Data mining highly multiple time series of astronomical observations

F. Huang
School of Information Technology, Deakin University,
221 Burwood Highway, Victoria 3125, Australia

Abstract

This is a case study of data mining a large data set of astronomical interest. Our first concern is the outliers apparently existing in the data set. We used a robust method to do curve fitting and identify outliers, and estimated the occurrence intensity of outliers. We find that the occurrence intensity of outliers varies considerably over time. Besides, we designed a test which led to rejection of the hypothesis that all observation series are independent of each other. Combining this fact with our estimation of the occurrence intensity of outliers we believe there are common factors transiently acting on many series of observations. Additionally, we analyse gaps in time series and summarise simple but possibly interesting characteristics of data from a methodological viewpoint of data mining.
Keywords: data mining, highly multiple time series, loess, MACHO project, nonparametric curve fitting, outliers.

1 Introduction

The MACHO Project is a collaboration between scientists at the Mt. Stromlo and Siding Spring Observatories, the Center for Particle Astrophysics at the Santa Barbara, San Diego, and Berkeley campuses of the University of California, and the Lawrence Livermore National Laboratory. The primary aim is to test the hypothesis that a significant fraction of the dark matter in the halo of the Milky Way is made up of objects like brown dwarfs or planets: these objects have come to be known as MACHOs, for MAssive Compact Halo Objects. The signature of these objects is the occasional amplification of the light from extragalactic stars by the gravitational lens effect. The amplification can be large, but events are extremely rare: it is necessary to monitor photometrically several million stars for a period of


Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

years in order to obtain a useful detection rate. For more details, see Cook et al., 1999, etc., and also the website: http://wwwmacho.anu.edu.au. Besides the primary aim, the MACHO project also enriched our knowledge on stars. For example, it was realized from the MACHO project that there are variable stars and non-variable stars, and it is important to estimate the period or frequency of a variable star. Quinn and Tompson (1990), Breiman (1995) and Hall, Reimann and Rice (1999) give various methods.

The dataset in this study contains 791 observation time series of blue spectral band of stars in the Large Magellanic Cloud. There are in total 2001 observation times spreading over 1566 days. While there is more than one observation value in some one-day periods for some stars, there are no observation values for some stars at some observation time. The values of the observation time, or the relative Julian date called in this paper, is Julian date minus 2449000.

In section 2 we note there is a long period when there is no observation data at all, and show that there are many outliers in the data set. We illustrate a method for identifying all outliers and estimating the occurrence intensity of the outliers over the observation period. In section 3, a statistical test is suggested which shows there are common factors acting transiently on many observation series among the total of 791 time series. Concluding remarks and comments are given in section 4.

2 Missing values and occurrence intensity of outliers

The first step in exploratory data analysis is to view data using graphical methods. By doing so for the MACHO data set, we find there are many periods when there are no observations at all. Figure 1 shows the longest period of no observations in the data set, which starts at the relative Julian date 319 and ends at 372. That is, for a period of 52 days there is no data for all series.

Later I learned that it was due to a fire accident occurred at the Mt. Stromlo Observatory.

Still by viewing the data through graphical tools, we find there are apparently many outliers in the data set. From the plot of all time series we find that there are particularly many outliers for many observation series in some periods, while just a few in most other periods. This observation lets us think that perhaps there are common factors acting on many observation series during some period. To get support from the data, we first identify out all outliers in all time series and get an estimate of the occurrence intensity of the outliers. We remark that we can only see a few very large outliers in the plot of all time series since the scales of the time series vary. A better view is to standardize all time series before plotting them together. To get an estimate of the occurrence intensity of the outliers we identify all outliers for each series and put them together to obtain the occurrence intensity. The method is illustrated below.

