.g.*, $q_o \sim 0$ and $H_o \sim 50 \mathrm{kms}^{-1}/\mathrm{Mpc}$. Such a set of parameters increases the volume element and ages of present-day galaxies relative to higher q_o and H_o models, providing larger numbers of faint counts and decreasing predicted evolutionary effects at a given redshift.

In the context of such models, recent galaxy surveys indicate an excess of blue counts by a B_J of 19.0^m, or a mean redshift of only $z \approx 0.1$, increasing substantially with magnitude (Maddox *et al.* 1990; Tyson 1988). Red counts also indicate an excess, though not of the same degree as the blue (Koo and Kron 1992). Perhaps more interestingly, however, the counts at brighter red magnitudes display very large variations between and even within individual surveys (Sebok 1986; Picard 1991a). Near-infrared (K band) counts, on the other hand, are less steep and appear to be more consistent with NE models (Cowie *et al.* 1990; Djorgovski *et al.* 1994), as does the apparent redshift distribution of galaxies (Broadhurst *et al.* 1988; Colless *et al.* 1990; Lilly and Gardner 1991).

A variety of physical effects have been proposed to account for these observations, including dramatic galaxy evolution at low redshift (Maddox *et al.* 1990), significant evolution of LF shape (Broadhurst *et al.* 1988), density evolution through galaxy mergers

and/or the disappearance of entire populations (Cowie, Songaila, and Hu 1991), large inhomogeneities in the number density of galaxies on scales of $(125h^{-1}\mathrm{Mpc})^3$ (Picard 1991a), or even a non-zero cosmological constant (Fukugita *et al.* 1990). Any of these effects represents a fundamental, and largely *ad hoc*, revision of current standard models of galaxy evolution, cosmogony, or cosmology. In the spirit of Occam's razor, it is natural that one should fully explore the consistency of the data with less intricate models before embracing any of these alternatives.

With this goal in mind, Koo and Kron (1992) investigated the possibility that when uncertainties in the observations and models are more fully taken into account, there is no need for exotic evolutionary scenarios to simultaneously account for the observed number counts, colors, and redshift distribution of galaxies. Their model significantly differs from most others not only in that it attempts to fit all these data at once, but in how they add the flexibility in the model to do so. They claim that in order to adequately represent the variety and range of colors observed at all magnitudes, one must allow for a rich variety of galaxy spectral types: *i.e.,* no small number of classes dominates. Indeed, given the wide variety of observed galaxy spectral energy distributions (SEDs), there is no obvious or agreed upon standard breakdown of fundamental SED types. Consequently, in their model, Koo and Kron try to fit the data by substituting one form of model complexity (more galaxy types) for other, more traditional ones (evolution of LFs or non-standard cosmology). They assert that allowing for a finer disaggregation of present-day galaxy types is consistent, even justified, by direct observation, unlike these other methods of adjusting the models to fit the data. Through trial and error, Koo and Kron adjust the non-evolving LF of each of their specified classes so as to best fit the data. Adopting a $q_o = 0$, they claim to establish a proof of concept, by producing a set of predictions which match the observations reasonably well with the addition of only relatively mild galaxy evolution.

Koo, Gronwall, and Bruzual (KGB,1993a) took the Koo and Kron model one step further, using a non-negative least squares optimization algorithm to find the best-fitting set of LFs for a set of eleven non-evolving galaxy spectral classes. Each class is characterized

by age and star formation rate. They assume $q_o = 0.05$ and $H_o = 50$ km sec^{-1}/Mpc^{-1}. With these 'optimal' LFs, their NE model matches the observations significantly better than the Koo and Kron model, indicating even less need for rapid or complicated evolution of galaxies or non-standard cosmology to explain the data.

Implicit in the KGB model is the assumption that our knowledge of the present-day LF of galaxies, especially at a fine level of color-class disaggregation, is sufficiently uncertain that it makes sense to float these, rather than any other, model parameters freely. Of course there are independent derivations of color-integrated and color-dependent LFs (*e.g.*, Loveday *et al.* 1992; Eales 1993; Lonsdale and Chokshi 1993; Metcalfe *et al.* 1991), which could provide at least some measure of 'reality check' on the LFs generated by the KGB model. Unfortunately, the other derivations display a significant amount of internal discrepancy. Consequently, allowing for the apparent uncertainty in alternative estimates, KGB claim that their LF solutions are generally consistent with independent observations.

KGB fit their model to a combination of observations from many heterogeneous sources. They note that differences in magnitude zero points, detection efficiency, measurement procedure, and random and systematic photometric errors are all important, but are best left for the model to account for rather than 'correcting' the data to some standard form in advance. However, to accurately model these effects requires detailed knowledge of the experimental procedure at a level seldom even realized by the observers themselves. Therefore, in their analysis to date, KGB do not explicitly take these effects for different surveys into account. Nonetheless, it is these uncertainties, in the random and systematic effects characterizing each survey, that fundamentally prevent us from more usefully constraining the models. Uniform and better understood data, even if not of significantly higher quality, are crucial.

1.2 The Digitized Second Palomar Observatory Sky Survey (DPOSS)

In 1985, the Oschin Schmidt 48-inch telescope at Palomar was dedicated to work on the Second Palomar Observatory Sky Survey (POSS-II, Reid *et al.* 1991). This photographic survey was prompted by the requirements of the space observatories, notably IRAS and

HST, as well as the general desire to provide a newer epoch survey of the northern sky to complement both the original POSS and the recent, higher-quality SERC/ESO surveys of the southern skies. POSS-II, which is more than 60% complete as of August, 1994, will eventually cover 894 fields spaced 5° apart in three passbands: blue (IIIa-J + GG 395), red (IIIa-F + RG610), and near-infrared (IV-N + RG9). The typical limiting magnitudes for point sources in the corresponding J, F, and N bands are 22.5^m, 21.5^m, and 19.5^m, respectively.

While the photographic survey is still under way, ST ScI and Caltech have already begun a collaborative effort to digitize the complete set of plates (Djorgovski *et al.* 1992; Lasker *et al.* 1992; Reid and Djorgovski 1993). So far, only a subset of the J, F, and N plates have been scanned and processed. Both the photographic survey and the plate scanning are estimated to be > 90% complete circa 1997. The resulting data set, the Palomar-STScI Digital Sky Survey (DPOSS), will consist of $\sim$ 3 TB of pixel data: $\sim$ 1 GB/plate, with 1 arcsec pixels, 2 bytes/pixel, 20340^2 pixels/plate, for all survey fields in all three colors. ST ScI will provide an astrometric solution for each plate accurate to within approximately 0.5 arcsec RMS over scales less than a degree. In conjunction with the plate survey, we are also conducting an intensive program of CCD calibrations using the Palomar 60-inch telescope, using the Gunn-Thuan *gri* bands. These CCD images serve both for magnitude zero-point calibration and object classification purposes. The plate scans, when complete, will be the highest quality set of digital images covering the entire northern sky produced to date.

In order to make most efficient use of these data, and to generally facilitate the exploitation of POSS-II, Caltech Astronomy and the JPL Artificial Intelligence Group have been engaged in a collaborative effort to integrate state-of-the-art computing methods for the scientific utilization of DPOSS. The traditional means of extracting useful information from imaging surveys is through the construction of object catalogs. Thanks to developments in the fields of pattern recognition and machine learning, in addition to raw computing power, it is now possible to reliably construct such catalogs objectively and automatically with a higher degree of accuracy than ever before. The result of our joint effort is the Sky

Image Cataloging and Analysis Tool (SKICAT), a suite of programs designed to facilitate the maintenance and analysis of astronomical surveys comprised of multiple, overlapping images.

1.3 The Palomar Northern Sky Catalog (PNSC)

The result of applying SKICAT to DPOSS will be the Palomar Northern Sky Catalog (PNSC), which when completed, is expected to contain $\gtrsim 5 \times 10^7$ galaxies, and $\gtrsim 2 \times 10^9$ stars, in three colors (photographic JFN bands, calibrated to CCD gri system), down to the limiting magnitude equivalent of $B \sim 22^m$, with star-galaxy classifications $\sim 90 - 95\%$ accurate down to the equivalent of $B \sim 21^m$. The catalog will be continuously upgraded as more calibration data become available. It will be made available to the community via computer networks and/or suitable media, probably in installments, as soon as scientific validation and quality checks are completed. Analysis software (parts of SKICAT) will also be freely available.

A small portion of the PNSC covering a region near the North Galactic Pole is already complete, providing an early indication of the scientific potential of the full catalog. In this paper, we report on the first detailed analyses performed using these data, in the form of galaxy counts in the J and F passbands. In short, we find the data to be of high enough quality and sufficiently well understood to provide useful new constraints to galaxy evolution models. In addition, their consistency with existing NE models provides yet more evidence that elaborate evolutionary scenarios or non-standard cosmology may not be necessary to account for the observations.

The single greatest known source of systematic error in our measured number counts is uncertainty in the instrumental to calibrated-magnitude transformation. In brief, the procedure we use to calibrate the survey is first to adjust each plate's instrumental magnitudes by an offset to match a survey-wide instrumental system. Next we apply a linear transformation to convert the survey instrumental J and F magnitudes to the Gunn-Thuan g and r system. For analysis purposes, we restrict our galaxy catalogs to $16^m < g < 20.5^m$ and $16.5^m < r < 19.6^m$, so as to remain within the well-calibrated, non-saturated, and

well-classified ($\sim$ 90% complete and $\sim$ 10% contaminated) portion of each catalog. In this magnitude range, we estimate that the systematic plate-to-plate RMS error in zero point offsets are under 0.10^m in g for J plates and below 0.05^m in r for F plates.

In the section that follows, we provide a more complete description of the plate and CCD data and the measurement procedures used in our analysis. Section 3 describes the methodology and consistency of our technique for photometrically calibrating the plates. In Section 4, we compare our counts with those of other recent Schmidt plate surveys in addition to theoretical models. In the final section, we discuss these results.

2 The Data

Our survey is derived from four POSS-II survey fields measured in both F and J passbands. In addition, we have obtained extensive CCD coverage of small fields within these plates. Below we provide characteristics of the photographic and digitized plate data, as well as the methods we used for detecting, measuring, and classifying plate objects. This is followed by a description of the CCD data and our measurement procedures for them.

2.1 Plate data

2.1.1 Photographic plates

The four POSS-II fields used in this study (numbers 380, 381, 382, and 442) were chosen for their proximity to the North Galactic Pole, where many previous galaxy surveys have been performed, and because they were the first digitized plates available in two colors and for which we had CCD coverage. The fields are depicted in Figure 1. Also noted in the figure are the locations of the CCD sequences obtained within these fields, indicated by the number of the Abell cluster on which they were centered (*e.g.*, A1694) and/or the CCD field number (*e.g.*, F6) from an *ad hoc* numbering system we adopted for our CCD fields.

The plate number, center location, approximate photographic sky density, exposure time, sky transmission quality, grade assigned to, and estimated limiting magnitude in g or r of each of our survey plates appears in Table 1. The grade reflects the quality of the plate in terms of depth, seeing, number of artifacts (*e.g.*, plane trails), etc., as judged by the POSS-II quality control staff. A and B-grade plates are automatically accepted in the POSS-II, while C-grade and lower observations are typically repeated. What we term J plates actually result from the combination of Kodak III-aJ emulsion with the GG395 filter, whereas F plates are from III-aF emulsion combined with the RG610 filter.

Each Oschin Schmidt plate is 14 inch square in size, corresponding to a $6.6° \times 6.6°$ field of view. The dashed circles in Figure 1 centered within each plate field have a diameter of 6° and enclose the relatively unvignetted portion of each plate. Tritton (1983) measured

the vignetting function of the U.K. Schmidt Telescope, which should be similar to that for the Oschin, and found that the vignetting correction was at most 0.03^m at a radius of 3°, rising to a level of 0.25^m in the plate corners. DeCarvalho (1994, priv. comm.) has begun a study of the vignetting function of POSS-II plates using DPOSS scans and verifies these results.

Tinney (1993), in his analysis of POSS-II F and N plates, was unable to detect vignetting effects in his catalogs, which were restricted to a 3° radius. We also confine our photometric analysis to objects within this radius (also avoiding the sensitometry spots) on each plate, with a minor exception related to field 442 noted below. This restriction would have to be relaxed in order to cover a continuous solid angle of sky using multiple plates. However, to use these data reliably would require an empirical estimate of the actual vignetting function for the Oschin Schmidt, which we will only be able to measure when a larger number of digitized plates are available. We therefore restrict, for the time being, our analysis to the unaffected portions of each plate until such an empirical correction is obtained. We also excluded regions surrounding bright stars, so as to avoid contaminating our catalogs with artifacts mistakenly classified as galaxies, or true galaxies with very poorly measured magnitudes due to stellar contamination.

The plate holder on the Oschin Schmidt is nitrogen flushed during each exposure to help assure uniform hypersensitization across the plate. In their photometric analysis of U.K. Schmidt plates, the APM group (Maddox, Efstathiou, and Sutherland 1990) found that plates observed in this fashion suffered much less variation in response across the field. Unfortunately, only eight of their plates were observed using this method; consequently, they had to go to significant effort to remove this large source of field variation in their survey. All of the POSS-II plates were obtained using nitrogen flushing.

Additional, non-vignetting field variations at the level of a few percent of sky are still present within individual DPOSS images and show up most clearly when analyzing binned versions of the plate scans. However, without a sufficient number of plate scans in hand, it is difficult to determine whether the observed sky background variations are additive or multiplicative in nature, due to zodiacal light, uneven emulsions, uneven hypering or

developing, or to a limited degree, actual Galactic (extragalactic?) sky background variations on scales less than a degree. For our analysis, like that of Tinney (1993) and Picard (1991b), we do not correct for these effects, preferring to wait until we have better understanding of their origin before taking them into account. Instead, we verify below that our plate-to-plate consistency, even while ignoring these effects, is within acceptable limits for our scientific purposes. Nonetheless we note that a better determination of the source of these background variations may be an interesting subject of future research using DPOSS.

2.1.2 Digitized scans

The plate scan data provided by ST ScI are in the form of images $23,040 \times 23,040$ pixels in size, scaled in arbitrary photographic density units. Each pixel is one square arcsecond in size with a dynamic range of two bytes. Each scan includes an image of the 16 sensitometry spots that appear in the southwest corner of each POSS-II plate. The first step in reducing the digitized plate data is to fit a characteristic curve, or so-called 'HD' curve, to these spot levels, providing a density to intensity transformation for the entire plate.

The mathematical formula we use to fit the measured plate densities (D) to relative intensities (I) is:

$$\log I = \frac{P(D)}{(D_S - D) \times (D_T - D)} \tag{1}$$

where $P(D)$ is a polynomial function of the density, and the saturation and toe densities, D_S and D_T, are those corresponding to fully exposed and unexposed portions of the plate, respectively. An example of such a fit for plate F442 appears in Figure 2. The polynomial coefficients, together with the toe and saturation values, establish the conversion applied to each pixel value.

There is a long history to efficiently modeling the HD curve. The method employed by ST ScI (Russel *et al.* 1990) in constructing their Guide Star Catalog, for example, involves a more complicated formula and averaging many plates together. By their own admission, however, they find the more complicated expression to be overkill for the linear part of the curve of most interest. In addition, we found considerable variation of the curve among different plates, requiring independent fits. We find the instrumental magnitudes

resulting from our HD fits to be extremely consistent from plate to plate, in the sense of only requiring a single zero point offset to match them. This provides, in our opinion, the most important test of the validity of our linearization scheme.

2.1.3 Object detection and measurement

The three most critical elements of plate processing are detection, photometry, and classification. By using the Faint Object Classification and Analysis System (FOCAS, Jarvis and Tyson 1979; Valdes 1982) for image detection and measurement, SKICAT, the system we designed to process and manage the DPOSS plate scans, is able to reach close to the faintest reliable limits of the plate scans, *i.e.*, down to a typical equivalent limiting B magnitude of $\sim 22^m$ for galaxies. In addition, by measuring quasi-asymptotic rather than isophotal magnitudes, using local sky estimates from annuli surrounding each object, and adapting the measurement thresholds within and across each plate to adjust for differences in sky level, noise, and pixel-to-pixel correlation, we are able to obtain very consistent photometry within and across plate boundaries.

SKICAT automatically analyzes each plate as a set of 13×13 overlapping 'footprint' images of 2048^2 pixels each. Not only is this approach computationally convenient, but it provides greater sensitivity to position-dependent plate effects. It also facilitates quality control via the systematic comparison of the overlap regions. SKICAT applies the FOCAS utilities to each of these footprints in order to construct the full plate catalog. First SKICAT robustly estimates sky and sky sigma values for each footprint, providing values that are quite accurate even when relatively large and bright sources exist in the image. Seeded with these values, the FOCAS detection and background estimation procedures are found to work well on the footprints. We were able to test the accuracy of this approach by applying it to the simulated plate images described in Appendix A. There we discuss how we created the simulations and how we were able to use them to optimize and asses our choice of FOCAS detection and measurement parameters.

The FOCAS detection algorithm works by tracking each image area above some threshold comprising some minimum number of pixels. In Appendix A, we describe in detail how

we determine and adjust this threshold in order to achieve uniform sensitivity within and between plates. The local sky brightness for each object feature is measured using the FOCAS 'sky' command, which calculates the median pixel value in an annular region surrounding each feature, avoiding pixels that are within the detection isophote of another feature. The accuracy and systematic effects of this sky measuring algorithm are likewise addressed in Appendix A.

After obtaining the sky estimate, additional attributes for each feature are measured using the FOCAS 'evaluate' routine. The total number of measurements number more than 30. Three different types of magnitudes are measured: aperture, isophotal, and 'total'. Each magnitude (m) is instrumental and computed according to:

$$m = 30.0 - 2.5 \log L$$

where L is the luminosity, or sky-subtracted integrated intensity for each measurement. The offset of 30.0 is arbitrary and was chosen to make the instrumental magnitudes approximate the final calibrated values within a magnitude or two. The aperture magnitudes are computed using a five arcsec radius. The isophotal magnitudes measure the sky-subtracted flux within the detection isophote. The so-called FOCAS total magnitudes are computed by 'growing' the detection isophote out a pixel at a time in all directions until the total area is at least twice the original, then calculating the sky-subtracted flux within that area. This magnitude is meant to provide a flux measurement less biased with respect to surface brightness profile, approximating something like an asymptotic or true total magnitude. The cost of decreased systematic error in this measurement is greater sensitivity to sky subtraction, and hence, increased random error (relative to isophotal or aperture magnitudes). Appendix A provides a detailed comparison of the accuracy of these three types of magnitudes for both stars and galaxies in the DPOSS scans. For the compelling reasons outlined there, we have chosen to use FOCAS total magnitudes in our analysis.

Each object was deblended using the FOCAS 'splits' command. Effectively, this routine runs the detection algorithm on every detected object, but using successively higher thresholds. 'Islands' detected at a given threshold are entered into the catalog as new

objects, and all attributes are remeasured for them. The 'parent's' flux is divided between the 'children' according to the ratio of isophotal fluxes obtained using the higher threshold. This process continues recursively until no more islands are detected.

Improvements can certainly be made to the deblending process so as to improve the quality of the photometry of the deblended objects, to better take deblending into account when matching overlapping plates, and to handle the extreme crowding conditions to be found in lower Galactic latitude POSS-II plates. Nonetheless, we find the present implementation to be more than sufficient even for detailed analyses of higher latitude plates, and that it at least represents a step above reduction without the use of deblending at all, as in the case of the APM survey.

The J2000 RA and Dec of the central pixel of each object is calculated using transformation coefficients provided by ST ScI. We have found these to provide $\sim$ 1.0 arcsec RMS accuracy after correcting for systematic deviations on scales less than about a square degree. In the future, ST ScI will provide more accurate plate solution coefficients that should provide better that 0.5 arcsec accuracy on larger scales.

2.1.4 Object classification

The accuracy of star/galaxy separation generally determines the effective limiting magnitude, in terms of scientific usefulness, of imaging surveys. This limit is, in very many respects, more important than the object detection limit in terms of its impact on the variety of programs for which the data may be used. For this reason, we concentrated a great deal of effort in evaluating the effectiveness of various object classification algorithms. A principal goal of SKICAT was to provide an effective, objective, repeatable, and examinable basis for classifying sky objects at levels beyond the limits of previously existing technology. A full description of our classification procedure is beyond the scope of this paper and will be published separately (Weir, Djorgovski, and Fayyad, in prep.). Here we provide just a sketch of our methodology and results.

Historical methods for classifying objects on plate scans would preclude the identification of the majority of objects in each DPOSS image, since they are too faint for traditional

recognition algorithms, or even manual inspection. These methods generally involve algorithms for separating stars from galaxies within some low dimensional but relatively well discriminating parameter space (*e.g.,* magnitude vs. first moment radius), or within a higher dimensional, but less discriminatory, space of attributes.

SKICAT's procedure for object classification improves upon historical techniques in two ways. First, it measures and utilizes a more powerful set of object attributes; second, it benefits from recent developments in machine learning that enable the computer to automatically determine near-optimal rules for distinguishing objects within high dimensional parameter spaces. In particular, SKICAT utilizes the GID3* and O-Btree decision tree induction software (Fayyad 1991; Fayyad and Irani 1992; Fayyad and Irani 1993), together with the Ruler system (Fayyad, Weir, and Djorgovski 1993) for combining multiple trees into a robust collection of classification rules. These algorithms work by using measurements of a training set of classified objects and inferring an efficient set of rules for accurately classifying each example. The rules are simply conjunctions of multiple "if...then.." clauses, which condition upon, in our case, any of eight different object parameters to determine an object's classification. The real advancement in using this type of classifier relative those used in most large-scale surveys to date is twofold: first, we are able to condition upon a larger and more diverse set of attributes; second, we allow the computer to decide what are the optimal number and form of the rules.

We also experimented with neural nets, and found their performance to be no better than that of decision trees, with the additional disadvantages of slow training and difficulty in interpreting their results (but see Odewahn *et al.* 1992 for a related work). Decision trees are constructed very quickly, and there is never a problem with convergence, unlike with neural nets.

We created separate sets of classification rules for objects from J and F plates. We used the CCD calibration data, described below, which generally have superior image quality, to construct the training sets used to train the plate object classifiers. Classifications derived from the CCD data, more reliable than "by eye" estimates from the plates themselves, were matched to plate measurements to form the training sets. For attributes we used a set of

robust, renormalized object parameters that we found to be distributed in a stable fashion within and across plates. These attributes included a variety of object brightness and shape parameters, in addition to measures of the fit of each object to a locally derived point spread function (PSF). By training the algorithms to classify based on these attributes, we were able to nearly completely remove the effect of PSF variation across a given plate, or even between different plates. Our average accuracy of star-galaxy classifications as a function of magnitude was determined from tests using independent CCD-classified plate data. In both the J and F bands, the accuracy drops below $\sim 90\%$ at about the same equivalent magnitude level, $B \sim 21.0^m$ (see Figure 3). This is $\sim 1^m$ above the plate detection limits, and nearly 1^m better than what was achieved in the past with similar data. This increase in depth effectively doubles the number of galaxies available for scientific analysis, relative to the previous automated Schmidt surveys.

2.2 CCD data

We are conducting a systematic program on the Palomar 60-inch telescope to obtain CCD sequences for use in conjunction with DPOSS. The CCDs are used for photometric calibration as well as for training data to construct the plate object classifiers described above. To date, we have concentrated these observations on Abell clusters and random fields within selected POSS-II fields in the North and South Galactic Caps. These fields were targeted for initial analysis due to their overlap of previous surveys (*e.g.,* that by Picard 1991b), and because they formed two large, contiguous mosaics covering the highest latitude plates in both the North and South. Higher latitude plates are of initial interest in such surveys because they suffer less from crowding effects and are, hence, easier to analyze.

The CCD sequences are being obtained using the Gunn g, r, i photometric system (Thuan and Gunn 1976), for calibrating the J, F, and N plates respectively. These CCD passbands were chosen to provide a reasonable match to the emulsion plus filter combinations of these plates (see Figure 4), and they do so better than any other standard CCD photometric system. The primary disadvantages of the Gunn system are that the standard stars are few, bright, and do not span a large range in color. Nonetheless, we

found the standards sufficient for calibrating our CCD data to the precision and accuracy necessary for our analysis of DPOSS. We, therefore, chose the Gunn system in order to reduce the importance of a color term when calibrating the plates to a CCD standard. The plate g magnitudes may subsequently be transformed to the more standard B_J passband using the relation

$$B_J = g + 0.39 + 0.37(g - r) \tag{2}$$

from Windhorst *et al.* (1991), which is roughly equivalent to $B_J \sim g + 0.5$ mag for a faint field field galaxy of average color.

The CCD exposures were typically 1800 seconds in g, 1200 in r, and 600 in i using an un-thinned Tektronix CCD (CCD11). This is a 1024^2 pixel device with an inverse gain of $\approx e^-$/ADU, read-out noise of 5 e^-, and a pixel size of 24μ, resulting in a field of view of $6.35' \times 6.35'$. Starting in September 1992, we began testing and using CCD16, which is a thinned version of the same Tektronix chip. The quantum efficiency of CCD 16 is twice that of CCD 11 in g and 1.6 times higher in r. We aimed for sufficient depth in our observations to allow for an SNR of at least 10 in the photometry of objects at the classification limit of the survey, or effectively 21.0^m in B_J.

