

# Irregularity analysis in time series data

Siu-Tong Au<sup>†</sup>, Rong Duan<sup>†</sup>, Wei Jiang<sup>‡</sup>

<sup>†</sup>*at&t Research Labs  
Florham Park, NJ* \*

<sup>‡</sup>*Department of Systems Engineering and Engineering Management,  
Stevens Institute of Technology, Hoboken, NJ*

September 2, 2006

Government and corporations nowadays collect time series data at lowest possible details such as by locations, parts, products, or even individuals. Most of data cleaning methods assume one known type of irregularity. This paper provide a framework for the situation that there are multiple irregularities hiding in large volumes of cross sectional time series and develops a data mining platform to capture these key irregularities one by one based on their importance. It attempts to automate how a data analyst looking at time series graphs when cleaning the data (but there are too many to look at). Clustering is applied to group time series with similar pattern, and the principal irregular component of the dominated time series group is extracted and adjusted. The platform continues to cluster, extract and adjust the next significant irregular components iteratively. Finally all these significant irregular components are summarized in graphic forms to help analysts to know the data better and faster before any analysis and modeling.

*Key words:* Change Points; Data Quality; Decomposition; Outliers; Regression Models

## 1 introduction

Government and corporations nowadays collect time series data at lowest possible details by such as locations, parts, products, or even individuals. However, advanced automatic data collection systems also generate noisy data due to many technical and social reasons. Current data quality problems cost U.S. businesses more than \$600 billion per year. Companies typically have problems with the quality of information that serves as the very foundation of their primary business applications. The direct costs of poor quality information have been estimated at from 10 to more than 20 percent of an organization's operating revenue. Beyond direct costs, poor quality information also creates tremendous opportunity losses. For example, inaccurate or inconsistent data can hinder a company's ability to understand its current-and future-business problems. This leads to poor decisions that can cause a host

---

\*©Copyright at&t Knowledge Ventures 2006. All rights reserved

of negative results, including lost profits, operational delays, customer dissatisfaction, internal systems failures, inaccurate business forecasts, ineffective sales and marketing efforts and much more.

Although the most successful manufacturers have rigorous and defined quality assurance and improvement programs that help them build better products faster and cheaper, many business companies typically have not taken the same approach to improving the quality of its data production and maintenance. Ballou [1] has identified four dimensions of data quality: accuracy, completeness, consistency and timeliness. The key of data quality is “fitness of use”, which the unfit data is referred to as “data irregularities” in this paper, and is critical in data analysis since poor quality data can easily bias the analysis if treated improperly.

This paper provide a generic framework for mining data irregularities from business users’ point of view. While visualization is an exceptional tools in finding the pattern of irregularities, most of data cleaning methods assume one known type of irregularity [2], [3], [4]. Here we are interested in the situation that there are multiple irregularities hiding in large volumes of cross sectional time series. The aim is to develops a data mining platform to capture these key irregularities one by one based on their importance. It attempts to automate how a data analyst looking at time series graphs when cleaning the data (but there are too many to look at). Clustering is applied to group time series with similar pattern, and the principal irregular component of the dominated time series group is extracted and adjusted. The platform continues to cluster, extract and adjust the next significant irregular components iteratively. Finally all these significant irregular components are summarized in graphic forms to help analysts to know the data better and faster before any analysis and modeling.

This paper is organized as follows. In section 2, we introduce the general framework and fundamental concept of the approach. In section 3, we show the experimental result with simulation data. Conclusion is in section 4.

## 2 General Framework

Traditionally, time series  $Y_i(t)$  are decomposed to trend component  $T_i(t)$ , seasonal component  $S_i(t)$ , cyclical component  $C_i(t)$  and irregular component  $I_i(t)$  [5].  $i$  is the time series index and  $t$  is the time point index.  $Y_i(t) = T_i(t) + S_i(t) + C_i(t) + I_i(t)$ . Trend  $T$  is a long term movement in a time series. It is the underlying direction. The seasonal component  $S$ , often referred to as seasonality, is the component of variation in a time series which is dependent on the time of year. It describes any regular fluctuations with a period of less than one year consist of a systematic pattern. Cyclical component  $C$  is cyclical fluctuation around the trend line, caused by irregular movements as in a typical business cycle of expansion and construction. The difference between a cyclical and a seasonal component is that the latter occurs at regular (seasonal) intervals, while cyclical factors have usually a longer duration that varies from cycle to cycle. Irregular components  $I$  is defined as left over when the other components of the series (trend, seasonal and cyclical) have been accounted for. In this paper we further decompose irregular component to outlier component  $O$ , change point component  $Ch$  and noise  $N$ . Outlier  $O$  is extreme fluctuations due to rare events, they are the points far from all other points. Change point  $Ch$  is a point where the structure of the data changed afterward. Noise  $N$  is defined as the piece which is left over

after all other component are account for.

For all different types of components, we generate a base component pool. Just like for trend , we can have linear trend  $Y_i(t) = \beta_1 x_i(t) + \beta_0$ , where  $\beta_0$  is intercept and  $\beta_1$  is slope. Quadratic trend  $Y_i(t) = \beta_2 x_i(t)^2 + \beta_1 x_i(t) + \beta_0$ , log-linear trend  $\ln(Y_i(t)) = \beta_1 x_i(t) + \beta_0$  , etc. For Outlier components, we can have  $Y_i(t) = a_t x_i(t)$ ,  $a_t$  is a binary code and for most of  $t$ ,  $a_t = 0$ , which means a one outlier component when only one  $a_t = 1$  and two outlier component when two  $a_t = 1$  and so on. Similar for change point components. we can have one change point component and two or more change point component.

In this paper, we propose a interactive graphical visualization method to group irregularities by its significance. For a large data set, we decompose the time series one by one and visualize the common irregular components for the whole set of data. Plot the data first and use the information retained from graphical method [6] to add the component by its significance and iteratively plot and measure the improvement till all the major irregularities have been modeled. The model improvement can be measured by the overall Goodness of Fit Statistics. For one time series  $Y_i$ , the current model is  $\hat{Y}_i$  and the candidate components are  $CC_1, CC_2, \dots, CC_n$ , using Sum of Squares Due to Error ( $SSE_i$ ) as Goodness of Fit Statistics here. This statistic measures the total deviation of the response values from the fit to the response values. The smaller the better. If all the components are considered,  $SSE_i$  should be close to 0.

$$\hat{Y}_i = CC_1 + CC_2 + \dots + CC_{n-1} + CC_n \quad (1)$$

$$SSE(Y_i) = \sum (Y_i - \hat{Y}_i)^2 \quad (2)$$

$$(3)$$

The global  $SSE$  is the sum of  $SSE(Y_i)$  for all time series.  $SSE = \sum_1^N (SSE(Y_i))$ .  $SSE$  will be measured each time adding a new component. The whole process will be stopped after  $SSE$  less than a threshold  $T$ .

First, calculate the  $SST_i$ , the sum of squares of treatment.  $SST_i = \sum (Y_i - \bar{Y}_i)^2$  of each time series.  $\bar{Y}_i$  is the mean of the time series  $Y_i$ . Plot the top  $N$  series with the largest  $SST(Y_i)$  from the whole data set and visualize the main common component which causes the variance, add that component  $CC_i$ .  $\hat{Y} = CC_i$ , and use  $SSE(Y_i)$  to measure the improvement. Label those series with top  $SSE(Y_i) < T$  as one class that have component  $CC_i$  as the most significant component. Calculate the global  $SSE$  with component  $CC_i$  build in model  $SSE(CC_i)$ . Plot the top  $N$  series from the data that is left have the largest  $SSE(CC_i)Y(i)$  and visualize the common component. Repeat this process till the global  $SSE$  is smaller than a threshold. This interactive visualization method give data analyst more control on the data and the unknown, hidden components might be inspected. After adding all components, visualization can tell if noise is left or some new components which have to be considered. The detailed steps are as follows:

- Define the candidate component pool and decompose the time series to component  $CC_1, CC_2, \dots, CC_n$
- Set a threshold  $T$
- For each time series, calculate  $SST(Y_i)$  and sort the time series based on the  $SST(Y_i)$  and plot the top  $N$  series

- Pick the first variation component  $CC_k$  based on the 1 page of the top series
- Add the key variation component for those unclassified time series and calculate the Sum of Square of Errors ( $SSE(Y_i)$ ). Sort the time series based on the  $SSE(Y_i)$
- Label the series as one class which have  $CC_k$  as a major component when  $SSE(Y_i) < T$
- Calculate the global  $SSE$  and plot  $N$  time series which have the largest  $SSE_{CC_i}(Y_i)$  and visualize the major common irregularities
- Repeat last three steps till  $SSE < 0.1 \times SST$

### 3 Experiment

In this section, we apply graphical visualization method on simulation data. There are 400 time series. 100 are linear trend plus Gaussian noise, 100 are linear trend plus Gaussian noise plus outlier, 100 are step function plus noise and the last 100 are step function plus noise plus outlier. The samples are like in Fig 1a. The candidate pool can be linear trend, 1-outlier, 2-outlier, 1-change point, 2-change point, noise.

First, decompose the time series.

- Trend: median filter to smooth the data and linear regression to get the linear trend
- Change point: define a base sequence  $B$  for series  $Y_i$ ,  $B = [111-1-1-1] \times median(Y_i)$ , use sliding window to calculate the correlation between  $B$  and  $Y_i$  and order the correlation coefficient. Pick the largest point as the 1-change point component and top 2 as the 2-change point component
- Outlier: detect outlier by 3-Sigma rule and order the outliers. Pick the largest point as the 1-outlier component and top 2 as the 2-outlier component
- Noise: the left over after trend, 1-change point component, 2-change point 1-outlier and 2-outlier

Set the threshold  $T = 0.5$ , Calculate and sort  $SST(Y_i)$  for each time sequence. Plot the top 50 to check which component contribute the most to the error. For the simulation data, as shown in Fig 1b, the change point is the most significant component. Add the change point and calculate  $SSE_{CC_i}(Y_i)$ , label  $Y_i \in class_{CC_i}$  if  $SSE_{CC_i}(Y_i) < T$  as shown in Fig 1c, those  $\hat{y}_i$  are the series with step function as a major component. Calculate the global  $SSE_{CC_1}$  after add component  $CC_1$ . Continue to check the top 50 with largest  $SSE_{CC_1}(Y_j)$ , as in Fig 1d, the most significant irregular component is outlier now. We add one outlier and calculate  $SSE_{CC_1, CC_2}(Y_j)$ , class  $Y_j \in class_{CC_1, CC_2}$  if  $SSE_{CC_1, CC_2}(Y_j) < T$  as a new class, as shown in Fig 1e. The process can be repeated till the global  $SSE < 0.1 \times SST$ . Which means 90% variance have been extracted and classified. The component which is left is almost noise, as shown in Fig 1f. Fig 2 shows the lift chart of global  $SSE$  when components are added gradually.

## 4 Conclusion

The quality of information that serves as the cornerstone of virtually every critical business process—customer intelligence, billing, accounting, inventory management, product development, marketing, sales, logistics—is typically an “unknown” in most organizations.

In this paper, we introduce a framework to decompose a time series into different components and iteratively group the components by its significance. An interactive graphical visualized method is proposed to evaluate the significant component. By using this approach, Data analyst will have more control of the process and the hidden components can be inspected graphically.

## References

- [1] D. P. Ballou and H. L. Pazer, “Modeling data and process quality in multi-input, multi-output information systems,” *Management Science*, vol. 2, pp. 150–162, 1985.
- [2] D. W. Scott, “Outlier detection and clustering by partial mixture modeling,” *COMPSTAT*, 2004.
- [3] D. Pena, “A new statistic for influence in linear regression,” *Technometrics*, vol. 47, p. 1, 2005.
- [4] R. D. Cook, “Detection of influential observations in linear regression,” *Technometrics*, vol. 19, pp. 15–18, 1977.
- [5] W. S. Cleveland, *Seasonal and Calendar Adjustment*. New York: North-Holland, 1983.
- [6] W. S. Cleveland, *The elements of Graphing Data*. California: Wadsworth, Inc, 1985.

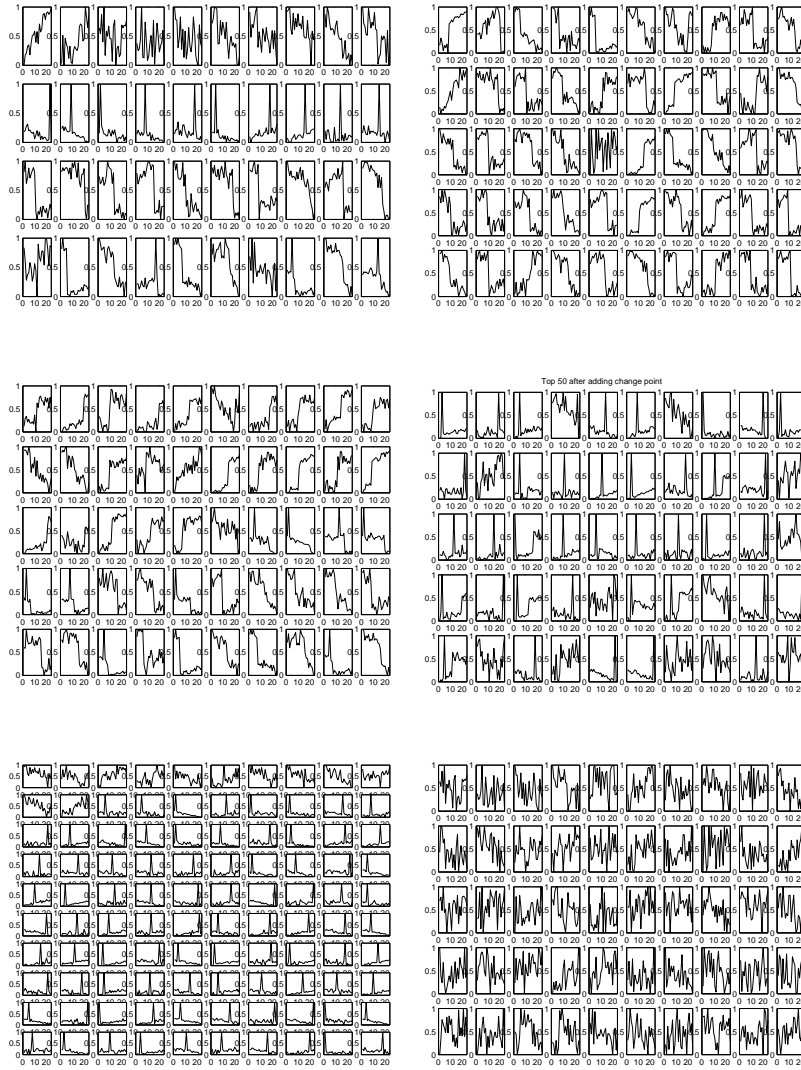


Figure 1: Simulation data: (a): sample data for 4groups (b):Top 50 the most has the largest  $SST$  (c):Sample series in change point class (d):Top 50 the most significant irregularity after adding change point.  $SST$  (e):Sample series in 1 outlier class. (f):The left unclassified series.

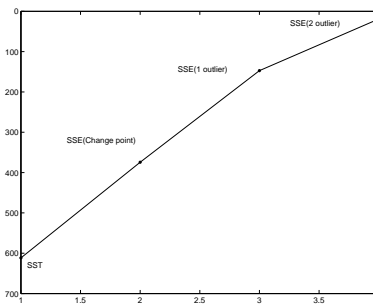


Figure 2: Lift chart after component added one by one