The Math Citadel

Some New Thoughts on Anomaly Detection

Rachel Traylor, Ph.D.
Co-founder/Chief Scientist
www.themathcitadel.com

Content

Introduction

- Definitions and Common Methods

Overview of Formal Time Series Analysis

- What's a Time Series?
- Outliers
- Types of Interventions

Fixing the issues

Why do you care?

What's an anomaly?

- Something "not normal"
- Something "too different"
- Something outside an acceptable variance or tolerance
- $\vdots$

Notice all of these are *relative* definitions

Common Methods and Examples

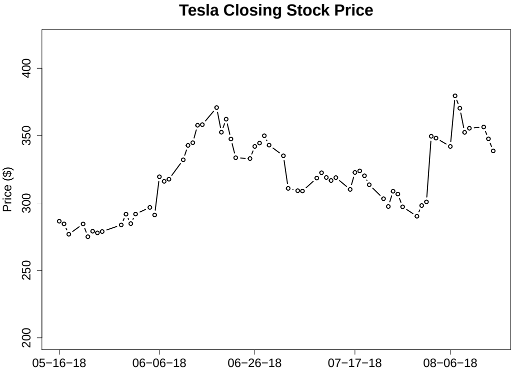
- Simplest (Etsy): 3σ away from the mean
- Netflix (Robust Anomaly Detection): Robust Principal Component Analysis that decomposes a data matrix into "background" and "anomalies"
- Twitter: Seasonal Hybrid ESD builds changes the metrics slightly of the Extreme Studentized Deviate test to include windowing
- Google Analytics: Bayesian model that requires 90 day training period for daily outliers, or 32 weeks for weekly outlier detection

All of these are essentially based on the same philosophy. They're uninterpretable (Netflix), inappropriate for time series (Etsy and Twitter), and slow (Google Analytics)

What is a time series?

Time Series

- a set of observations x_t recorded at specific times t
- Statistics: a realization of some random process X_t



What is a Time Series?

Classical Decomposition Model

$$X_t = m_t + s_t + Y_t$$

where

- m_t is a function known as a trend component,
- s_t is a function with a known period d
- Y_t is a stationary random noise component

Note

This isn't so different from regression, except the error terms behave very differently.

Modeling the Noise

ARIMA–Autoregressive Integrated Moving Average

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = W_t + \theta W_{t-1} - \cdots \theta_q W_{t-q}$$

where $\{W_t\}$ is white noise

Our goal is to estimate this process to understand how the noise terms affect the model.

Modeling in Practice

- Very difficult with real data
- a bit of an "art" using the autocovariance function

Where do outliers come in?

If Y_t is an ARMA process, we can have different events influencing the series at specific points in time to produce a new series.

$$Y_t^* = Y_t + \omega \frac{A(B)}{G(B)H(B)} I_t(t')$$

Types of Interventions

Additive Outlier

$$\frac{A(B)}{G(B)H(B)} = 1$$

This is the "true outlier" – affects only a single point in time, and never occurs again.

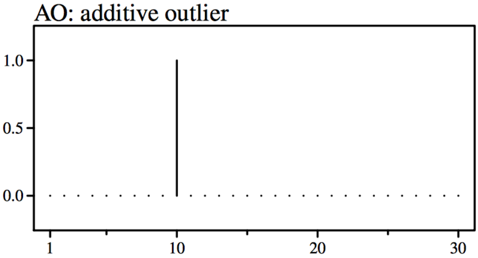


Figure: TSOutliers R package documentation (https://www.jalobe.com/doc/tsoutliers.pdf)

Types of Interventions

Level Shift

$$\frac{A(B)}{G(B)H(B)} = \frac{1}{1-B}$$

This represents a permanent vertical shift in the time series.

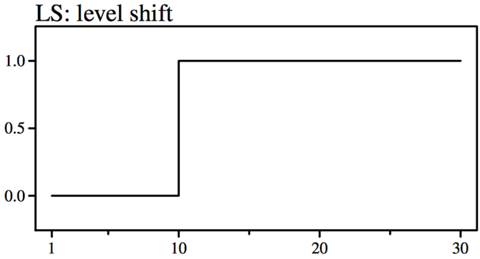


Figure: TSOutliers R package documentation (https://www.jalobe.com/doc/tsoutliers.pdf)

Types of Interventions

Temporary Change

$$\frac{A(B)}{G(B)H(B)} = \frac{1}{1-\delta B}$$

This represents a permanent shift in the time series that decays over time with a rate δ.

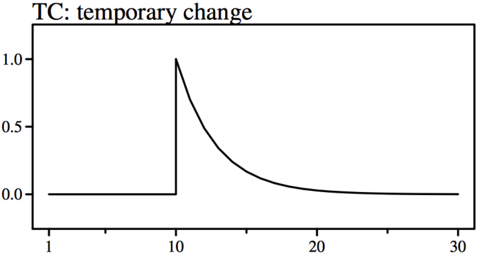


Figure: TSOutliers R package documentation (https://www.jalobe.com/doc/tsoutliers.pdf)

Types of Interventions

Innovation Outlier

$$\frac{A(B)}{G(B)H(B)} = \frac{\Theta(B)}{\alpha(b)\phi(B)}$$

This represents an initial impact with effects lingering over observations (not just time). Effect may increase or decrease with time

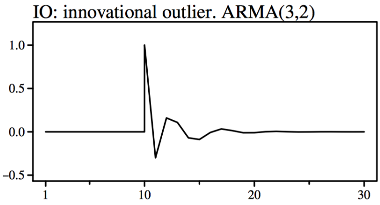


Figure: TSOutliers R package documentation (https://www.jalobe.com/doc/tsoutliers.pdf)

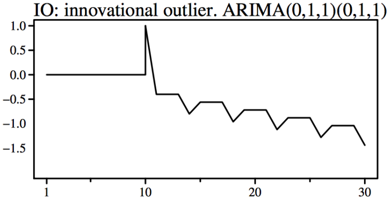


Figure: TSOutliers R package documentation (https://www.jalobe.com/doc/tsoutliers.pdf)

Good model, Difficult Implementation

Issues

- Two loops model data and detect outliers
- In practice, creates a "stuck" state for certain types of datasets

What do we do about it?

Enter the Fuzzy Numbers

- 2013 paper discusses the transformation of financial time series aggregated in the form of Japanese Candlesticks transformed into fuzzy numbers
- Created a generalized Fuzzy AR model that simplified TS estimation with better predictive ability

Japanese Candlesticks

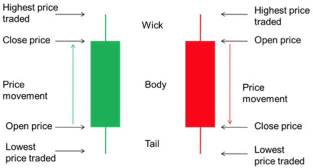


Figure: Japanese Candlesticks

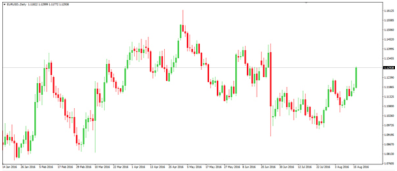


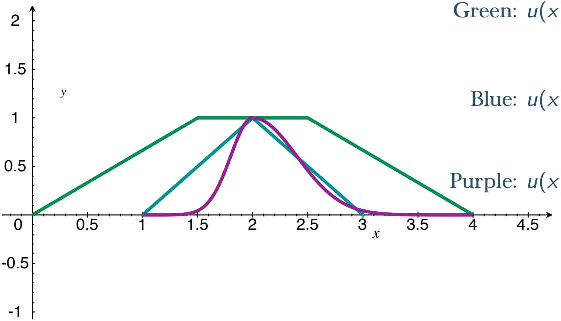
Figure: Candlestick Chart

Source: Green Box Markets

Fuzzy Numbers

What is a fuzzy number?

- An extension of the notion of a fuzzy set, with membership function on $[0, 1]$ instead of $\{0, 1\}$



Green: $u(x) = \begin{cases} \frac{2x}{3}, & 0 \leq x \leq 3/2 \\ 1, & 3/2 \leq x \leq 5/2 \\ \frac{2(4-x)}{3}, & 5/2 \leq x \leq 4 \end{cases}$

Blue: $u(x) = \begin{cases} x - 1, & 1 \leq x \leq 2 \\ 3 - x, & 2 \leq x \leq 3 \end{cases}$

Purple: $u(x) = \begin{cases} \exp\left(-\frac{(x-2)^2}{2 \cdot 0.2^2}\right), & 1 \leq x \leq 2 \\ \exp\left(-\frac{(2-x)^2}{2 \cdot 0.4^2}\right), & 2 \leq x \leq 4 \end{cases}$

Figure: Different types of a Fuzzy 2

From Data to Fuzzy Candlesticks

Summarizing without losing information

- Clump data (secondly to minutely, hourly to daily, etc)
- Transform into fuzzy candlestick by estimating the functions

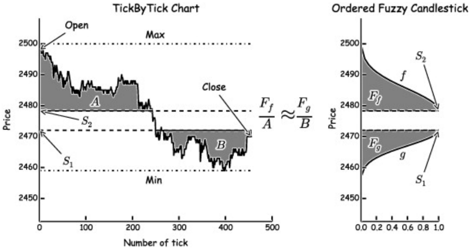


Figure: Transforming Data into an Ordered Fuzzy Candlestick

Source: Modeling and forecasting financial time series with ordered fuzzy candlesticks. A. Marzalek and T. Burczynski, *Information Sciences* (2014), Vol. 273,pp 144-155

Fuzzy Autoregressive Model

Generalization of Autoregressive Models

$$\bar{X}_t = \bar{\alpha}_0 + \sum_{i=1}^{p} \bar{\alpha}_i \bar{X}_{t-i} + \bar{\epsilon}_t$$

Results

- better predictive results for highly volatile financial time series than standard AR models (Marszalek and Murczynski, 2014)

Our Work: Fix Intervention Analysis

- generalize fuzzy AR into fuzzy ARIMA
- generalize intervention definitions
- Improve TSO algorithm

Why do you care?

- The ability to model more simply (faster)
- The ability to detect *and* classify issues
- Real-time anomaly detection at the edge

Note: The patentable IP for this project is not currently claimed and is available.

Questions?