On photometric nights, we would observe from 10 to 12 different standard stars at a range of air masses and color. On non-photometric nights with adequate ($< 2.5''$) seeing, we would take longer exposures in each passband, following up with shorter exposures of the same field on photometric nights in order to calibrate them. In the analysis presented in this thesis, we have only used CCDs obtained on nights recorded as apparently photometric in the observing log book. We subsequently verified the consistency of the photometry for each of these nights by examining the residuals of the standard stars, requiring that they demonstrate no temporal trends and have a standard deviation below 0.03^m. Every night that we recorded as clear at the time of observation, and that we have reduced to date, has met these criteria.

We reduced the CCD data using the standard CCDRED facility within IRAF. The procedures include debiasing, edge-trimming, and flat-fielding. In order to achieve a flat-field variation of less than 1% across each field, we followed a three-step process: division

by a normalized image of the illuminated dome (dome flat), to account for pixel-to-pixel variations; division by a blurred, dome-flattened twilight sky image (sky flat), to take out large-scale variations; and a blurred, dome and sky-flattened average of the deep exposures taken during the night, to take out the remaining large-scale variations. The latter averages were derived by normalizing each exposure by its sky brightness and ignoring values in the image stack deviating more than 2.5 standard deviations from the mean for that pixel (sigma clipping).

We calibrated the CCD observations independently each night using the IRAF APPHOT package. We typically took three exposures of each standard star per frame, averaging the aperture magnitudes to provide a mean and standard error per observation. Each night we solved for the maximum likelihood values of the coefficients A_r, A_g, B_r, B_g, C_r, and C_g in the system of equations:

$$\begin{aligned} r &= r_{inst} + 2.5 \log t_r + A_r + B_r \sec z_r + C_r(g - r) \\ g &= g_{inst} + 2.5 \log t_g + A_g + B_g \sec z_g + C_g(g - r), \end{aligned}$$

where r_{inst} and g_{inst} are the instrumental magnitudes, t_r and t_g are the exposure times, and z_r and z_g are the airmasses at which the observations were made. Applying these coefficients, we measured a standard error typically less than 0.02^m in g and r for our calibrated standard stars each night.

As in the case of plate images, we measured FOCAS total magnitudes from the CCDs. The surface brightness threshold applied for both object detection (with a minimum area requirement of six contiguous pixels above the threshold) and isophotal magnitude measurement was 24.6 magnitudes per square arcsecond in both g and r. This value represented an approximate average of the plate thresholds determined *after* reducing and bootstrap calibrating them using a threshold corresponding to simply a constant number of standard deviations above the sky. Our estimate of the calibration uncertainty in the resulting CCD galaxy catalogs down to a magnitude limit of 20.5^m in g and 19.5^m in r, derived by comparing independent observations of the same fields, is approximately 0.05^m per CCD.

We found FOCAS's built-in classifier to provide very accurate results on the CCDs down

to the plate detection limit, which is our magnitude limit of interest. We were, therefore, able to let FOCAS automatically classify each object, with just a follow-up check by eye, producing excellent quality data without the need for much human interaction or more sophisticated classification algorithms.

3 Plate Calibration

The method we use to photometrically calibrate the plate data is a two step process, described in detail in Appendix B. Briefly, the steps consist of first transforming the plate magnitudes onto a common instrumental system (we find that a simple offset for each plate suffices), then linearly transforming the instrumental F and J magnitudes to r and g, respectively. We demonstrate the accuracy of our calibration procedure with plate-to-plate comparisons of calibrated magnitudes and number counts.

3.1 Calibrated plate-to-plate magnitude comparisons

As one check on the consistency of our plate photometry, we compared calibrated plate magnitudes with one another in the four plate overlaps. Figures 5 and 6 plot r and g magnitude differences vs. mean magnitudes for these regions. Tables 2 and 3 quantify these results. In the magnitude range $15.0^m < r < 19^m$, the mean offset is -0.003^m with standard deviation 0.039^m. In the range $14.5^m < g < 19.5^m$, the mean difference is 0.008^m with standard deviation 0.045^m. These results are consistent with error estimates based on comparing calibrated plate to CCD magnitudes, which imply a systematic plate-to-plate RMS error in zero point offsets of under 0.10^m in g for J plates and below 0.05^m in r for F. The non-systematic RMS error in a single plate measurement, as measured using both plate/CCD and plate/plate overlaps, is approximately 0.15^m in r and 0.21^m in g.

3.2 Internal consistency of galaxy counts

As an additional check on the consistency of our photometric calibrations, we compared galaxy number counts, $A(m)$, for each of the plates in our four survey fields, as depicted in Figure 7. Of particular note is the consistency of the level and slope of the counts between plates of a given passband, especially relative to previous surveys (*e.g.*, Sebok 1986; Picard 1991a).

In Figure 8 we plot the average of the counts from the four survey plates in each band (solid line) versus the counts resulting from alternative plate-to-CCD calibration transformations. The dotted lines surrounding the solid line represent the average result of

adjusting the slope of the linear calibration by one empirically-estimated standard deviation both up and down. After adjustment, in both cases, we re-derived a best-fitting intercept corresponding to that slope. We believe these dotted lines bracket our true uncertainty in the average counts due to plate-to-CCD calibration uncertainties. The differences observed between individual plates in Figure 7 are readily explained due to magnitude-zero point errors at the level implied by this uncertainty and Poissonian counting statistics. Large scale structure presumably, at some level, also accounts for some variation.

As an additional check on the systematic effects of the plate-to-CCD magnitude transformation process, we compare the counts derived after applying both a simple offset (zeroth order) and cubic calibration transformation. As noted in Appendix B, the offset transformation produces magnitudes very similar to those from linear calibration, hence, the implied counts are very similar. On the other hand, the cubic transformation results in significantly different results in the r band, yielding a slope difference of 0.05 mag/dex. We reject the cubic transformation on theoretical grounds (if the HD curve is fitted properly, the appropriate magnitude transformation should be linear) and empirically, largely because of the instability it produces in stellar color-magnitude diagrams. Had we ignored or never investigated the latter effect, however, we might very well have followed the standard practice of previous surveys of fitting high order calibration curves, in order to account for anticipated residual nonlinearities in the plate data. Figure 8 highlights the point that such seemingly unimportant details as choice of polynomial order can result in large systematic errors in scientifically relevant measurements several steps down the reduction chain. The appearance of such fragility in the results should have an appropriately cautionary effect on those who would attempt to infer too much from these or similar data.

In summary, we have tested and applied a photometric calibration technique for the POSS-II scan data which involves using plate overlaps to establish a zero point offset to an instrumental standard. 'Global' CCD transformation functions are then applied to convert instrumental J and F magnitudes to Gunn g and r, respectively, for all of the plates. Using this procedure, it appears one may be able to achieve consistent and reliable photometry over a large portion of the survey without unreasonably many CCD calibration sequences.

In fact, provided a full side of a plate overlaps a well calibrated plate, our analysis indicate that one can calibrate that plate using the overlap alone to within a zero-point uncertainty of $0.05^m - 0.1^m$. To achieve a similar uncertainty for a single plate using CCD data alone would require the equivalent of an order of three CCD fields per plate, as we have used here. Because the plates may be accurately transformed to a uniform instrumental system and, in turn, *all* their overlaps with CCDs combined to infer a single calibration curve, one should be able to effectively pursue a strategy of obtaining only a few CCD sequences per many plates, provided the plates are relatively contiguous.

4 Comparisons with Other Surveys and Theoretical Models

In Figure 9 we plot our r-band counts against model predictions and measurements by Picard (1991a) and Sebok (1986) from independent scans of Palomar Schmidt IIIa-F plates. The slope of our counts between $17.0^m < r < 20.0^m$ is 0.52 with a formal uncertainty of 0.01. The discrepancy between our measurements and others is fairly large for objects brighter than 17^m and compared to Picard's Northern counts, in particular. The latter counts are from a survey of eight plates not more than 20 degrees from our own. We have no explanation for this discrepancy, but note that the internal consistency of the counts among plates within that survey is poor relative to ours, indicating either the effects of significant physical variation in these fields or photometric zero-point or classification uncertainties. A clearer understanding of the source of this inconsistency awaits the availability and analysis of the same plates within DPOSS.

The model predictions in Figure 9 are those from the NE model by KGB discussed in Section 1, a mild evolutionary version of the same (Koo, Gronwall, and Bruzual 1993b), and a model closely approximating that by Guiderdoni and Rocca-Volmerange (GRV, 1990, provided by Gronwall, priv. comm.). The KGB evolutionary model incorporates the same spectral classes and LFs as the NE model, but with mild luminosity evolution of a subset of the spectral classes according to the evolutionary tracks of Charlot and Bruzual (1991) and Bruzual and Charlot (1993).

What is most surprising and illustrative in Figure 9 is the exceptionally high consistency between the DPOSS counts and the KGB evolution and NE models. The predicted counts were taken directly from their models without any renormalization or magnitude zero-point adjustment. While the KGB models were constructed so as to fit the existing data, we note that our results were not in their sample, and that previous bright measurements (viz. Picard 1991a and Sebok 1986), given their dispersion, would not seem to have restricted the models' predictions to any precise level. We can infer that consistency with other data sets, namely bright and faint counts in the same and different passbands, colors, and

redshift distributions, significantly influenced the model predictions shown. Hence, in the context of these models, the fact that our counts match the predicted counts so well is an indication of the consistency of our measurements with these other, diverse data samples.

In contrast, a comparison of our counts with the GRV model indicates an increasingly excessive number of galaxies relative to the NE hypothesis. We note that unlike the KGB models, we did normalize the GRV model to our counts at $r = 17^m$, as they had not been previously scaled for consistency with any data. This model is an example of traditional galaxy distribution synthesis models, which include a number of galaxy *morphological* types, not color classes, and pre-defined Schechter LFs for each type. It is these models to which previous researchers have compared their counts and postulated the existence of excess galaxies at faint magnitudes.

In Figure 10 we plot the measured differential number counts in g_J from this survey, the APM southern survey of SERC/ESO plates, the predictions of no evolution models by KGB, GRV, Ellis (1987), and the mild evolution model of KGB. The upper panel reflects a conversion of the B_J magnitudes of all the non-DPOSS counts to g using the transformation equation 2 from Windhorst *et al.* (1991) assuming a mean galaxy color $(g - r)$ of 0.3^m, which we measure within the usable magnitude range of our survey. This transformation roughly implies $B_J \sim g + 0.5^m$. The lower panel is the result of horizontally shifting all non-DPOSS counts until the DPOSS and APM counts are normalized at $g = 17.0^m$, corresponding to a transformation of $B_J \sim g + 0.7^m$. We note that any transformation we apply is only very roughly approximate, as there is a significant color term implied by the differences in the plate IIIa-J, CCD g, and B_J bandpasses. As a further indication, we point out that in Bruzual (1992) the average galaxy color at low redshifts in his models is $(g-r) \sim 0.5^m$, implying, according to Windhorst *et al.* (1991), $B_J \sim g+0.6^m$. Accordingly, we believe that the uncertainty in the magnitude zero-point of our counts relative to the others is approximately 0.2^m.

This uncertainty results in particular difficulty when trying to compare our color measurements to those of the model. In Figure 11 we plot our $(g - r)$ colors versus the transformed KGB NE predictions for $(B_J - R_F)$. The color transformation from the model's

$(B_J - R_F)$ to the data's $(g - r)$ system, which may be off by as much as 0.3^m, is derived by attempting to match actual star color distributions in the two systems.

Due to the zero point uncertainties in the blue counts, we are able to infer less from the consistency of these preliminary g measurements with either theory or other data, awaiting the production of model counts simultaneously in the B_J and g passbands. However, we note that the slope of our g counts between $17.0^m < g < 20.0^m$ is 0.49 with a formal uncertainty of 0.01, in excellent agreement with the slope of the APM counts, 0.50 ± 0.01, in the equivalent magnitude range. Again, we also find that in comparing both our measurements and APM's with the latest NE models, we find much closer agreement than was found relative to older NE models (*e.g.*, Maddox *et al.* 1990).

KGB explain some of the discrepancy between their NE model and traditional ones' predictions as being due to the fact that these other models generally do not account for the wide dispersion of galaxy colors within a galaxy morphological class; these traditional models tend to over-predict red galaxies. Some previous models also assume a single LF for each type, or luminosity functions for blue galaxies which tend not to even match the local number density estimates. Although their NE model admittedly fails to sufficiently account for all of the observations to which they try and fit (*e.g.*, the model under-predicts faint blue galaxy counts by $\sim 40\%$ by $B_J = 24^m$), it is nonetheless able to relatively consistently predict counts over a sufficiently large magnitude range as to suggest that mild evolution and proper accounting of the systematic errors in the data sets could account for the remaining discrepancies, without the need for evolutionary or cosmological exotica. The KGB evolution model plotted in Figures 9 and 10 reflect an early first attempt to include some degree of evolution in their model, producing a marginally better fit to our data in both colors, and more significant improvement at fainter levels (Gronwall priv. comm.).

5 Discussion

We describe the first scientific results using the Digitized Second Palomar Observatory Sky Survey. We have measured $A(m)$ in two passbands from DPOSS galaxy catalogs derived from an approximately 100 squared degree region centered near the North Galactic Pole. The IIIa-J and IIIa-F data were calibrated to the Gunn-Thuan g and r CCD photometric system using internal overlaps and overlaps with a set of CCD sequences distributed across the plates. Our estimated zero point uncertainty for the combined set of four plates in each band is $\sim 0.05^m$ in r and $\sim 0.10^m$ in g. The measured differential counts as a function of magnitude, both in level and slope, are very consistent from plate to plate, helping to confirm the consistency of our plate-to-CCD photometric calibration technique.

In both the blue and red passbands, our measured counts agree well with the no evolution predictions of KGB, and less so with comparable empirical measurements, especially at brighter magnitudes. As in comparable previous surveys, we do not find good agreement between our measurements and the predictions of traditional galaxy NE models. However, in light of the most recent KGB models and our consistency with them, we believe these initial DPOSS results provide additional empirical verification of the plausibility of their hypothesis: that recent and/or extreme galaxy evolution or non-standard cosmology is not demanded by the data at this time.

Further refinements of galaxy evolution models must include a detailed accounting of the detection and measurement process in order to compare all the observations on a consistent basis and provide a more conclusive comparison of model predictions with the data. For example, as a note of caution, we refer to Figure 17, which demonstrates the significant difference in measured differential number counts that result just from applying different methods of photometry on the same data, in this case simulated images from our survey. Although these detailed simulations suggest that systematic biases in our measured galaxy counts are negligible within our catalog's estimated 90% completeness limit, we nonetheless fail to fully take into account, for example, the effect that different distributions and forms of galaxy surface brightness profiles would have on observed counts.

A fully comprehensive model might, for example, provide for a distribution of surface brightness profiles as a function of galaxy spectral class. A particular survey's surface brightness detection and measurement thresholds might then be appropriately taken into account when comparing model predictions to observations.

All of these *caveats* withstanding, we nonetheless consider the KGB model the relevant null hypothesis for explaining our data, believing it to be the most consistent and comprehensively calibrated NE model produced to date. As we claim to have firm estimates of the completeness and photometric accuracy of our results, the consistency (or lack thereof) of our observations with the model helps provide an additional check on whether mild galaxy luminosity evolution remains a valid means of explaining the data. Gronwall (priv. comm.) is currently in the process of re-optimizing the non-evolving LFs of the KGB model in the context of our data, the results of which shall be forthcoming.

As more DPOSS data, more accurate calibration, and the possibility of predicting counts and colors in the g passband are achieved in the near future, we believe that one will be able to place significantly more restrictive constraints on galaxy evolution models at bright magnitudes. Far more difficult is the remaining task of coming to understand and model the idiosyncratic systematics of each data set to which one should compare. In this respect, the PNSC and DPOSS should prove to be far more amenable than many other sources, due to the good statistics (tens of millions of galaxies), uniform quality, and well-understood properties of the data. These initial results provide a glimpse of the scientific potential of the full data set when it becomes available.

This work was supported at Caltech in part by NASA AISRP contract NAS5-31348, the Caltech President's fund, and NSF PYI award AST-9157412, and at JPL under a contract with NASA. The POSS-II is partially funded by grants to Caltech from the Eastman Kodak Co., the National Geographic Society, the Samuel Oschin Foundation, NSF grants AST 84-08225 and AST 87-19465, and NASA grants NGL 05002140 and NAGW 1710. We acknowledge the efforts of the POSS-II team at Palomar, the scanning team at STScI, and the SKICAT team at JPL, most especially Joe Roden.

A Appendix - Plate Photometric and Detection Sensitivities

In order to optimize the plate reduction procedure and understand its sensitivity to various image characteristics, we created simulated images that matched the digitized plate scans as accurately as possible. In particular, we wished to study detection efficiency, photometric accuracy, and photometric consistency as a function of magnitude type, object profile, isophotal threshold, image seeing, and image noise. While we had a choice of which type of magnitude to measure (*e.g.,* aperture or isophotal), the other characteristics are simply observable, but variable. Through careful simulation, our hope was to better understand the systematic effects in our DPOSS catalogs resulting from known variations in these image qualities.

A.1 Simulation quality

Our plate image simulations were constructed using the ARTDATA package within IRAF. The tasks GALLIST and STARLIST were used to construct a list of objects used to populate a 2048^2 simulated footprint image. The random object lists were created assuming a uniform spatial distribution and a power law luminosity function. We attempted to match the quality and object density of plate J380, as it was as representative a plate from the survey as any. For galaxies in the g band apparent magnitude range 15^m to 19^m, a set of 60 galaxies with a power law slope of 0.35 ($L \propto 10^{0.35m}$) was found appropriate, while 5200 objects with a power law slope of 0.6 (Euclidean) was used for the range 19^m to 23^m. The galaxy profiles were all exponential disks with half-power radii uniformly distributed between +50% and −50% of a canonical size for a given magnitude, specified by a maximum of 15.0 arcseconds for the brighter sample and 2.90 for the fainter objects. Each had random inclination, i, ranging uniformly from 0° to 90°, with axial ratios given by

$$a/b = \sqrt{0.99\sin(i)^2 + 0.01}.$$

An internal absorption coefficient was also applied based on the inclination (see the IRAF ARTDATA/GALLIST documentation for details). A minimum redshift of 0.02 and 0.126

was assigned to the bright end of each magnitude range. Within GALLIST, object redshifts are assumed to be proportional to the luminosity distance, or the square root of the apparent luminosity, and are used to compute the mean apparent sizes of the galaxies according to $z/(1+z)^2$, the cosmological redshift factor for angular diameters.

The random star lists were constructed assuming a uniform spatial distribution and a power law luminosity function with slope 0.2, which we found to provide the best fit to our data. A total of 800 stars in the g magnitude range 15.0^m to 22.5^m were found to match the measurements of plate J380.

Galaxy bulges and ellipticals were not included in these simulations. As their profiles fall somewhere in between stars and exponentials, we nonetheless believe these simulations sample the relevant extremes in the data. The fact that the exponential disks were constructed with randomly generated half-power radii also assures that the images sample a distribution of different surface brightness profiles.

Noiseless images containing stars and galaxies were created using the MKOBJECT task within ARTDATA. To simulate the effects of seeing, depending on the simulation, we convolved the image with either a bivariate Gaussian:

$$I(r) = \exp[-\ln(2)(\frac{r}{r_o})^2],$$

or Moffat (1969) function:

$$I(r) = [1 + (2^{1/\beta} - 1)(\frac{r}{r_o})^2]^{-\beta},$$

where I is object intensity, r is the radial distance from the object center, r_o is the half-intensity radius scale parameter, and β is the Moffat parameter, which we take to be 2.5. As we show below, the choice of point spread function (PSF) form ultimately made very little difference for our purposes. The half-intensity radius of PSF we applied was 2.7 arcseconds, closely matching the width measured on plate J380.

Next we ran one of our own routines for adding noise to the image. The choice of an appropriate noise level was complicated by the fact that the noise is correlated from pixel to pixel in actual plate images. In the subsequent section, we describe how we used

simulated images to determine how best to adjust our detection thresholds to compensate for this correlation. We simulated this effect by adding signal dependent and independent random Gaussian noise to the image before convolving it with a small blurring kernel. The latter was achieved by running the IRAF task GAUSS on the noisy image using a pixelated Gaussian distribution of standard deviation 0.51 pixels and sampled out to four standard deviations. As a final step we crudely simulated the effects of saturation by cropping all pixels values to some maximum level.

After numerous iterations of adjustments to the many parameters involved, we managed to construct a set of simulated plate images that match the real data well. Figures 12 and 13 demonstrate that the ensemble distribution of object shapes and sizes in J380 and the simulated data are very closely matched. In particular, the scatter of objects is quite similar, indicating consistent noise properties between the two. A further test of the correspondence between the real and simulated data, especially at the noise level, is depicted in a plot of number counts as a function of magnitude (Figure 14).

A.2 Effect of correlated pixel noise on detection

For optimal sensitivity, the FOCAS detection algorithm applies a threshold equal to some number of estimated standard deviations (sky sigma) above the locally estimated sky. SKICAT provides FOCAS with a robust value for the sky sigma, individually derived from statistics for each footprint. However, because of spatially varying pixel-to-pixel correlation within each plate scan, using the same multiple of sky sigma as the threshold for all footprints would not result in the same detection sensitivity.

To compensate for this effect and approach a common level of sensitivity between and within plates, we sought to derive a factor by which to scale the measured sky sigma so as to make it correspond to approximately one standard deviation in an *unblurred* version of each footprint. To establish this scaling factor as a function of measured blur, we used one of our simulated footprint images matching the average noise[1] and object number statistics

[1]The appropriate level of uncorrelated, Gaussian random noise was determined in an iterative fashion. First, we found a Gaussian kernel which, when convolved with the image, produced a degree of blur, as measured by the pixel-to-pixel correlation, closely approximating that of an average footprint. We then found that noise amplitude which, after convolution, resulted in a measured sky sigma closely matching

of real footprints, then we convolved it with a series of Gaussians of different width. Given the convolution kernel, the appropriate scale factor is simply the square root of the inverse of the sum of squares of the normalized kernel elements. By measuring the pixel-to-pixel R^2 for each image, we are able to empirically derive a mapping from measured (square) correlation to scale factor. We found a sixth order polynomial to provide a good fit to the relation. We also established the relation using a blank simulated sky image and derived virtually identical results, lending confidence in the robustness and accuracy of our correlation estimation procedure.

We chose 2.5 times this scale factor times the estimated sky sigma as our detection threshold in plate instrumental intensity units. We also required every object to comprise at least six contiguous pixels. We used the built-in FOCAS pre-detection blurring function, which is simply a five by five array of linearly increasing weights from each edge to the center. The FOCAS detection algorithm works by convolving the image with this kernel, then searching for contiguous pixels with values greater than the locally estimated sky by the specified detection threshold. To adjust for the convolution, which is meant to improve the sensitivity of the detection algorithm, the detection threshold is scaled by the square root of the inverse of the sum of squares of the normalized kernel elements. Note this is the same blurring correction we applied earlier to account for the correlation induced by the scanning process.

Our choice of detection parameters, in particular our scaling correction for pixel-to-pixel correlations, results in relatively consistent sensitivity as a function of plate quality, as evidenced by the relative uniformity of object density we detect from footprint to footprint and plate to plate. Our choice of threshold, minimum area, and pre-detection blurring were chosen after extensive tests on both real and simulated images, establishing some feel for the trade-off between completion (percentage of real objects detected) and contamination (percent of detected objects which are not real). On simulated images, this combination of parameters resulted in an average FOCAS detection isophote corresponding to roughly 2.0 times the *uncorrelated* sky sigma, which is sufficiently far into the noise as to pick up

that of an average footprint.

every object readily detectable by eye. It also resulted in what we considered a manageable number of detections per footprint and plate, in excess of the density saved in previous Schmidt plate surveys.

A.3 Sensitivity to magnitude type and sky subtraction

The next set of tests using the simulated images were to determine what type of magnitude provides the most reliable estimate of true star and galaxy magnitudes for our data. The four types we tested were aperture magnitudes, using a 10 arcsecond diameter; isophotal magnitudes, measured using an isophote approximately 2.0 (uncorrelated) sigma above the local sky; FOCAS total magnitudes, measured by growing the isophote out until it encompasses twice the isophotal area used in the detection process; and Gaussian 'corrected' magnitudes of the sort used in the APM survey (Maddox, Efstathiou, and Sutherland 1990). The latter correspond to total integrated magnitudes assuming a Gaussian profile fit to each galaxy. Given a measured threshold intensity, t, isophotal area, A, and isophotal magnitude, m, Maddox, Efstathiou, and Sutherland (1990) show that the difference between total and isophotal magnitudes can be given by a parameter ϵ, where $m_{tot} = m + 2.5\log_{10}(1+\epsilon)$ and

$$\frac{At}{10^{m/2.5}} = \epsilon \ln(1 + \frac{1}{\epsilon}). \tag{3}$$

A quadratic approximation may be applied to invert this expression and solve for ϵ as a function of A, m, and t. The error introduced by this approximation is negligible. Maddox, Efstathiou, and Sutherland (1990) use this type of magnitude to attempt to remove the effect of plate-to-plate variations in their isophotal magnitudes due to varying threshold isophotes. They justify the use of Gaussian profiles based on the fact that the underlying true profiles of faint galaxies are blurred by seeing into approximately Gaussian form. Bright objects, on the other hand, have a small correction factor, so the profile assumption is not important.

We measured and computed these different types of magnitudes for every galaxy detected in our simulated footprint data. Plots of true minus measured magnitude as a function of true magnitude and magnitude type appear in Figure 15. As we expect, the

aperture magnitudes tend not to measure all of the flux of brighter, hence generally larger, galaxies, although they provide less biased estimates at fainter levels. The isophotal magnitudes are also systematically biased too faint, simply due to the use of an isophotal threshold. FOCAS total magnitudes, on the other hand, seem to provide a reasonably unbiased estimate of actual magnitudes as a result of extending the measurement threshold in a profile-dependent way. The Gaussian total magnitudes derived from correcting the isophotal magnitudes, however, systematically overcompensate for the thresholding effect, resulting in magnitudes severely biased in the direction of being too bright.

