4.3 FUZZY IMAGE PROCESSING APPLIED TO TIME SERIES ANALYSIS

R. Andrew Weekley[1]*, Robert K. Goodrich[1,2], Larry B. Cornman[1]
[1]National Center for Atmospheric Research**, Boulder, CO
[2]University of Colorado, Department of Mathematics, Boulder, CO

1. INTRODUCTION

The analysis of times series data plays a fundamental role in science and engineering. An important analysis step is the identification and classification of various features in the data. Quality control can be viewed as a subclass of general feature identification and classification, for example, differentiating between a true signal and a contaminating signal. Many algorithms exist for the quality control of time series data, such as Fourier or wavelet analysis, as well as robust and standard statistics. (Abraham and Ledolter 1983, Priestly, 1981 and Barnett and Lewis, 1977). However these algorithms are applicable only when certain assumptions are satisfied, such as stationarity, and a relative few number of outlier (less than 50% of the data). Unfortunately there are times when an instrument is failing and the assumptions of the standard methods are violated, and hence are not applicable. However, a quality indicator is still needed. Typically the image processing of a human is used to identify and mitigate failure mode data. Human analysts are adept at feature identification and classification, nevertheless in many applications it is desired to have an automated algorithm that performs this role. In this paper, a machine-intelligent algorithm that mimics the feature classification and identification processing of the human analyst is presented and applied to the quality control of time series data. In cases where there are a large number of outliers - a situation that is problematic to most algorithms - this algorithm is able to classify the desired signal as well as the outliers.

For example, consider the time series in Figure 1 consisting of anemometer measurements of wind speed in ms^{-1} as a function of time in seconds. Aside from a few isolated outliers, it is straightforward to see that this is good quality data, and that most quality control algorithms would not have difficulty in diagnosing the outliers. On the other hand, the data in Figure 2 is certainly more complex, showing similar regions of good data as with the previous figure, but intermixed

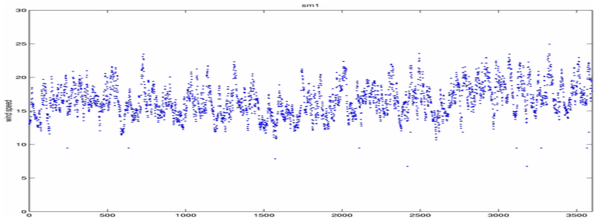

Figure 1. Nominal anemometer time series data. The vertical axis is wind speed in m s^{-1}. The horizontal axis is time in seconds.

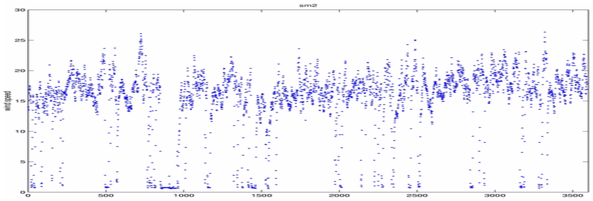

Figure 2. A nearby anemometer time series in a failure mode. Here a nut holding the anemometer head has worked loose.

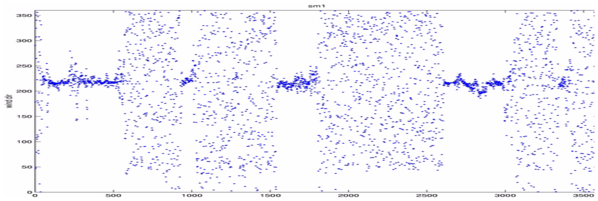

Figure 3. Data from a spinning anemometer. The vertical axis is wind direction measured in degrees from north in a clockwise direction, the horizontal axis is time in seconds.

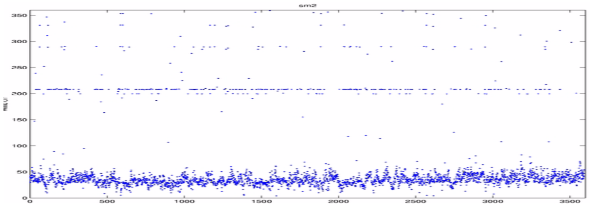

Figure 4. Data from an anemometer in a failure mode caused by a bad transistor.

* *Corresponding author address:* R. Andrew Weekley, NCAR/RAP, P.O. Box 3000, Boulder, CO 80307-3000.
e-mail: weekley@ucar.edu

** The National Center for Atmospheric Research is sponsored by the National Science Foundation.

with numerous outliers. This data is from an anemometer located three meters from the one whose data is shown in Figure 1, and from the same time period. In this situation, standard methods might work well on certain segments of this time series, but other sections -- such as between 800 and 1000 seconds -- might cause problems. However, except for some of the points close to the “good data,” a human analyst would not have much difficulty in discerning the good from the bad data (in fact, this anemometer was having a mechanical failure where the strong winds vibrated, then loosened, the nuts holding the anemometer in place).

Consider the additional cases shown in Figure 3 and Figure 4. From video footage, it has been observed that certain wind frequencies can excite normal modes of this type of anemometer’s wind direction head and can cause the device to spin uncontrollably. Data from such a case can be seen in Figure 3, where the vertical axis is wind direction measured in a clockwise direction from North. The horizontal axis is again time measured in seconds. Between about 500 seconds and 1000 seconds the wind direction measuring device is spinning and the data becomes essentially a random sample of a uniform distribution between about 50 degrees to 360 degrees. The true wind direction is seen intermittently at about 225 degrees, which is in general agreement with the value from another nearby anemometer. Figure 4 shows the wind direction at another time distinct from that in Figure 3; the true wind direction is around 40 degrees. Notice the suspicious streaks in the time series data near 200 degrees, as well as other spurious data points. Again standard time series algorithms would have a difficult time with these two examples, however it is straight forward for the human analyst to identify both the suspect and good data. This illustrates another aspect of the time series problem beyond identifying points as outliers, that is, *classifying* the nature of the outliers.

It is difficult to create a single algorithm that can detect and identify many different types of data quality problems. Given that the human analyst is able to quality control data in a pathological case, motivated the development of a multi-component, fuzzy logic machine intelligent algorithm, the Intelligent Outlier Detection Algorithm (IODA). IODA incorporates cluster analysis, fuzzy image processing, local and global analysis, correlation structure, as well as *a priori* knowledge when available, and returns a quality control index (confidence value) between 0 and 1 that indicates the reliability of the data. In this paper the techniques to accomplish such processing are given in the context of anemometer time series data. It is important to note that these techniques could be modified to accommodate time series data from other instruments with different failure processes, or failure modes.

2. FUZZY IMAGE PROCESSING

When creating a fuzzy logic algorithm, the characteristics and rules a human expert might use to partition the data (Zimmerman, 1996) into a classification set C must be determined. Various tests are devised to ascertain whether a certain datum should be in C. The result of each test are transformed into values between 0 and 1 by ***membership functions***. A membership function value near 1 for some test T is interpreted to mean that the test indicated that it was likely that the datum should be classified as in C, and a value near zero would indicate the datum is not likely to belong to C. These membership functions are combined in many different ways by fuzzy logic rules to obtain a final membership function. A threshold is set and the points with final membership values above this threshold are defined to be in C. If a threshold is not used the value of the combined membership function may be used to indicate the degree to which the datum belongs to C (a confidence value). This final membership function is called the fuzzy set C since the membership value indicates the degree to which a datum belongs to C. This methodology allows for a multiplicity of tests and membership functions. In the case of fuzzy image processing, a membership value is assigned to each point from an analysis of an image (Chi, Hong and Tuan, 1996). This value can be thought of as a height and by interpolation a surface. See for example Figure 6 where a cold (blue color) represents a higher height in the surface.

3. INTELLIGENT OUTLIER DETECTION ALGORITHM (IODA)

Suppose the data from Figure 2 is broken into overlapping subregions using a sequence of running windows. For each data window, an estimate of the probability density function (i.e. a normalized histogram) is calculated (note that the window size must be selected large enough to contain enough data to give reasonable estimates of the density functions, but small enough to reflect the local structure of the non-stationary time series). These histograms can be plotted as a waterfall

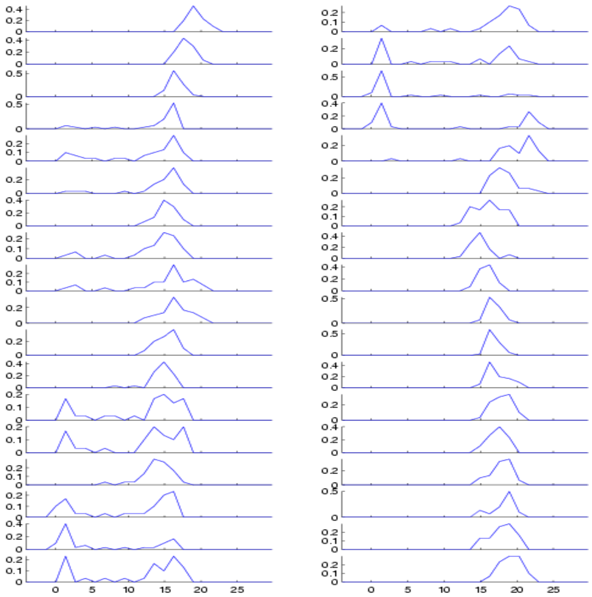

Figure 5. Plot of stacked time series histograms. The feature at the left represents the drop out data caused by the loose nut.

plot, or stacked histograms, as shown in Figure 5. The histogram for the first time window is shown in the bottom left, the plots then run up the left column as a function of time and continue from the bottom right plot to the top right. These stacked histograms can also be plotted as a contour image, left to right as a function of time, as shown in Figure 6. The image in Figure 6 is the first fuzzy interest map used in IODA, and is called the ***histogram field*** (Figure 6 is a plot of the entire hour of data, whereas Figure 5 is of only the first 555 data points). The contour plot in Figure 6 can be thought of as a surface, where the color of the contour represents the height of the surface above each point in the time-wind speed plane. The dark blue colors represent a larger height, whereas the dark red color represents a height near zero.

It is natural for a human to see large clumps of blue in Figure 6. These region in the image can be encircled using a contour algorithm and define concentrations or clusters of points as shown in Figure 6. Here the cluster boundaries that surround the data points of the cluster are shown in blue. These clusters are found in the histogram field by selecting a threshold for the contour algorithm, if a lower threshold is selected, the clusters grown in size and connect. A sequence of clusters can be found by incrementally lowering the contour threshold, or by “lowering the water”. This expression is related to the idea that the contour is a set of mountain peaks and the threshold level

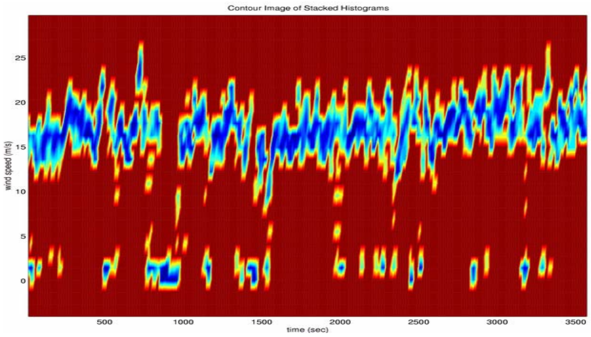

Figure 6. The data from Fig. 5 is plotted from left to right. The horizontal axis is time in seconds. The vertical axis is wind speed in m s $^{-1}$. The color represents the height of this histogram with blue a higher height and red a height near zero.

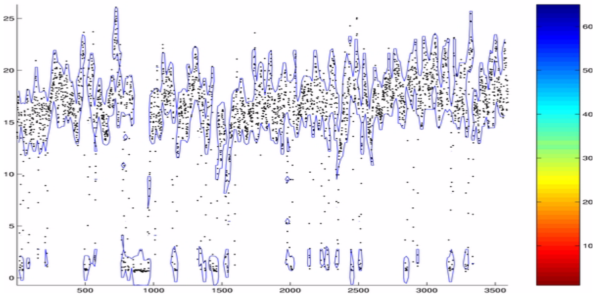

Figure 7. A threshold is selected and a contour plot is drawn around the data above the threshold. This breaks the data into several clusters.

represents a water level. As the level is lowered, the peaks become connected by ridge lines.

Notice these blue clumps do not contain all the data points in the original time series, i.e., there is ***cluster data*** and ***non-cluster data***, however by inspection, the analyst combines these local clusters into larger scale ***features***. For instance in Figure 6, a human expert might group the large clusters centered around 17 m/s^{-1} into a feature and the others near 1 m/s^{-1} into a second feature. In actuality there are three tasks at hand: characterize/categorize the clusters, group the clusters into features and characterize/categorize the non-cluster data. In this way, the “good” and “bad” data are distinguished. For example, in this case there are three cluster categories, atmospheric clusters, failure mode clusters, and unknown clusters. ***Atmospheric clusters*** contain points that fit an expected model, ***non-atmospheric clusters*** are clusters that do fit an expected model for a particular failure mode. Unknown clusters are neither atmospheric or non-atmospheric. The notion of atmospheric, non-atmospheric clusters are quantified using fuzzy logic algorithms. These fuzzy logic

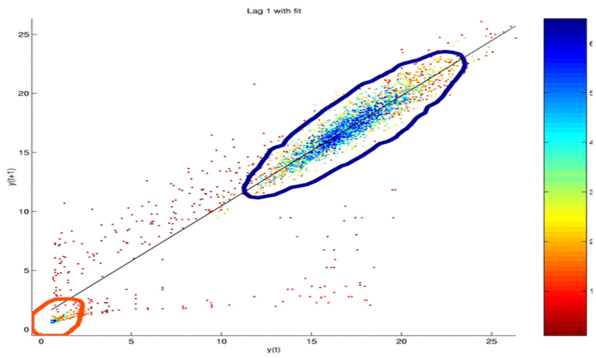

Figure 8. A plot of the ordered pairs (y(t), y(t + 1)). The vertical and horizontal axes are wind speed in m s $^{-1.}$

algorithms require *a priori* knowledge of the characteristics of the atmospheric signal as well as knowledge of the failure modes. One of the strengths of fuzzy logic algorithms is the ease in which new characteristics can be added, when needed or discovered. It is important to note that not all of the data points will clearly exhibit the *a priori* expected behavior. Again, this natural ambiguity in classifications is handled by fuzzy logic methods.

3.1. Classifying clusters

In general, clusters found using the above techniques will, by definition, group data structures together. Clusters can be classified according to whether they are consistent with a known problem or failure mode, such as the loose nut scenario, or result from nominal atmospheric data, i.e., data with expected statistics. One such characterization is auto-correlation as defined by lags of data. Consider the scatter plot of y(t) vs. y(t+1), or lag 1, for the loose nut case, shown in Figure 8. Here y(t) is the wind speed at time t. The color for each point is simply the geometric mean of the *initial confidence* for the points y(i) and y(i+h) is given by:

$$C_{i,\, i+h} = \sqrt{C_i \cdot C_{i+h}}$$

The solid black line is the confidence-weighted linear best fit to the data (here h=1). Notice in the lag scatter plot there are two distinct groups of data, the atmospheric data centered near 18 m s^{-1}, and the drop out data centered near the origin. The fact there are two groups of data in the lag plot indicates that the data is not stationary, and can be used later to separate the histogram clusters into stationary groups of clusters. It is possible to define a confidence-weighted auto correlation. Thus in lag(1) space, pairs of points that fit the expected model should cluster around a line with a slope close to one. A $\rho = \rho(1)$ (the sample correlation) value close to zero indicates a poor fit and a value near one indicates an excellent fit. In fact, ρ squared represents the percent of variation in *y*(i+1) explained by the fit.

Similar techniques can be applied to the lag plot to find clusters of points, that were applied to the time series data. An image can be created by calculating a local density using a tiling of overlapping rectangles. A contour algorithm can be applied to the resulting image and clusters of data can be found (other data clustering techniques would work well for this problem as well). Again the water can be lowered and new clusters can be found for each contour threshold, and a value for ρ can be calculated for each of these new clusters. The largest cluster that has a large value for ρ, and contains points that are close to the best fit line is then selected, as shown in Figure 8. This is done by a fuzzy algorithm. The color of the contour surrounding the lag points represents the atmospheric score given to that cluster. Where a cool color is a high score and a warm color is a low score. Notice, the high scoring cluster in Figure 8 contains the data from the time series plot that a human would probably classify as atmospheric. The pairs of points given by y(i) and y(i+1), in this lag cluster, are termed “atmospheric” since they fit the expected auto correlation model of atmospheric data. It is now possible to calculate an atmospheric score for each cluster in time series space.

3.2. Additional Membership Values

Numerous additional membership values can be calculated. For instance a membership value can be calculated by recursively fitting the time series with straight lines. Suppose a segment of data is fit with a line, and an quality of fit for the data is calculated (such as ρ squared). If the quality indicator is too low, the best fit line is bisected and new fits are calculated for the two new sets of data. Fits are calculated until the quality indicator is good enough or there are to few points in the fit. A ***local best fit*** confidence can then be calculated by how far a point is from the fit. Another similar membership value can be calculated from the lag plot, and the atmospheric lag cluster. Specifically the number of sigma a point is from the best fit line can be calculated given the variance in the atmospheric lag cluster data. A ***lag cluster nsigma*** confidence can then be calculated using an appropriate membership function such as e^{-x^2}.

3.3. Final confidence calculation and the Final Feature

Recall from the Figure 6 that there were multiple clusters in the primary mode data. These clusters can be combined into a single large cluster or a ***final feature*** that spans the entire time interval. The idea is to partition the clusters into meta-clusters, or cluster the clusters. In the histogram field shown in Figure 6, both the good clusters (centered on 17 ms^{-1}) and the dropout clusters (near zero) appear as bright blue regions. Recall that there were two clusters of data in the lag plot (Figure 8), if the time series data were stationary then there should only be a single lag cluster, i.e., the points in the lag-1 plot would be distributed along the one-to-one line. The fact there are two such clusters, can be used to partition the histogram clusters into meta-clusters that belong to the same stationary feature, that is by determining which data points occur in which cluster in both (Figure 8 and Figure 6). For instance, suppose all the points which are in the large cluster near the center of Figure 8 are given a large ***stationary lag cluster membership value***. Next the ***stationary histogram cluster membership value*** is calculated by finding all the histogram clusters with a point which has a large stationary lag cluster membership value. Consequently all clusters centered on 17 ms^{-1} in Figure 6 will be given a large stationary lag cluster membership value, and hence belong to the same stationary feature.

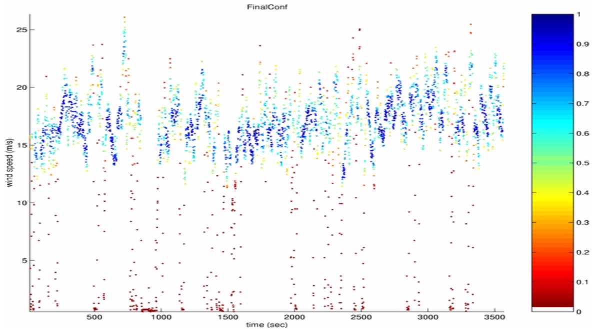

Figure 9. The data from Figure 2 is shown where the color of each point is determined by the combined membership function. A blue color has a membership near 1, and a red color has a membership near zero.

A ***combined membership value*** or ***final confidence*** for each point can be calculated (Figure 9) by a fuzzy combination of all the individual membership values, i.e. the histogram membership value, the stationary feature membership value, the local best fit value, the lag cluster nsigma value, and so on. Notice that the combined membership value correctly gives a low confidence to the data dropouts, and the spurious points that fall between the dropouts and the primary signal. Notice the confidence value performs well in Figures 10 and 11 as well. Figure 10 is the final confidence values for the data from Figure 3. Notice most of the points from time periods when the anemometer was spinning are given low membership values (red). Figure 11 is the final confidence values given for the data in Figure 4. Notice the points in the suspicious streaks in the time series data near 200 degrees are given low membership values (red). These confidence values in Figures 9, 10 and 11 do identify most of the outliers. Simulations and human verifications have been done, and IODA does well in classification of outliers. These studies will be reported elsewhere.

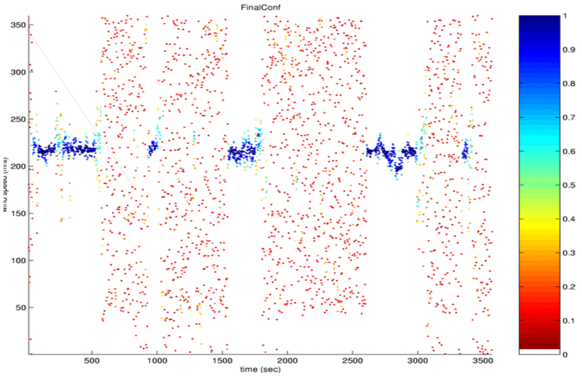

Figure 10. The data from Figure 3 is shown with blue colors representing points with high membership values.

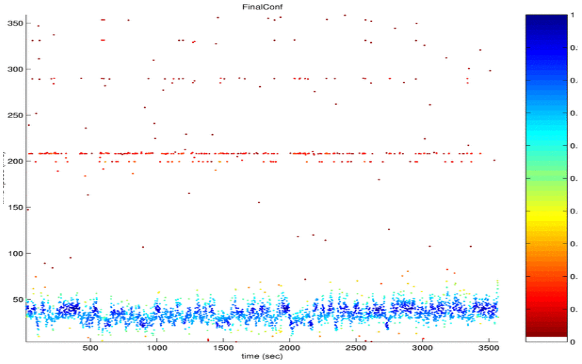

Figure 11. The data from Figure 4 is shown. Notice the streaks in the data are given a low membership value.

4. ADDITIONAL APPLICATIONS

The data in Figure 12 is the vertical wind as measured by an aircraft flying in the vicinity of Juneau. As before the color of the points indicates the confidence in the data. There are two categories of data in this plot, semi-continuous data

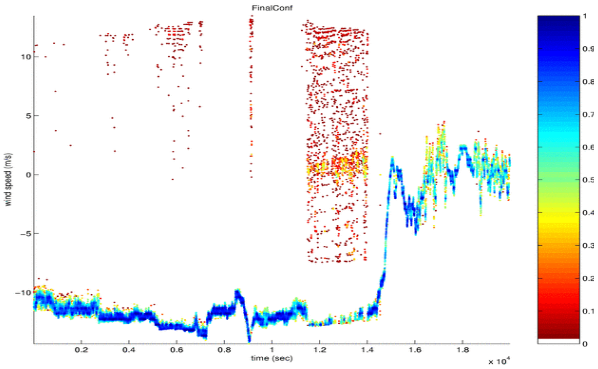

Figure 12. An example of aircraft wind speed time series data is shown. The vertical axis is in m s $^{-1}$ and the horizontal axis is in seconds. High membership values are blue.

(mostly cool colors), and disperse data (disperse warm points centered on 0 ms^{-1}). On a gross scale the confidence values assigned to the data correspond to what a human expert might give the points. However upon closer inspection there are low confidence points intermixed with the semi-continuous data. These halo points are not well auto correlated and hence are given a low atmospheric membership value. The confidence values shown in Figure 12 were calculated by IODA without any modification or tuning of parameters, and it is important to note that during the development of IODA such an example was never explicitly considered (although individual aspects were, such as the near uniform appearance of the disperse data). This example is encouraging evidence that the principles and implementation of IODA can be applied to time series data from sources other than anemometers.

Figure 13 is the result of a simulation. Points were selected at random from an interval of time with few outliers. The selected data points were then replaced with values selected from a uniform distribution with the same range as the nominal anemometer data. As can be seen from Figure 13 the underlying time series is discernible to the human eye (blue). Furthermore the confidences assigned to the time series again roughly corresponds the what a human might assign to the data (given infinite patience). Such examples are seen in lidar data in the presence of a weak signal return.

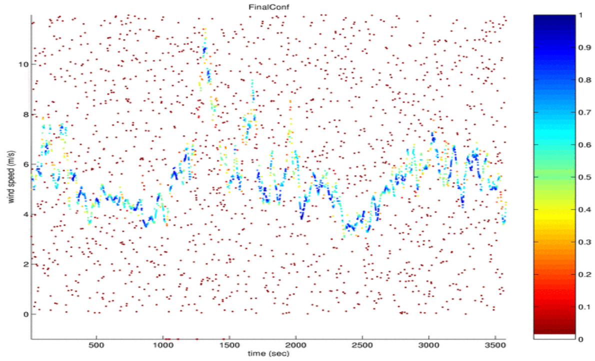

Figure 13. A simulation of anemometer data with uniform background noise. Notice the signal is given by high membership values (blue).

5. CONCLUSIONS

We have studied the use of fuzzy image processing techniques to find outliers in time series data. Even though IODA was developed using anemometer data, and their specific failure modes, these techniques should apply to other data sets as well. Specifically, if the time series is approximately stationary over the analysis time window considered, the range of the correlation coefficients for nominal data is understood in the first few lag spaces, the outliers are not well correlated with the true data, and the statistical properties of the outliers are known, then such an analysis should perform well. That is, if a human expert can separate the true data from the outliers in these and other images, then there is hope to construct fuzzy modules to separate the data from the outliers.

6. REFERENCES

Abraham, Bovas and Johannes Ledolter, 1983: Statistical Methods for Forecasting. J. Wiley and Sons, New York, pp. 445

Barnett, V. and T. Lewis, 1977: Outliers in Statistical Data, (3rd Edition), John Wiley and Sons, New York, pp. 584

Chi, Z., Y. Hong and P. Tuan, 1996: Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition, Wold Scientific, Singapore, pp. 225

Priestley, M. B., 1981: Spectral Analysis and Time Series, Academic Press, New York, pp 890.

Zimmerman, H. J., 1996: Fuzzy set theory and its applications, Kluwer Academic Publishers, Boston, pp. 435

7. ACKNOWLEDGEMENTS

We thank Brant Foote for providing funding through the National Center for Atmospheric Research, Research Applications Program opportunity funds. We also thank Corinne Morse and Inger Gallo for their editorial assistance.