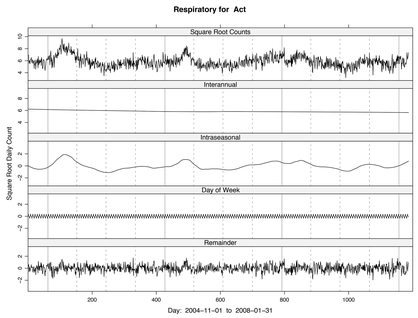
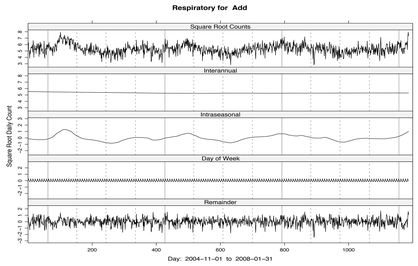
STL and local regression for modeling disease surveillance counts

David Anderson, Ryan Hafen, Ross Maciejewski, William S. Cleveland, David S. Ebert, Mourad Ouzzani, Shaun J. Grannis, and Ahmad Abusalah

Purdue University Regional Visualization and Analytics Center

Introduction





The two plots above show examples of the seasonal decomposition of time series by loess (STL) for two Indiana hospitals (called "Act" and "Add" to protect their anonymity). The raw data are the square root of the daily counts of Emergency Room (ER) patients with a respiratory chief complaint. STL has been used to break the data down into four components. The interannual component contains long term trends taking place over the course of several years. The intraseasonal component contains short term trends taking place over the course of a few months. The Day of Week component is a stable seasonal component and contains effects due to the day of the week. The remainder, what is left after the other components are removed, can be modeled as Gaussian white noise. The four components sum up to the square root counts. On the plot, the x–axis is the day, the y–axis is the square root daily count, the solid vertical lines are on the start of a year (January 1), and the dashed vertical lines are at the start of quarters (April 1, July 1, and October 1).

Purpose

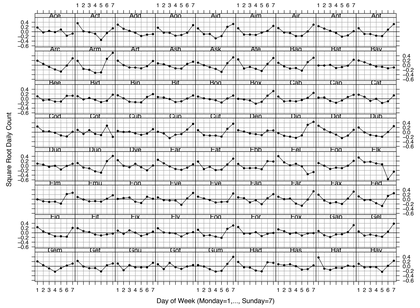
The STL decomopsition serves two purposes. The first is to provide public health officials with advanced visualization tools that can be used to monitor and track disease activity. The second is to lay the foundation for automatic alert procedures via control charts. Several one–sided procedures for real–time control charting can be implemented as natural extensions of STL. These can be used to detect potential disease outbreaks, as well as to detect possible incidents of bioterrorism. A key advantage of the division into components is that it allows for a clear picture of the factors effecting the counts, and prevents day of the week, intraseasonal, and interannual trends from influencing and hindering alert procedures.

The Indiana Public Health Emergency Surveillance System (PHESS)

We are working in collaboration with the Indiana State Department of Health. Our data consist of daily counts of chief complaints from Emergency Rooms at 72 Indiana hospitals that are part of PHESS. Data is collected at the hospitals and sent to the Indiana University School of Medicine, where it is processed and sent to the Purdue University Regional Visualization and Analytics Center (PURVAC). The chief complaints are divided into eight classifications by a Bayesian classifier (CoCo). Data for the first hospitals goes back to November 2004, and new hospitals have come on line continually since then.

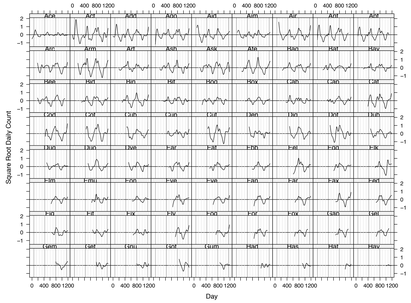
Properties of the components

The day of week component



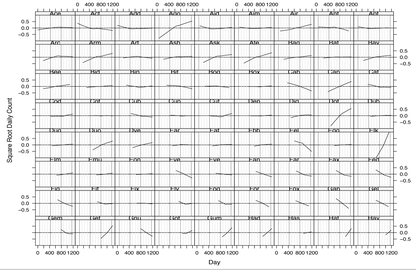
The above graph shows the day of the week component for respiratory square root counts for all 72 hospitals. The general pattern is U shaped, with a high on Monday, a decline through Friday, and a rise over the weekend. For some hospitals, the day of the week component is very high over the weekend, for others it is considerably lower. The day of the week component is small but significant, and it can not be ignored. It allows public health officials to know on what days hospitals are busiest and to plan accordingly. It also allows, for example, the number of patients at an ER to be recognized as unusually high for a Friday even if the same number of patients wouldn't have been unusually high for a Monday.

The intraseasonal component



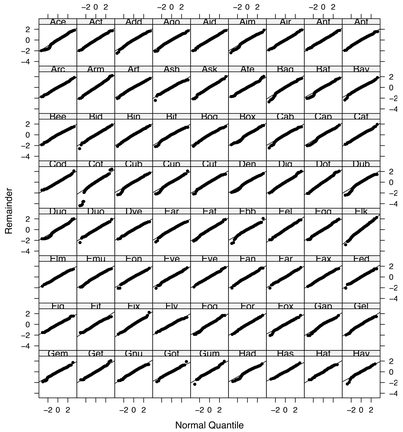
The intraseasonal component for respiratory square root counts for the hospitals is shown above. For most hospitals, the seasonal flu shows up clearly in the intraseasonal component each winter. In particular, for many hospitals there is a particularly strong peak in early 2005, and in some hospitals a double peak in late 2006 and early 2007. Careful monitoring of the intraseasonal component can be used to obtain real–time information on a flu season while it is still in progress.

The interannual component



The interannual component (shown above for respiratory square root counts) varies from hospital to hospital. Some hospitals have a rising or declining number of patients over time, while others remain fairly constant.

The remainder



Numerous diagnostic plots verify that the remainder is Gaussian white noise. For example, the above Normal probability plots show that the distribution is well approximated by a Gaussian distribution.

The square root transformation

We take the square root transformation because it makes the statistical properties across the hospitals much simpler. The raw data result in a remainder which is non–Gaussian and in which the standard deviation depends on the mean of the series. In contrast, the square root transformation leads to a remarkable improvement; the variance is stabilized, comparison between hospitals is enabled, and the distribution of the remainder is much closer to Gaussian. The square root transformation allows for a more parsimonious model, since without it adjustments would have to be made to account for the unequal variability between hospitals.

Next step: control charts

We are working to build control charts for disease detection, where variability is based on the remainder component. A variety of charts will be used: some for rapid detection of events local in time, and others for detection of gradually developing events.

Acknowledgements

This project was funded by the PURVAC Center of Excellence by a grant from the Department of Homeland Security, Science and Technology Directorate, Office of University Programs. Their support is greatly appreciated.

We would also like to thank our collaborators: The Indiana State Department of Health (ISDH) and the Indiana Network for Patient Care (INPC).

For further information

Please contact David Anderson: anders91@purdue.edu
More information on this and related projects can be obtained at http://www.purvac.org.