In Figure 16 we plot the average and standard deviation in one magnitude bins of the difference between true and measured magnitudes for isophotal, FOCAS, and Gaussian total magnitudes, measured for both stars and galaxies in our simulations. Note that by a true g magnitude of 20.5^m, the Gaussian total magnitudes of galaxies have the smallest scatter, but are systematically bright by nearly 0.3^m. Isophotal magnitudes are biased by approximately 0.1^m, but in the opposite direction. For both stars and galaxies, the FOCAS total magnitudes have the least bias across the fainter and more relevant magnitude ranges. Hence, we choose to use FOCAS total magnitudes when analyzing the photometry from DPOSS.

To further test the scientific relevance of the choice of magnitude type, we computed the differential galaxy counts measured from our simulated data using each of the different magnitude types. We plot these results in Figure 17. The solid line indicates the true number of objects used to create the data. Note that using both the isophotal and FOCAS total magnitudes result in measurements that trace the actual counts fairly well out to a magnitude of approximately 20.5^m in g. Due to the bias in the Gaussian total magnitudes we computed, however, those number counts are significantly inflated above truth at faint levels. A large number of faint objects are artificially shoved to the left, boosting the counts of objects in the range $19.5^m < g < 21.0^m$. We are unable to assert that an effect of just this sort helps account at least in part for the excess counts observed by the APM group in their survey (Maddox *et al.* 1990), as we have not attempted to simulate and explore these effects using their data (which consist, *e.g.*, of photographic densities and

not fitted intensities). Nonetheless, after performing these tests, we were convinced of the merit of using FOCAS and not Gaussian total magnitudes for our survey.

In addition to magnitude type, quality of sky subtraction is one of the most determining factors of accurate photometry. In this capacity, we found the FOCAS sky estimation routine to perform superbly. In Figure 18 we plot the mean and standard deviation of the error in the measured sky intensity of the simulated data as a function of magnitude, in units of the image's (uncorrelated) sky noise sigma. Note that the mean sky error is well below a tenth of the sky sigma, and after accounting for average object areas, results in an average magnitude error well under 0.01^m down to a g of 20^m, rising only to 0.02^m by $g = 22^m$. To verify that our simulated data matched the real data well enough to make this test relevant, we compared the measured sky values in our simulated data with measurements from plate J380 (Figure 19). Just as in Figures 12 through 14, we found excellent agreement between the two distributions. Note, however, that we have not simulated the effects on sky subtraction of crowding of the sort expected at low Galactic latitudes, where more specialized techniques will be required. Our simulations only verify that the FOCAS local sky estimation routine performs quite well for images such as those in high latitude DPOSS fields.

A.4 Sensitivity to detection threshold, seeing, and noise

The primary reason for using something like a Gaussian corrected magnitude is to help remove the effect of expected image variations, such as in the surface brightness of measurement thresholds. Ideally, of course, one would like to use the same surface brightness threshold, in terms of calibrated magnitudes per square arcsecond, when measuring all plates. This presents the quandary, however, of knowing the photometric calibration of the plate prior to it ever being reduced. In the end, one must just choose a consistent means for determining the isophotes, and afterwards try to account for the resulting variations in the actual levels. We sought to quantify how well FOCAS total magnitudes hold up to these sorts of varying plate effects, as well as determine the sensitivity of our detection method to these variations.

We found that detection efficiency and measured FOCAS total magnitudes vary in an expected manner as a function of detection and measurement threshold, seeing, and noise. These variations are at the level expected due to noise considerations, and they are generally not too significant until $g > 21.4^m$. These results are illustrated in Figures 20 through 25. These figures demonstrate the effect of varying each factor (threshold, seeing, and noise) by an amount at the limit of what is expected (or found to date) in the actual survey. For each factor, we plot the detection efficiency and measured magnitude error (the offset relative to the truth) for stars and galaxies as a function of that factor (Figures 20, 22, and 24). We also plot the consistency of measured magnitudes from image to image assuming different levels of that factor (Figures 21, 23, and 25).

Of particular note is the consistency of stellar and galaxy magnitudes as a function of different isophotal thresholds out to reasonably faint magnitudes. We find that out to a g magnitude of 20.5^m, the average offset due to a threshold difference of 0.2^m is less than 0.025^m, which is well within the systematic offset we actually measure from plate to plate. This means that varying thresholds are not the principal contributor to plate-to-plate variations in zero points in our data, but rather these variations are more likely due to a composite of many factors, including seeing, plate sensitivity variations, as well as thresholds. This justifies our choice of using FOCAS total magnitudes as opposed to Gaussian magnitudes, which are meant to explicitly remove the effects of varying thresholds. We found that the consistency of Gaussian magnitudes is, in fact, better out to fainter magnitudes. However, this consistency is achieved at the cost of significant bias in the measured magnitudes, as demonstrated in Figure 15. We believe the error we would introduce in attempting to remove this bias, which is very difficult to measure for *real* data, would far exceed the error and inconsistency resulting from using FOCAS total magnitudes, so we do not attempt it.

Our tests also indicate that different levels of seeing have moderate effects on detection efficiency for both stars and galaxies, though the effects on absolute and relative magnitudes are much more pronounced for stars than galaxies. Varying the seeing width from $3.0''$ to $3.6''$ results in a relative stellar magnitude offset of about 0.07^m by $g = 20.5^m$, while it is

less that 0.03^m for galaxies. Changing the shape of the PSF from a Gaussian to a Moffat profile of the same width has virtually no effect on the measured galaxy magnitudes, but has an effect on the order of 0.05^m for stellar magnitudes out to a g of 20.25^m.

Different realizations of noise at the same and higher levels have little effect on detection efficiency and magnitudes out to the plate classification limit. The false detection rate, or catalog contamination, however, rises dramatically at the faint end with just a 20% increase in noise level. This observation helped motivate our choice of a noise-dependent, rather than constant surface brightness, detection and measurement threshold, helping to keep catalog contamination at a reasonably constant level. Using a constant surface brightness threshold would require knowledge of a plate's photometric calibration *before* processing it, which is rather difficult to achieve. In any case, because of the variability in pixel-to-pixel noise correlation we measured within plates, we were largely limited for practical reasons to using the scaled-number-of-sigma-above-sky threshold described in detail in Section A.2 above. Our attempts at using a constant surface brightness threshold resulted in catalogs of such variable depth and contamination, as a result of varying noise correlation, as to render them useless. Instead we chose to live with varying threshold isophotes, verifying through these simulations that the systematic effects on magnitudes are within acceptable limits.

In summary, systematic galaxy magnitude offsets due to expected variations in threshold isophote, image seeing, and noise appear to be below 0.05^m down to our classified galaxy 90% completeness limit of $\sim 20.25^m$ in g. The effects of these factors on detection efficiency and catalog contamination qualitatively meet our expectations, and have virtually no affect on catalogs out to the classification limit. Expected variations in image seeing and noise do play major roles in determining the plate detection limit, however.

A.5 Object profile sensitivity

We also performed a limited number of tests to determine the sensitivity of FOCAS detection and total magnitudes to galaxy profile. As our simulated images were created only using exponential disks, we were unable to test the sensitivity to an exhaustive set

of profiles. The galaxy parameters we tested against were axial ratio (a/b) and normalized half-light radius at a given magnitude (r_{norm}), which varied from -50% to $+50\%$ in our simulations. Figure 26 indicates that there are no significant systematic variations as a function of these galaxy shape parameters out to magnitudes of interest. Differences in these parameters do affect the accuracy of faint magnitudes for large values of both quantities, however (see Figure 27). These would be galaxies with relatively flat profiles. This figure plots galaxies with true magnitude less than or equal to 22.0^m in g. Out to our galaxy catalog completeness limit of 20.25^m, the average offset is only about 0.2^m. Nonetheless, the sensitivity of DPOSS galaxy photometry to variations in object profile is a subject which should demand careful attention in the future.

B Appendix - Photometric Calibration

B.1 Instrumental plate-to-plate calibration

The initial step in establishing a uniform magnitude system for all of our survey plates was to compile a list of all objects detected and classified as a galaxy within pairs of adjoining plates. We limited the list to objects detected within 2.9° of the center of each overlapping plate, insuring that the photometry would be minimally affected by vignetting effects. Figure 28 plots the difference in instrumental magnitude as a function of mean magnitude for the galaxies in the overlap between F380 and F381. When performing the density to intensity transformation before processing each plate, we attempt to scale the average sky value of each plate to the same level so that the instrumental magnitudes for each plate tend to be quite close. They are also scaled to roughly match their calibrated values within a magnitude or two. Note that for $F_{inst} > 15.0^m$, the difference in magnitudes between plates appears to consist almost exclusively of a DC offset term, with no higher polynomial terms obviously necessary to express the relation. The same holds true for the other plate overlaps measured in this study, including J band overlaps, as will be quantified below. The simple, zero order nature of this relation implies that for unsaturated galaxies, our method of linearizing the plate densities is consistent from plate to plate, and our choice of isophote and magnitude type are consistent from plate to plate. It also demonstrates the relatively high photometric uniformity within the plates resulting nitrogen flushing during each exposure.

A simple test to verify the adequacy of a simple zeroth order offset between plates is to measure their consistency for a mutually overlapping ring of three or more plates. In our survey, three such fields (380, 381, and 442) were available for both F and J. To obtain a statistically meaningful number of objects in the 380/442 overlap, we had to relax the radial distance restriction to 3.1°. Otherwise, the lists of overlap objects were generated exactly as described above.

For each band, we measured the average magnitude offset between each of the three catalog pairs. We restricted the estimate to within the mean magnitude range of 16^m to 19^m

for F_{inst} and 16^m to 20^m for J_{inst}. We then solved for the least-squares best fit zero point offsets between field 380, our chosen standard, and fields 381 and 442. We used the three pairwise estimated offsets as our measurements and the three plate closure requirement as a constraint. The original pairwise offsets and those obtained after subtracting our least-squares fits appear in Table 4. The mean offset between fitted pairs in the magnitude ranges quoted above is less than 0.01^m in both colors, with a standard error of 0.018^m in F_{inst} and 0.036^m in J_{inst}.

Given the high degree of consistency resulting from this matching procedure, we applied these fitted offsets, transforming all magnitudes to the field 380 instrumental standard. For field 382, which does not overlap 380, we simply offset relative to the fitted field 381.

B.2 Absolute calibration using CCD data

In our final stage of calibration, we combined matching CCD and plate measurements from all four plates in order to establish the plate-to-CCD photometric calibration curves. We did so by fitting zero, first, and third order polynomials to the lists of J and F magnitudes (in the instrumental field 380 system) vs. calibrated Gunn g and r magnitudes, respectively. Once again, we restricted the lists to objects with a maximum distance from a plate center of 2.9°. The fitted data points and their residuals after applying the linear calibration transformation appear in Figures 29 through 32. The calibration accuracy on a per plate basis is also reflected in Table 5, while the averages across all plates appear in Table 6.

Our empirical estimates of the uncertainties in the linear transformation offsets, based on independent calibrations of each plate, are approximately 0.025^m in r and 0.05^m in g, though the formal uncertainties are about a factor of two smaller. The empirically-derived uncertainty in the slope of the transformations is approximately 0.015 for both g and r, dominating the calibration uncertainty in net effect after adjusting the offset to achieve an optimal fit using the different slope. We explicitly test for the effect of this uncertainty on our measured counts in section 3.2.

Our choice of a linear plate-to-CCD magnitude transformation was largely driven by our presumption that most nonlinearities in the faint magnitude ranges of interest should

have been taken into account by our fit to the HD curve, as described in Section 1. We suspected that any attempt to account for additional nonlinearity might involve fitting into the noise. This suspicion was largely confirmed by plots of stellar color-magnitude diagrams as a function of which transformation we applied (see Figure 33). Note that the two distinctive stellar ridges display a high degree of nonlinearity when one applies the cubic transformation. As this result is in contradiction with the expected distributions within color-magnitude diagrams for galaxies, we chose to remove the cubic polynomial from consideration. Instead, we chose to apply the linear transformation for both g and r, as it resulted in a marginally better fit than the zero point offset and still a reasonable stellar color-magnitude diagram.

As our error analysis in Tables 5 and 6 reveals, after the initial zero point adjustment is applied to each plate, a single transformation function converting instrumental to calibrated magnitudes applies consistently well across multiple plates, with standard zero point errors relative to CCD photometry of less than 0.05 magnitudes in r and 0.10 in g. We note, however, that this calibration process is only valid for the portion of each plate not significantly affected by vignetting. DeCarvalho (1994, priv. comm.) is in the process of mapping out the POSS-II vignetting function using a large sample of DPOSS scans, which should allow us to extend the area of each plate suitable for photometry.

References

Broadhurst, T. J., Ellis, R. S., , and Shanks, T. 1988, *MNRAS*, **235**, 827.

Bruzual, G. 1992, in *Cosmology and Large-Scale Structure in the Universe*, ed. R. R. de Carvalho, A.S.P. Conf. Ser. #24, 54.

Bruzual, G. and Charlot, S. 1993, *APJ*, **405**, 538.

Charlot, S. and Bruzual, G. 1991, *APJ*, **367**, 126.

Colless, M., Ellis, R. S., Taylor, K., and Hook, R. N. 1990, *MNRAS*, **205**, 1287.

Cowie, L. L., Gardner, J. P., Lilly, S. J., and McLean, I. 1990, *ApJL*, **360**, L1.

Cowie, L. L., Songaila, A., and Hu, E. M. 1991, *Nature*, **354**, 460.

Djorgovski, S., Lasker, B., Weir, N., Postman, M., Reid, I., and Laidler, V. 1992, *BAAS*, **24**, 750.

Djorgovski, S. G. *et al.*1994, *ApJL*, submitted.

Eales, S. 1993, *ApJ*, **404**, 51.

Ellis, R. 1987, in *Observational Cosmology, IAU Symp. 124*, ed. A. Hewitt, G. Burbidge, and L. Z. Fang, (Dordrecht: Reidel), 367.

Fayyad, U. 1991. Ph.D. thesis, EECS Dept. The University of Michigan.

Fayyad, U. and Irani, K. 1992, in *Proceedings of the Tenth National Conference on Artificial Intelligence AAAI-92, San Jose, CA.*

Fayyad, U. and Irani, K. 1993, in *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, France*, Morgan Kauffman, in press.

Fayyad, U., Weir, N., and Djorgovski, S. 1993, in *Proceedings of AAAI-93 Workshop on Knowledge Discovery in Databases, Washington D.C.*, ed. G. Piatetsky-Shapiro, AAAI/MIT Press, 1.

Fukugita, M., Takahara, F., Yamashita, K., and Yoshii, Y. 1990, *ApJ*, **361**, L1.

Guiderdoni, B. and Rocca-Volmerange, B. 1990, *A&Ap*, **227**, 362.

Jarvis, J. and Tyson, J. 1979, *SPIE Proc. on Instrumentation in Astronomy*, **172**, 422.

Koo, D., Gronwall, C., and Bruzual, G. 1993a, *ApJL*, **415**, L21.

Koo, D., Gronwall, C., and Bruzual, G. 1993b. priv. comm.

Koo, D. and Kron, R. G. 1992, *Ann. Rev. Astron. Astrophys.*, **30**, 613.

Lasker, B., Djorgovski, S., Postman, M., Laidler, V., Weir, N., Reid, I., and Sturch, C. 1992, *BAAS*, **24**, 741.

Lilly, S. J.and Cowie, L. L. and Gardner, J. P. 1991, *ApJ*, **369**, 79.

Lonsdale, C. J. and Chokshi, A. 1993, *ApJ*, **105**, 1333.

Loveday, J., Peterson, B. A., Efstathiou, G., and Maddox, S. J. 1992, *ApJ*, **390**, 338.

Maddox, S., Efstathiou, G., and Sutherland, W. 1990, *MNRAS*, **246**, 433.

Maddox, S., Sutherland, W., Efstathiou, G., Loveday, J., , and Peterson, B. 1990, *MNRAS*, **247**, 1p.

Metcalfe, N., Shanks, T., Fong, R., and Jones, L. R. 1991, *MNRAS*, **249**, 498.

Moffat, A. 1969, *A&Ap*, **3**, 455.

Odewahn, S., Stockwell, E., Pennington, R., Humphreys, R., and Zumach, W. 1992, *AJ*, **103**, 318.

Picard, A. 1991a, *AJ*, **102**, 445.

Picard, A. 1991b. Ph.D. thesis, California Institute of Technology.

Reid, I. *et al.*1991, *Publ. Astron. Soc. Pac.*, **331**, 465.

Reid, N. and Djorgovski, S. 1993, in *Sky Surveys: Protostars to Protogalaxies*, ed. B. T. Soifer, A.S.P. Conf. Ser. #43, 125.

Russel, J. L., Lasker, B. M., McLean, B. J., Sturch, C. R., and Jenker, H. 1990, *AJ*, **99**, 2059.

Sebok, W. 1986, *ApJSupp*, **62**, 301.

Thuan, T. X. and Gunn, J. 1976, *Pub. Astron. Soc. Pac.*, **88**, 543.

Tinney, C. 1993. Ph.D. thesis, California Institute of Technology.

Tritton, S. 1983, *The U.K. Schmidt Telescope Unit Handbook*, (Edinburgh: Royal Observatory Edinburgh).

Tyson, J. A. 1988, *AJ*, **96**, 1.

Valdes, F. 1982, *SPIE Proc. on Instrumentation in Astronomy IV*, **331**, 465.

Windhorst, R. *et al.*1991, *ApJ*, **380**, 362.

Figure 1: The initial set of DPOSS survey fields, analyzed in this paper. The dashed lines centered on each field outline the portion of the plate not suffering significantly from vignetting effects. The small labels within each field prefaced by an 'A' or 'F' designate the location of CCD sequences centered upon Abell clusters or random fields, respectively. The North Galactic Pole is indicated by a large dot in the lower middle of the plot.

Figure 2: The parametric form in Equation 1 is used to approximate the transformation function from the measured densities of the 16 plate sensitometry spots to relative intensities.

Figure 3: The accuracy of our star/galaxy separation technique is depicted by the completeness (fraction of galaxies classified as such) and contamination (fraction of non-galaxies classified as galaxies) measured within galaxy catalogs from four survey plates, using independent CCD sequences as the source of true classifications.

Figure 4: The relative transmission of the IIIa-J, IIIa-F, and IV-N plate plus filter combinations and the Gunn-Thuan g, r, and i system. The filter tracings were provided by S. Djorgovski and J. Smith (priv. comm.).

Figure 5: The difference in calibrated r magnitude vs. average magnitude for galaxies in the overlaps between four plate fields.

Figure 6: The difference in calibrated g magnitude vs. average magnitude for galaxies in the overlaps between four plate fields.

Figure 7: r and g band galaxy counts in our four fields. The sharp fall-offs at the faint end are due to truncation of the catalog, by construction, beyond the reliable classification limit, rather than the intrinsic plate detection limit.

Figure 8: Galaxy counts as a function of the plate-to-CCD transformation function. The thick solid line in each graph reflects the differential number counts resulting from our standard linear calibration of the plate magnitudes to the Gunn-Thuan standard. The dotted lines surrounding them reflect the counts derived by altering the slope of the transformation by one standard deviation. The dashed and dashed-dotted lines are the counts resulting from the application of best-fitting zeroth and third order transformations, respectively.

Figure 9: The measured differential number counts from this survey (solid line, with dots extending beyond the 90% completeness / 10% contamination limit), Picard's (1991) survey of POSS-II plates in the North and South Galactic Caps, and Sebok's (1986) survey of earlier-generation Palomar Schmidt plates. The other lines are the predictions of no evolution models by Koo, Gronwall, and Bruzual (KGB 1993a) and Guiderdoni and Rocca-Volmerange (GRV 1990), and the mild evolution model of Koo, Gronwall, and Bruzual (KGB 1993b).

Figure 10: The measured differential number counts from this survey (solid to dotted line) and the APM (Maddox *et al.* 1990) southern survey of SERC/ESO plates. The other lines are the predictions of no evolution models by KGB, GRV, and Ellis (1987), and the mild evolution model of Koo, Gronwall, and Bruzual (1993b). The upper panel reflects a conversion of the B_J magnitudes of all the non-DPOSS counts to g using the transformation in equation 2 from Windhorst *et al.* (1991) assuming a mean $(g-r)$ of 0.3^m, as measured in our data. The lower panel is the result of horizontally shifting all non-DPOSS counts until the DPOSS and APM counts are normalized at $g = 17.0^m$.

Figure 11: The distribution of galaxy colors in three magnitude intervals in g. The histograms are from the four DPOSS fields in our survey. The solid line is the no evolution model prediction of KGB. The color transformation from the model's $(B_J - R_F)$ to the data's $(g-r)$ system, which is approximate and may be off by as much as 0.3^m, is derived from matching star color distributions in the two photometric systems.

Figure 12: A comparison of measured object areas as a function of FOCAS total magnitude for our simulated plate image vs. a section of the scanned plate J380.

Figure 13: A comparison of measured object intensity weighted first moment radii as a function of FOCAS total magnitude for our simulated plate image vs. a section of the scanned plate J380.

Figure 14: A comparison of the number of objects detected in each of nine magnitude bins for our simulated plate image vs. two different sections of the scanned plate J380. One section was taken from the center of the plate, the other from the top.

Figure 15: The true minus measured magnitude as a function of true g magnitude as a function of true magnitude and measured magnitude type. The solid lines connect average values in one magnitude bins.

Figure 16: The average and standard deviation in one magnitude bins of the difference between the true and measured magnitudes as a function of magnitude type.

Figure 17: Differential galaxy counts measured from our simulated data using three different magnitude types. The solid line indicates the actual number of objects used to construct the data.

Figure 18: The sky measurement error, and standard deviation thereof, of the simulated plate data as a function of FOCAS total magnitude in units of the image's sky pixel-to-pixel sigma prior to pixel blurring. The resulting magnitude errors are negligible ($< 0.01^m$) below $g = 20^m$.

Figure 19: A comparison of measured sky values (in arbitrary intensity units) as a function of FOCAS total magnitude for our simulated plate image vs. a section of the scanned plate J380.

Figure 20: The effect of varying the isophotal threshold on detection efficiency and magnitude accuracy.

Figure 21: The average magnitude offset for stars and galaxies measured using different isophotal thresholds.

Figure 22: The effect of varying the seeing shape and width on detection efficiency and magnitude accuracy.

Figure 23: The average magnitude offset for stars and galaxies measured on images with varying seeing shapes and widths.

Figure 24: The effect of using the same image but different noise realizations, one of the same level, another 20% higher, on detection efficiency and magnitude accuracy.

Figure 25: The average magnitude offset for stars and galaxies measured on images with different noise.

Figure 26: Detection efficiency as a function of galaxy shape as measured by the normalized half-light radius, r_{norm}, and axial ratio, a/b.

Figure 27: The accuracy of measured magnitudes as a function of the same galaxy shape parameters as in Figure 26 out to a true magnitude of 22.0^m.

Figure 28: The difference in instrumental F magnitude as a function of mean magnitude for the galaxies in the overlap between F380 and F381. The dashed line at a difference of 0.276^m indicates the offset we obtain from a simultaneous least squares optimization of the offsets between F380, F381, and F442 in the magnitude range $15^m < F_{inst} < 19^m$.

Figure 29: We fit zero, first, and third order polynomials to measured plate and CCD red magnitudes from a combined list of objects in the four indicated plate fields. We chose the linear calibration for its theoretical appeal and because it produced the most reasonable stellar color-magnitude diagrams.

Figure 30: We fit zero, first, and third order polynomials to measured plate and CCD blue magnitudes from a combined list of objects in the four indicated plate fields. As for r, we chose to use the linear calibration.

Figure 31: The differences between measured plate and CCD magnitudes of galaxies after calibrating the instrumental plate magnitudes to the r system.

Figure 32: The differences between measured plate and CCD magnitudes of galaxies after calibrating the instrumental plate magnitudes to the g system.

Figure 33: The $g_J - r_F$ color of stars vs. r_F magnitude in Field 380 for three different methods of plate to CCD magnitude calibration. Note that the two stellar ridges take their expected linear form only in the case of zeroth and first order (offset and linear, respectively) transformations.

Field	Plate	RA (1950)	Dec	Dens	Exp (mins)	Trans	Grade	m_{lim}
J380	1744	12^h24^m	35°	1.48	80	Hazy	A	21.81
F380	3847	12^h24^m	35°	1.51	100	Cloud	B	21.39
J381	3116	12^h48^m	35°	1.56	60	Clear	B	21.81
F381	2353	12^h48^m	35°	1.22	90	Clear	B	21.12
J382	1790	13^h12^m	35°	1.13	65	Clear	A	21.81
F382	2268	13^h12^m	35°	1.28	90	Clear	A	21.39
J442	3131	12^h39^m	30°	1.87	75	Clear	A	21.81
F442	3068	12^h39^m	30°	1.28	75	Cloud	B	21.12

Table 1: Plate number, center location, approximate photographic sky density, exposure time, sky transmission quality, grade, and approximate limiting magnitude of the survey plates in our four field region.

Plates	15.0 < r < 19.0		15.0 < r < 20.0		14.5 < g < 19.5		14.5 < g < 20.5	
	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}
380/381	0.031	0.172	0.019	0.241	0.035	0.373	0.024	0.345
380/442	-0.009	0.166	-0.015	0.225	-0.055	0.201	-0.032	0.327
381/382	-0.056	0.173	-0.001	0.256	0.004	0.160	0.002	0.252
381/442	0.022	0.324	0.008	0.306	0.046	0.534	0.026	0.412

Table 2: Average offsets and standard deviations between calibrated g and r magnitudes within four plate overlap regions in two magnitude ranges.

