OJPHI

ISDS 2015 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Enhancing EpiCenter Data Quality Analytics with R

Andrew Walsh*

Health Monitoring, Pittsburgh, PA, USA

Objective

To demonstrate the broader analytical capabilities available by making the R language available to EpiCenter reporting

Introduction

The EpiCenter syndromic surveillance platform currently uses Java libraries for time series analysis. Expanding the data quality capabilities of EpiCenter requires new analysis methods. While the Java ecosystem has a number of resources for general software engineering, it has lagged behind on numerical tools. As a result, including additional analytics requires implementing the methods *de novo*.

The R language and ecosystem has emerged as one of the leading platforms for statistical analysis. A wide range of standard time series analysis methods are available in either the base system or contributed packages, and new techniques are regularly implemented in R. Previous attempts to integrate R with EpiCenter were hampered by the limitations of available R/Java interfaces, which were not actively developed for a long time.

An alternative bridge is via the PostgreSQL database used by EpiCenter on the backend. An R extension for PostgreSQL exists, which can expose the entire R ecosystem to EpiCenter with minimal development effort.

Methods

The PL/R extension version 8.3 was installed in PostgreSQL 9.2 using R version 3.2.1. Gaussian and Poisson regression models were fit using the base glm function. Negative binomial regression models were calculated using the R package MASS version 7.3.42.

Regression models were fit using covariates calculated from dates - day of week, hour of day, days since start of time series, and periodic variables with an annual period. Model fits were compared using root mean squared error (RMSE) and median average deviance (MAD). Out-of-range values were defined as observations outside the 99.9% confidence interval defined by the model distribution.

Predictions for periods with outages were generated from the date-based covariates and models fit to other data.

Results

A total of 16,028,901 emergency department (ED) registrations from 415 hospitals were collected from July 1, 2014 to June 30, 2015. Hospitals were grouped on whether known or obvious data quality issues existed in their data (N=71) or not (N=344). Model performance was assessed on data from hospitals without apparent data quality issues. Gaussian regression models were fit with no covariates, approximating EpiCenter's moving average analysis method. Poisson and negative binomial regression models were fit using date-based covariates.

The Gaussian models had an average RMSE of 2.89; for the Poisson and negative binomial models it was 2.09. The MAD for the Gaussian models was 2.26; for the Poisson and negative binomial models it was 1.26.

Out-of-range values were generated comparing observations to the 99.9% confidence intervals calculated from the model fit. All detected issues were assumed to be false alerts, as no known or obvious data quality issues existed in this data. Figure 1 shows the number of false alerts relative to ED volume. False alerts from the Gaussian model decreased with increasing volume, while false alerts from the other two models showed the opposite trend.

Detection of known outages in the data from the 71 hospitals with known issues showed a similar performance profile between the modeling options.

Conclusions

The PL/R extension for PostgreSQL provides a convenient option for extending the data quality analytics of EpiCenter. By adding the resources of the R environment, new techniques can be implemented and deployed flexibly with minimal development effort. Future work will focus on integrating these methods into the main EpiCenter workflows.

ED volume data is not always well modeled with a Gaussian distribution, particularly at smaller facilities. Regression models can account for the structure in the data, such as the day-of-week effect, and also more accurately reflect the true distribution of the data, improving precision in detecting data quality problems. While the analysis presented here makes some over-broad simplifying assumptions (e.g. there are almost certainly unknown and subtle data quality issues in the data which was assumed to be reliable for the purposes of quantifying false alerts), it does demonstrate the advantage of expanded analytical capabilities.

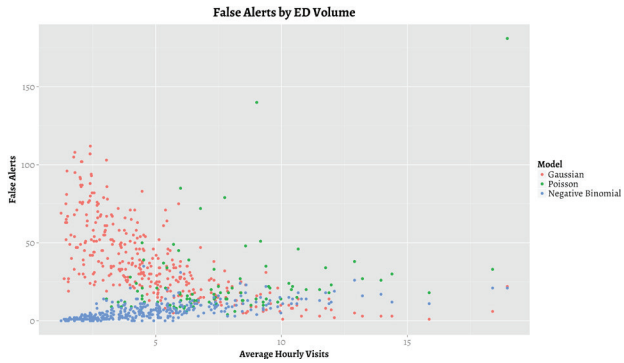

Figure 1

Keywords

R; data quality; EpiCenter; syndromic surveillance

Acknowledgments

We wish to thank the New Jersey, Ohio, Pennsylvania and Wyoming Departments of Health for funding support and data for this work.

***Andrew Walsh**
E-mail: andy.walsh@hmsinc.com