Knowl Inf Syst (2007) 11(2): 137–154
DOI 10.1007/s10115-006-0026-6

**Knowledge and Information Systems**

REGULAR PAPER

**Sabyasachi Basu · Martin Meckesheimer**

# Automatic outlier detection for time series: an application to sensor data

Received: 12 April 2005 / Revised: 14 October 2005 / Accepted: 14 January 2006 /
Published online: 22 August 2006

**Abstract** In this article we consider the problem of detecting unusual values or outliers from time series data where the process by which the data are created is difficult to model. The main consideration is the fact that data closer in time are more correlated to each other than those farther apart. We propose two variations of a method that uses the median from a neighborhood of a data point and a threshold value to compare the difference between the median and the observed data value. Both variations of the method are fast and can be used for data streams that occur in quick succession such as sensor data on an airplane.

**Keywords** Time series · Outliers · Jaccard coefficient · Simulation · Sensor data

## 1 Introduction

Time series data are observations collected sequentially over time. Consider as examples the daily stock price for a firm, the number of defects per hour in a manufacturing plant, the hourly yield from a chemical process, the altitude of an airplane every second, and so on. In all these instances, a measure (stock price, number of defects, etc.) is associated with a time stamp (day, hour, minute, etc.) as it is collected. When observations are collected at frequent time intervals, large data sets in the form of time series are generated. The analysis of time series permits extracting information from the data and gaining insight into the dependence of observations. Efficient data processing and extraction of relevant information from these data are needed in many businesses. For example, in the area of financial services, large amounts of data are used to monitor credit card usage and

S. Basu (✉) · M. Meckesheimer
The Boeing Math Group, The Boeing Company, P.O. Box 3707, MS 7L-22,
Seattle, WA 98124-2207, USA
E-mail: {sabyasachi.basu, martin.meckesheimer}@boeing.com

detect fraudulent activity. In the manufacturing sector, data are used to detect machine failure and process shifts and trends. An important example from the automotive and aerospace industries is the use of data to detect abnormal operating conditions of a vehicle or an aircraft and improve safety.

In most applications, large amounts of real-time data from multiple sources need to be processed and synthesized efficiently to provide relevant information to engineers, researchers, accident investigators, operators, and many other users. This becomes complicated if the behavior of the series varies over time. However, for most time series, the values close together in time are more correlated with each other than those that are separated in time. It is important to take into consideration the fact that time series observations are often affected by unusual events or disturbances that create spurious effects in the series and result in extraordinary patterns. These unusual values (or outliers) in the data have adverse effects on understanding the properties of the time series. Identification of outliers is very important in many fields that deal with time series data since they can contain information that may lead to an intervention of a process and prevent failures or abnormal operating conditions. Thus, there is a need for effective and efficient methods for outlier detection in real time.

The processing of Flight Data Recorder (FDR) signals is a big challenge. The data are used to monitor aircraft systems. Information extracted from these data is used by maintenance engineers to prevent failures and by investigators to determine the causes of an accident. The data are obtained in the form of signals that are captured by hundreds of sensors with different time stamps at various sampling frequencies. These signals may be correlated and noisy; in addition, missing data points as well as unusual values are common. In order to supply engineers, pilots, and other users with efficient analysis tools, there is a need to extract high quality information from these data. For this purpose, the signal from the sensor must be preprocessed to obtain data in form of a time series (that is, a sequence of observations with a time stamp). Then, analysis and modeling strategies are applied to detect and treat "unusual" values, identify changes in phase, and analyze unaligned correlation. Figure 1 provides a global picture of the model building and model usage stages of the process. The ultimate goal is to extract a signature from a clean data set (model building) and use this signature for detection of anomalies of a process or system (model usage). Zhang et al. [5, 6] point out the need for information enhancement of low quality data for analysis.

In most cases, the time series needs to be cleaned to reduce noise, and impute missing values. The sequence of tasks involved in extracting information from a signal is illustrated in Fig. 2.

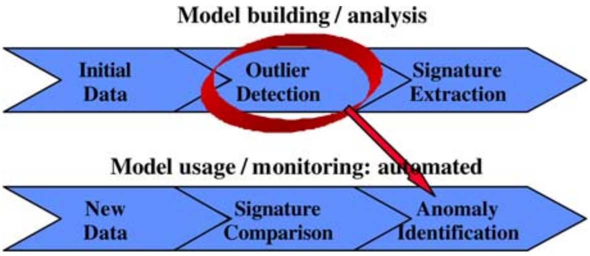

**Fig. 1** Model building and usage

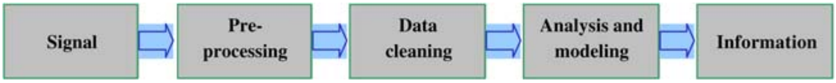

**Fig. 2** Process for extracting information from a signal

The focus of this article is to investigate the requirements for an efficient data cleaning method that will facilitate the analysis and modeling of signals obtained from multiple sensors.

The article is organized as follows. In Sect. 2, we provide some background on time series analysis and give an overview of the common methods for detecting outliers in time series. In Sect. 3, we describe an automated method for data cleaning that does not rely on a specific time series model. In Sect. 4, we discuss an experiment to evaluate the data cleaning method based on two FDR signals, and in Sect. 5 we draw conclusions from this study and point to further research ideas.

## 2 Time series analysis

### 2.1 Definitions

A *time series* is a set of observations collected sequentially in time. A time series of $N$ successive observations $y_1, y_2, \ldots, y_N$ can be regarded as a sample realization from an infinite population of such time series generated by a stochastic process, which can be stationary or nonstationary. The understanding of the structure and dependence of observations of a single (*univariate*) or several (*multivariate*) time series is achieved through *time series analysis*. The information obtained from time series analysis can be applied to forecasting, process control, outlier detection, and other applications, such as those mentioned in the introduction. In general, a univariate time series may be decomposed into two parts: the *signal* part, $f(y_{t-k}, \ldots, y_1)$, which is a function of the past values of the series, and the *noise* part, $a_t$, which is a sequence of independent and identically distributed (*iid*) variables. The basic general linear time series model can be written as

$$y_t = \sum_{k=1}^{\infty} \pi_k y_{t-k} + a_t,$$

where $k$ is the lag, or the distance between two observations, and $\pi$ are the weights of the model. For this basic model, a weakly stationary time series has the following properties:

(a) $E(y_t) = \mu$ is constant for all $t$,
(b) $\mathrm{Var}(y_t) = \sigma^2$ is constant for all $t$, and
(c) $\mathrm{Cov}(y_{t-k}, y_t) = \gamma_k$ depends only on the separation lag $k$ and not on $t$.