Magnitude range	$\overline{m}_{Offset}$ mean	$\overline{m}_{Offset}$ sigma	σ_{Offset} mean
$15.0 < r < 19.0$	-0.003	0.039	0.209
$15.0 < r < 20.0$	0.003	0.014	0.257
$14.5 < g < 19.5$	0.008	0.045	0.317
$14.5 < g < 20.5$	0.005	0.027	0.334

Table 3: Average offsets and standard deviations between calibrated g and r plate magnitudes across four field overlaps in two magnitude ranges. The offset means and sigmas are computed using the overlap $\overline{m}$ measurements listed in Table 2. The mean σ_{Offset} values are derived from the σ measurements from the same table.

Plates	Δ F_{inst} Original			Δ F_{inst} Fitted		
380 - 381	0.289	±	0.227	0.019	±	0.208
380 - 442	0.131	±	0.201	-0.012	±	0.197
380 - 381	-0.119	±	0.283	0.009	±	0.280

Plates	Δ J_{inst} Original			Δ J_{inst} Fitted		
380 - 381	0.198	±	0.342	0.021	±	0.342
380 - 442	-0.039	±	0.311	-0.026	±	0.311
380 - 381	0.160	±	0.406	0.024	±	0.407

Table 4: The mean and standard deviation of the measured difference in instrumental magnitudes of galaxies in the indicated plate overlaps, before and after applying fitted plate offsets. The measurements were obtained over the magnitude range $15^m < F_{instr} < 19^m$ and $16^m < J_{inst} < 20^m$.

Plate	15.0 < r < 19.0		15.0 < r < 20.0		14.5 < g < 19.5		14.5 < g < 20.5	
	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}	$\overline{m}_{Offset}$	σ_{Offset}
380	0.039	0.237	0.049	0.218	0.008	0.192	-0.043	0.273
381	-0.034	0.159	0.018	0.196	0.042	0.194	0.022	0.223
382	0.027	0.129	0.034	0.168	0.084	0.336	0.071	0.286
442	-0.021	0.195	-0.042	0.230	-0.106	0.320	-0.144	0.385

Table 5: Average offsets and standard deviations between calibrated plate magnitudes and corresponding CCD magnitudes in g and r for fields 380, 381, 382, and 442 in two magnitude ranges.

Magnitude range	$\overline{m}_{Offset}$ mean	$\overline{m}_{Offset}$ sigma	σ_{Offset} mean
$15.0 < r < 19.0$	0.003	0.036	0.180
$15.0 < r < 20.0$	0.015	0.040	0.203
$14.5 < g < 19.5$	0.007	0.081	0.261
$14.5 < g < 20.5$	-0.024	0.093	0.292

Table 6: Average offsets and standard deviations between calibrated plate magnitudes and corresponding CCD magnitudes in g and r across four fields in two magnitude ranges. The offset means and sigmas are computed using the plate by plate $\overline{m}$ measurements listed in Table 5. The mean σ_{Offset} values are derived from the σ measurements from the same table.

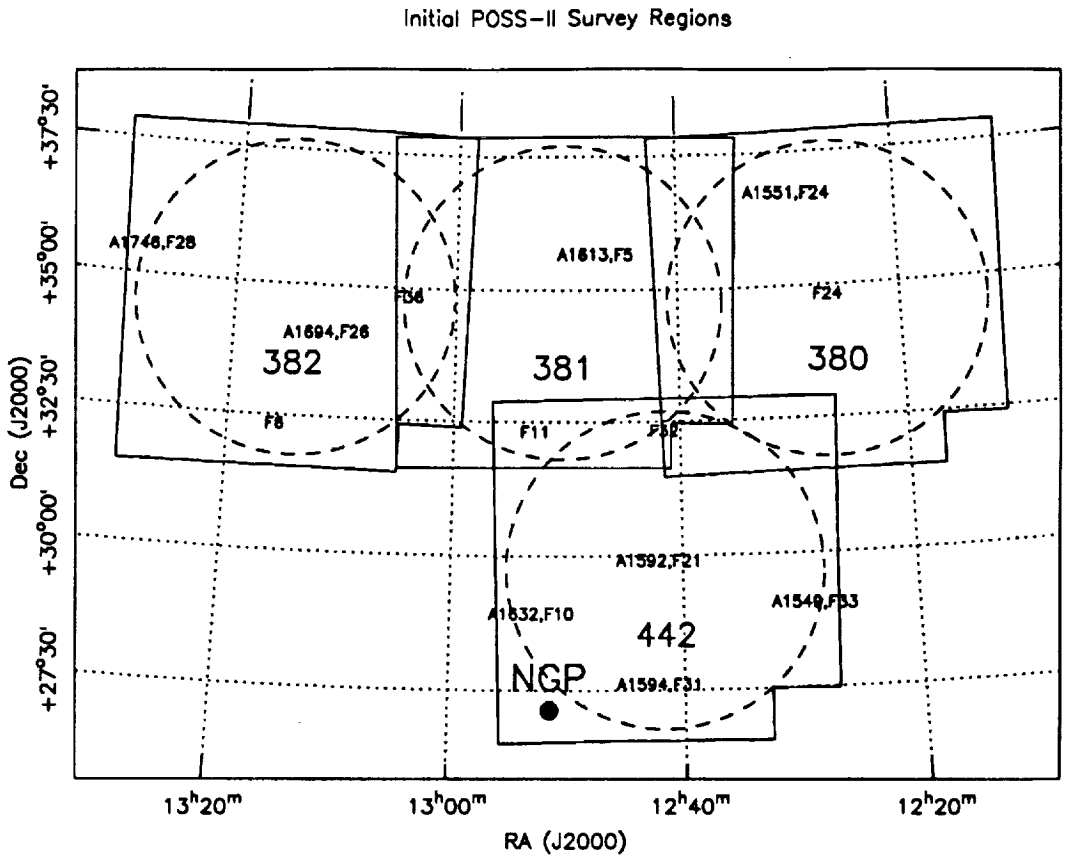
Initial POSS–II Survey Regions
Dec (J2000)
+37°30'
+35°00'
+32°30'
+30°00'
+27°30'
A1551,F24
A1746,F28
A1613,F5
F56
F24
A1694,F26
382
381
380
F8
F11
F32
A1592,F21
A1548,F33
A1632,F10
442
NGP
A1594,F31
13h20m
13h00m
12h40m
12h20m
RA (J2000)

Figure 1:

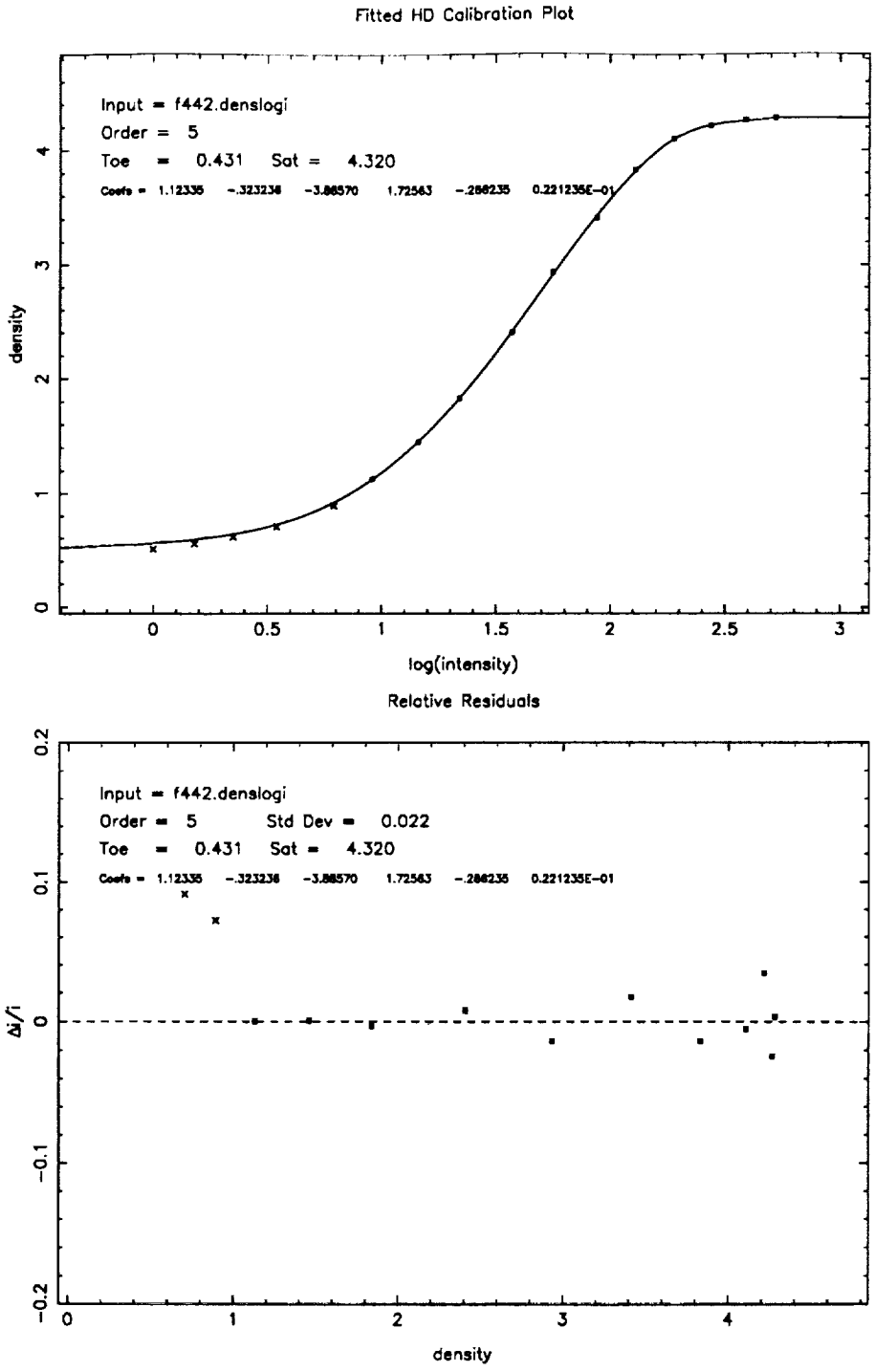
Fitted HD Calibration Plot
Input = f442.denslogi
Order = 5
Toe = 0.431 Sat = 4.320
Coefs = 1.12335 -.323238 -3.88570 1.72563 -.286235 0.221235E-01
density
log(intensity)
Relative Residuals
Input = f442.denslogi
Order = 5 Std Dev = 0.022
Toe = 0.431 Sat = 4.320
Coefs = 1.12335 -.323238 -3.88570 1.72563 -.286235 0.221235E-01
Δi/i
density

Figure 2:

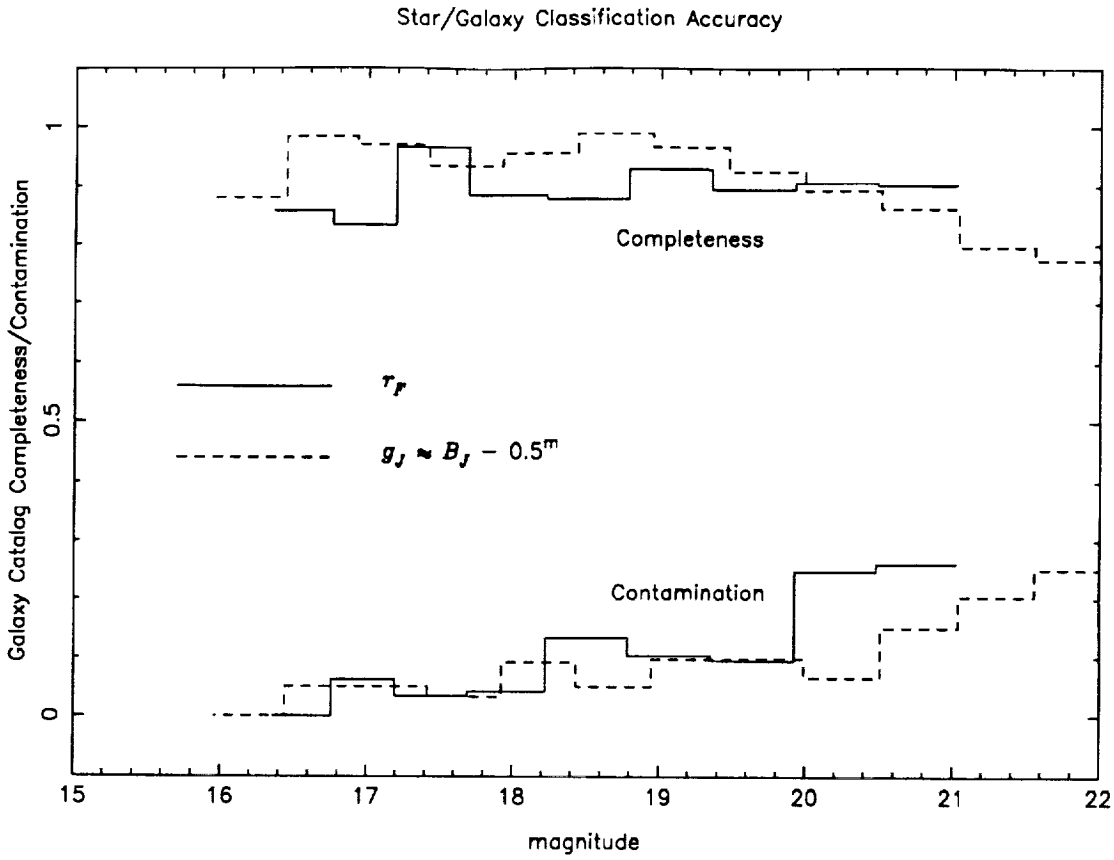
Star/Galaxy Classification Accuracy
Galaxy Catalog Completeness/Contamination
0
0.5
1
Completeness
r_F
$g_J \approx B_J - 0.5^m$
Contamination
15
16
17
18
19
20
21
22
magnitude

Figure 3:

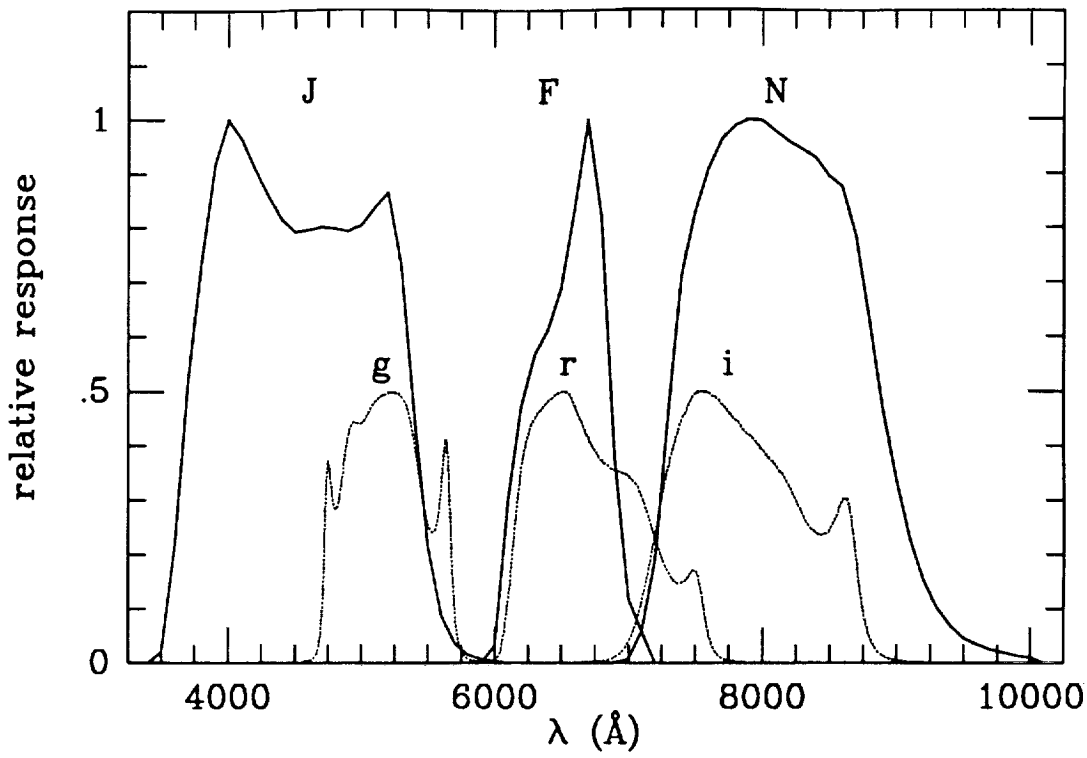

Figure 4:

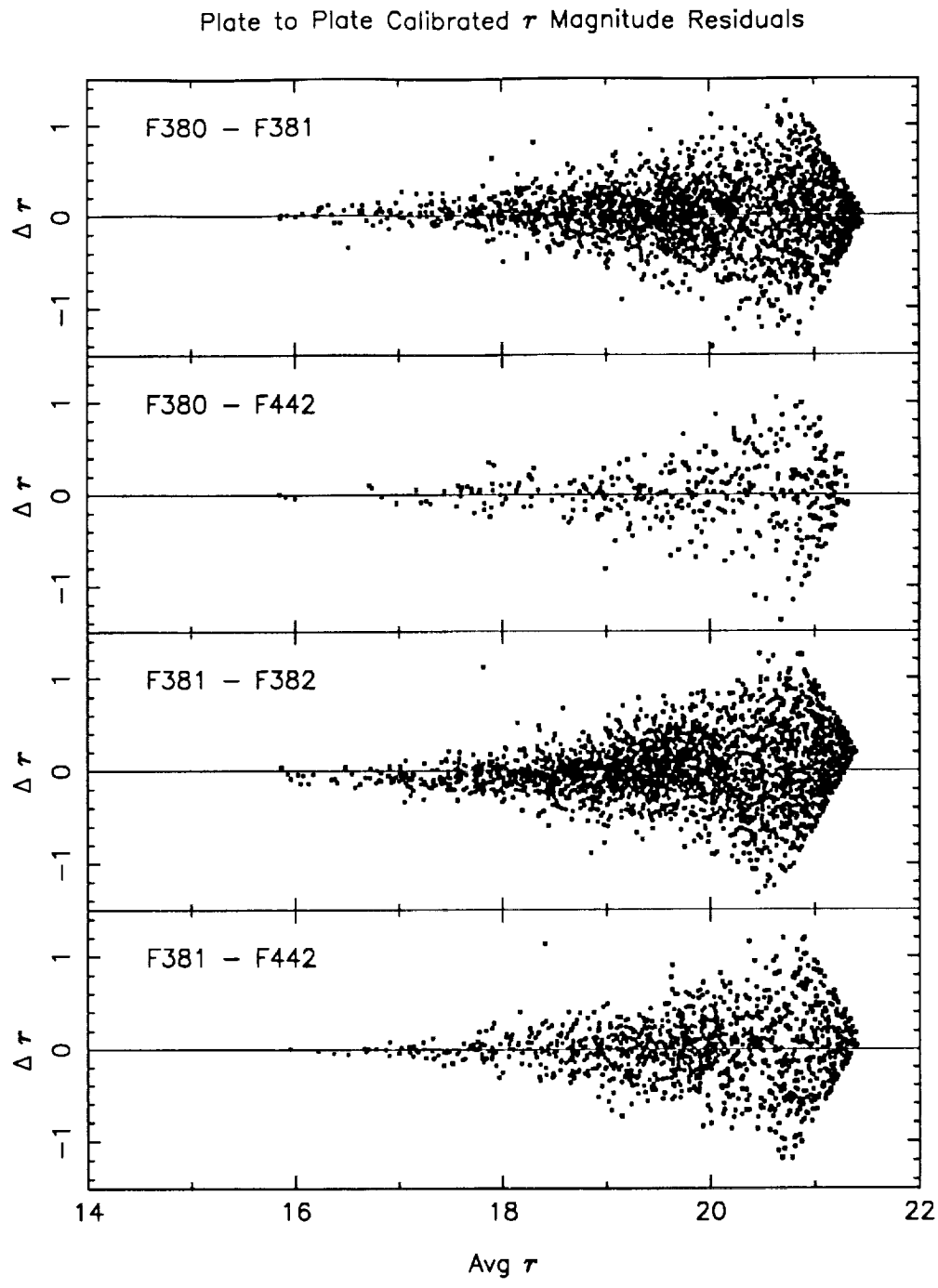
Plate to Plate Calibrated τ Magnitude Residuals
F380 – F381
F380 – F442
F381 – F382
F381 – F442
Δ τ
1
0
−1
14
16
18
20
22
Avg τ

Figure 5:

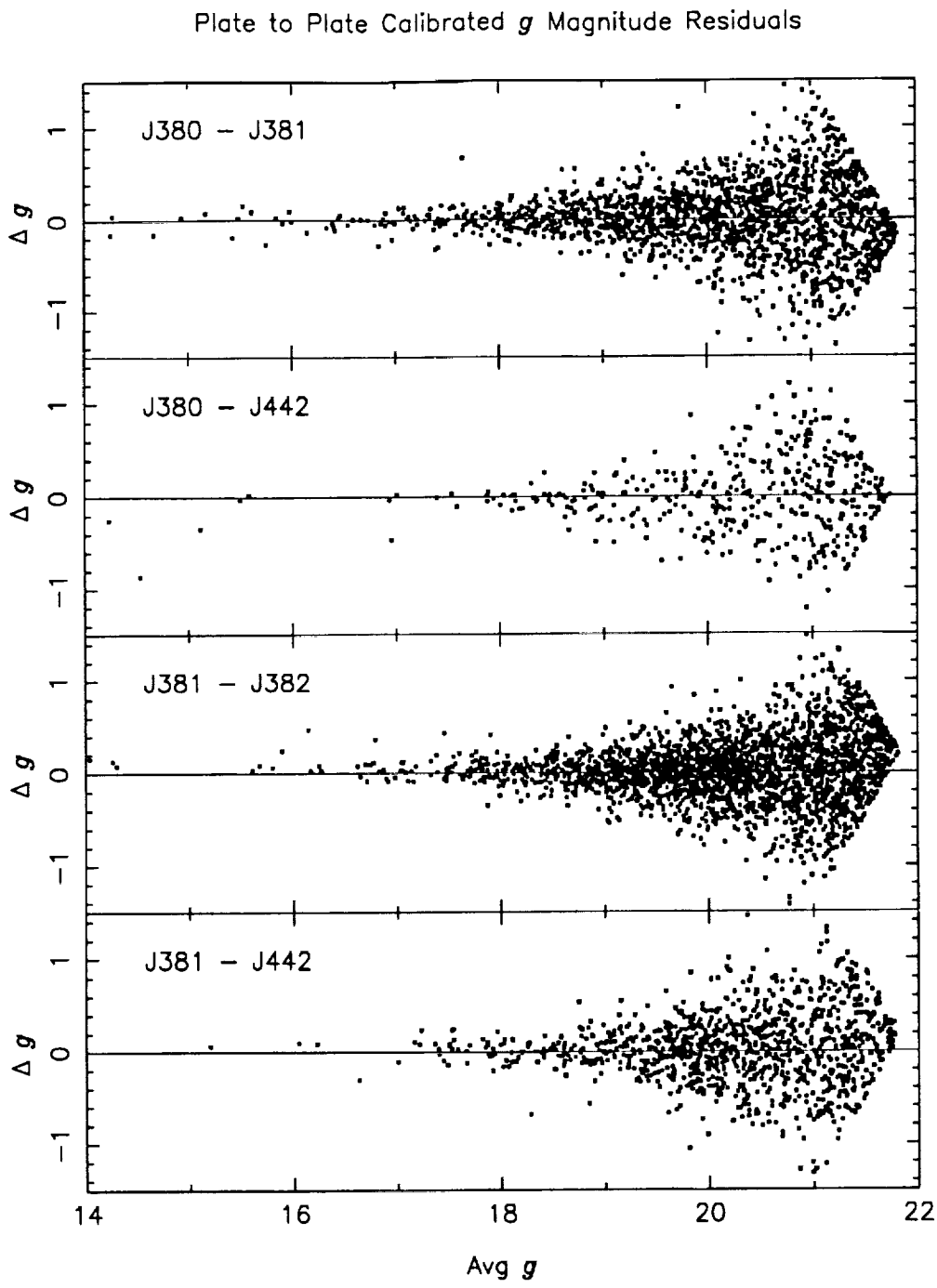
Plate to Plate Calibrated g Magnitude Residuals
J380 – J381
J380 – J442
J381 – J382
J381 – J442
Δ g
1
0
−1
14
16
18
20
22
Avg g

Figure 6:

