Quality Assessment of KOA HIRES Extracted Spectra

R. Goodrich, J. Mader, H. Tran, B. Berriman, D. Ciardi, A. Laity, T. Barlow

V1.1, 2008 July 31

1 Introduction

Extraction of spectra from HIRES raw images is an important step towards producing science from the HIRES data. Proper extraction relies on an intimate knowledge of the goals of the science, and an excellent understanding of the instrument, how the data were obtained, and what products are desired from the extraction. With this in mind, KOA strongly recommends that archive users extract the spectra themselves from the raw data.

Nevertheless, the archive provides an extracted spectra "browse product," useful for getting a better idea of the type of data that might be extracted from the raw images. The archive does this by running an automated data reduction pipeline (DRP) that uses standard techniques and the extraction program MAKEE in order to produce spectra of individual orders from the 2-D raw images.

Such an automated pipeline cannot hope to properly extract 100% of the orders. This is for various reasons:

- Sometimes raw images, or subsets of them, were never intended to allow spectral extraction. This could be the case, for instance, when an observer is interested in the red parts of the spectrum, and uses a slit length that is optimal for those wavelengths, while allowing orders in the blue to overlap.
- It may be unclear to any but the original P.I. what objects were expected to be extracted. This can occur if there are two or more objects on the slit at the same time. Which object was intended for reduction?
- The reduction tool may make assumptions about the science object that are wrong in some cases. For instance, MAKEE if optimized for point sources. A spectrum of an extended object that fills the slit would require significantly different extraction techniques.
- The original observer may have failed to take appropriate calibration data. This sometimes occurs when the data were not meant for science (e.g. a test exposure to get a quick subjective feel for the target's spectrum), or the observer simply made a mistake.

When the DRP runs on the entire set of HIRES data, it will sometimes fail to produce any output. When it does produce output, that output may or may not be of high quality. In this latter case, it is of use for the archive to provide some indication of the level of quality of the extracted spectrum. With roughly 1,000,000 orders in the archive, it would be extremely expensive to manually inspect and grade each order. Hence, we have set up a process to provide automated grading of each order, as we now discuss.

2 Goals

As mentioned, we wish to provide some system of grading for each of the roughly 1,000,000 orders in the archive. We have chosen to represent this, at least conceptually, as assigning grades A–D and F to each order.

A grading system should be statistically accurate enough to provide some reasonable level of value to the user. In this case we would like to provide value to the end user, i.e. the archive user. We would like to be able to provide some indication of the quality of the spectral extraction. However, as the highest priority, we would like to indicate which extractions are particularly bad. We choose this as the highest priority in part because the natural assumption of an archive user is likely to be that the products are all good, and we want to manage expectations in situations where that assumption is not true.

Hence, our top goal is:

- Automatically identify, at 90% or better, extractions which contain no sign of the underlying, good data, but instead are dominated by extraction problems. We would grade these "F."

Our secondary goals is:

- Avoid misclassifying more than 2% of the excellent extractions, defined as those that show no sign of extraction problems, but appear to be highly accurate representations of the underlying data, as poor extractions. We would grade these "A."

While we could define many more goals, such as accurate classification of spectra that show some extraction problems, but clear signs of the underlying data, the goals we have set above are significantly complicated and open-ended that we will concentrate on meeting them.

3 Technique

Without some "ground truth" to reference, e.g. a pre-existing spectrum of the object that can be compared to the KOA extraction, determination of the quality of an extraction is necessarily somewhat subjective. Since HIRES represents state-of-the-art equipment on one of the largest telescopes in the world, we are unlikely to find existing data of comparable quality, hence we make the assumption that we must proceed without "ground truth" to guide us.

Nevertheless, an expert can look at a HIRES spectrum and provide a reasonable assessment as to what is real and what is an extraction artifact. Visual grading such as this is extremely resource intensive, and visual grading of a million spectra would take prohibitively long.

Hence we search for an algorithm that can be used to classify data, and that can be automated to rapidly provide classifications for all of the extracted orders. In order to develop this algorithm we have followed these steps:

- Select a number of observations of different types of objects, taken with different observing techniques, to provide a representative sample of conditions under which KOA is expected to produce extracted spectra.
- An expert visually grades those spectra. This provides a "training set" that is used to test potential classification algorithms. It also provides a trusted data set which is used to train other "graders," and provides examples of different levels of extraction fidelity for archive documentation purposes.

- A set of parameters is developed that may have some bearing on the extraction quality. For example, the number of errors generated in the log file during the extraction, or the number of pixels rejected during profile or sky fitting, etc.
- The values of these parameters for the training set are compiled, and correlations between the visual grade and the parameter values are sought.
- From such correlations, algorithms are developed and tested, with the goals stated in section 2 being the primary criteria.
- Once a suitable algorithm is found, it is automated, and run on all million orders in the archive.
- A second, random sample set of data is selected, to be used as a confirmation, and a better statistical measure of the accuracy of the classification algorithm. While the training set was somewhat deliberately chosen to represent all aspects of data taken with HIRES, it does not represent a statistical sample, since some types of data will be rarely taken, and other types (e.g. planet hunters) will be very common in the database. The random sample is meant to be a less biased, statistical sample.
- A second round of visual grading of the random sample is used to compare to the results of the classification algorithm, and statistics are compiled from that. These are the statistics that represent the true rate of correct and incorrect classifications in the archive.
- Assuming that the results from the classification algorithm are satisfactory (meeting the goals given in section 2), those grades are then used in the archive to provide guidance to archive users on the quality of the extraction.

We now give details about these various stages, and provide some results of the algorithm development.

3.1 The training set

3.1.1 Sample selection

The goal of defining the training set was to sample most of the types of data and styles and techniques of observing that are encountered in the HIRES data. It is not intended to be an unbiased statistical sample of the data that closely matches the frequency with which the various data and techniques are represented in the archive. It is rather meant to sample the entire range of data.

In order to accomplish this, we chose a wide range of observers. Since observers tend to take a series of data of similar type, and use similar techniques for their data collection, this seems like a good proxy. We chose one night from each of more than 30 observers, ending up identifying 32 nights in this manner.

Since the number of observations in these 32 nights was still prohibitively large, we further reduced the number of CCDs by choosing the first two observations of each night. This has the further advantage of often (not always) sampling a science target and a

brighter, calibration standard, again representing a range of what might be observed during each night.

We used all three CCDs for each of the selected observations, since the different CCDs represent different types of data, such as different interorder separation, and different flux levels.

3.1.2 Visual grading

Our final training set consisted of 2839 orders. These were visually graded by one of us (Goodrich), using a Web interface constructed and streamlined specifically for this purpose.

Grades of 1, 3, or 5 were assigned to each order. These were defined as follows:

- Grade 1: extractions that show no evidence of any extraction problems, but appear to be an excellent representation of the true underlying data. In a lettered scheme these would be grades "A."
- Grade 3: extractions that show some combination of the underlying data and extraction problems. These would be graded B, C, or D in a lettered scheme.
- Grade 5: extractions that show no evidence of the underlying data, but are dominated by extraction problems. These would be graded "F."

Note that grade 3 encompasses a range of different extraction qualities. This was done purposely for two reasons: to limit the complexity of the resource-intensive visual grading stage, and to avoid the problem of trying to define the differences between grades B–D. A consequence of this decision is that while grades 1 and 5 are very clearly defined, orders of grade 3 can be either nearly "perfect" extractions, close to grade 1, or barely passable extractions, close to grade 5. Given the stated goals (section 2) and the complexity of developing an appropriate algorithm even for those limited goals, this seems like a prudent decision.

We did, however, leave room for dissecting grade 3s into finer grades by not using the digits 2 and 4. Should it be deemed necessary to provide finer grading, we can use those two digits to represent the "B" and "D" grades, respectively.

In a small fraction of the cases it was impossible to grade an order using the graphics provided on the Web page. (Those graphics are identical to what will be provided on the archive user interface.) Usually this was because a single ion hit caused the autoscaling feature of the plotting to plot on such a scale that the rest of the order was unresolved on the plot. Hence no grade was assigned, and instead we recorded a question mark ("?") in the database.

3.1.3 Results

Of the 2839 orders visually graded, the numbers of grades assigned were:

Table 1. Results of visual grading of the training set

Grade	Number	Percentage[1]
1	2498	88
3	191	7
5	143	5
?	7	n/a

1. The percentage does not include the ungraded ones marked "?".

The majority of the extractions, nearly 90%, are excellent. Relatively few are grade 5 (or "F"). Recall that roughly ¼ of the CCDs do not make it successfully through the MAKEE reduction process, and hence do not provide any output products (other than a log file). However, when output products are produced, MAKEE provides good quality.

3.2 Parameters for automation

Using the training set, we scanned through log files looking for clues that an automated classification algorithm might use to correctly grade an extraction. Once the list was gathered (Table 2), scripts were written to extract the parameters from the log files and into a database, where they could be further analyzed. Note that at this stage only the parameters from the training set are required.

Table 2. Parameters for grading algorithm

Parameter name	Description	Source
errors	Number of errors encountered in fitting the object profile.	Log
priSky	The number of pixels rejected from the first pass of the sky fitting.	Log
adjSky	The number of pixels effected in adjacent columns	Log
addSky	Additional pixels rejected in a second pass	Log
highObj	One side of object window	Log
lowObj	Other side of object window	Log
pse0		Log
cosmic	Number of "cosmic rays" rejected.	Log
traceWidth	Total width of the trace.	Log

3.3 Correlations

In general the technique used to search for correlations was to sort by some parameter, usually in descending order so that large values of the parameter are at the top, and then look for evidence that this parameter provides a good discriminant between grades 1 and

5. Alternatively, sorting by grade, and seeing whether grade 5s showed any anomalous values of the parameters could be used. In this way the most promising parameters were identified efficiently.

Further analysis was done by plotting either the distribution function or the cumulative distribution function of the parameter. This was done separately for each grade. Clear differences between the functions for different grades indicates a good discriminant, and the value at which that parameter separates the grades.

An example of a distribution function is shown in Figure 1.. The number of errors in the log file for the order is the abscissa, and the frequency of the different grades is the ordinate. These errors are generated during a fit to the object profiles, and represent an inability of the reduction tool to properly identify the pixels containing flux, and potentially confuse the object and sky pixels. Note in this figure that grades 1 and 3 are not well separated, but for N(errors) $\geq$ 5, this parameter provides a reasonable discriminant between grades 1+3 and grade 5.

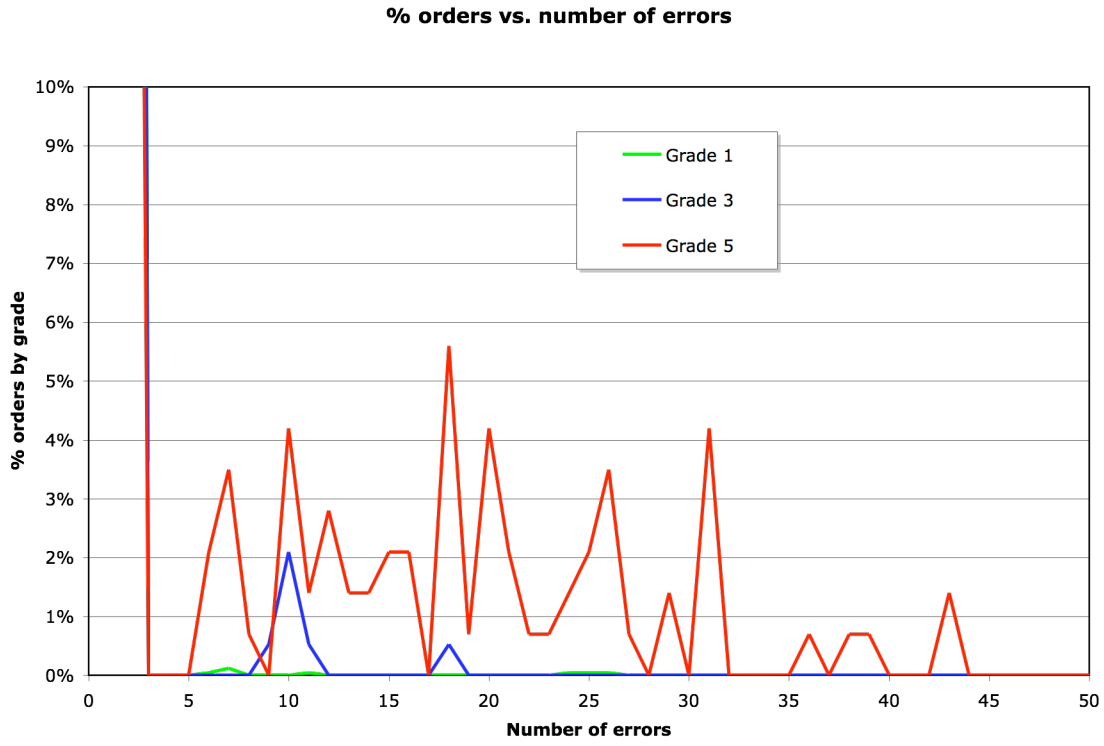

Figure 1. Distribution function for the number of errors in the log file for the three grades.

When using cumulative error distribution functions, such as shown in Figure 2, we looked for regions where the y-axis separation of the different grades (in particular grades 1+3 and grades 5) is large.

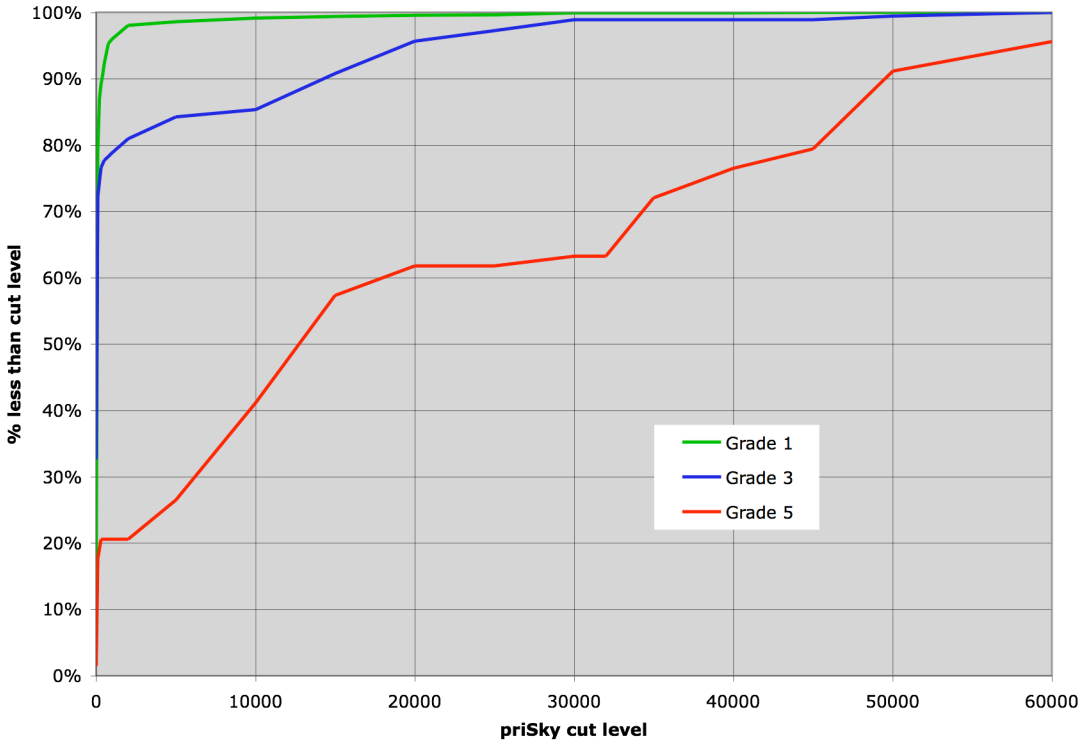

Figure 2. Cumulative distribution function for the parameter priSky.

In Fig. 2 if you look at the separation between the green line and the red line at an x-value of 10,000, you can see that roughly 60% (100% minus the y-axis value) of the grade 5s have priSky greater than this value, but only <1% of grade 1s have a value at least this high. If you look at priSky ≥ 20,000, only 40% of the grade 5s are found, and also fewer grade 1s. The equivalent of not using this parameter is priSky = ∞, i.e. at the rightmost end of the entire cumulative distribution. (The plot as shown does not extend that far.)

3.4 Classification algorithm

As a result of the search for discriminants between grades 1 and 5, we produced the following algorithm:

1. Orders with N(errors) ≥ 5 were classified as grade 5.
   This correctly identified 52% of the orders with a visual grade of 5, and incorrectly identified only 0.16% of the visual grade 1s as grade 5. This corresponds to only 4 of the 2498 grade 1 orders being classified as grade 5. (This part of the algorithm also classified 7 grade 3s, or 3.7%, as grade 5s.)
2. Of the remainder [those with N(errors) < 5], those orders with priSky > 4000 were also classified as grade 5.

The combination of these two criteria classified 90% of the visual grade 5s correctly, and misclassified only 1.6% of the grade 1s. The algorithm thus meets our goals (section 2) *for the training set*.

Note that looking at the parameters for some of the orders visually graded 5 show nothing that would indicate that the extraction was poor. Hence 100% accuracy of an automated classification algorithm should not be expected.

3.5 The random sample

In order to independently verify the performance of the chosen classification algorithm, we generated a random sample of 64 observations to grade via the algorithm and by separate visual grading. Note that the random sample may include previously graded observations, since they both draw from the same parent population.

The random sample contained 2912 orders, similar in size to the training set. Results are shown in Table 3.

Table 3. Visual and automated grading of the random sample

Grade	Visual grading	Automatic grading	% misgraded
1	2515	2793	1.7
3	144	n/a	n/a
5	154	119	64.3
?	7	n/a	n/a

The "% misgraded" column for grade 1s in Table 3 is calculated as the percentage of visual grade 1s that were graded 5 by the automated algorithm. Similarly, the "% misgraded" for grade 5 is calculated as the percentage of visual grade 5s that were automatically graded as 1s. Our goals (section 2) were that the first number be less than 2%, and the last number be less than 10%. While the first goal is met, the automatic QA system misgraded nearly 2/3 of the grade 5s.

In visually grading the random sample, it was noticed that the misgraded orders tend to be the lowest numbered (bluest) orders on the CCD. Profile plots provided by MAKEE showed a tendency for only half of the profile to be shown in these orders. So there is some hope that if the visual cue that there is a potential extraction problem can be converted into an automatic step, an updated QA algorithm may be able to perform better.

4 Conclusions

Visual QA of the KOA extracted spectra is resource intensive, so an automated solution was sought. An initial sample was chosen to represent a wide range of HIRES observation types, and development of an algorithm from this sample met our goals of <2% misclassified good extractions, and <10% misclassified poor extractions.

However, when applied to a random sample of HIRES spectra, the QA algorithm did not meet our goal of <10% misclassified poor extractions, erroneously grading 64% of those orders as "pass." The random sample better represents the archive contents in terms of the *numbers* of each type of observation.

Visual cues are available for many poorly extracted orders, so more work on the QA algorithm could well provide more reliable grading.

5 Appendix: methodology behind classification algorithm

There is a wide variety of techniques that can be used to develop an automated QA system. We discuss some of the theoretical aspects below.

5.1 Parameters as a multidimensional space.

We identified a number of parameters that could have some bearing on the quality of the order extraction, including lines printed to the log file and fit parameters saved in the FITS files and headers. We essentially treat this parameter set as a multidimensional parameter space, and ask what part of this space is occupied by the good extractions and the poor extractions. Assuming that the two groups are well separated in this multidimensional space, in principle the groups can be identified algorithmically, and an automated QA system employed.

5.2 Cut on a single parameter (plane perpendicular to that axis)

The simplest way of dividing the parameter space is to make a cut using a single parameter (e.g. "$X > 5$" and "$X \leq 5$"). This represents a plane perpendicular to that axis (the X axis in the example).

Multiple parameters with independent cuts can also be made, representing planes perpendicular to different parameter axes. This is the approach taken in the analysis above.

5.3 Linear combinations of parameters (plane at an arbitrary angle in N-space)

Cuts that involve multiple dependent parameters are a further refinement (e.g. "$3*X+4*Y > 17$") represent planes at arbitrary angles in the multidimensional parameter space. Again, multiple planes can be used to refine the volume separation.

5.4 Higher order functions. (curved surfaces in parameter space)

The next logical refinement is to remove the restriction for linear combinations. So higher order, curved surfaces could be defined in the parameter space (e.g. "$2*X + 4*Y^2 > 2.3$").

5.5 Principal Component Analysis

Another potential technique is to apply principal component analysis (PCA). Essentially treating each order as a series of parameter values, a PCA might identify the characteristics of a good *vs*. a poor extraction.