In words, these properties describe a time series that is not subject to significant changes due to different process phases and can therefore be modeled with the basic general linear time series model. For a detailed discourse in time series, please refer to Box et al. [1].

## 2.2 Outliers in time series

An *outlier* can be defined as a data point in a time series that is significantly different from the rest of the data points. Outliers are unusual observations that affect the analysis of the data and therefore must be treated with caution.

There are different types of outliers that can occur in a time series. An *additive outlier* is a measurement error at time $T$, $1 \leq T \leq N$, caused by factors outside the system. For example, a machine breakdown or human error when recording the data could result in an additive outlier; an additive outlier does not affect the trend of a process. Another type of outlier is the *innovative outlier*, which is caused by some change in a process or system. The main difference between an additive and an innovative outlier is that the latter indicates the beginning of a new trend in the process, which will eventually return to normal. Finally, *level shifts* imply a permanent change in the process mean (a change from stationary to non-stationary process). For example, a change of system state such as a change from climb to cruise of an aircraft. It is important to consider the possibility of different outlier types occurring at the same period. For an observed process, there may be one or more of these outliers present which makes the task of detecting them even more difficult.

The two main issues associated with outliers are detecting the outlier and deciding what to do once it has been detected. Outlier detection involves identifying the time of occurrence, which may not be known, as well as recognizing the type of outlier. Chang et al. [2] proposed an approach for outlier detection using likelihood ratios to analyze what, if any, will be the most likely type of outlier at each data point in the time series. Once an outlier has been detected, it can be removed by estimating a model in which the outlier is incorporated [4].

The main issue with traditional strategies for outlier detection is that they are model dependent. That is, once the type of outlier is known a model that incorporates the outlier can be estimated. This may not always be practical when dealing with large time series obtained from signals that record processes that are subject to frequent changes. In this case, the resulting time series may be highly nonstationary in which case a model-dependent implementation for outlier detection may not be applicable, since there is no standard model that will fit the entire time series. In addition, time series obtained from signals may be extremely noisy due to inaccurate sensor readings. Finally, it may be desirable to perform analysis of the signal in real time to monitor a process, detect changes, and make adjustments. Therefore, there may not be time to fit a model to incorporate outliers.

An automated approach for outlier detection in real time requires a strategy that relies on a generic method (model independence) and is computationally inexpensive (for real-time analysis). In the following sections we describe a method that can be applied to remove noise from data obtained from sensor data that are collected in quick succession. The method removes outliers (or signal noise, in this case) from the signal, which can then be used for more accurate real time process analysis.

## 3 Data cleaning

A method for cleaning data involves two aspects. The first aspect is identifying which data points in a time series are outliers (outlier detection). The second aspect addresses the issue of what to do with a data point that has been identified as an outlier. This can include data imputation which refers to the replacement of these identified outliers with reasonable values. It is important to take into consideration the fact that an outlier may have two interpretations: it can either be noise in the signal, or it can be an indication of an anomaly for a specific reason. In the present study, we are concerned with detecting outliers that are anomalous to its neighboring values regardless of the reason. Two variations of the method are presented in the following subsections.

### 3.1 Two-sided median method for cleaning noisy data

The proposed approach for cleaning noisy data obtained from a sensor signal is to use the median value of a neighborhood of data points to determine whether a particular data point is an outlier. Given a time series $y_1, y_2, \ldots, y_N$, define a neighborhood of points $\eta_t^{(\kappa)} = \{y_{t-\kappa}, \ldots, y_{t-1}, y_{t+1}, \ldots, y_{t+\kappa}\}$, where $2\kappa$ is the size of the neighborhood window, starting at $t-\kappa$ and ending at $t+\kappa$. In order to clean the data, compute the median in the neighborhood of points,$m_t^{(\kappa)}$, and compare the median to $y_t$. Then calculate the absolute value of the difference between $m_t^{(\kappa)}$ and $y_t$ and compare it to a specified threshold, $\tau$. If $|y_t - m_t^{(\kappa)}| < \tau$, keep $y_t$; if $|y_t - m_t^{(\kappa)}| \geq \tau$, then label $y_t$ as an outlier and replace $y_t$ with $m_t^{(\kappa)}$ to obtain a clean time series $y_1^*, y_2^*, \ldots, y_N^*$. The median method approach for data cleaning is illustrated with a simple example in Fig. 3 with a neighborhood window width of $\kappa = 3$. We will label this method as the *two-sided median method.*

The shaded area covering the neighborhood of data points indicates the threshold ($\pm\tau$). In this simplified example, $y_7$ is an outlier in the current neighborhood of points and would be replaced by the median value $m_7^{(3)}$ for analysis and

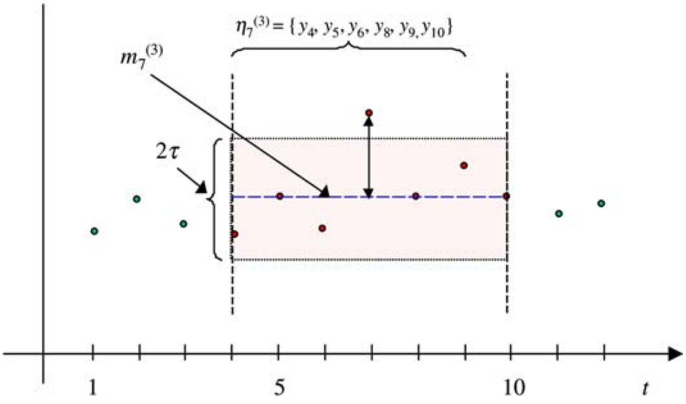

**Fig. 3** Median method for data cleaning

modeling of the time series. Note that Pearson [3] used a similar method but used the median from the whole dataset and not a local median to obtain the limits.

Note that using the median method with data from before and after a particular point in time delays detection of these outliers until $\kappa$ more data points are observed.