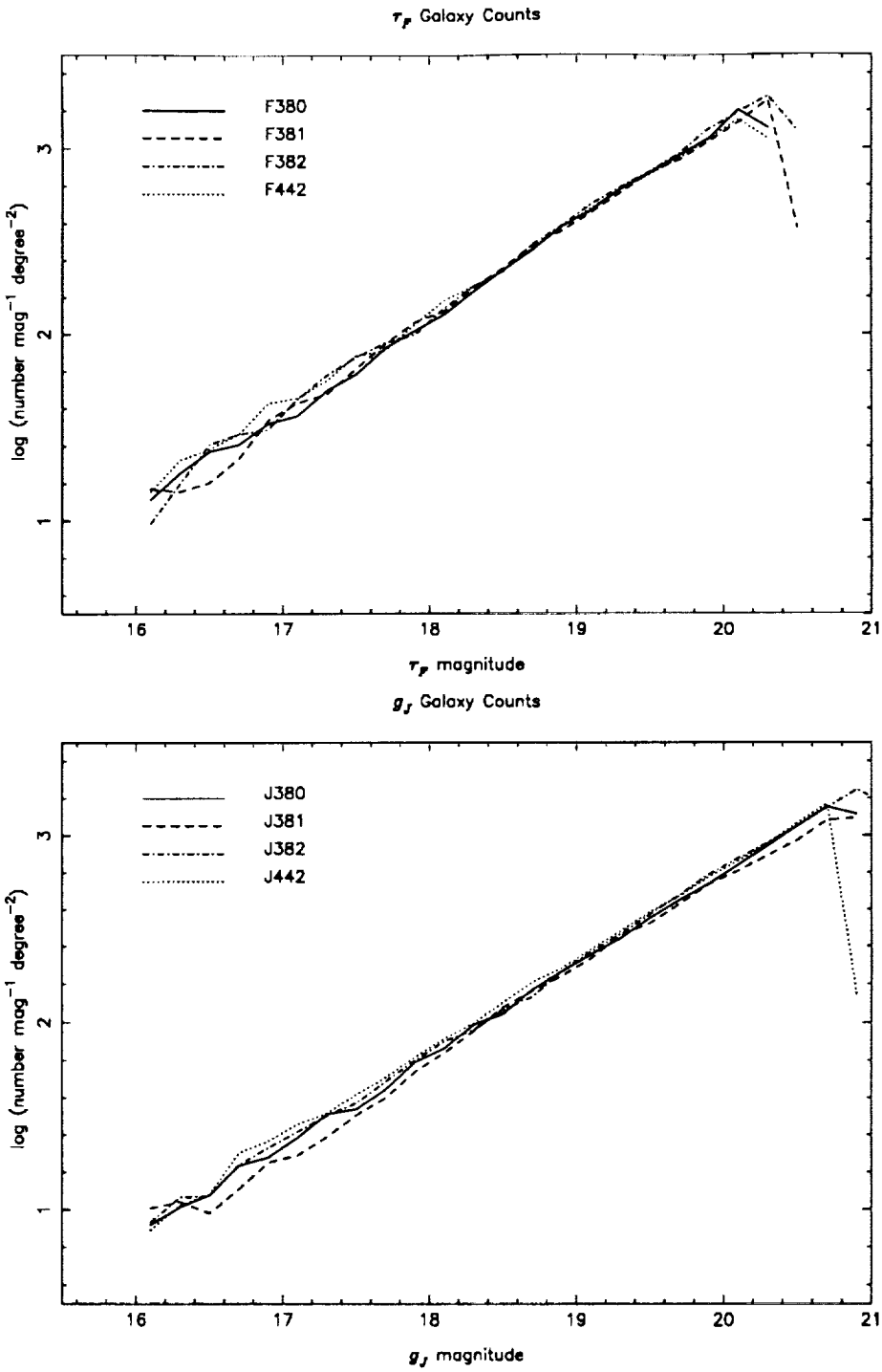
τ_F Galaxy Counts
F380
F381
F382
F442
log (number mag^{-1} degree^{-2})
τ_F magnitude
g_J Galaxy Counts
J380
J381
J382
J442
log (number mag^{-1} degree^{-2})
g_J magnitude

Figure 7:

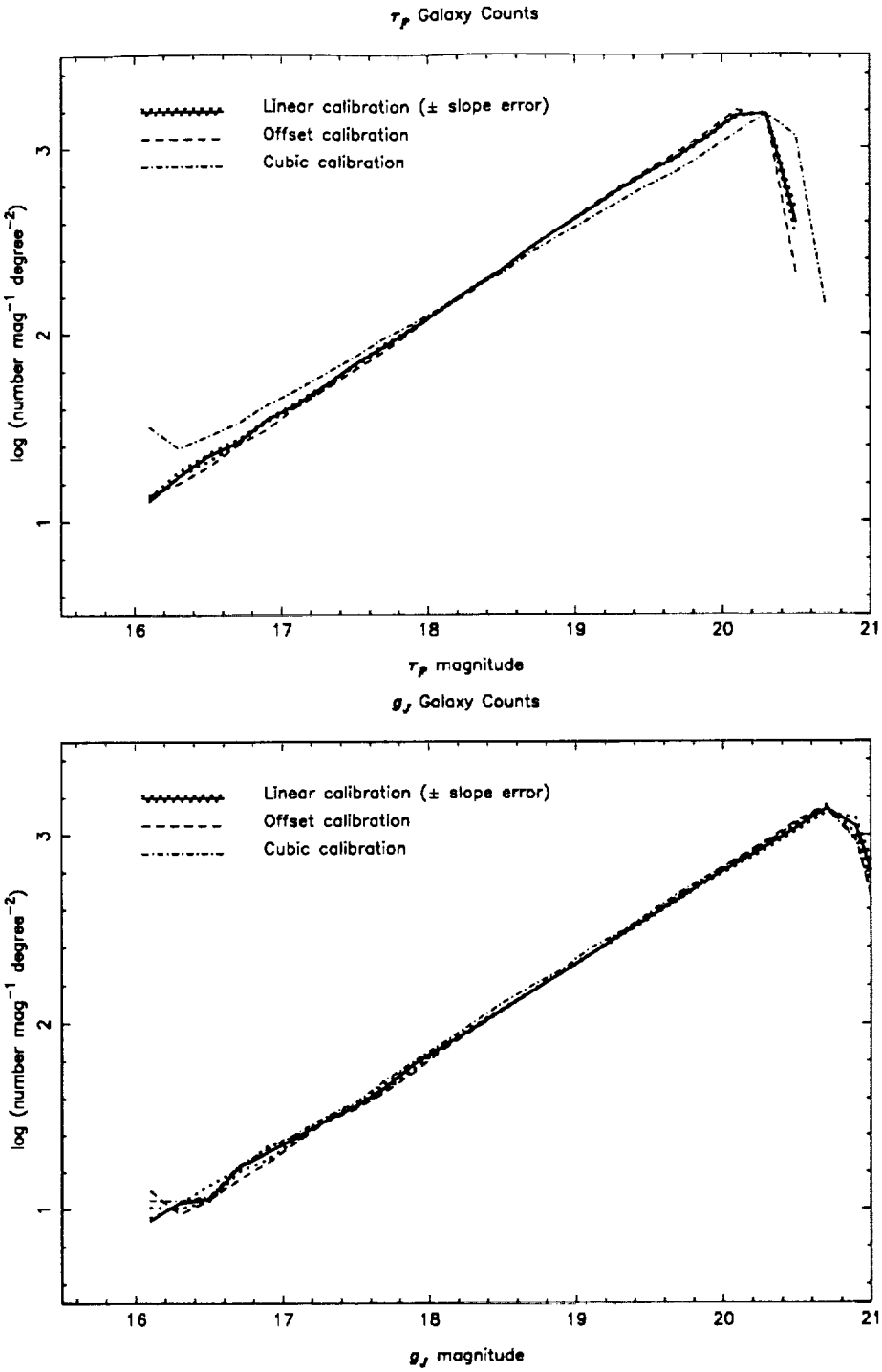
r_F Galaxy Counts
Linear calibration (± slope error)
Offset calibration
Cubic calibration
log (number mag^{-1} degree^{-2})
1
2
3
16
17
18
19
20
21
r_F magnitude
g_J Galaxy Counts
Linear calibration (± slope error)
Offset calibration
Cubic calibration
log (number mag^{-1} degree^{-2})
1
2
3
16
17
18
19
20
21
g_J magnitude

Figure 8:

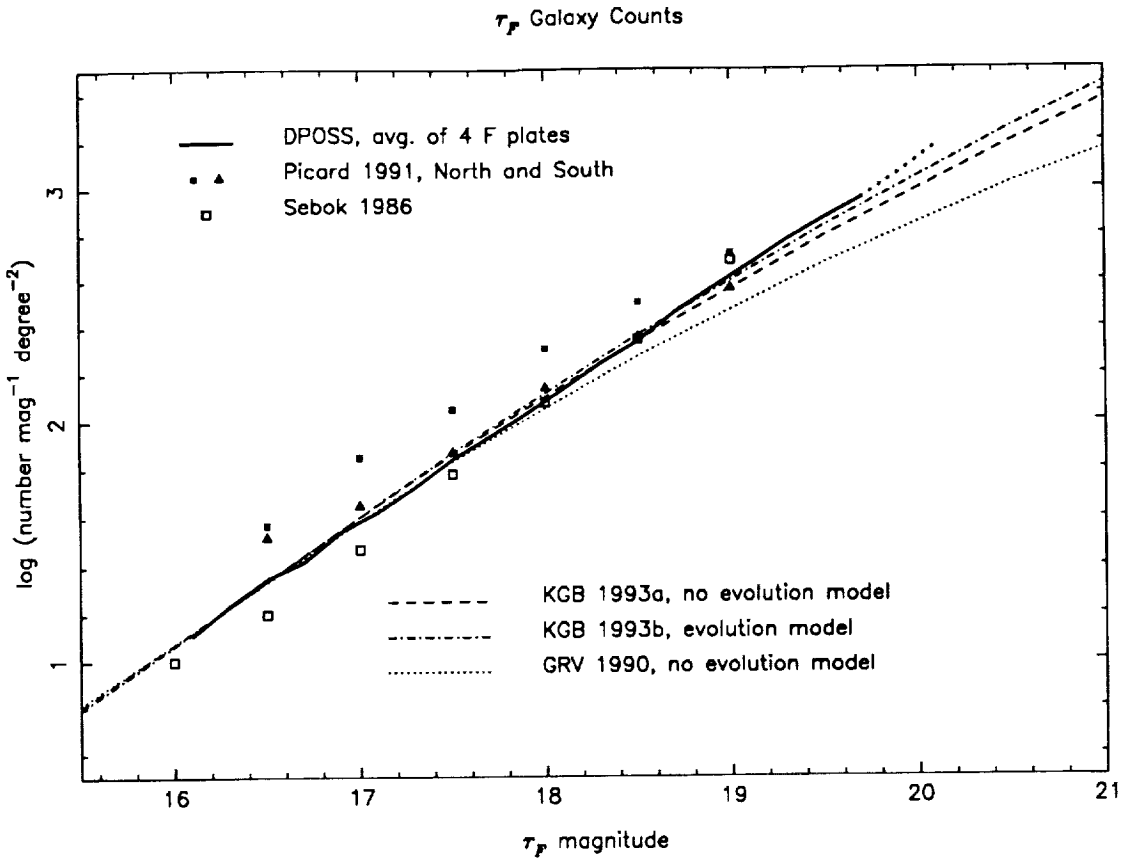
r_F Galaxy Counts
DPOSS, avg. of 4 F plates
Picard 1991, North and South
Sebok 1986
KGB 1993a, no evolution model
KGB 1993b, evolution model
GRV 1990, no evolution model
log (number mag^{-1} degree^{-2})
r_F magnitude
1
2
3
16
17
18
19
20
21

Figure 9:

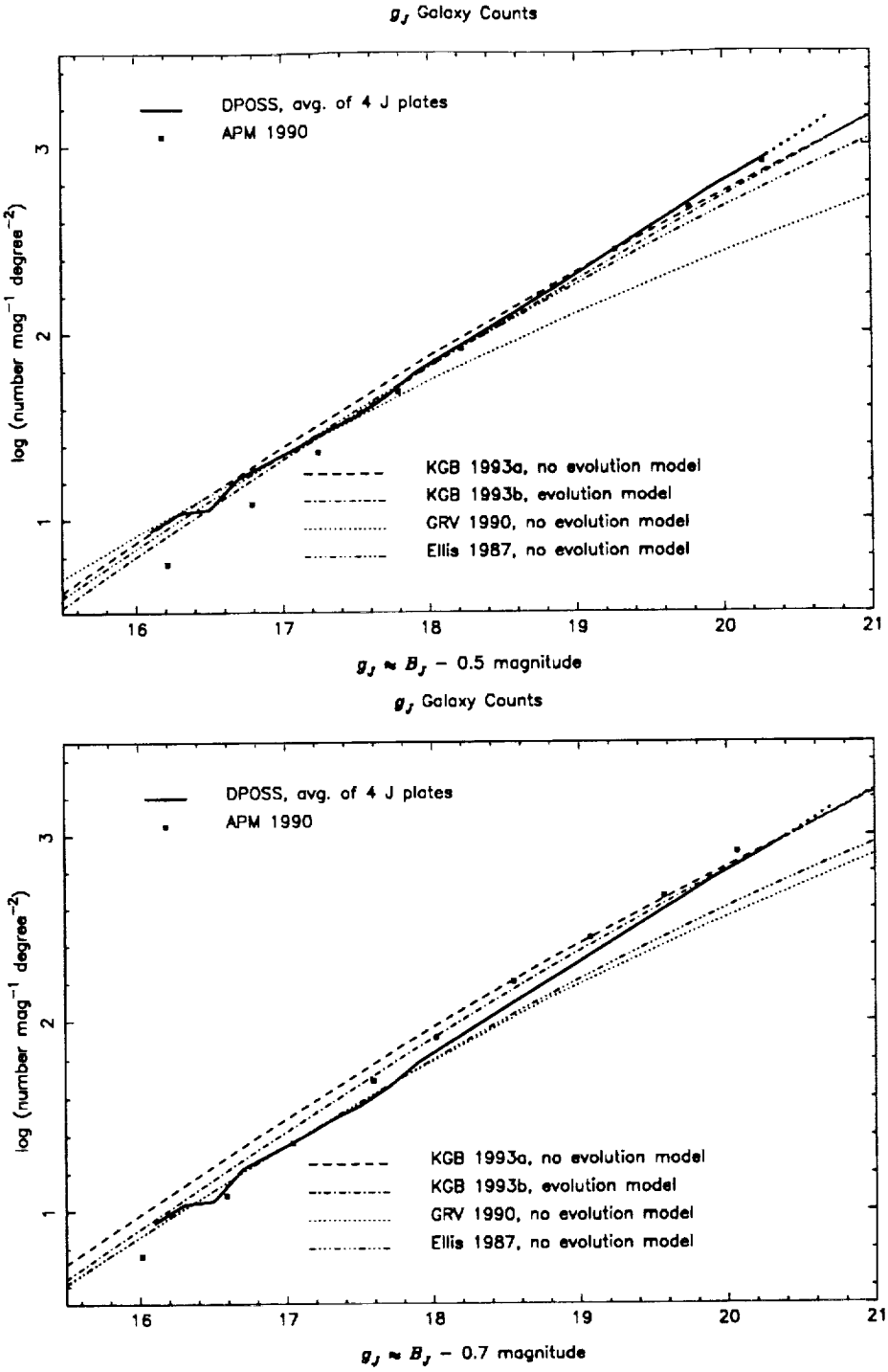
g_J Galaxy Counts
DPOSS, avg. of 4 J plates
APM 1990
KGB 1993a, no evolution model
KGB 1993b, evolution model
GRV 1990, no evolution model
Ellis 1987, no evolution model
log (number mag^{-1} degree^{-2})
$g_J \approx B_J$ − 0.5 magnitude
g_J Galaxy Counts
DPOSS, avg. of 4 J plates
APM 1990
KGB 1993a, no evolution model
KGB 1993b, evolution model
GRV 1990, no evolution model
Ellis 1987, no evolution model
log (number mag^{-1} degree^{-2})
$g_J \approx B_J$ − 0.7 magnitude

Figure 10:

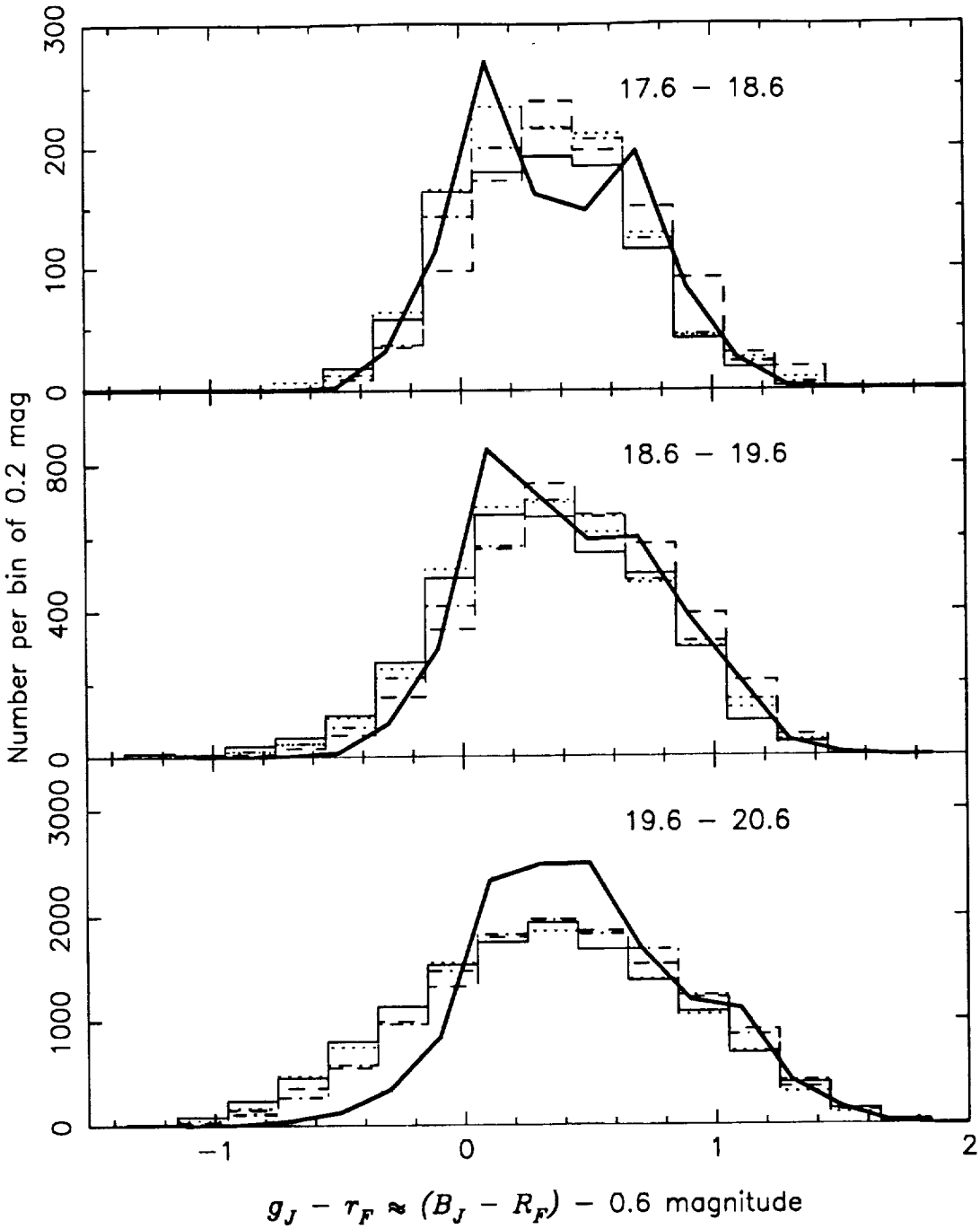
17.6 – 18.6
18.6 – 19.6
19.6 – 20.6
Number per bin of 0.2 mag
0
100
200
300
0
400
800
0
1000
2000
3000
−1
0
1
2
$g_J - r_F \approx (B_J - R_F)$ – 0.6 magnitude

Figure 11:

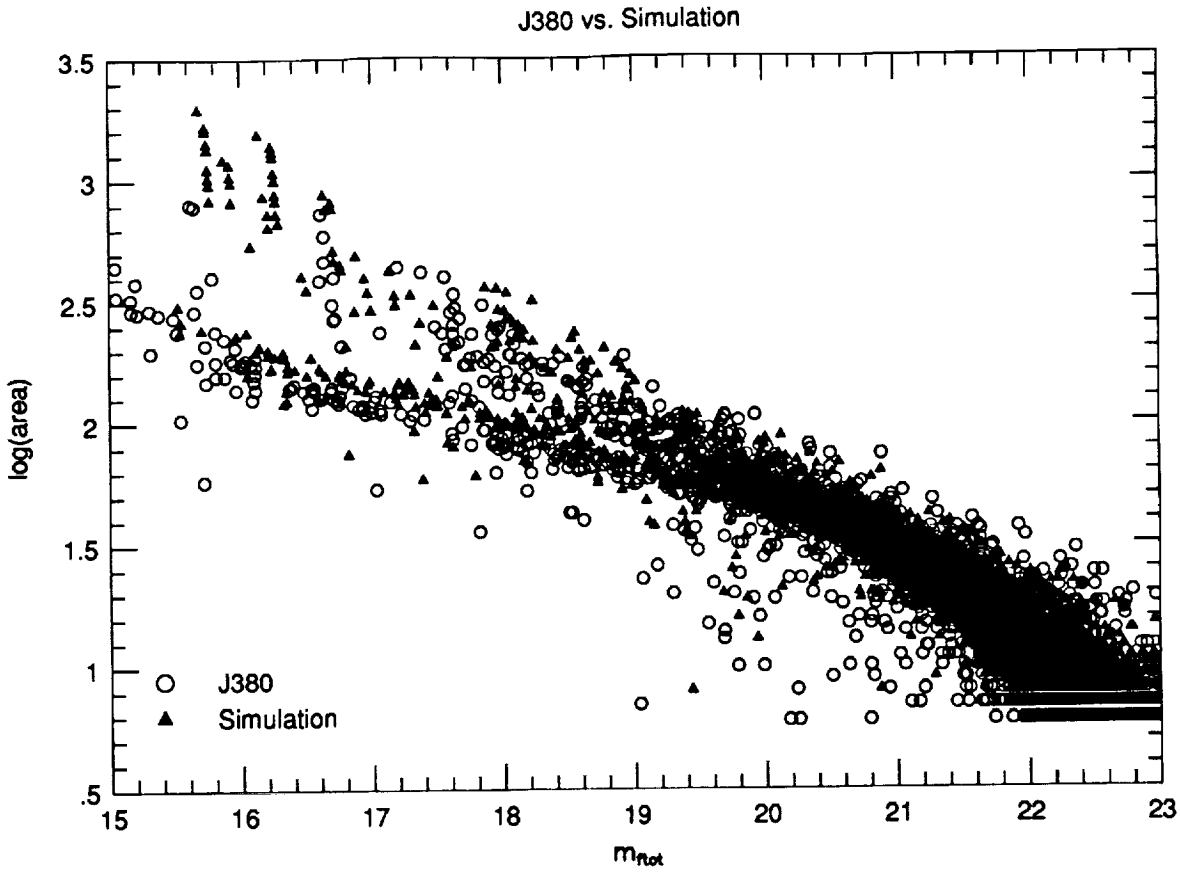
J380 vs. Simulation
log(area)
3.5
3
2.5
2
1.5
1
.5
15
16
17
18
19
20
21
22
23
m_{rot}
J380
Simulation

Figure 12:

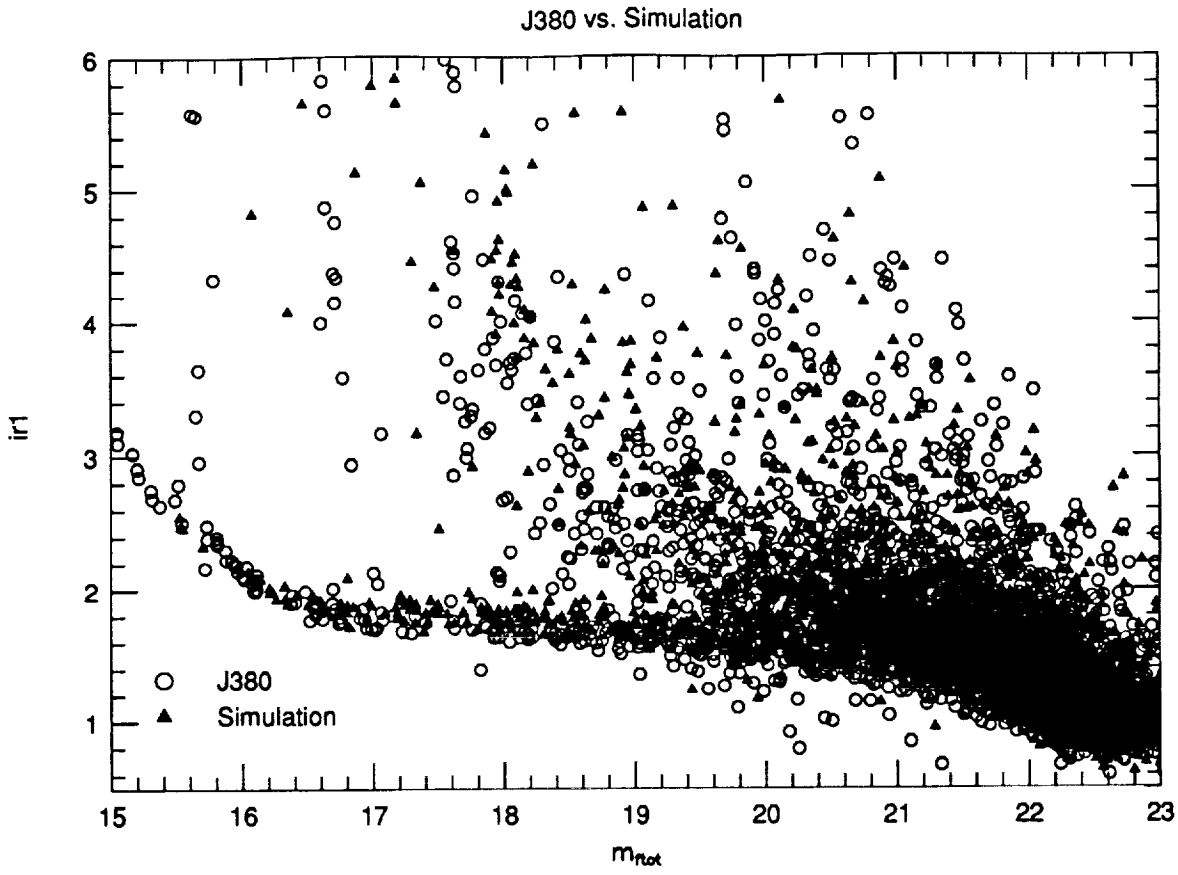
J380 vs. Simulation
ir1
6
5
4
3
2
1
15
16
17
18
19
20
21
22
23
m_{rtot}
J380
Simulation

