Using Change Point Detection for Monitoring the Quality of Aggregate Data

Ian Painter*, Julie Eaton and Bill Lober

University of Washington, Seattle, WA, USA

Introduction

Data consisting of counts or indicators aggregated from multiple sources pose particular problems for data quality monitoring when the users of the aggregate data are blind to the individual sources. This arises when agencies wish to share data but for privacy or contractual reasons are only able to share data at an aggregate level. If the aggregators of the data are unable to guarantee the quality of either the sources of the data or the aggregation process then the quality of the aggregate data may be compromised.

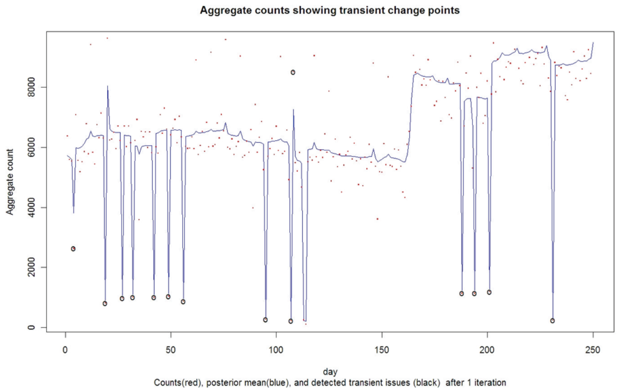
This situation arose in the Distribute surveillance system (1). Distribute was a national emergency department syndromic surveillance project developed by the International Society for Disease Surveillance for influenza-like-illness (ILI) that integrated data from existing state and local public health department surveillance systems, and operated from 2006 until mid 2012. Distribute was designed to work solely with aggregated data, with sites providing data aggregated from sources within their jurisdiction, and for which detailed information on the un-aggregated 'raw' data was unavailable. Previous work (2) on Distribute data quality identified several issues caused in part by the nature of the system: transient problems due to inconsistent uploads, problems associated with transient or long-term changes in the source make up of the reporting sites and lack of data timeliness due to individual site data accruing over time rather than in batch. Data timeliness was addressed using prediction intervals to assess the reliability of the partially accrued data (3). The types of data quality issues present in the Distribute data are likely to appear to some extent in any aggregate data surveillance system where direct control over the quality of the source data is not possible. In this work we present methods for detecting both transient and long-term changes in the source data makeup.

Methods

We examined methods to detect transient changes in data sources, which manifest as classical outliers. We found that traditional statistical process control methods did not work well for detecting transient issues due to the presence of discontinuities cause by long term changes in the source makeup. As both transient and long-term changes in source makeup manifest as step changes, we examined the performance of change point detection methods for monitoring this data. These methods have been previously used for detecting changes in disease trends in data aggregated from Distribute (4). Following Kass-Hout (4), we used the Bayesian change point estimation procedure of Barry (5) as implemented in the R package BCP (6). We examined both offline and online detection using time series held at a constant lag.

Results

We found that transient problems could be detected offline as neighboring change points with high posterior probability. When multiple outliers exist close together, detection can be improved by iteratively removing flagged data points and re-running the change point detection on the reduced data. Following the removal of outliers, remaining change points indicated long-term changes. To enable real-time monitoring for data quality problems we modified this offline detection process to in addition flag individual change points (rather than pairs of change points) detected in the most recent 5 days.



Keywords

Data Quality; Surveillance; Changepoint methods; Distribute

Acknowledgments

We would like to thank the Markle Foundation and the Centers for Disease Control for providing funding for the Distribute project.

References

1. Olson DR, et al. Applying a New Model for Sharing Population Health Data to National Syndromic Influenza Surveillance: DiSTRIBuTE Project Proof of Concept, 2006 to 2009. PLOS Currents Influenza. 2011 Sep 12.
2. Painter I, et al. How good is your data? 2011 ISDS Conference Abstract. Emerging Health Threats Journal 2011, 4
3. Painter I, et al. Generation of Prediction Intervals to Assess Data Quality in the Distribute System Using Quantile Regression. JSM proceedings, Section on Statistics in Defense and National Security. 2011 Dec.
4. Kass-Hout TA, et al. Application of change point analysis to daily influenza-like illness emergency department visits. JAMIA. 2012 Jul 3.
5. Barry D, Hartigan JA. A Bayesian analysis for change point problems. J Am Stat Assoc 1993;35:309–19.
6. Erdman C, et al. bcp: An R package for performing a Bayesian analysis of change point problems. Journal of Statistical Software 23(3). 2007.

*Ian Painter
E-mail: ipainter@uw.edu