## 3.2 One-sided median method for cleaning noisy data

If we want to identify outliers with only past data, we can use a simple modification of the two-sided median method and look at the first difference of the observed series $y_t$, $z_t = y_t - y_{t-1}$. Define $\tilde{m}_t^{(y)} = \text{median}\{y_{t-2\kappa}, \ldots, y_{t-1}\}$ and $\tilde{m}_t^{(z)} = \text{median}\{z_{t-2\kappa}, \ldots, z_{t-1}\}$. Calculate $\tilde{m}_t^{(2\kappa)} = \tilde{m}_t^{(y)} + \kappa \cdot \tilde{m}_t^{(z)}$ as the predicted value for $y_t$, and compare it with $y_t$. If $|y_t - \tilde{m}_t^{(2\kappa)}| < \tilde{\tau}$, keep $y_t$; if $|y_t - \tilde{m}_t^{(2\kappa)}| \geq \tilde{\tau}$, then label $y_t$ as an outlier and replace $y_t$ with $\tilde{m}_t^{(2\kappa)}$. Note that, if there are $\kappa$ outliers in a row, it may mean that the process has changed at first occurrence of these consecutive outliers and needs to be investigated. We will label this method as the *one-sided median method*.

The data cleaning method can be used to detect sudden jumps in the time series. Note that the window width and threshold values are determined somewhat arbitrarily and may depend on the signal. In some cases, the window width and threshold values may vary for different parts of the signal. One might also use information from the actual signal or process to determine appropriate values for window width and threshold (for example, the percent deviation from the mean signal level). However, this may not always be possible, in particular when dealing with a non-stationary process. In Sect. 4, we discuss a computational study to gain more insight into this aspect of the data cleaning method.

## 3.3 Sample signals

In order to illustrate the data cleaning methodology we selected two different signals obtained from a flight data recorder (FDR): altitude of the aircraft, roll angle. The following paragraphs illustrate what the two signals look like before and after applying the data cleaning methods. There have been many studies on the analysis and interpretation of FDR data. A partial list of references is given at the end the article. However, these studies assume that the data are free from any nonsensical outlier, and the only anomalies are due to real phenomena. For all the data cleaning methods, we used $\kappa = 3$. The value of $\tau$ was determined based on engineering knowledge.

### *3.3.1 Altitude*

The signal for *altitude of the aircraft* (see Fig. 4) is obtained from a sensor that measures the altitude in feet. The recording begins at a time zero with the aircraft in a climb. We observe six landings and five takeoffs and five distinct cycles with takeoff, climb, cruise, approach and landing. Although the different phases of flight are clearly distinguishable, there is a considerable amount of noise and

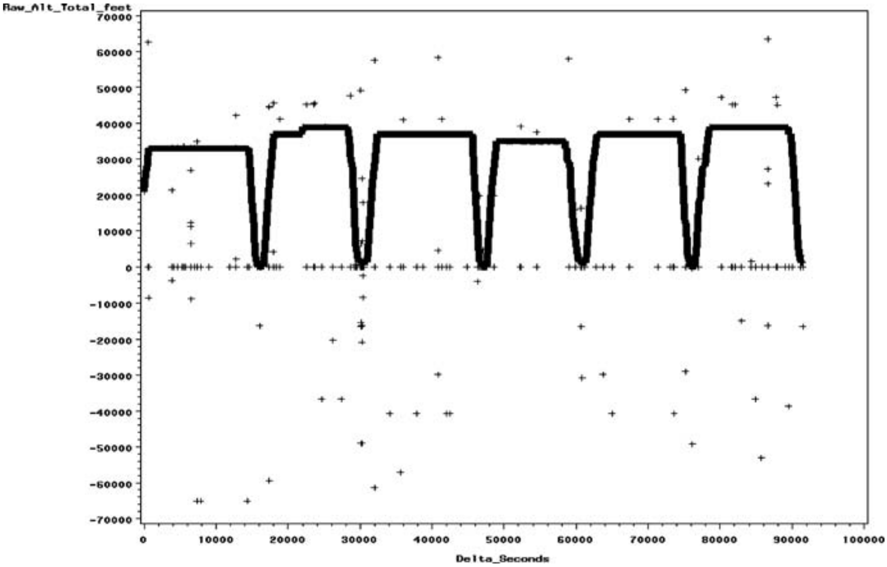

**Fig. 4** Raw signal data for total altitude

erroneous data present in the signal, which may be due to inaccurate sensor readings.

To apply the data cleaning method using both median methods, we used $\kappa = 3$ and $\tau = 25$. Results for the two-sided and one-sided methods are shown in Fig. 5.

The altitude signal is an example where one might use different window width and threshold values for different parts of the signal, corresponding to different flight stages. For instance, during take-off or landing altitude changes more rapidly than during cruise. Also, during cruise a change in altitude similar to one that is normally observed during take-off and landing, may indicate an outlier. Therefore, a larger threshold value is more appropriate for take-off and landing.

### *3.3.2 Roll angle*

The signal for *roll angle* (see Fig. 6) is obtained from a sensor that records the movement of an aircraft about its longitudinal axis.

In straight and level flight, the roll angle is zero. Although the signal for roll angle is quite different from the altitude signal we can still identify the same cycles as before, since more variation in the signal can be observed during initial climb and final approach for each of the five distinct cycles, as the pilot makes more adjustments immediately after takeoff and before landing. Note that there are several data points that are clearly outliers, since roll angles of ±100° are not realistic. Figure 7 shows the clean signal for roll angle using $\kappa = 3$ and $\tau = 5$ using the two-sided and one-sided median methods. Again, the resulting (cleaner) signal captures the changes in the roll angle and deletes the outliers using both methods.

Note that both methods generally perform quite well in detecting outlying values. The choice of method may be determined by the application at hand. If we

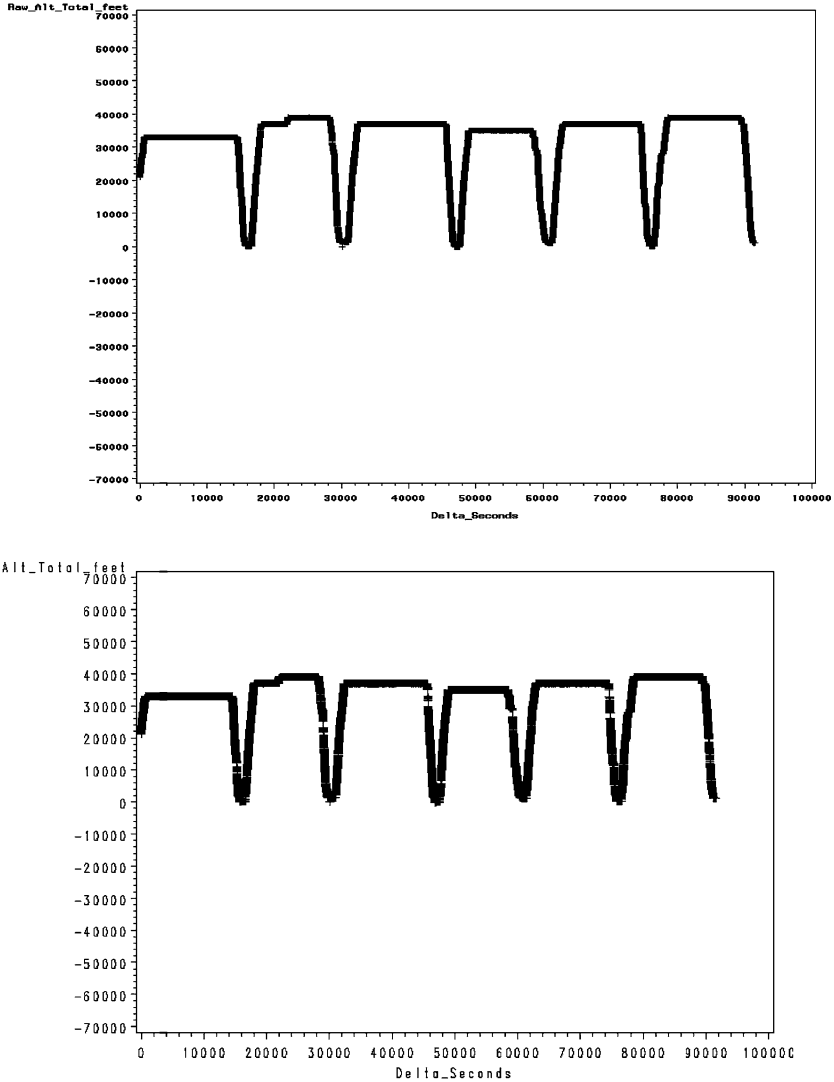

**Fig. 5** Cleaned data for altitude: two-sided method (*top*) and one-sided method (*bottom*)

want to perform real time detection, then the one sided method may be more appropriate. On the other hand, if the analysis is performed at a later time, then the two sided method may be more appropriate.

## 4 Data cleaning experiment

The two FDR signals shown in the previous section illustrate that the data cleaning methods are promising, and can successfully remove a considerable amount of noise from the signal. However, we also note that there are several factors that

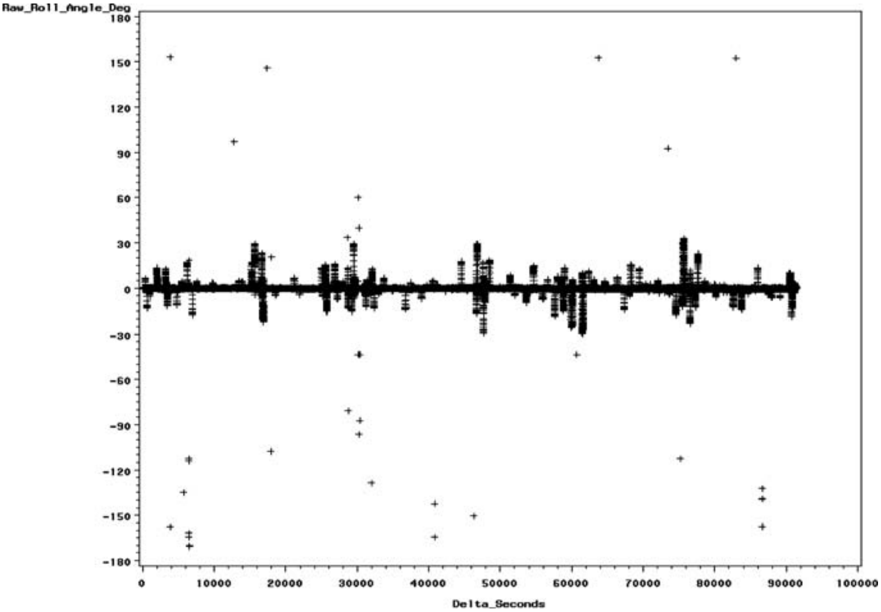

**Fig. 6** Raw signal for roll angle

must be taken into account when applying the proposed data cleaning methods to these signals. As we observed, data from multiple sensors can look very different and requires the data cleaning methods to be flexible and adaptable. In addition, some valid data points may have been detected as outliers and imputed in error. The objective of the computational experiment described in this section is to test the effectiveness of the proposed data cleaning methods and identify potential areas of improvement. For this experiment, we use the two signals introduced in the previous section. For the experiment, we will only discuss the method where we look at both sides of a point in time to determine outlying values.

The design variables for the computational experiments are:

(a) The *window width* ($\kappa$) of the median method, which controls how many neighboring points are included in the calculation of the median. The larger the window width, the more points are included to compute the median. We vary $\kappa$ from 3 to 31 in increments of two; that is, $\kappa = \{3, 5, \ldots, 31\}$.

(b) The *threshold* ($\tau$) for classifying a data point in the signal as an outlier. The threshold is used to compare the median value with the actual data point at time $t$, that is, $|m_t^{(\kappa)} - y_t|$. If the difference between the two is larger than the threshold, the data point is considered an outlier. The appropriate threshold value varies from one signal to another and requires knowledge about the signal. We vary the $\tau$ for the two signals as follows:

- Total altitude: from 10 to 200 ft in increments of 10 ft; i.e., $\tau_{\text{alt}} = \{10, 20, \ldots, 200\}$.
- Roll angle: from 1 to 10° in increments of 1°; i.e., $\tau_{\text{roll}} = \{1, 2, \ldots, 10\}$.

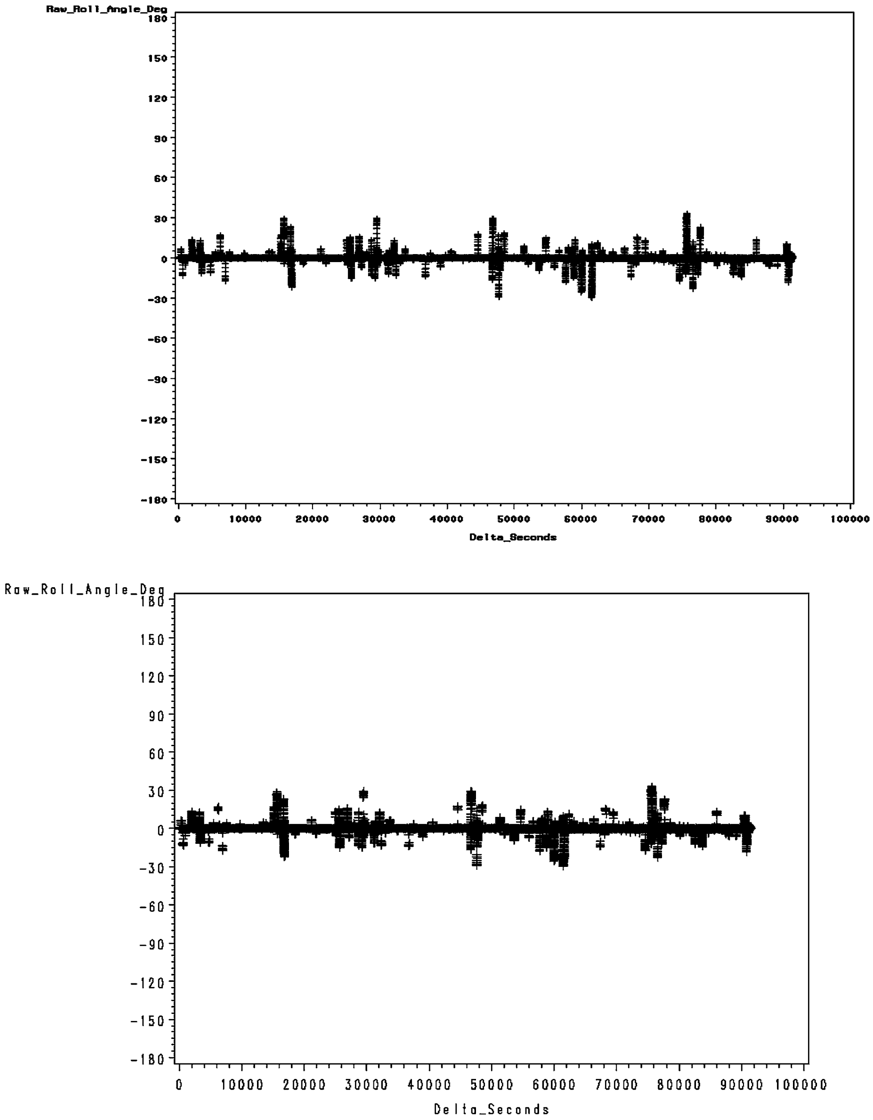

**Fig. 7** Cleaned data for roll angle: two-sided method (*top*) and one-sided method (*bottom*)

In addition, it will be necessary to define what we mean by outlier for each of the two signals:

- Total altitude: any value outside a range from 0 to 40,000 ft.
- Roll angle: any value outside a range from −45 to +45°.

For the purpose of this computational experiment, we consider the truth as being non-outlier if it is within the range and an outlier if it is outside the range. This is a simplification for the purpose of the experiment and we understand that we may mislabel some of the values. For example, in the altitude data, if the value

**Table 1** Assessment of two-sided median method

| | Data cleaning method | |
|---|---|---|
| | Not outlier | Outlier |
| Truth | | |
| Not an outlier | *Clean data points* ($A$) Data points that are not outliers and are not identified as outliers by the data cleaning method. | *False positives* ($B$) Data points that are not outliers, but are identified as outliers by the data cleaning method. |
| Outlier | *False negatives* ($C$) Data points that are outliers, but are not identified as outliers by the data cleaning method. | *Outliers* ($D$) Data points that are outliers and are identified as outliers by the data cleaning method. |

at a given time is 1,000 ft and at the next time point which is a second later is 30,000 ft, then one of them is obviously an outlier. But, in the above designation, none of them will be designated as an outlier. This may lead to more false positives by identifying an outlier, when the "truth" as defined above will not call the value an outlier.

In order to assess the effectiveness of the two-sided median method, under the above assumptions, we can classify the results from our experiment into four categories (see Table 1).

The categories in Table 1 correspond to the four possible outcomes of one experimental run, which consists of using the two-sided median method with a particular combination of window width ($\kappa$) and threshold value ($\tau$). Category $A$ is equivalent to an ideal situation in which a signal is free of noise and/or erroneous data. Categories $B$ and $C$ are undesirable, because the two-sided median method is not able to distinguish between noise and signal.

Our interest is in maximizing the number of times the two-sided median method correctly identifies outliers (Category $D$) and minimizing the number of points false positives (Category $B$) and false negatives (Category $C$). To achieve this, Jaccard's coefficient can be used to assess the performance of the two-sided median method by counting the number of data points in each category. The measure of goodness is computed using the following expression:

$$J_{(\kappa,\tau)} = \frac{D_{(\kappa,\tau)}}{B_{(\kappa,\tau)} + C_{(\kappa,\tau)} + D_{(\kappa,\tau)}} = \frac{1}{1 + \frac{B_{(\kappa+\tau)}+C_{(\kappa+\tau)}}{D_{(\kappa+\tau)}}}.$$

We include only categories $B$, $C$, and $D$, since the clean data points ($A$) outnumber the other categories by a large factor, and this would not allow us to observe any differences clearly. Note that the Jaccard ($J$) coefficient is inversely proportional to the false detection and directly proportional to the correct detection of the outliers. However, it puts equal cost for both false negatives and false positives. For this experiment, the larger the coefficient, the higher the effectiveness of the two-sided median method, because we want to both maximize the number of correctly identified outliers and minimize the number of false positive and false negatives. Note that because of the definition of an outlier for this experiment, we may have larger number of observation in category $B$, thus underestimating the Jaccard coefficient.

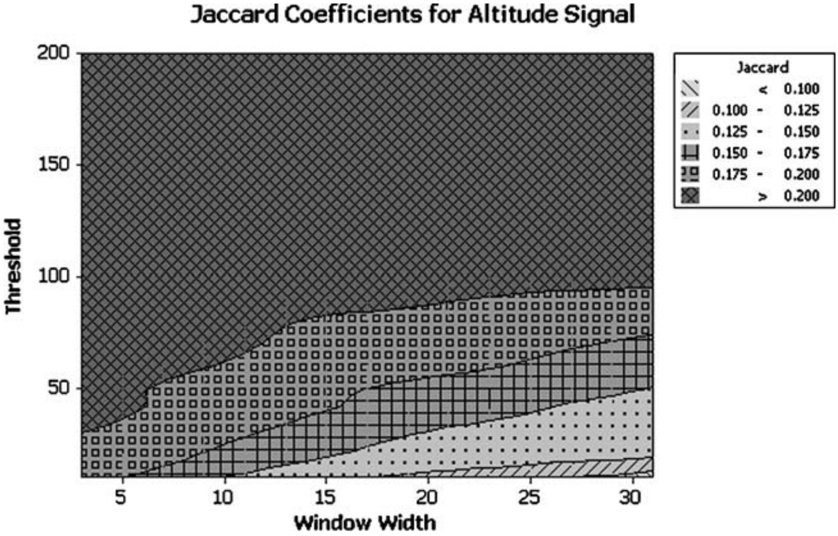

**Fig. 8** Contours of the Jaccard coefficients for the altitude signal

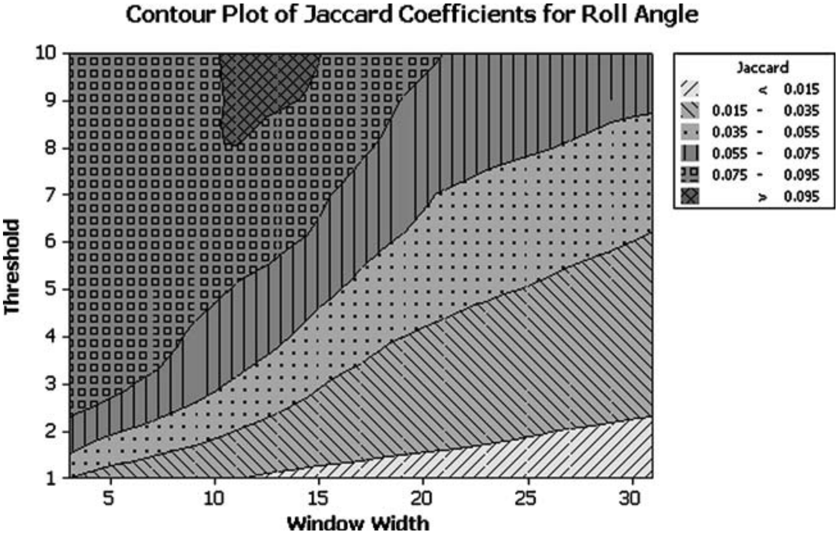

**Fig. 9** Contours of the Jaccard coefficients for the roll angle signal

Figures 8 and 9 show contour plots of the Jaccard coefficients obtained for our computational experiments, where coefficient $J_{(\kappa,\tau)}$ corresponds to one combination of window width ($\kappa$) and threshold value ($\tau$).

The general observation that a smaller window size results in higher Jaccard coefficients for a wider range of threshold values can be made for all two signals. For example, in Fig. 8 we observe that high Jaccard coefficients are obtained for a large range of values. However, the general trend indicates that the narrower the window width (say, $\kappa = \{3, 5\}$) the wider the range of threshold values that still result in high coefficient values. This is useful when the threshold value for the signal is not well known.

Finally, we noted that during the experiments we did not observe any false negatives for the roll signal. The main reason for observing false negative counts for the altitude signal was that there were long series of negative altitudes values with the difference of consecutive altitude values being within the threshold. The two-sided median method may correctly identify the first few outliers (at most $2\kappa$), but then may not be able to detect any differences after that. Since there are fewer (or no) consecutive outliers in the other three signals, there are fewer (or no) false negative counts. In general, we expect that:

– Reducing the threshold value reduces the number of false negatives (that is, outliers that were not identified by the method) and increase the number of false positive. In other words, more data points will be identified as outliers.
– Reducing the window width ($\kappa$) will lead to more false negatives and less false positives.

## 5 Simulation experiment

To understand the effect of the degree of correlation and the effect of window size and threshold values, we performed a small simulation experiment. For the simulation, we generated time series from an AR(1) model [1]

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

where $\phi$ is the first order autocorrelation parameter and $\varepsilon_t$ are iid with mean 0 and variance $\sigma_\varepsilon^2$. The $Y_t$ is stationary with mean 0 and variance $\sigma_\varepsilon^2/(1-\phi^2)$ and $\mathrm{Cor}(Y_t, Y_{t-k}) = \phi^k$. If $\phi$ is zero, then $Y$'s are uncorrelated.

For our simulation experiment, we only consider $\phi > 0$, because it reflects the case where nearby values are similar, a typical situation we observe in real data. The different parameters that we varied in the simulation experiment are:

- Value of $\phi$ which ranged from 0.0 (0.1) 0.9
- Threshold value ($\tau$) from 2.0 (0.5) 6.0
- Window width ($\kappa$) from 3 (1) 6.

Other parameters that we did not vary were:

- Value of $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2 = 1$
- Length of the time series = 1000
- Number of Outliers: 30 groups of 2 consecutive outlying values
- Number of Simulations for each combination: 100

The outliers were generated for the 60 values using the following rule:

- Generate a time series of length 1000 using the appropriate value of $\phi$
- Generate a random number, $x$, between $-0.5$ and 0.5 by generating a random number between 0 and 1 and then subtracting 0.5
- Calculate $x_1 = \exp(3 \times \mathrm{abs}(x)) + 3$

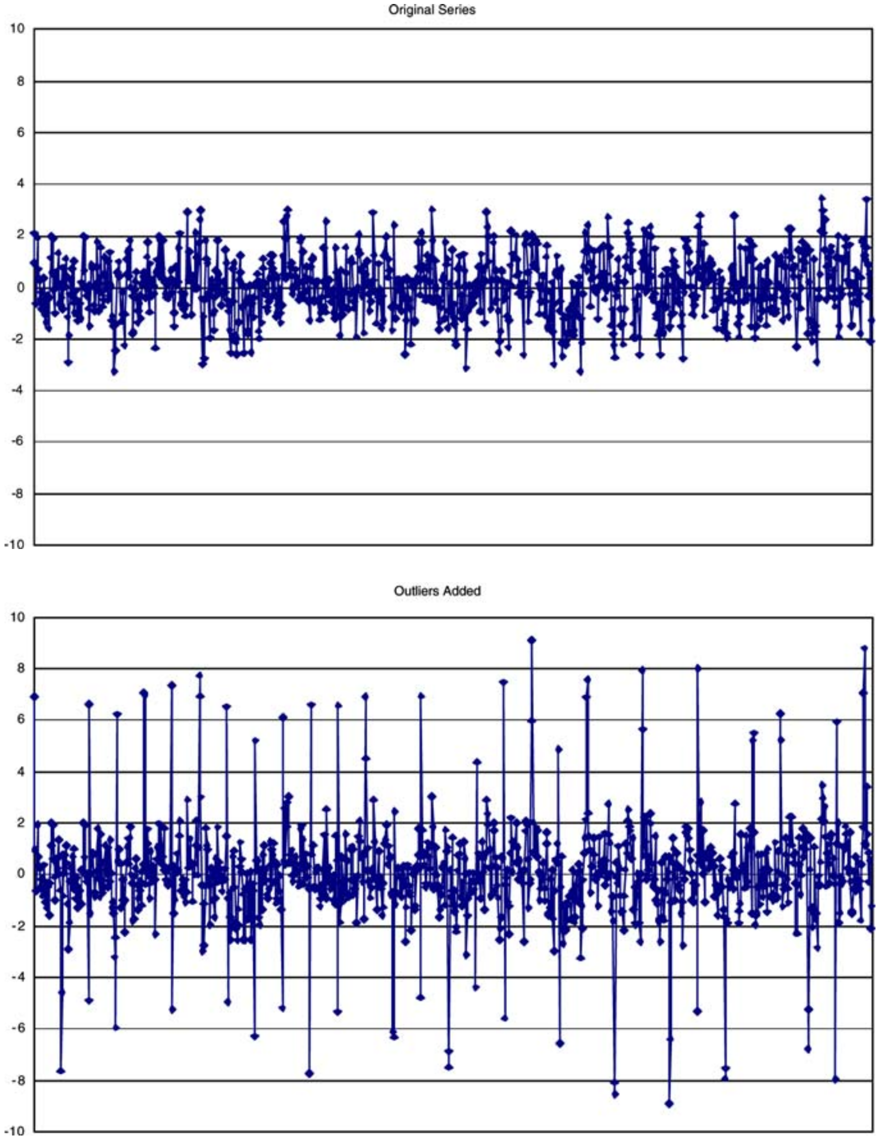

**Fig. 10** **a** Time series plot of one of the generated series with $\phi = 0.5$. **b** Time series plot for the series in (a), with 60 outliers added

- If $\text{mod}(t, 33) = 0$ or 1, then modify $y_t$ as follows: $y_t = y_t + \text{sign}(y_t)\, x_1 \sigma_\varepsilon$, where mod is the modulo operator and $\text{sign}(x) = \text{sign of } x$. Note that we use 30 in the mod operator because it generated the required number of outliers.

Figure 10a, and b gives the time series plot of one of the series generated for the simulation. Figure 10a shows the original time series and Fig. 10b shows how the series looks after the 60 outliers were added.

Figure 11 shows the comparison of Jaccard coefficients for different values of window size and cutoff value for $\phi = 0.5$. Note that the values do not change

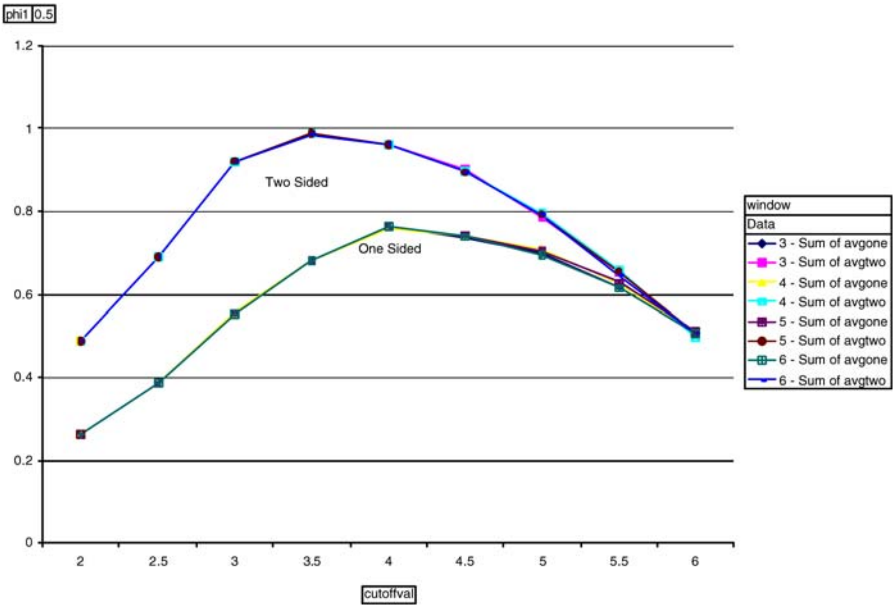

**Fig. 11** Jaccard coefficients for one sided and two sided methods for $\phi = 0.5$

for different window sizes and it is similar for all values of $\phi$. Therefore, results from different window sizes will be combined hereafter. Note that the Jaccard coefficients for the two sided method are consistently higher than for the one-sided method. The coefficients increase at first as the window size increases, but then the coefficients decrease after a certain window width.

Figure 12 shows the contour plots of the average Jaccard coefficients for the one and two sided methods. The averages were computed over 100 simulation runs. Note that the Jaccard coefficients are generally higher for the two-sided method, except for very large values of $\phi$ and cutoff value ($\tau$). In general, a medium value of $\tau$, around 4, leads to the highest values of the Jaccard coefficient for both the methods. This leads to the conclusion that the proper choice of the cutoff value is critical for a balance of false positives and false negatives, the contribution of both are reflected in the Jaccard coefficient.

## 6 Summary and conclusions

The objective of this article was to investigate efficient data cleaning methods that will facilitate the analysis and modeling of signals obtained from multiple sensors. Data cleaning is especially important for the analysis of sensor data, which can contain a significant amount of noise. We proposed one-sided and two-sided median methods that provide fast, automated, and model independent outlier detection and can be used for signal data cleaning. We then performed a computational study using two FDR signals to understand the effects of threshold and window width parameter values that are required by the method. We also looked at a small simulation study from an AR(1) model with different autocorrelation structure,

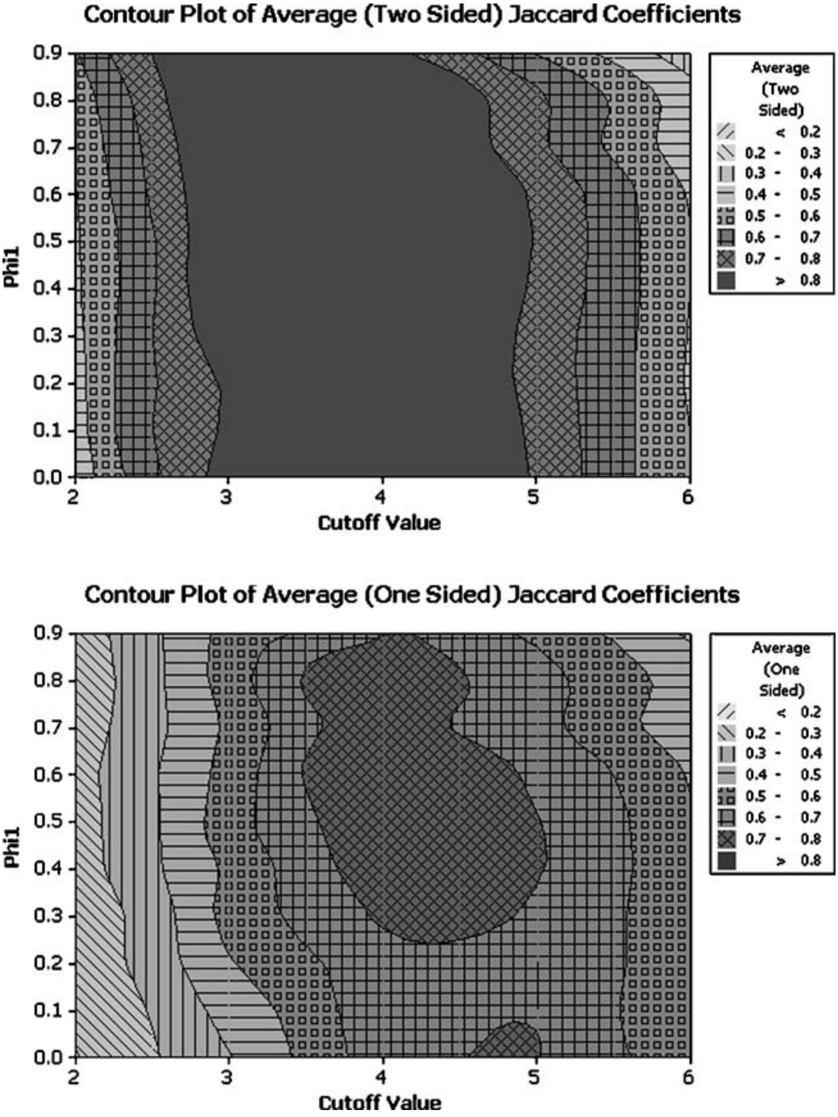

**Fig. 12** Average Jaccard coefficients: two-sided (*top*) and one-sided (*bottom*) methods

window widths and threshold values. The results indicate that the proposed data cleaning methods are promising but offer room for improvement. In particular, the data cleaning methods need to be improved to account for situations in which a series of outliers occur in the data. When this happens, the current implementation detects the first few outliers correctly, but cannot detect a string of outliers longer than the window width, since it confuses the sequence of outliers of same value from the actual signal. Furthermore, it is important to keep in mind that removing an outlier may not always be a good choice, since the outlier could provide important information about the process or system. These methods can be used to identify unusual values and the expert can decide whether the detected unusual value is an outlier or an anomaly that is a real phenomena. Finally, we observed that the proposed methods are sensitive to the selection of an appropriate threshold value; it is therefore required to have some knowledge about the signal. Based on our computational experiments, we recommend using a small window width to maximize the efficiency of the proposed data cleaning methods, especially for cases in which there is little information about how to set the threshold value.

**Acknowledgements** The work was funded by the Data Mining Core Technology program within The Boeing Company. We would like to thank Rod Tjoelker for supporting this research. We would also like to thank the editor and the referees for helpful comments.

## References

1. Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control, 3rd edn. Prentice-Hall, Englewood Cliffs, NJ
2. Chang I, Tiao GC, Chen C (1988) Estimation of time series parameters in the presence of outliers. Technometrics 30:193–204
3. Pearson RK (2002) Data mining in the face of contaminated and incomplete records. In: Second SIAM conference on data mining, Arlington, VA
4. Peña D (2001) Outliers, influential observations, and missing data. In: Peña D, Tiao GC, Tsay RS (eds) A course in time series analysis. Wiley, New York, pp 136–170
5. Zhang S, Zhang C, Yang Q (2003) Data preparation for data mining. Appl Artif Intell 17:375–382
6. Zhang S, Zhang C, Yang Q (2004) Information enhancement for data mining. IEEE Intell Syst March/April 12–13

**Martin Meckesheimer** has been a member of the Applied Statistics Group at Phantom Works, Boeing since 2001. He received a Bachelor of Science Degree in Industrial Engineering from the University of Pittsburgh in 1997, and a Master's Degree in Industrial and Systems Engineering from Ecole Centrale Paris in 1999. Martin earned a Doctorate in Industrial Engineering from The Pennsylvania State University in August 2001, as a student of Professor Russell R. Barton and Dr. Timothy W. Simpson. His primary research interests are in the areas of design of experiments and surrogate modeling.

**Sabyasachi Basu** received his Ph.D. in Statistics from the University of Wisconsin at Madison in 1990. Since his Ph.D., he has worked in both academia and in industry. He has taught and guided Ph.D. students in the Department of Statistics at the Southern Methodist University. He has also worked as a senior marketing statistician at the J. C. Penney Company. Dr. Basu is also an American Society of Quality certified Six Sigma Black Belt. He is currently an Associate Technical Fellow in Statistics and Data Mining at the Boeing Company In this capacity, he works as a researcher and a technical consultant within Boeing for data mining, statistics, and process improvements. He has published more than 20 papers and technical reports. He has also served as journal referee for several journals, organized conferences, and been invited to present at conferences.