Figure 13:

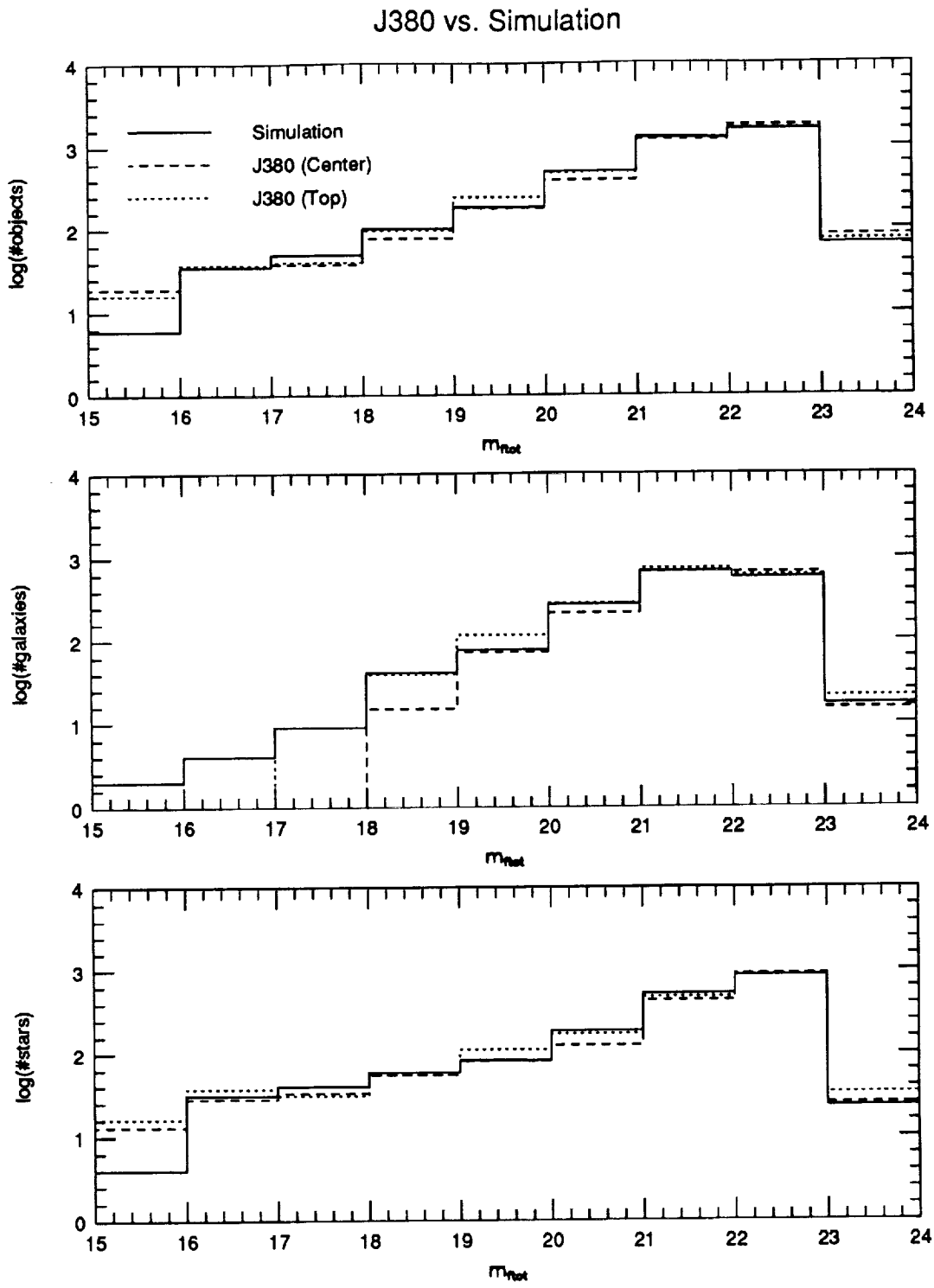
J380 vs. Simulation
Simulation
J380 (Center)
J380 (Top)
log(#objects)
log(#galaxies)
log(#stars)
m_{tot}
0
1
2
3
4
15
16
17
18
19
20
21
22
23
24

Figure 14:

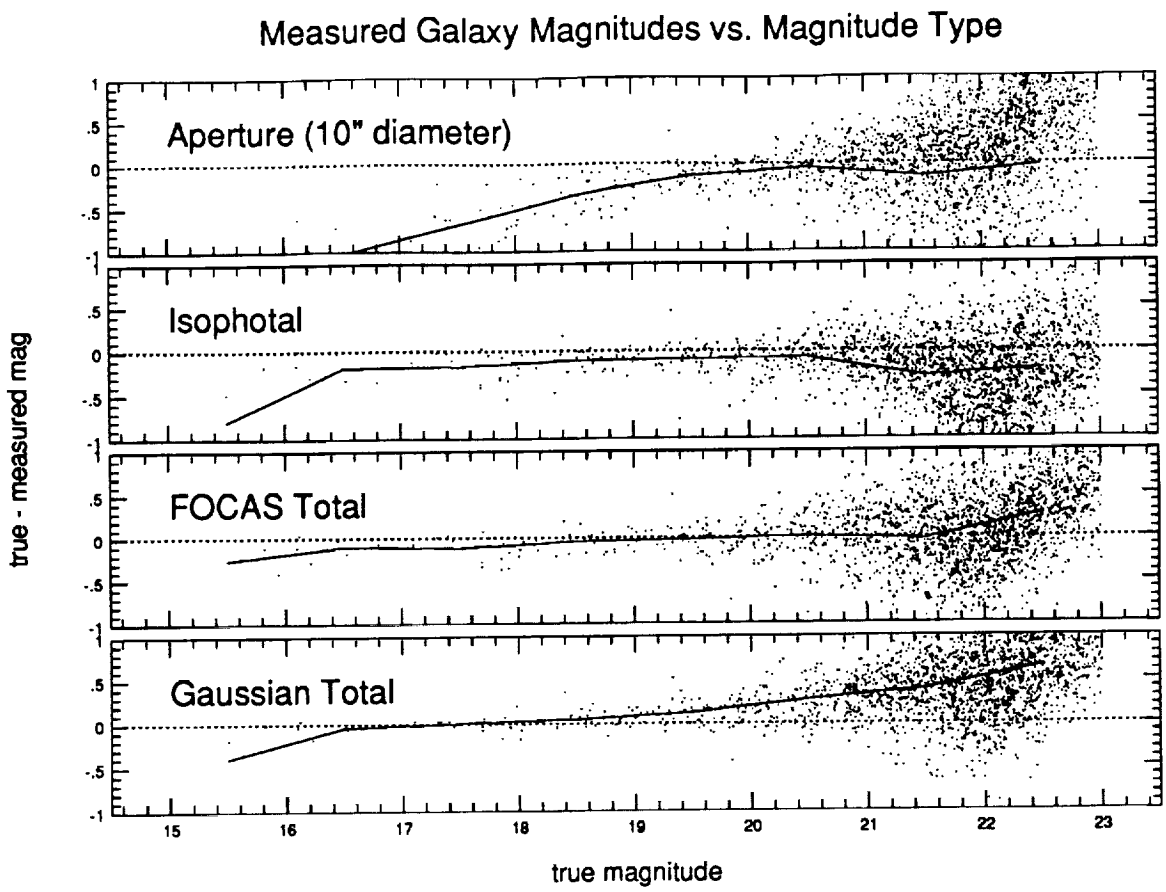
Measured Galaxy Magnitudes vs. Magnitude Type
Aperture (10" diameter)
Isophotal
FOCAS Total
Gaussian Total
true - measured mag
true magnitude
1
.5
0
-.5
-1
15
16
17
18
19
20
21
22
23

Figure 15:

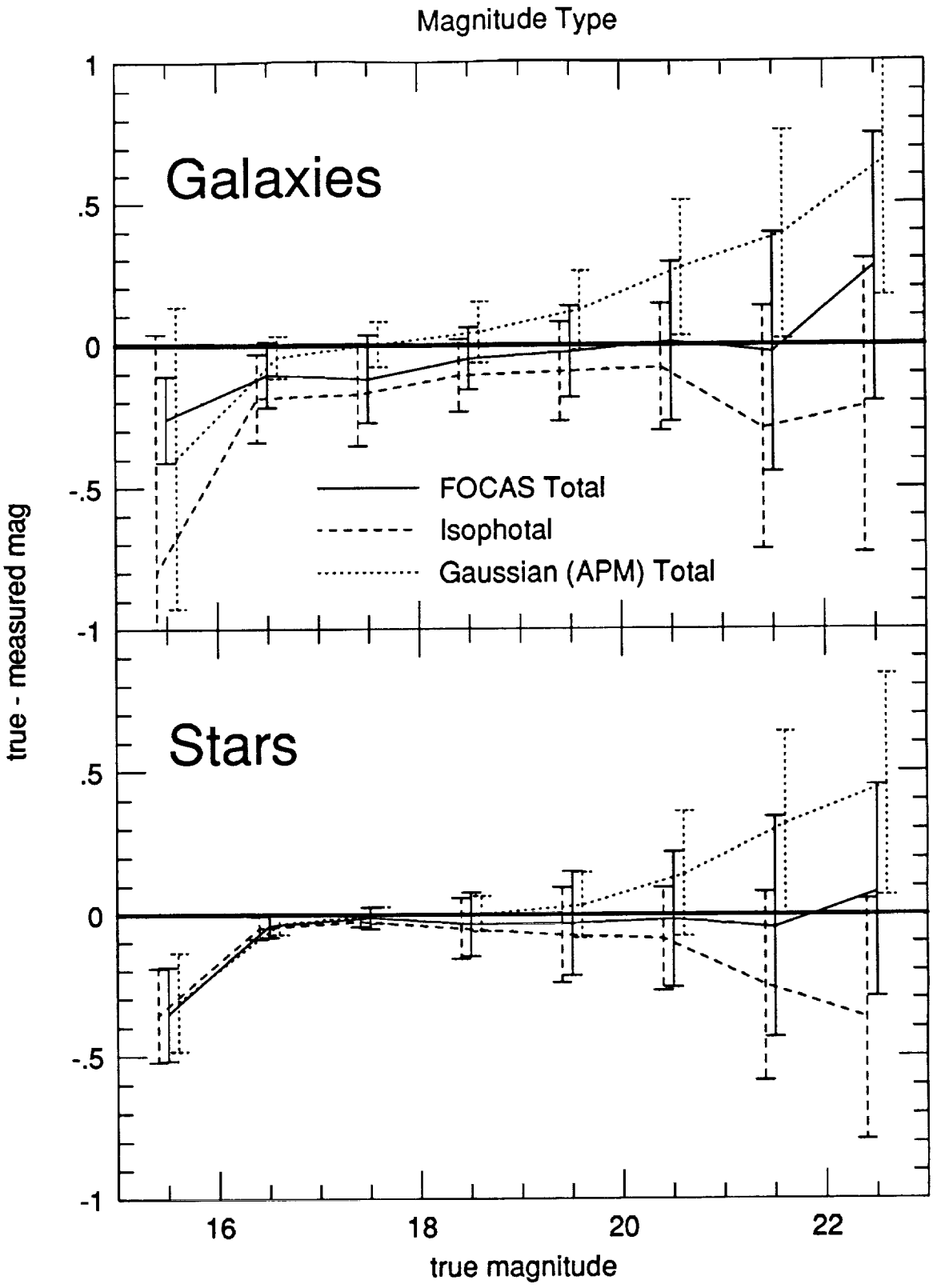
Magnitude Type
Galaxies
Stars
FOCAS Total
Isophotal
Gaussian (APM) Total
true - measured mag
true magnitude
1
.5
0
-.5
-1
16
18
20
22

Figure 16:

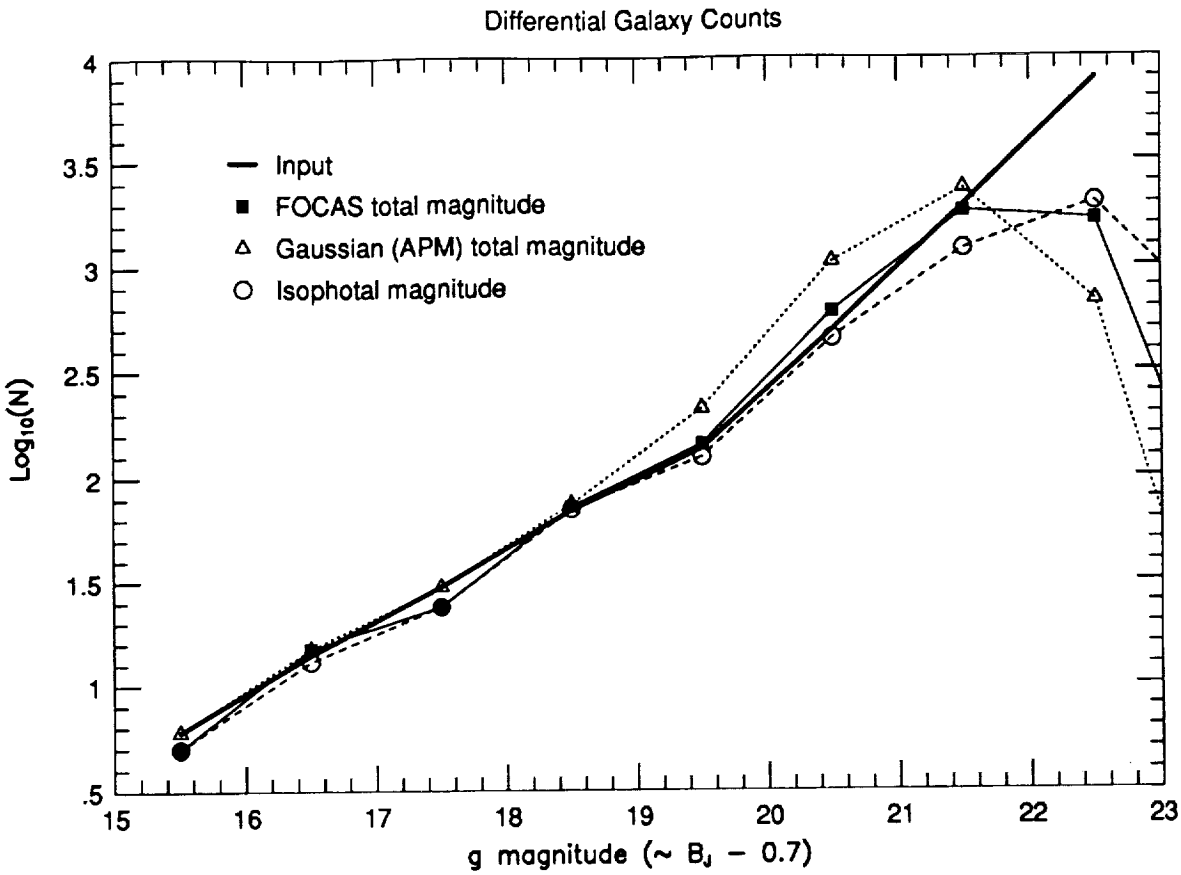
Differential Galaxy Counts
Input
FOCAS total magnitude
Gaussian (APM) total magnitude
Isophotal magnitude
Log10(N)
g magnitude (~ Bj − 0.7)
4
3.5
3
2.5
2
1.5
1
.5
15
16
17
18
19
20
21
22
23

Figure 17:

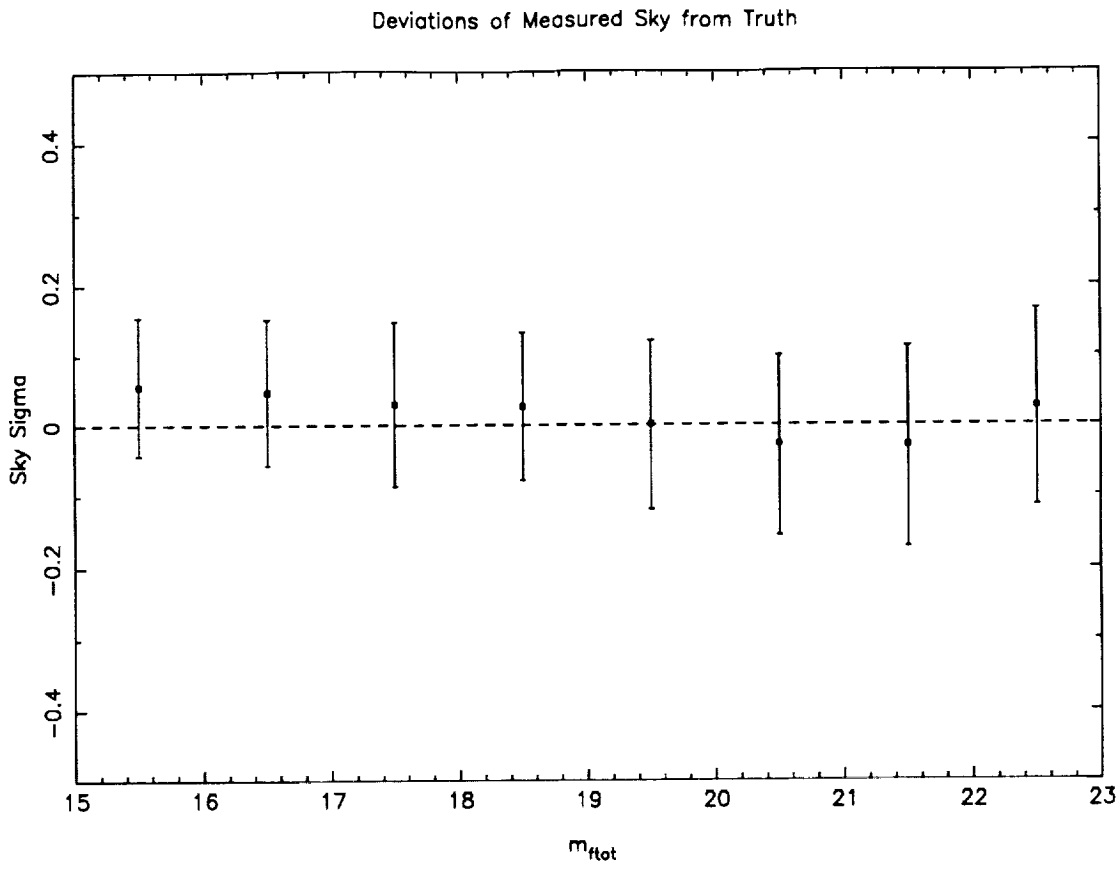
Deviations of Measured Sky from Truth
Sky Sigma
0.4
0.2
0
−0.2
−0.4
15
16
17
18
19
20
21
22
23
m_{ftot}

Figure 18:

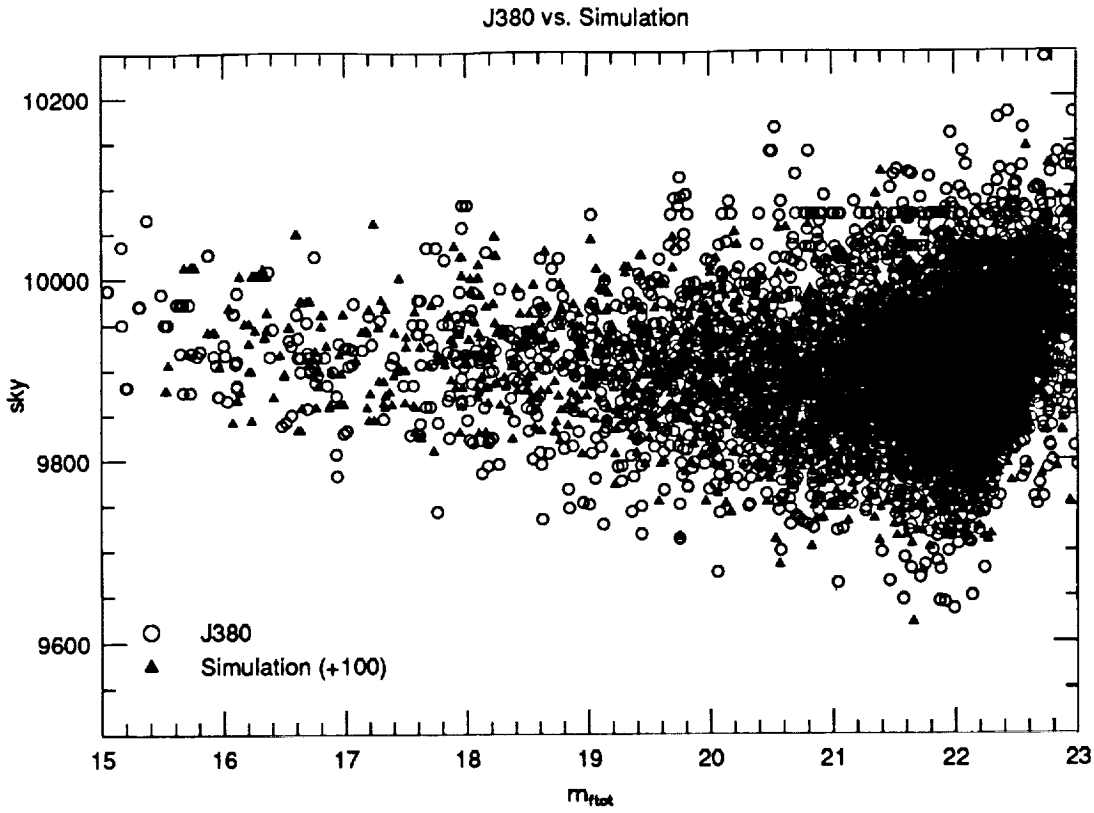
J380 vs. Simulation
10200
10000
9800
9600
sky
J380
Simulation (+100)
15
16
17
18
19
20
21
22
23
m_{ftot}

Figure 19:

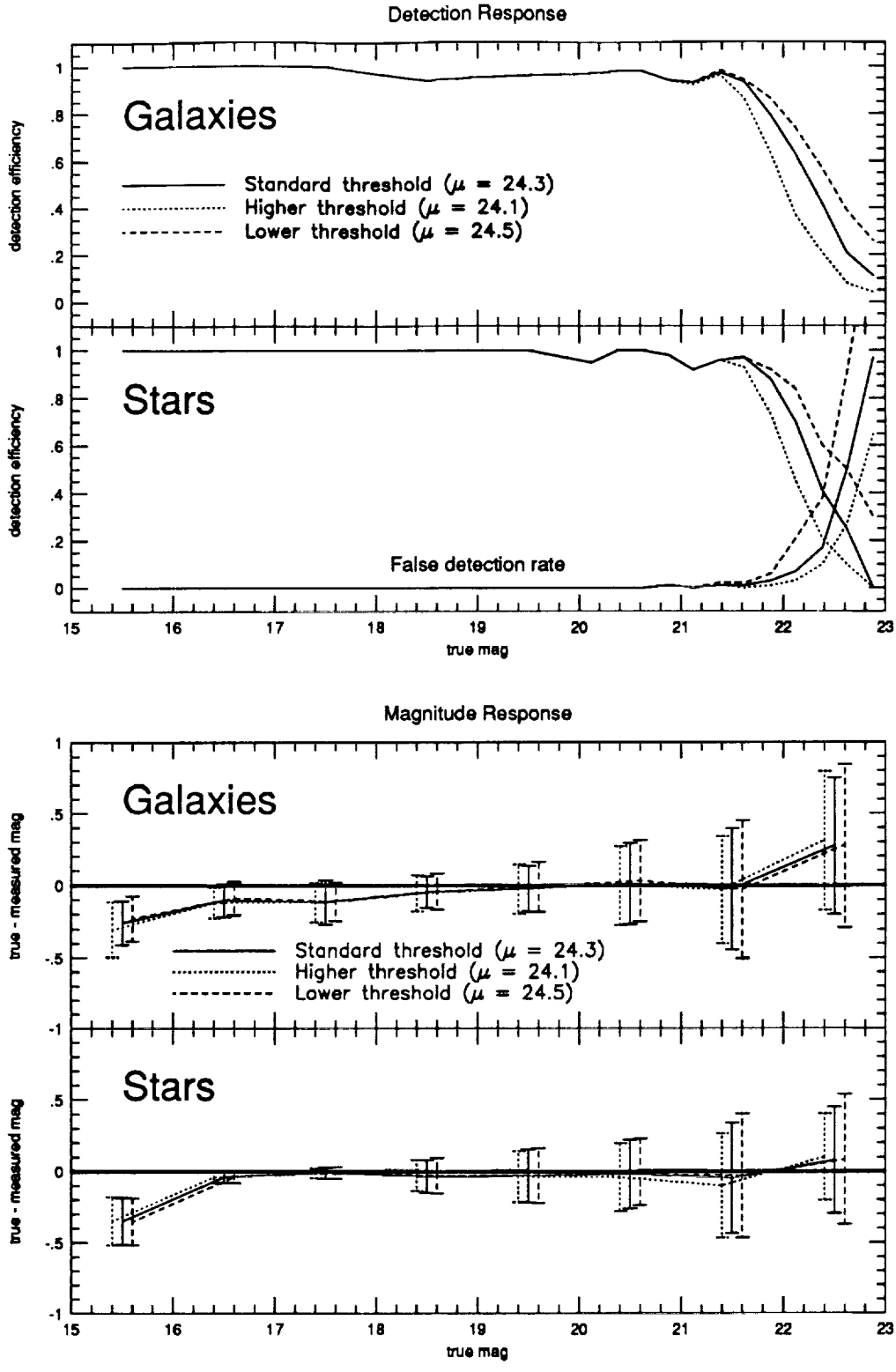
Detection Response
Galaxies
Standard threshold (μ = 24.3)
Higher threshold (μ = 24.1)
Lower threshold (μ = 24.5)
detection efficiency
Stars
False detection rate
true mag
Magnitude Response
Galaxies
Standard threshold (μ = 24.3)
Higher threshold (μ = 24.1)
Lower threshold (μ = 24.5)
true - measured mag
Stars
true mag