From Figure 2 we can see that the observation series could be fitted by non-parametric curve fitting method if there were no outliers. In our case there are many outliers; a good curve-fitting approach helps identify outliers, while existence of outliers affects the curve fitting. One possible approach is to identify

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

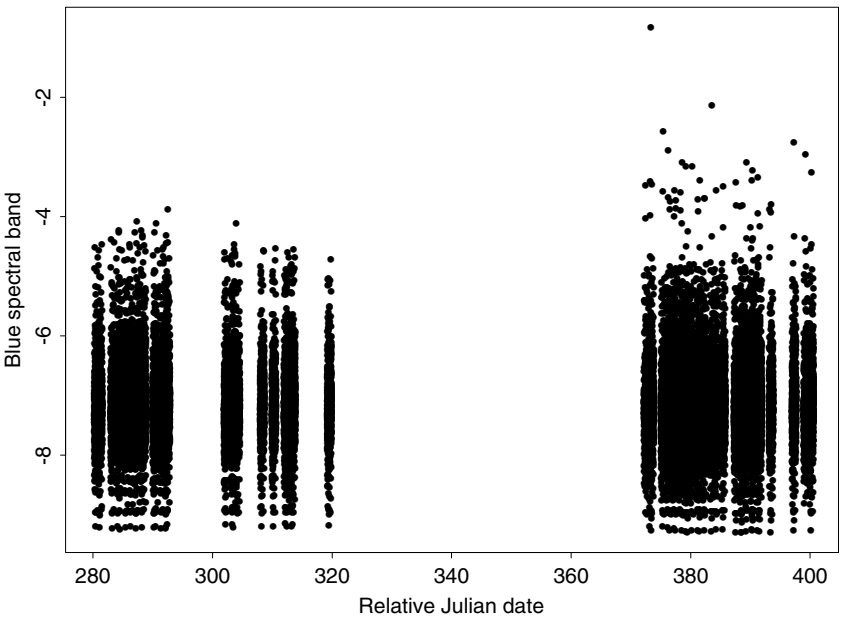

Figure 1: The longest period of no observations in the data set.

outliers using methods like Cook's distance (see Cook and Weisberg, 1980, p.497) in a parametric or locally parametric model. However, this approach is potentially very expensive in computing time, and so we took another approach. We used robust methods for curve fitting. In the statistical literature there are several tools to do this. In the case of parametric fitting, we can choose to use L1 fit, Huber's robust fit, etc. We used a robust local regression procedure called *LOESS* (see Cleveland and Grosse, 1991; Chambers and Hastie, 1992; Fan and Gijbel, 1996; etc.) in this research.

The model assumption for the local regression method *LOESS* is $y_i = g(x_i)+\epsilon_i$, for $i = 1, ..., n$, where g is a smooth function and ϵ_i are independent and identically distributed noise with a symmetric distribution. The method of *LOESS* is to approximate function g locally by, say, a polynomial of order 2, $g(x) \approx a_0 + a_1(x - x_i) + a_2(x - x_i)^2$ for x in a neighborhood of x_i. Estimation of a_0, a_1 and a_2 leads to an estimation of $g(x_i)$, $\hat{g}(x_i)$. The function call in SPLUS *(UNIX Version 5)* is *loess(X, Y, distribution=symmetric, span=myspan, degree=2),* where X is the series of observation times in relative Julian date and Y is the corresponding observed values of the blue spectral band. The assumption of a symmetric distribution of the noise leads to robust fitting (see Chambers and Hastie, 1992 for details). The values of span, $myspan$, were taken to be 0.02 for most series, and 0.1 or 0.5 for some other series, which were adjusted manually. Then we obtained

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

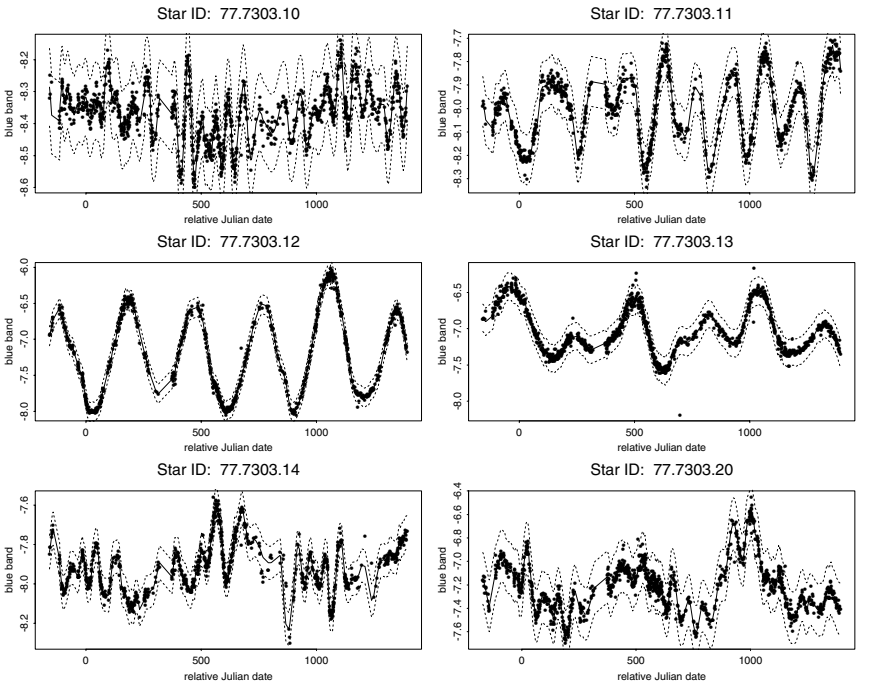

Figure 2: Fitted curves for the first six stars and corresponding cut curves for outliers.

$y_i = \hat{g}(x_i) + \hat{\epsilon}_i,$ for $i = 1, ..., n$, where $\hat{g}(x_i)$ is the fitted value at x_i and $\hat{\epsilon}_i$ is the residual of the fitting. Since we used a robust method we find the fitting is good although there are many outliers in the data. To determine the outliers, we first calculated an estimator of the standard deviation of the residual errors by a robust procedure *MAD* (see Hampel et al, 1986), $\hat{\sigma} = 1.4826 * median_i(\hat{\epsilon}_i - median_i(\hat{\epsilon}_i))$, then determined y_i as an outlier if $|\hat{\epsilon}_i| > cut * \hat{\sigma}$, where cut takes values of $6, 12, 18, 24, 30, 36$. The six plots in Figure 2 show the original data, the fitted curve $(x_i, \hat{g}(x_i))$ and curves $(x_i, \hat{g}(x_i) \pm 6\hat{\sigma})$ for the first six observation time series. We can see both the curve fitting and cutting out of outliers is good, although the *cut* values of 6 might be conservative for some observation values (i.e., too large to cut out some observation values to be outliers).

Figure 3 illustrates histograms of the numbers of outliers over the observation period for $cut = 6, 12, 18, 24, 30, 36$, from which we can see that outliers especially very wild outliers come to many observation series simultaneously, which leads us think that there may be common factors acting on many observation series. We propose a statistical test of this hypothesis in the next section.

We have also tried other non-parametric or parametric methods and the results are very similar. Methods we have tried include local linear non-parametric fitting (i.e., use the Splus command $loess(X, Y, distribution = symmetric, span =$

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

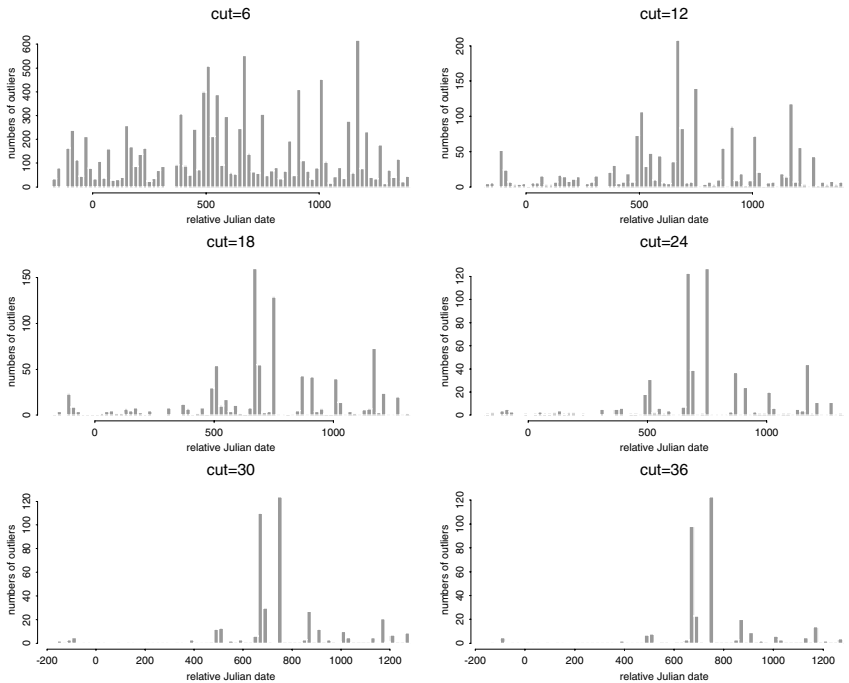

Figure 3: Histograms of numbers of outliers corresponding to $cut = 6, 12, 18, 24, 30$ and 36, respectively.

$myspan, degree = 1$) instead), L1-norm and Huber's robust fitting using multiple frequency trigonometric model (see formula (2.3) in J.Reimann, 1994), etc.

3 A statistical test that common factors exist

In this section we design a test for the hypothesis that common factors acting on many observation series exist. The idea of the test is to use the residual errors obtained from the non-parametric fitting illustrated in the previous section. If the fitting is good, which we believe to be the case, then the residual errors of each series would be like white noise and independent of each other, if the series were independent of each other. So the null hypothesis is that all series are independent of each other, and we reject this as follows.

We suppose there are values at each observation time for each series, so we have $(y_{s,t})_{m\times n}$, $(x_{s,t})_{m\times n}$, $(g_s(x_{s,t}))_{m\times n}$, $(\epsilon_{s,t})_{m\times n}$ and $(\hat{\epsilon}_{s,t})_{m\times n}$, where $m = 791, n = 2001$. But only some but not all of $(y_{s,t})_{m\times n}$ are observed, and the corresponding subset values of $(\hat{\epsilon}_{s,t})_{m\times n}$ are available. Define $\hat{T} := \max_{t=1,\ldots,n} \sum_{s=1}^{m} I(\hat{\epsilon}_{s,t} > 0)$. Although not all values of $\hat{\epsilon}_{s,t}$ are available, we know $\hat{T} \geq 587$ from available values of $\hat{\epsilon}_{s,t}$ calculated from robust curve fitting (see Figure 4).

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

By the large sample theory for non-parametric curve fitting (see Cleveland and Devin, 1988; Chambers and Hastie, 1992; Fan and Gijbels, 1992, 1996; etc.), the asymptotic distribution of $\hat{\epsilon}_{s,t}$ has zero median. So, under the null hypothesis, $P(\hat{\epsilon}_{s,t} > 0) \approx P(\epsilon_{s,t} > 0) = 1/2$, $P(\max_{t=1,\ldots,n} \sum_{s=1}^{m} I(\hat{\epsilon}_{s,t} > 0) \geq k) \approx P(\max_{t=1,\ldots,n} \sum_{s=1}^{m} I(\epsilon_{s,t} > 0) \geq k)$, and thus $P(\hat{T} \geq 587) \approx P(T \geq 587) = p_{587,791,2001} \approx 0$ (cf. Appendix), which leads to rejection of the hypothesis. Note that the approximation made by large sample theory is reasonable, since m is as large as 791.

Figure 4 is the plot of

$$\left(\sum_{s=1}^{m} I(y_{s,t} \text{ is observed and } \hat{\epsilon}_{s,t} > 0), \sum_{s=1}^{m} I(y_{s,t} \text{ is observed and } \hat{\epsilon}_{s,t} \leq 0) \right)$$

for $t \in (600, 800)$, from which we can see the most deviated point (587,106) which occurred at the relative Julian date 790.9633.

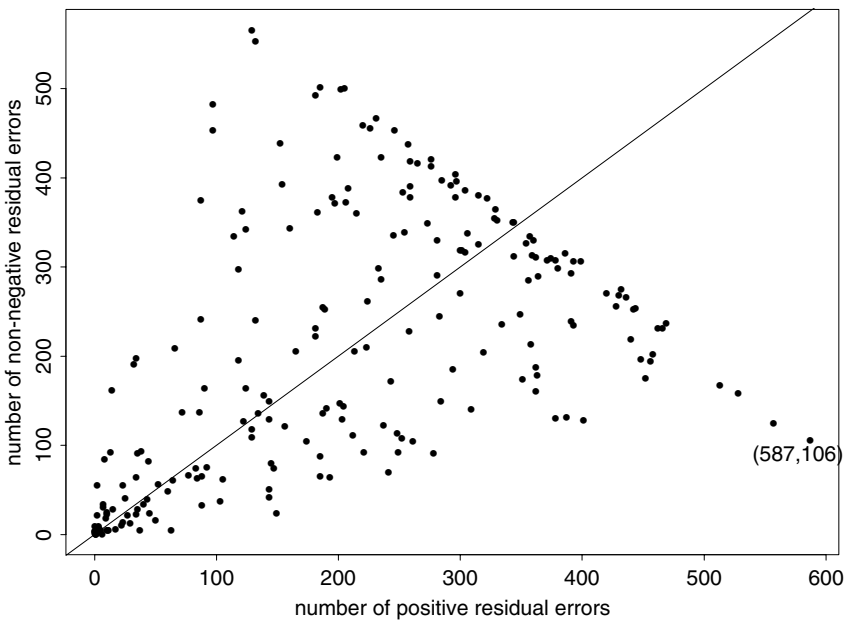

Figure 4: Plot of numbers of non-negative residual errors over numbers of positive residual errors.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)

4 Concluding remarks

This article is a case study of highly multiple and irregularly spaced time series. We study all 791 series together, trying to find common things among them. While common missing data periods are easy to see, concepts like occurrence rate of outliers are based on multiplicity of time series and might not be easy to calculate. We give a relatively fast method to determine outliers in a huge dataset of highly multiple time series, and a statistical test to see if common factors acting on many time series exist.

A surprisingly long common period during which there are no data at all for all series lead us learn that there was a fire accident at the observatory and the dataset is a combination of observations before and after the fire accident. With the assistance of statistical tools we also find that the occurrence intensity of outliers varies considerably day by day, i.e., many observation series have many outliers on some days while just a few during other days, and there are common factors acting transiently on many observation series. If the common factor could be thought something in the solar system yet to determine, we can get some checking points through the calculation in this article. For example, on relative Julian date 790.9633 there are many outliers in many observation series, so astronomers could look particularly at that day to find if there were some special things then. It is my understanding that any conclusions on the remote stars could only be reached after those common factors happened in the solar system have been separated out using statistical methods. That is the reason we stress the existence of common factor.

The main statistical tools used are: robust non-parametric curve fitting and robust estimation of variance; asymptotic theory of non-parametric fitting and fundamental principle of hypothesis testing in statistical theory.

Approaches which might be worth summarizing from the viewpoint of *data mining* might be: 1. find the longest common period of missing data, or no data, to see if it is surprisingly long; 2. very fast algorithm to identify outliers in a huge data set of highly multiple time series; 3. test hypotheses using statistical methods.

This is just a first report of our research, leaving many topics for further study. For example, separation of the common factors, more strict treatment on the theory part, etc., would be considered.

Acknowledgements

Thanks go to Profs. Peter Hall and Markus Hegland for their suggestions and revisions of the paper. Thanks also go to Peter Milne, Graham Williams and Bill Clarke for helping me access the data.

Appendix. Probability in the independent case

Let $(\epsilon_{s,t})_{m\times n}$ be a random matrix which corresponds to the noises of m time series of length n. We suppose the array random variables are independent of each other, with zero medians. Let $T := \max_{t=1,\ldots,n} \sum_{s=1}^{m} I(\epsilon_{s,t} > 0)$ be the theoretical

statistic. Given an integer $k \in \{1, ..., m\}$, we wish to compute the value of the probability $p_{k,m,n} := P(T \geq k) = P(\sum_{s=1}^{m} I(\epsilon_{s,t} > 0) \geq k$ for some $t \leq n)$. In fact,

$$
\begin{aligned}
&P(\textstyle\sum_{s=1}^{m} I(\epsilon_{s,t} \geq 0) \geq k \text{ for some t}) = 1 - P(\sum_{s=1}^{m} I(\epsilon_{s,t} > 0) < k \text{ for all t}) \\
&= 1 - \textstyle\prod_{t=1}^{n} P(\sum_{s=1}^{m} I(\epsilon_{s,t} > 0) < k) = 1 - \prod_{t=1}^{n}(1 - P(\sum_{s=1}^{m} I(\epsilon_{s,t} > 0) \geq k)) \\
&= 1 - (1 - P(\textstyle\sum_{s=1}^{m} I(\epsilon_{s,1} > 0) \geq k))^n = 1 - \exp(n \log(1 - P(\sum_{s=1}^{m} I(\epsilon_{s,1} > 0) \geq k))) \\
&\approx 1 - \exp(-nP(\textstyle\sum_{s=1}^{m} I(\epsilon_{s,1} > 0) \geq k)) \approx nP(\sum_{s=1}^{m} I(\epsilon_{s,1} > 0) \geq k) \\
&= nP(\tfrac{\sum_{s=1}^{m} I(\epsilon_{s,1} > 0) - m/2}{\sqrt{m/4}} \geq \tfrac{k - m/2}{\sqrt{m/4}})) \approx nP(\eta \geq \tfrac{k - m/2}{\sqrt{m/4}})
\end{aligned}
$$

where η is a random variable having a standard normal distribution. Corresponding to the size and values of our data analysed in section 3, the probability is $p_{588,791,2001} \approx 0$. Since m is as large as 791, and the probability value of the standard normal distribution corresponding to $p_{588,791,2001}$ is so small, the approximations made in the above reduction do not significantly affect the value of $p_{588,791,2001}$ by central limit theorem and Taylor expansion.

References

[1] Chambers, J.M. and Hastie, T.J., *Statistical Models in S,* pp. 309-376, 1991.

[2] Clarke, B. and Hegland, M., Identification and classification of interesting variable stars in the MACHO database, *CTAC'99,* Canberra, 1999.

[3] Cleveland, W.S. and Devlin, S.J., Locally-weighted regression: an approach to regression analysis by local fitting, *J.Am.Statist.Assoc.,* **83,** pp. 596-610, 1988.

[4] Cleveland, W.S. and Grosse, E., Computational methods for local regression. *Statistics and Computing,* **1,** pp. 47-62, 1991.

[5] Cook, K.H. et al, Variable stars in the MACHO collaboration database, *Astrophysics Journal,* **34,** pp. 345-386, 1999.

[6] Cook, R.D. and Weisberg, S., Characterization of an influence function for detecting influential cases in regression. *Technometrics,* **22,** pp. 495-508, 1980.

[7] Fan, J. and Gijbels, I., Variable bandwidth and local linear regression smoothers. *Annals of Statistics,* **20,** pp. 2008-2036, 1992.

[8] Fan, J. and Gijbels, I., *Local Polynomial Modelling and Its Applications,* Chapman & Hall, London, 1996.

[9] Hall, P., *Biometrika* Cenetary: Nonparametrics. *Biometrika,* **88(1),** pp. 143-165, 2001.

[10] Hall, P., Reimann, J. and Rice, J., Nonparametric estimation of a periodic function. *Biometrika,* **87(3),** pp. 545-557, 1999.

[11] Hampel, F., Rousseeuw, P., Ronchetti, E. and Stahel, W., *Robust Statistics,* John Wiley & Sons, New York, 1986.

[12] Quinn, B.G. and Thomson, P.J., Estimating the frequency of a periodic function. *Biometrika,* **78,** pp. 65-74, 1991.

[13] Reimann, J.D., *Frequency Estimation Using Unequally-Spaced Astronomical Data.* Ph.D thesis, University of California at Berkeley, 1994.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)