The AAVSO Data Validation Project

Kerriann H. Malatesta
Sara J. Beck
Gamze Menali
Elizabeth O. Waagen
AAVSO Headquarters, 25 Birch Street, Cambridge, MA 02138

Received October 27, 2005; accepted October 27, 2005

Abstract In 2002, NASA awarded the AAVSO a two-year grant to error check over 9.5 million observations in the AAVSO International Database from the founding of the organization in 1911 through 2001 and to make these observations available to researchers and the public. The purpose of the project was a systematic search of the data to look for potential and obvious errors, and to investigate and rectify any problems. In this paper, a project description is given and future data validation plans are reported.

1. Introduction

The AAVSO International Database is now the home for over 12.5 million observations of over 5,000 known and suspected variable stars contributed by over 6,000 observers worldwide since 1911. As the largest and highest quality digital database of variable star observations available, thousands of researchers, educators, and students have used the wealth of information contained in the database for both professional and personal projects. It is an impressive testament to the work and dedication of its observers that the AAVSO Headquarters fulfills thousands of requests for data annually. While the AAVSO maintains a strict quality-control policy to ensure error-free data for such requests, there never had been an organized and systematic review of all the data in the AAVSO International Database.

In 2002, NASA awarded the AAVSO a grant to "validate" observations for 4,922 stars in the AAVSO International Database from 1911 through 2001—a project estimated to take two years to complete. The goal of the project was not to produce "pretty" light curves free of scatter, but to flag observations so far removed from other observers' measurements that they would negatively affect the analysis of data to a statistically significant level.

A very important part of the data validation project was to find, investigate, and resolve discrepant observations. Out of the over 9.5 million observations validated, 633,126 (6.04%) were considered to be discrepant, and 441,879 (70% of these) were repaired by correcting data fields to match the observer's original report (see Table 1). Fortunately, many of these data problems in the AAVSO International Database no longer occur since computer hardware and software have become much more sophisticated than they were in the early days of computerization.

Flagging the remaining unresolved discordant points was left to the discretion of the experienced AAVSO Headquarters staff members performing the validation, subject to strict rules. In no case were any of the data deleted, removed from the archives, or adjusted to fit the light curve without justification from the original observer report. While discrepant points are not routinely disseminated, they remain in the permanent archives with a letter code designating them as such and are available upon request.

As a result of the Data Validation Project, less than 2% of the data covered by the project remain as unresolved discordant observations.

Table 1. Corrections made to the AAVSO International Database during the Validation Project.

Category		*Number of Observations*	*Percent of Validated*
Total Validated Observations		10,475,060	100.00%
Total Discrepant Observations (pre-validation)		633,126	6.04%
Total Data Repairs			
Designation	47,252		
Star Name	153,831		
Julian Date	15,017		
Magnitude	11,862		
Comment Field	39,843		
Observer Initials	18,599		
Other—unspecified*	155,475		
		441,879	4.22%
Total Remaining Discrepant**		191,247	1.83%

** Prior to 1994, the kind of repair made to an observation was not specified, just that a repair was made.*

*** No typographical error was found*

2. Pre-validation

The initial months of the project were used to: assemble a knowledgeable data validation team, with each member referred to as a "validator"; establish a list of stars to validate; develop a standardized set of rules and procedures; create programs to assist with the validation process; and upgrade hardware, such as installing non-glare monitors for each team member and securing a new server to house the final database.

The data validation team consisted of the following AAVSO Headquarters staff: Janet A. Mattei, Director and Project Principal Investigator; Elizabeth O. Waagen, Interim Director and Interim Project Principal Investigator and Validator; Rebecca

Pellock, Project Team Leader and Validator; Sara Beck, Katherine Davis, Kerriann Malatesta, Gamze Menali, and Sarah Sechelski, Validators. Each team member was selected for their knowledge of how the data were processed, familiarity with thousands of AAVSO observers, experience with accessing original observer reports, ability to research current and archival variable star charts, and for their expertise in variable star behavior. In addition, Aaron Price and Michael Saladyga provided hardware, programming, and processing help.

The stars chosen for validation were of many types, with the heaviest concentrations coming from the pulsating (66%) and the eruptive (23%) variables. The stars included in the project were then divided by class and distributed for validation amongst the team members. Eclipsing binary, RR Lyrae star, and comparison star data, as well as selected stars with complex histories or light curves, were not included in the project.

3. AAVSO International Database format

The basic format of the AAVSO International Database is one record per observation of a star. Each record includes identification of the star and of the observer, as well as the date and time of observation, magnitude, and any pertinent comments. The database is organized by designation, with observations of a star in increasing order of time. Table 2 describes the basic format of an AAVSO data record.

Table 2. Basic format of an AAVSO International Database record.

Desig.	*Name*	*JD*	*Mag.*	*Comments*	*Observer*
1544+28A	R CRB	2453012.6472	6.0	M	WEO

Desig. — *approximate 1900 position; letter may be added to distinguish from other stars with same numerical designation*
Name — *name of variable star as appears in AAVSO International Database*
JD — *date and time of observation in Julian Date and GMAT*
Mag. — *visual magnitude, unless band (*V, B, R, I, *etc.) given in column 54 of comments*
Comments — *one- or multi-letter codes for comments about the observation (observing conditions, uncertainty, CCD/PEP/photographic, band, etc.)*
Observer — *AAVSO Observer Initials, unique to each observer*

Note: for a detailed description of the AAVSO data format, see the AAVSO website http://www.aavso.org

4. Resolving designation, name, comment field code, and observer initial problems

Since the AAVSO International Database is currently organized by designation, data with erroneous designations wind up grouped with the wrong star. The first step in validating data involved checking the database for digitization errors in

non-corresponding variable star names and designations, and for problems in the comment and observer initials fields. Each team member was assigned a list of variable stars, and given a file of such discrepancies found by a computer program that identified potential problem points. The team members then referred to original observer reports to investigate and resolve all discrepancies.

4.1. Designation

Historically, most designation errors in the AAVSO International Database were caused by mismatches between star names and their designations as written by the observer. In some cases, the star name was written along with the designation for a different star or without a designation at all and it was up to Headquarters staff to determine the observer's intent. Designation errors were also introduced into the database in the early days of keypunching through the data entry software itself. The keypuncher could duplicate a star's name and designation with the touch of one key, while at the same time, for a while, the verification program had a flaw that could prevent the error from being caught. One other source of designation errors was a special program that was used to digitize all of the reports from a few prolific observers. This program, written without Headquarters' knowledge by a data entry technician who worked off-site, assumed that the star name given was correct and automatically changed the designation to match. Thus, if the star name was incorrectly written or misread, the designation assigned to that observation would be wrong.

In recent years, most data entry programs (e.g. WEBOBS or PCOBS) automatically link star names with designations. This has eliminated discrepancies, but it also means that there is no way to double check that the observer has specified the star they really intended to report.

The primary method for detecting designation errors was by viewing the light curve of a star and flagging discrepant observations. Data records with designation problems were often wildly off the mean so they were very noticeable. Each flagged observation was then looked up in the original paper reports and the discrepancy resolved by checking to see whether the observation fit better with the star named or the designation given.

In the special case where designation errors were caused by the program that automatically changed mismatched designations to agree with the star name, a more systematic approach to finding and correcting the problem could be employed. By plotting the raw data (in the order entered by the keypuncher) for a single observer as record number versus designation, spikes in the curves appeared wherever a designation was out of order. The records at these spikes were then checked in the original reports to see if a mistake had been made. See Figure1 for an example of such a plot.

Once a designation error was detected and resolved, the record had to be corrected in the archives. This was done using a program Headquarters has designed for such purposes and that ensures that the bad record is deleted and the corrected record is merged into the archives in the correct sorting order.

4.2. Star name

Star name errors were primarily caused by typographical errors on the part of the observer and/or keypuncher. Since a name error does not affect the position of a data point on the light curve of a star, these errors had to be detected using a program that searched for name/designation discrepancies.

In some cases, the error was obvious and could be corrected using a program that automatically corrects the star name to match the designation. This happens on a fairly routine basis even today, as monthly data are processed for merging into the archives. Other star name problems were not so easily resolved because it might not be obvious whether the star name or the designation was correct. For these observations, the original report had to be scrutinized to determine the observer's intent. Sometimes it helped to view the light curves of both stars to see where the observation fit best.

4.3. Comment field

Most changes to the comment fields (both the single-letter comment codes and the multi-character comment codes) were made in response to standardization of the codes and a change in the placement of the multi-character codes. In the case of step-magnitude information, some of it was corrupted by punch card reading errors. For the most part, comment problems were found by running programs on the database that looked for non-standard entries. Once the original report was consulted, the comment field was corrected accordingly.

4.4. Observer initials

Observer initial problems were mostly caused by card read errors which occurred during the punch card era at the AAVSO. These card read errors usually resulted in a number or symbol replacing one or more letters of an observer's initials. For example, "HE" sometimes wound up as "8E" in the archives. Observer initial problems were also introduced occasionally when a data entry technician misread or mistyped the observer initials written on a report or when the observer used something other than their official AAVSO initials.

Most observer initial problems were easily found by comparing the data archives with the master list of every assigned set of observer initials. The AAVSO never re-issues observer initials, so each observer's initials are unique to him or her.

Once records with unknown observer initials were identified, the problem could usually be resolved fairly easily by combing through the monthly folder covering the JD in question, looking for a report containing the actual observations. The reports within a folder are filed alphabetically, so unless the first letter of the initials was wrong, the report could usually be found right away. For pre-1961 data, for which the original reports are filed by observer rather than date, or in cases where the first initial was the problem, other techniques had to be employed. One of these involved examining the data of a star before and after the problem report to see if any of the people commonly observing that star had initials similar to the incorrect

ones. Clues as to the location and thus identity of an observer could also be gleaned from the decimal part of the JD.

No matter which method was used to resolve the problem, it was always carefully confirmed that all of the observations in the archives matched those in the original report. Usually, when one observation in a report had the wrong initials, the entire report did, too, and the entire report had to be corrected. Once all the problems were investigated against the original report, the validator then corrected the data archives to reflect their findings.

5. Rules for visual scrutiny

In the next step of validation, the data in the archives were subjected to an intense visual quality-control inspection through the context of each star's light curve. Here, the team members utilized their knowledge of variable star behavior to look for any suspicious data points. Since the team consisted of several members, rules were established to avoid subjective validation styles. The overall rules were divided into two categories: 1) lookup rules, whereby the validator was instructed to check the digitized observation against the observer's original report, and 2) editing rules, whereby the team member followed specific instructions to ensure editing homogeneity among validators.

5.1. Lookup rules

Comparing pages of discordant observations with the original reports to check for keypunching errors is a laborious task, but it was necessary for thorough and objective validation of the data. In order to spend time efficiently, guidelines had to be established as to which discrepant observations were to be compared to the original report. The team members agreed to refer to the original paper reports to check for JD and magnitude problems if:

- there was a systematic shift in time from the mean curve for an obvious string of data points
- any points fell within the seasonal observing gap at a date not reached by other observers
- there was an obvious misreport of a magnitude
- there was any report of bright data points prior to the discovery of a nova or supernova

5.2. Editing rules

The purpose of viewing the light curve was to check for potential problems, such as JD and magnitude errors, and to flag any remaining discrepant points. No observations were deleted or altered to match the light curve at this or any other point. All observations, both good and discrepant, remain in the permanent AAVSO

International Database. The difference between the good and discrepant points is an assigned validation letter code that is accessed by the AAVSO when extracting the data from the database for distribution.

When viewing the light curve, the team was instructed to disregard any flags placed by a previous staff member. The general overall philosophy was that if the validator had some doubt about deciding if an observation was good or discrepant, the point would be considered "good." Any unresolved discordant observation was flagged as discrepant only if it was:

- a fainter-than observation that fell below the mean curve
- a fainter-than observation that was substantially brighter than the star's known maximum magnitude
- a positive observation that fell outside a 2-magnitude spread about the mean
- an unfiltered CCD observation of a red star

The only exceptions to these rules were made for those portions of the light curves published in *AAVSO Monographs*, in that such data were left untouched by the validator.

6. Visual scrutiny

Following these rules, the data underwent two rounds of visual scrutiny. The purpose of the first round was to check for any discrepant observations; investigate and correct all JD and/or magnitude problems as compared with the original observation report; and to flag any truly discordant observations. The second round was to double-check for oversights.

6.1. Julian Date

There were two major causes of JD errors. During the punch-card era (1967–1981), it was common practice to use a program card in the keypunch machine that would automatically enter the first four or five digits of the JD at the press of a button. This saved a lot of time for the data entry staff, but occasionally introduced errors of 100 or 1,000 days when the hundreds' or thousands' place changed and the card wasn't replaced. The other major cause of JD errors stemmed from observers using a JD calendar from the wrong year or the wrong month. More than once it was discovered that the calendar card had been flipped over by mistake, thereby causing a six-month JD error. Generally, JD errors are first suspected in the database when the light curve of a star (particularly a Mira) is plotted and a bright observation is seen at minimum or a faint observation is seen at maximum. The discrepant observation should still fall within the normal magnitude range of the star.

As with other kinds of data errors, it was imperative that the observer's report be located and inspected. This could be a bit tricky with a JD error, since the post-1967 original reports are filed by date and those reports with true JD errors will not generally be found where one would think they should be. At this point it became

a bit of a guessing game. By examining the light curve of a star, it was possible to make an educated guess of what the JD of the discrepant point should be, bearing in mind that the 100- or 1,000-day errors were most common. The report files were then checked for the date of the suspected JDs until the report was found. Sometimes, in order to narrow the date search a bit, it helped to find out what other stars the same observer followed, and check those light curves for discrepant points with the same erroneous JD. With more than one star to work with it was easier to guess correctly what the JD should be. Figure 2 gives an example of a light curve with a JD error before correction (2a) and after (2b).

6.2. Magnitude

Magnitude problems in the AAVSO International Database were most commonly caused by typographical errors. During the era of handwritten reports, it was sometimes difficult to decipher the handwriting of the observer and many magnitude errors were introduced as a result. Also not uncommon were the keypunch mistakes that could occur when the data entry technician inadvertently typed in the magnitude of an adjacent star in the report, easily caused by a little slippage of the straight edge used as a place keeper.

Magnitude errors were usually detected by viewing the light curve of a star, flagging discrepant points and looking them up in the original reports. If a typographical error existed, it was obvious from what was written on the report. When it appeared as though the magnitude error was caused by a "ruler slip," it was a good idea to check the surrounding observations in a report to ensure that the problem wasn't more widespread.

6.3. The final check

Once JD and/or magnitude errors had been clarified and truly discrepant points flagged as such, the data then underwent a second round of viewing by the same validator to look for any obvious oversights. This second phase was performed at least one day later than the first to avoid eye fatigue. In some cases, such as for semiregular stars with complex light curves, the second viewing of the data was done by another member of the validation team to ensure adherence to the validation regulations. Any remaining questionable or problematic stars underwent a third round of visual scrutiny, performed by a more senior technical staff member, including the Director. These stars often had conflicting observations, complicated light curves, or complex histories in which the problems spanned either part of or the whole light curve.

7. The stamp of approval

Once all the stages of validation were complete for a given star, each point in the dataset was marked with a letter code indicating its status as validated. At this point, the data were released for download from the AAVSO web site.

8. Accessing the validated data

Since the initial data release via the AAVSO web site in 2003, over 4,000 downloads of validated data have been made. The validated data for each star may be viewed graphically online through the AAVSO Light Curve Generator, or may be downloaded as a data file.

In July 2004 the AAVSO launched a data usage report program which emails a monthly report to each observer whose data have been downloaded via the online data download tool, saying which observations have been downloaded and how many times. However, this report only covers data downloaded immediately via the web site; it does not include other types of data requests. For example, if the requester asks for data that are not yet validated, they are not included in the report. Other forms of data requests such as e-mail, telephone calls, and postal mail are also not included. Online data requests constitute around 50% of the official data requests received by AAVSO Headquarters.

Official data requests are usually requests for files of individual observations in standard format. However, unofficial requests, in the form of light curves or learning the current visual status of a star, are received through the light curve generator and the Quick Look file. We receive more than 700 requests per day through those two sources combined.

9. Conclusion and plans for the future

The Data Validation Project has undoubtedly been a major achievement in the history of the AAVSO. Completed on-budget and within the two-year deadline and taking 9,324 staff hours to accomplish, the high-quality data provided by AAVSO observers are now raised to an even higher standard, as detectable problems with name, designation, JD, magnitude, etc., have been investigated and eliminated. In addition, this monumental achievement has significantly cut down the number of data requests needing to be filled manually, and has made the accessibility of data, in most cases, nearly instantaneous via the web. At the completion of the project in September 2004, nearly ten million observations contributed by over 6,000 observers worldwide were made available via the AAVSO web site.

The AAVSO is continuing to tout the praises of its observers and resulting database by working with the National Virtual Observatory (NVO) to make the data available through the NVO framework and useable with NVO-developed tools. The NVO connection will make the valuable observations made by AAVSO observers even more accessible to both the professional and amateur communities, particularly to those not familiar with the AAVSO's rich database.

As a follow-up to the original DataValidation Project, work is presently underway to continue the validation of data from January 2002 to the most recent month of archived data, making the dataset for a given star as complete and up-to-date as possible. Future plans also include the validation of stars not included in the original

project, such as eclipsing binaries, RR Lyrae stars, suspicious comparison stars, and more.

10. Acknowledgements

The AAVSO Data Validation Team would like to extend its most sincere and heartfelt thanks to the late Dr. Janet A. Mattei, Director of the AAVSO from 1973 until her passing in 2004. Janet's vision and passion for variable stars shone through her leadership of the AAVSO and continues to radiate through the generations of professional and amateur astronomers that she inspired. On behalf of Janet and the association, we would like to thank each and every observer for their tireless dedication and shared adoration for variable stars, and for their contribution, no matter how big or small, to the AAVSO International Database.

The AAVSO gratefully acknowledges NASA grant NAG5-12602 for providing funding for the AAVSO Data Validation Project.

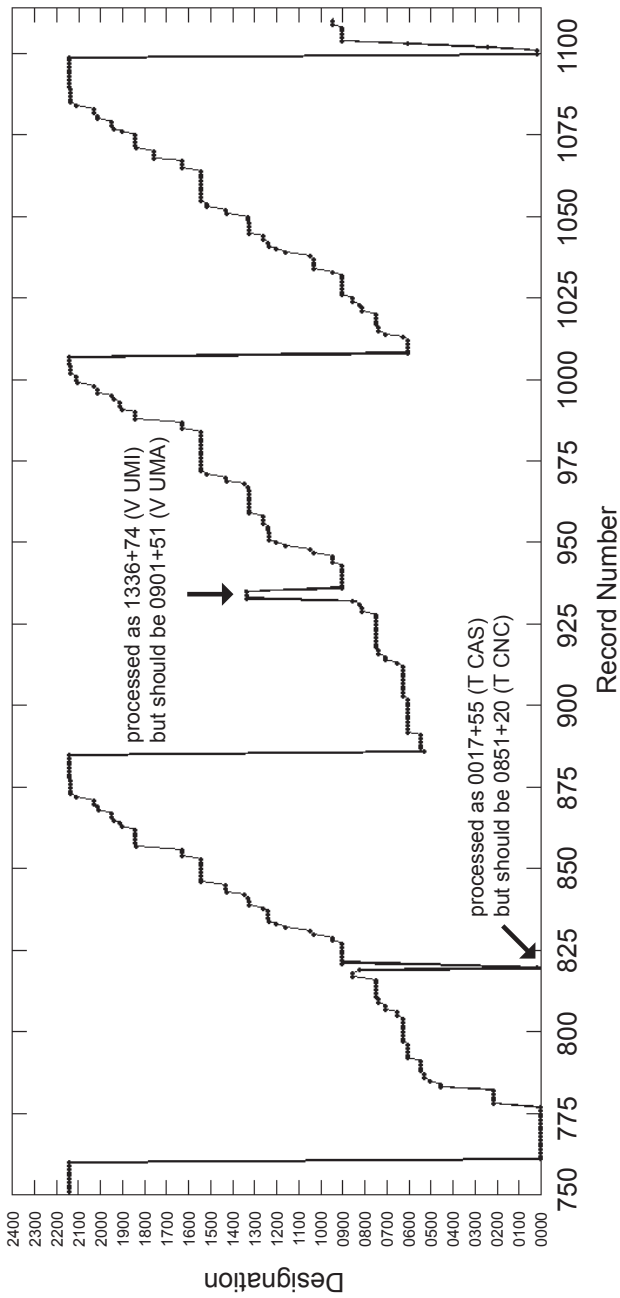

Figure 1. Finding designation problems in AAVSO data from a single observer (L—Giovanni Lacchini). In special cases, the validator could refer to the original raw data files as keypunched by the data entry technician. Since they were entered in order of designation, a plot of designation as a function of record number, as shown here, often revealed "spikes" where star name/designation problems occurred.

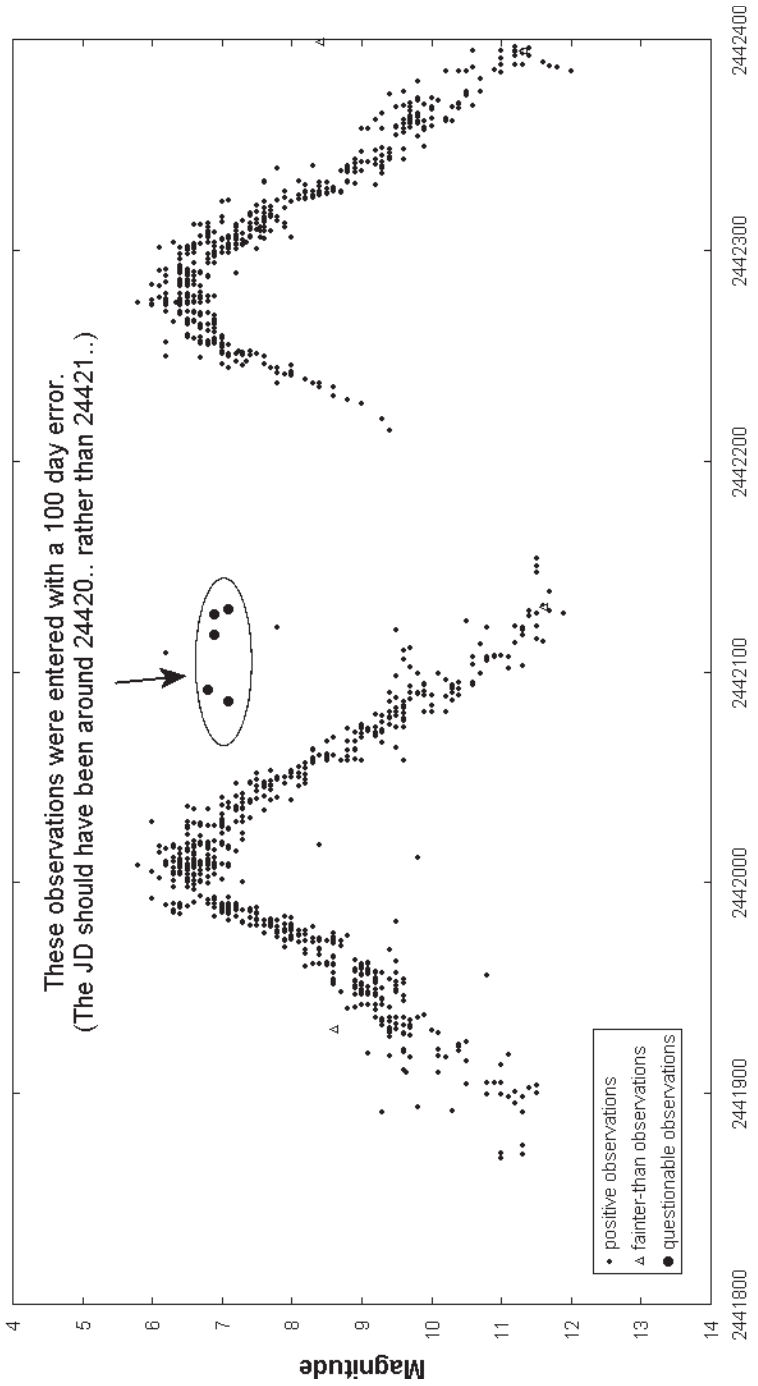

Figure 2a. R Trianguli light curve sample before validation data correction—questionable data points are circled.

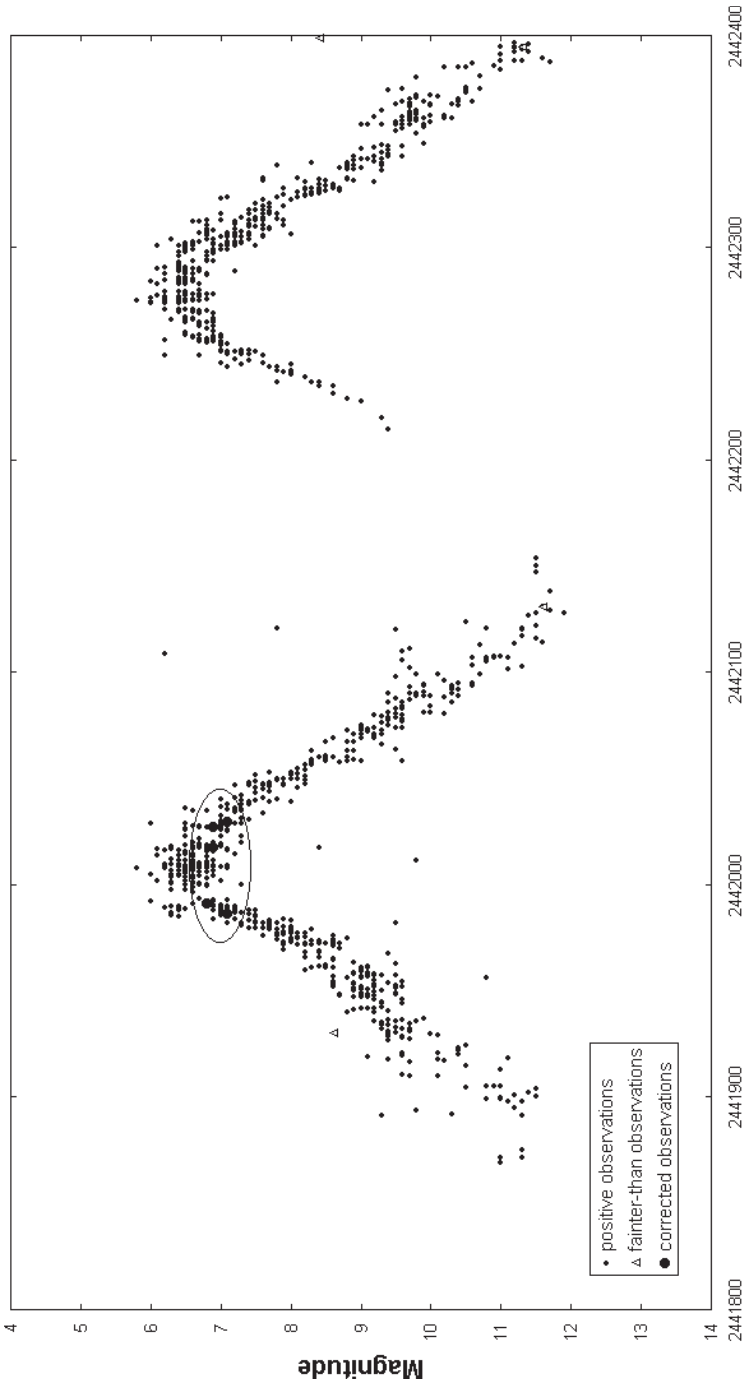

Figure 2b. R Trianguli light curve sample after validation data correction—the corrected data points are circled. (The rest of the light curve has not been validated.)