Figure 20:

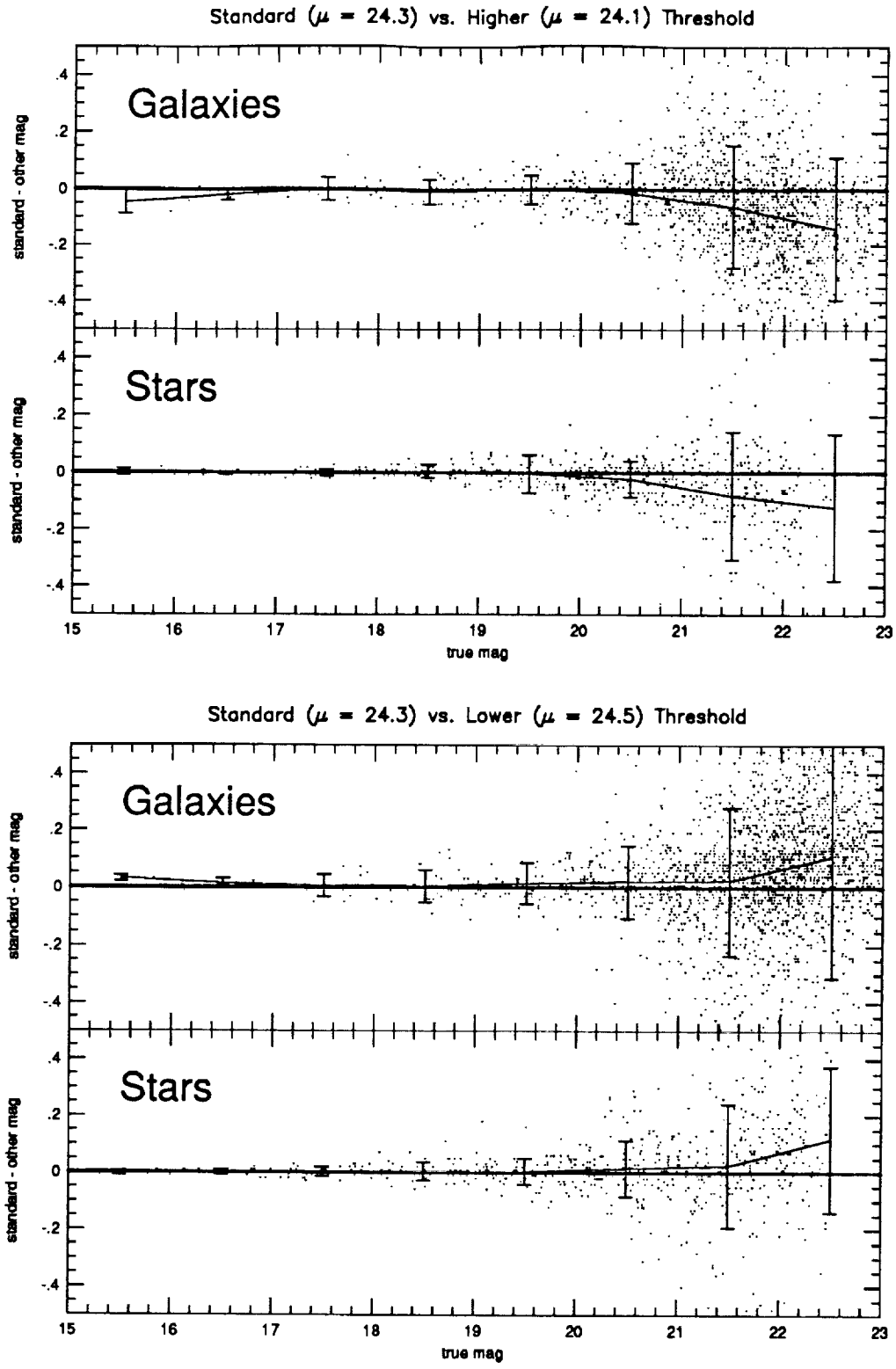
Standard (μ = 24.3) vs. Higher (μ = 24.1) Threshold
Galaxies
Stars
standard - other mag
true mag
Standard (μ = 24.3) vs. Lower (μ = 24.5) Threshold
Galaxies
Stars
standard - other mag
true mag

Figure 21:

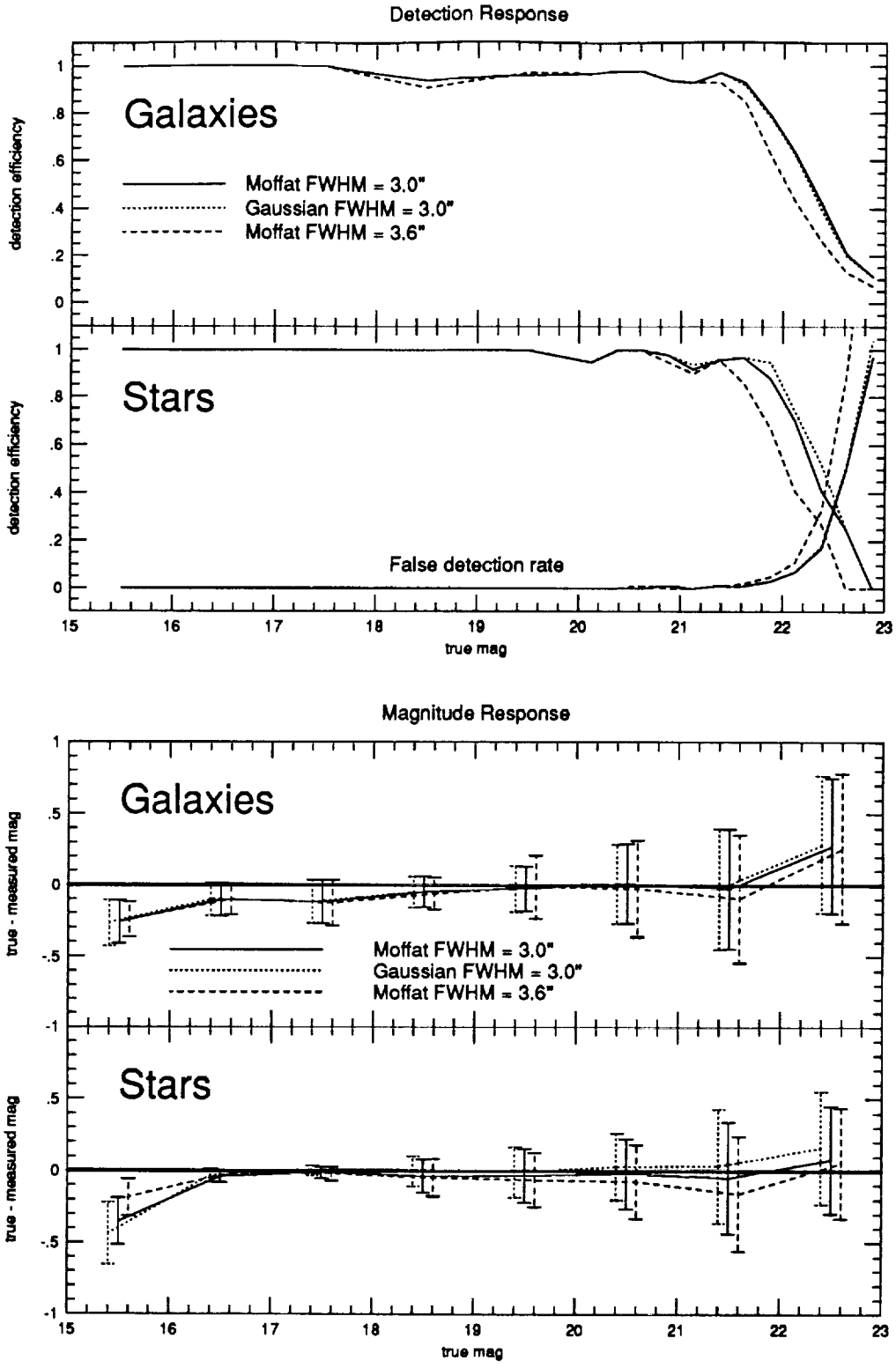
Detection Response
Galaxies
Moffat FWHM = 3.0"
Gaussian FWHM = 3.0"
Moffat FWHM = 3.6"
detection efficiency
Stars
False detection rate
true mag
Magnitude Response
Galaxies
Moffat FWHM = 3.0"
Gaussian FWHM = 3.0"
Moffat FWHM = 3.6"
true - measured mag
Stars
true mag

Figure 22:

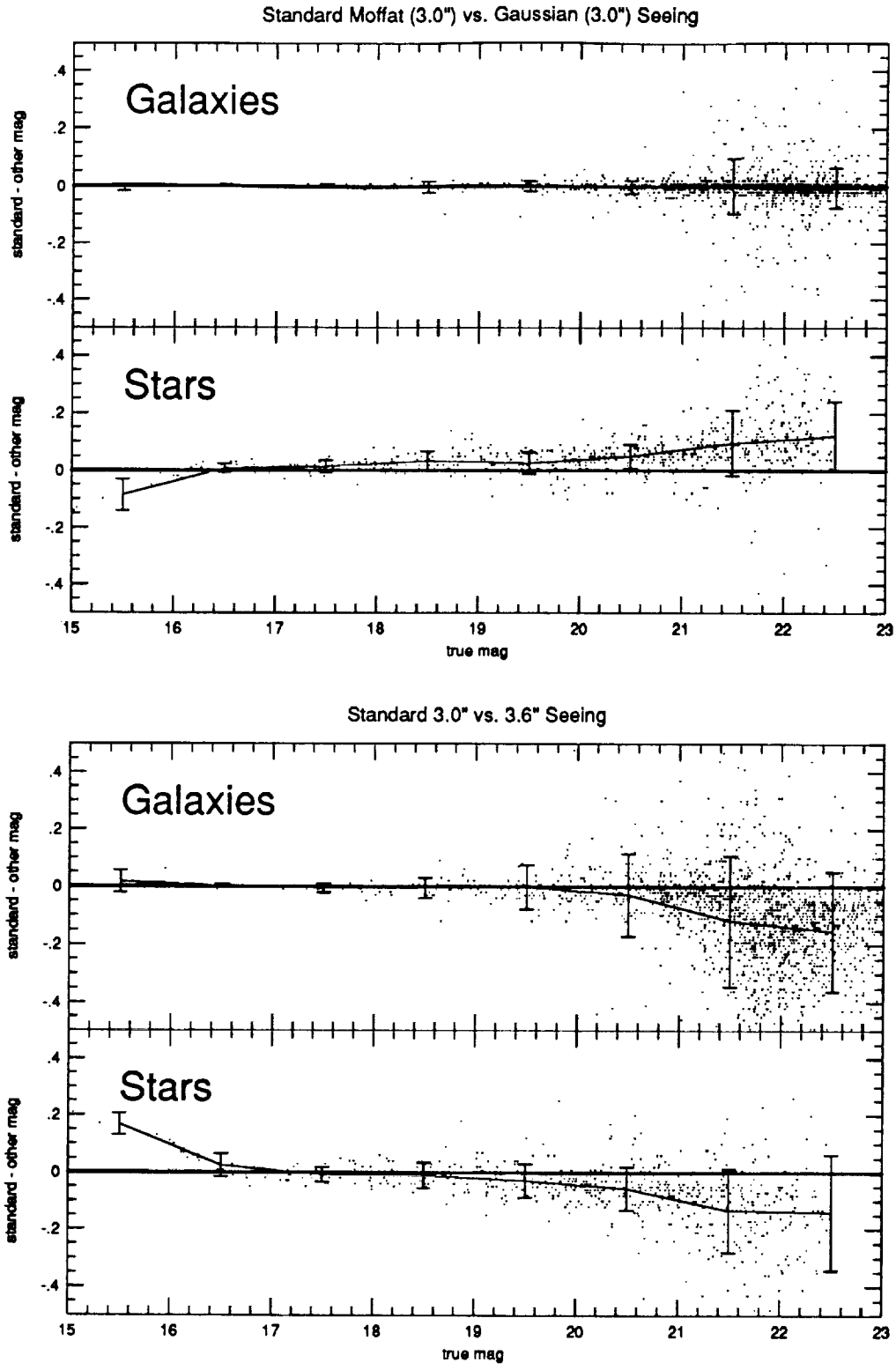
Standard Moffat (3.0") vs. Gaussian (3.0") Seeing
Galaxies
Stars
standard - other mag
true mag
Standard 3.0" vs. 3.6" Seeing
Galaxies
Stars
standard - other mag
true mag

Figure 23:

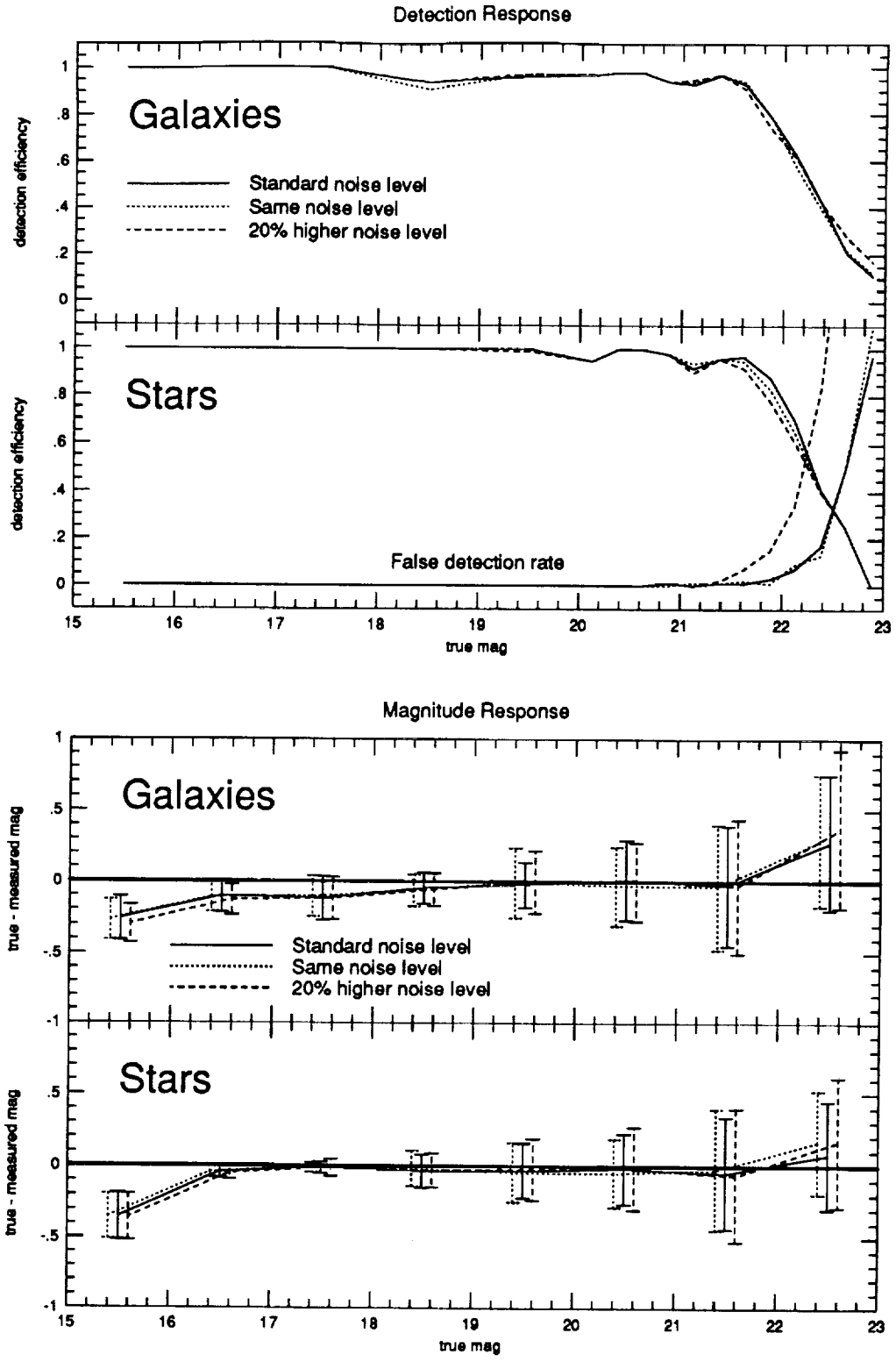
Detection Response
Galaxies
Standard noise level
Same noise level
20% higher noise level
detection efficiency
Stars
False detection rate
true mag
Magnitude Response
Galaxies
Standard noise level
Same noise level
20% higher noise level
true - measured mag
Stars
true mag

Figure 24:

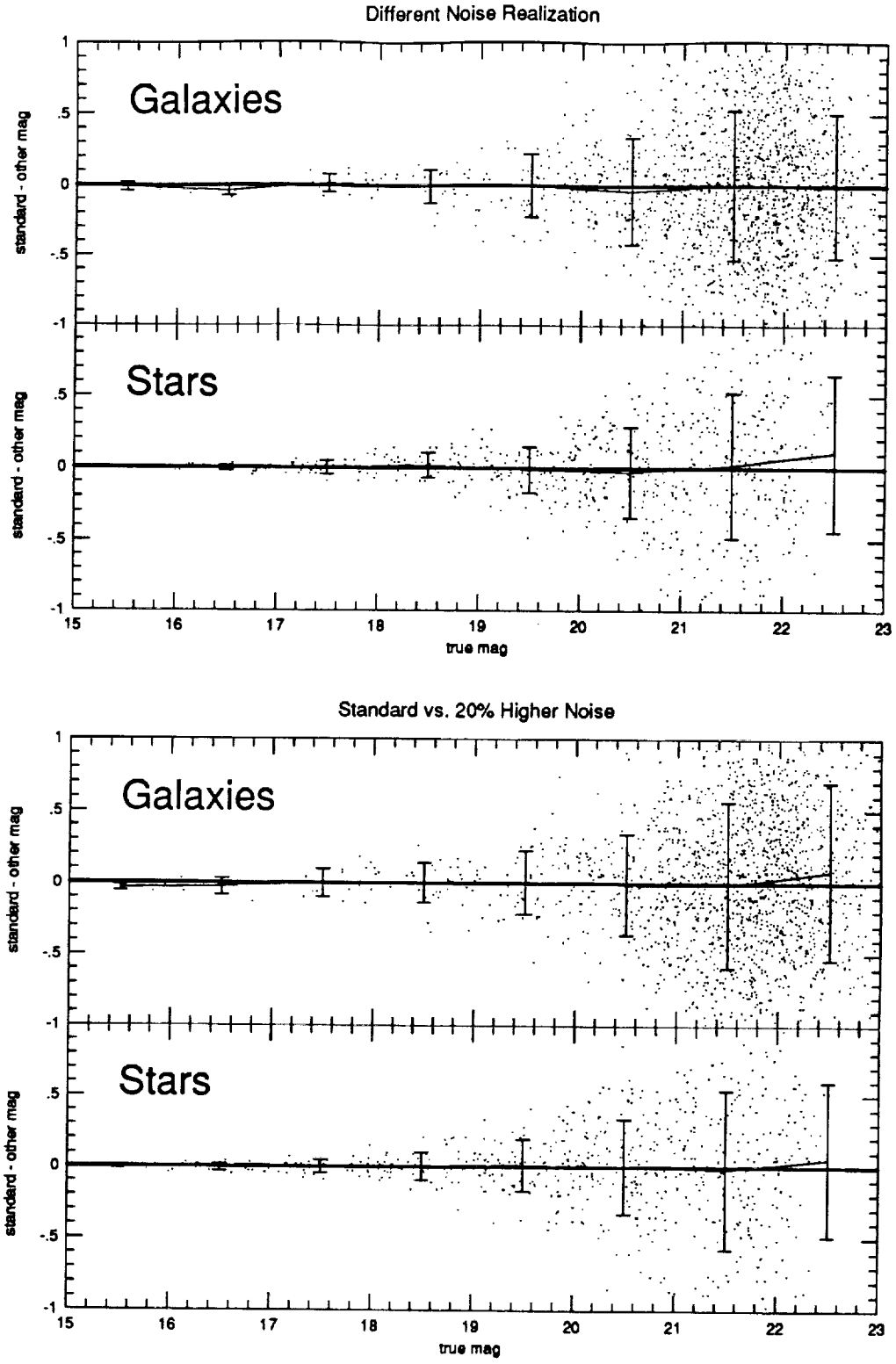
Different Noise Realization
Galaxies
Stars
standard - other mag
true mag
Standard vs. 20% Higher Noise
Galaxies
Stars
standard - other mag
true mag

Figure 25:

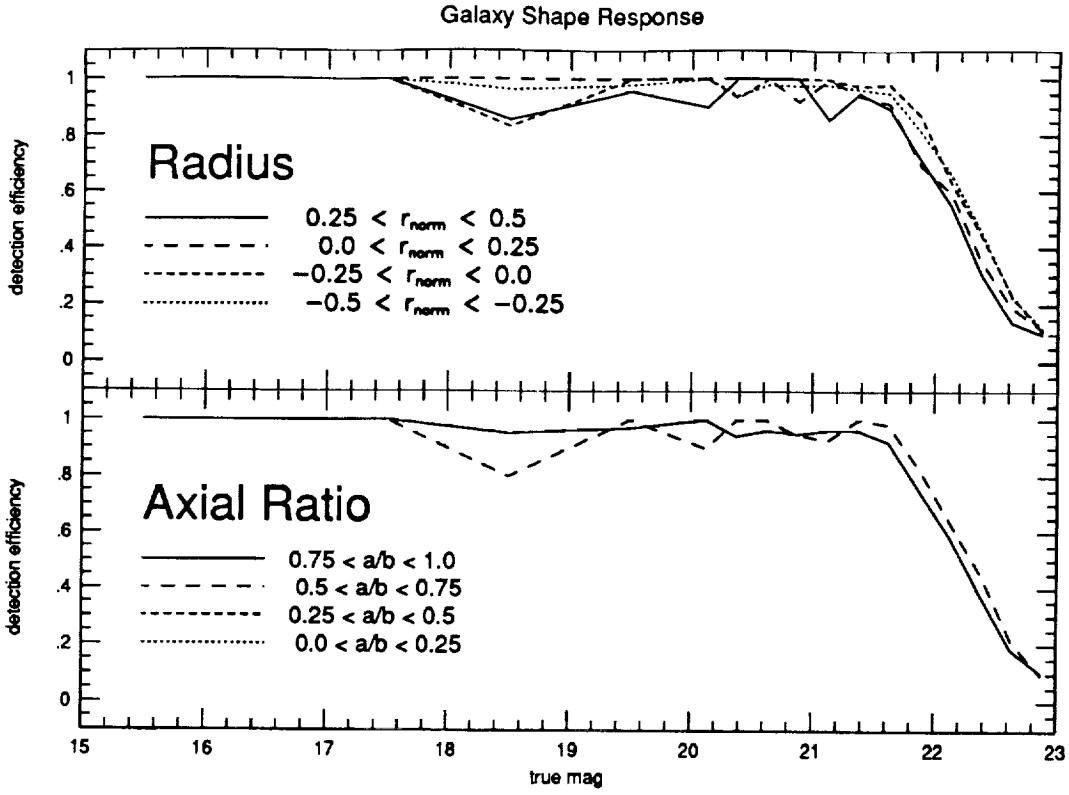
Galaxy Shape Response
Radius
0.25 < r_norm < 0.5
0.0 < r_norm < 0.25
−0.25 < r_norm < 0.0
−0.5 < r_norm < −0.25
Axial Ratio
0.75 < a/b < 1.0
0.5 < a/b < 0.75
0.25 < a/b < 0.5
0.0 < a/b < 0.25
detection efficiency
detection efficiency
true mag
0
.2
.4
.6
.8
1
15
16
17
18
19
20
21
22
23

Figure 26:

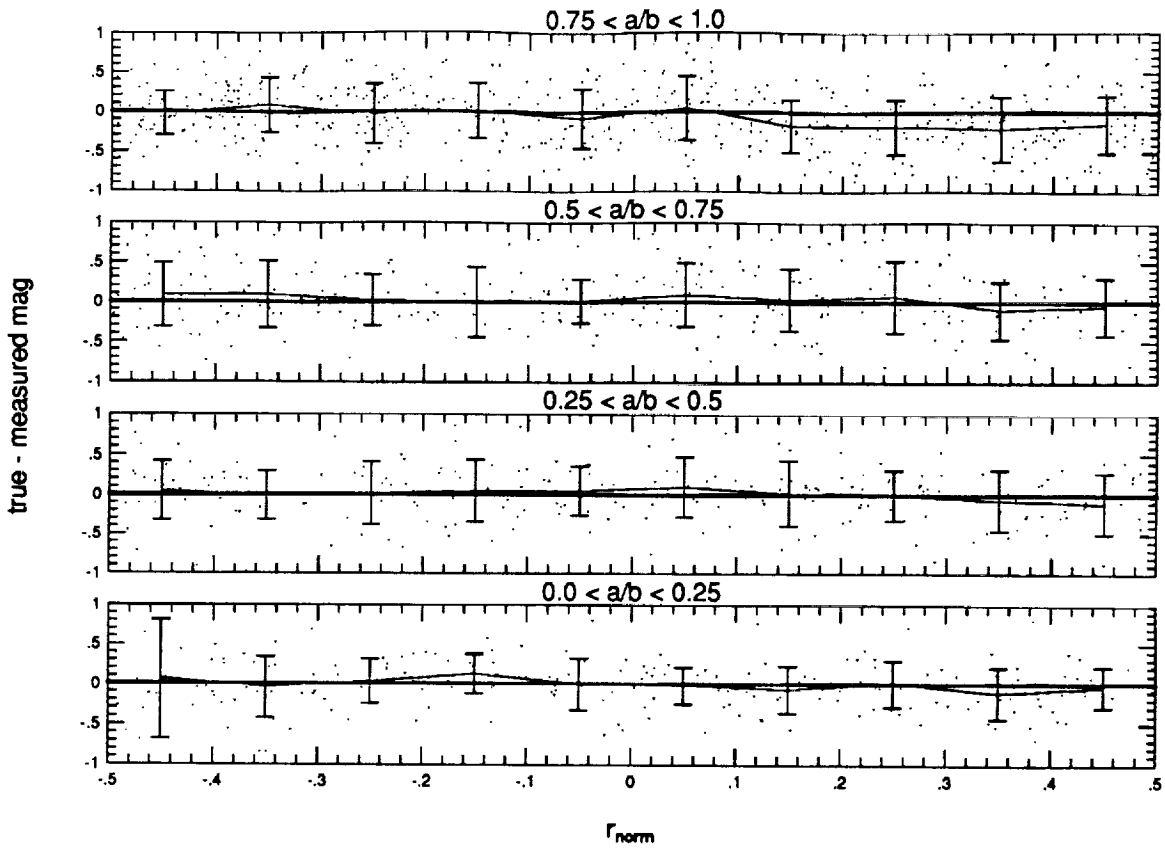
0.75 < a/b < 1.0
0.5 < a/b < 0.75
0.25 < a/b < 0.5
0.0 < a/b < 0.25
true - measured mag
r_norm
1
.5
0
-.5
-1
-.5
-.4
-.3
-.2
-.1
0
.1
.2
.3
.4
.5

Figure 27:

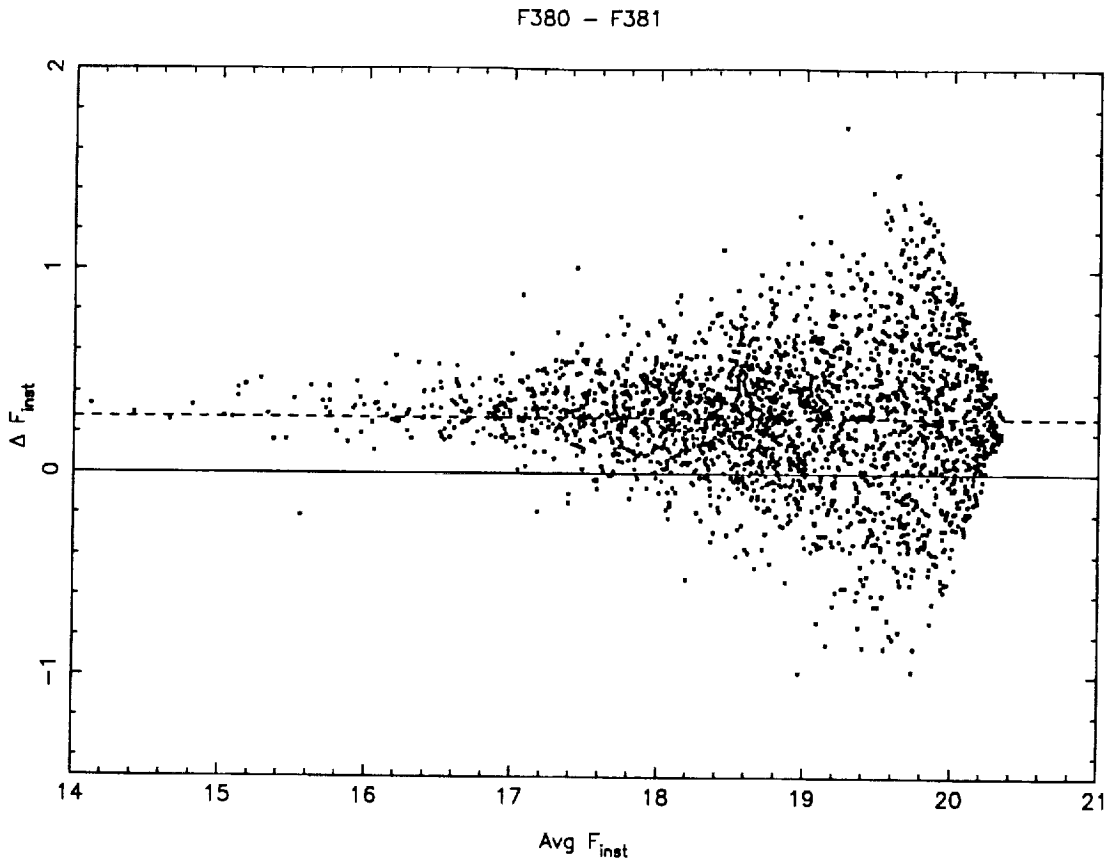

Figure 28:

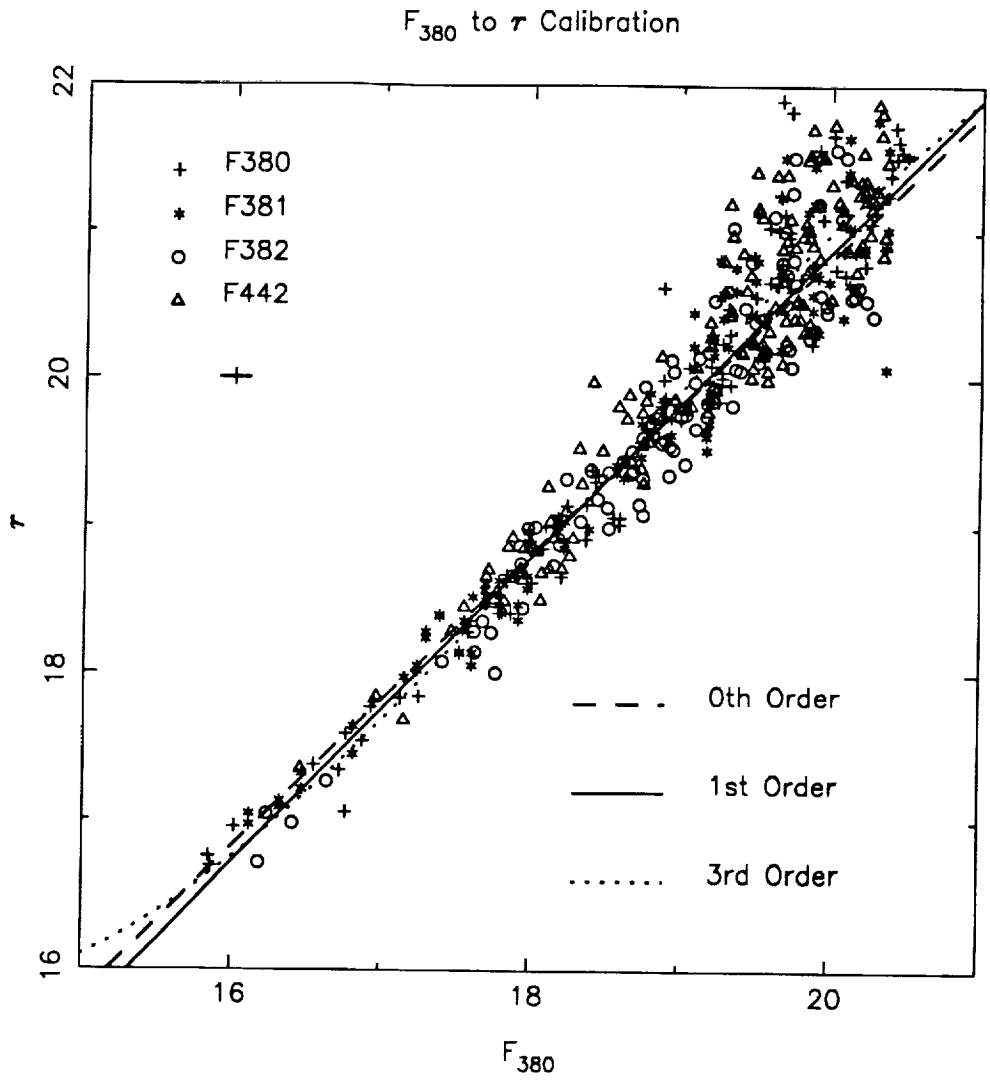
F_{380} to r Calibration
F380
F381
F382
F442
0th Order
1st Order
3rd Order
22
20
18
16
16
18
20
r
F_{380}

Figure 29:

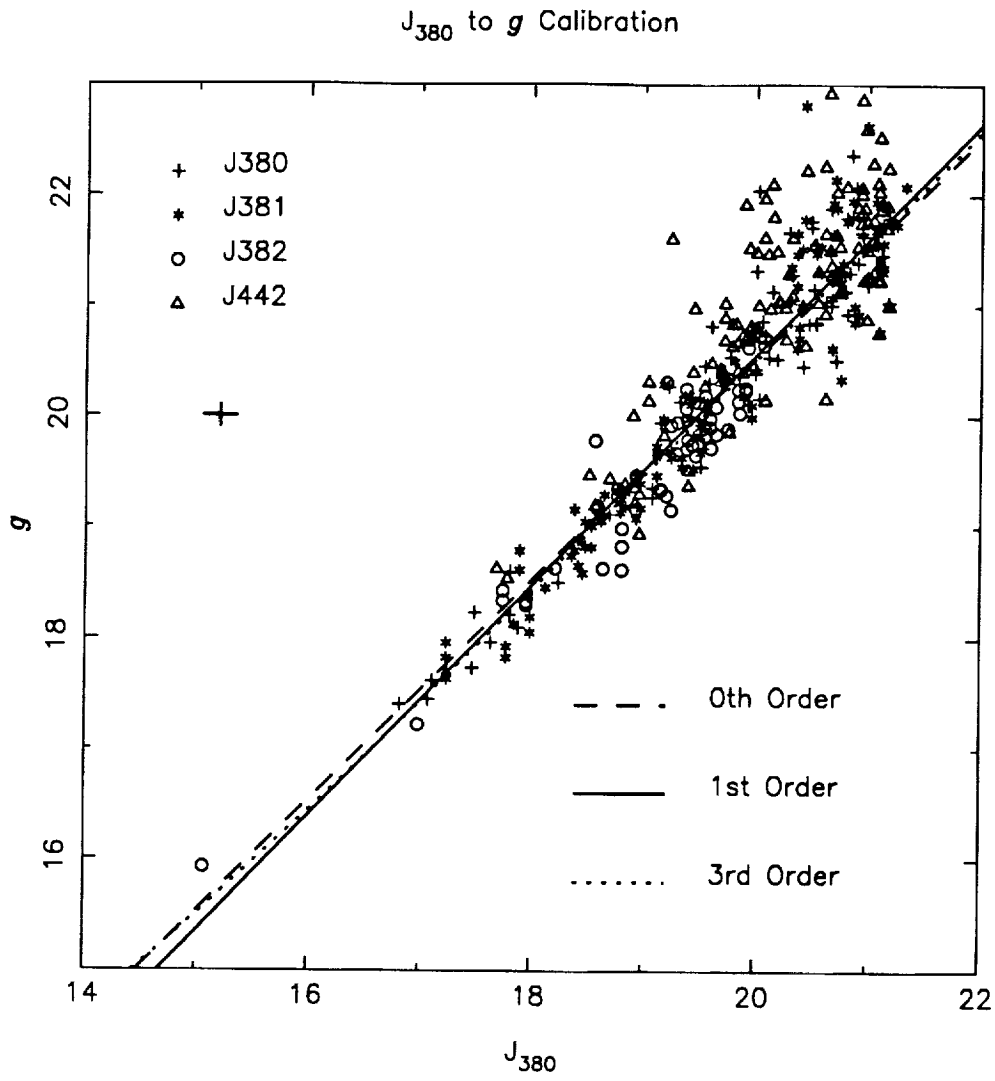
J_{380} to g Calibration
J380
J381
J382
J442
0th Order
1st Order
3rd Order
g
J_{380}
14
16
18
20
22

Figure 30:

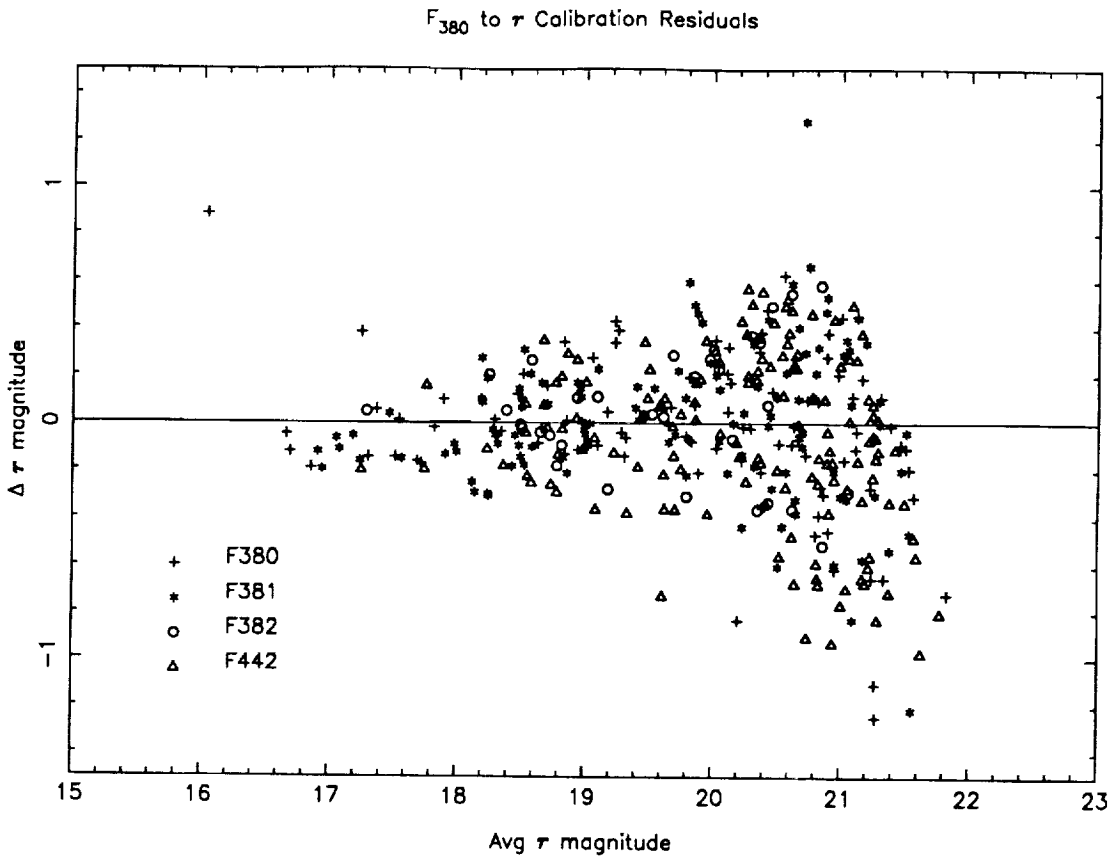
F_{380} to r Calibration Residuals
Δ r magnitude
Avg r magnitude
F380
F381
F382
F442
1
0
-1
15
16
17
18
19
20
21
22
23

Figure 31:

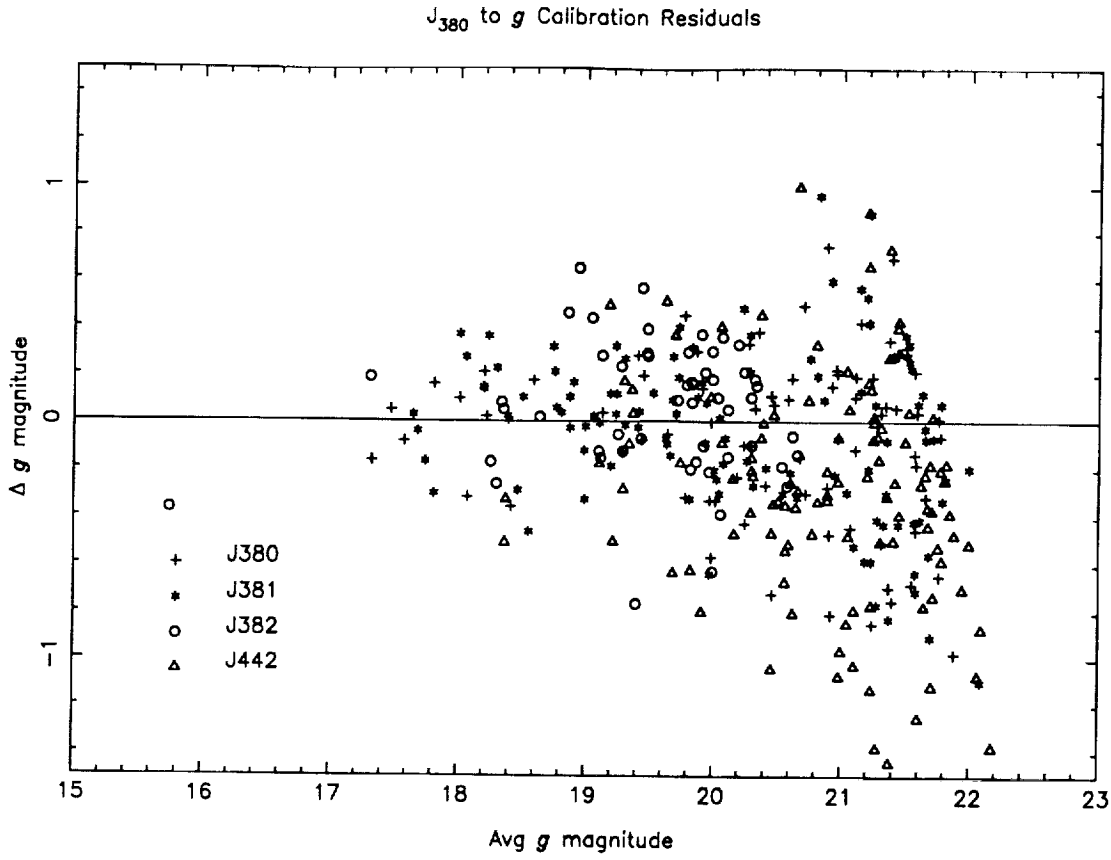
J_{380} to g Calibration Residuals
Δ g magnitude
Avg g magnitude
J380
J381
J382
J442
1
0
−1
15
16
17
18
19
20
21
22
23

Figure 32:

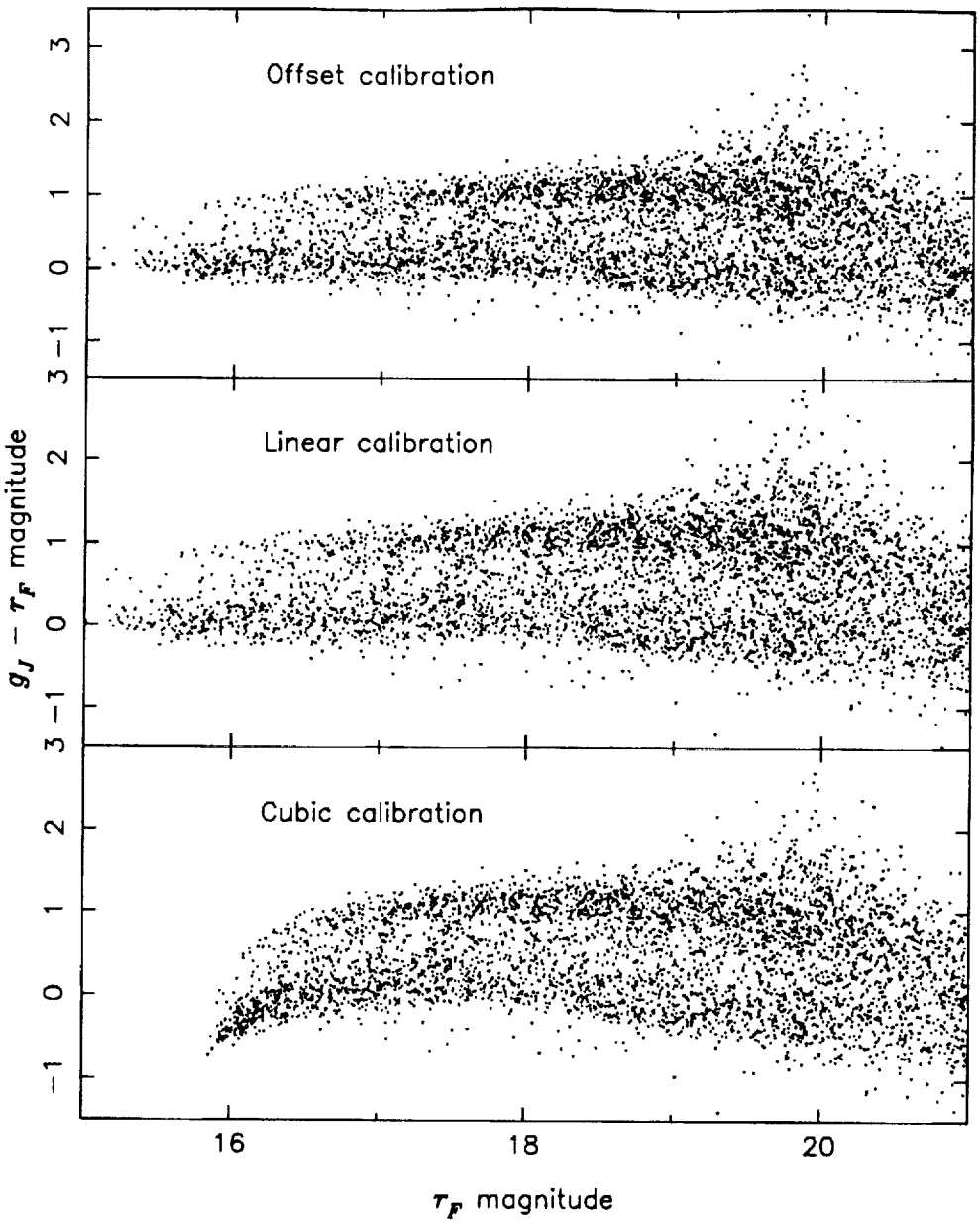
Offset calibration
Linear calibration
Cubic calibration
$g_J - r_F$ magnitude
r_F magnitude
16
18
20
3
2
1
0
−1

Figure 33:

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 12, 1994	Contractor Report

4. TITLE AND SUBTITLE

Multivariate Statistical Analysis Software Technologies for Astrophysical Research Involving Large Data Sets

5. FUNDING NUMBERS

930

6. AUTHOR(S)

S. G. Djorgovski

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

California Institute of Technology
1201 E. California Boulevard
Pasadena, CA 91125

8. PERFORMING ORGANIZATION REPORT NUMBER

NAS5-32337
(- 5555-32
NAS5-31348

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration - OSSA
Washington, D.C. 20546-0001

Universities Space Research Association
10227 Wincopin Circle, Suite 212
Columbia, MD 21044

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

CR-189393

11. SUPPLEMENTARY NOTES

Technical Monitor: J. Hollis, Code 930

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified–Unlimited
Subject Category 82
Report is available from the NASA Center for AeroSpace Information, 800 Elkridge Landing Road, Linthicum Heights, MD 21090; (301) 621-0390.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We developed a package to process and analyze the data from the digital version of the Second Palomar Sky Survey. This system, called SKICAT, incorporates the latest in machine learning and expert systems software technology, in order to classify the detected objects objectively and uniformly, and facilitate handling of the enormous data sets from digital sky surveys, and other sources. The system provides a powerful, integrated environment for the manipulation and scientific investigation of catalogs from virtually any source. It serves three principal functions: image catalog construction, catalog management, and catalog analysis. Through use of the GID3* Decision Tree artificial induction software, SKICAT automates the process of classifying objects within CCD and digitized plate images. To exploit these catalogs, the system also provides tools to merge them into a large, complex database which may be easily queried and modified when new data, or better methods of calibrating or classifying become available. The most innovative feature of SKICAT is the facility it provides to experiment with and apply the latest in machine learning technology to the tasks of catalog construction and analysis. SKICAT provides a unique environment for implementing these tools for any number of future scientific purposes. Initial scientific verification and performance tests have been made using galaxy counts and measurements of galaxy clustering from small subsets of the survey data, and a search for very high redshift quasars. All of the tests were successful, and producced new and interesting scientific results. Attachments to this report give detailed accounts of the technical aspects of the SKICAT system, and of some of the scientific results achieved to date. We also developed a user-friendly package for multivariate statistical analysis of small and moderate-size data sets, called STATPROG. The package was tested extensively on a number of real scientific applications, and has produced real, published results.

14. SUBJECT TERMS

Artificial intelligence, Sky surveys, Multivariate analysis, Databases, Object classification

15. NUMBER OF PAGES

205 (including attachs)

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited