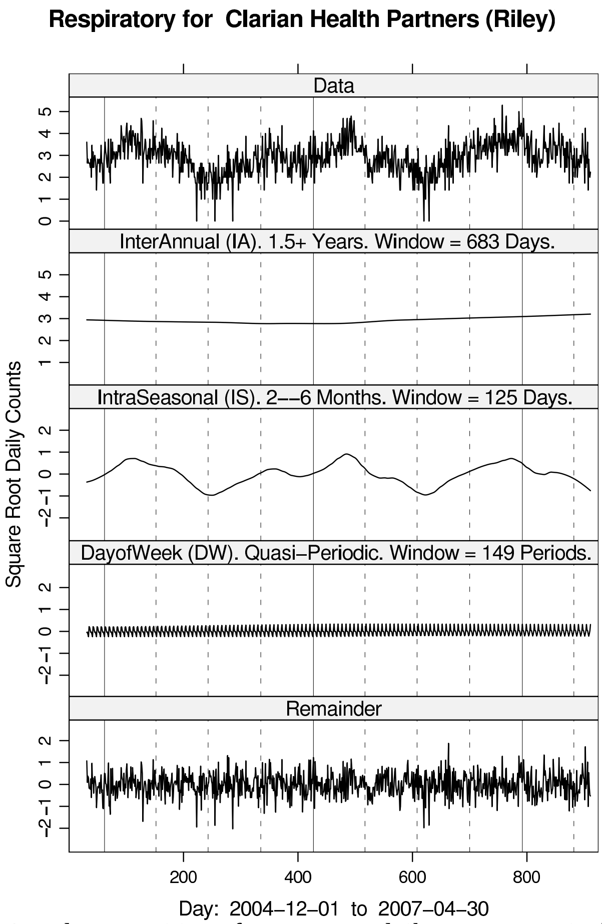
STL and Local Regression for Modeling Disease Surveillance Counts

David E. Anderson, Cheng Zheng, Ross Maciejewski, Ryan Hafen, William S. Cleveland, David S. Ebert, Mourad Ouzzani, Purdue Regional Visualization & Analytics Center, W. Lafayette. Shaun J. Grannis, Regenstrief Inst. & Indiana U. School of Medicine, Indianapolis.



STL decomposition of square-root daily counts. Vertical scales have same units/cm. Solid vertical line is Jan. 1.

Objective: Use the STL local-regression (loess) decomposition procedure and transformation to model the univariate time-series characteristics of chief-complaint daily counts as a first step in a time and spatial modeling. Develop visualization tools for model display and checking.

Background: Numerous methods have been applied to the problem of modeling temporal properties of disease surveillance data; the ESSENCE system contains a widely used approach *(1)*. STL *(2)* is a flexible, well-proven method for temporal modeling that decomposes the series into frequency components (see figure). A periodic component like DW can be exactly periodic or evolve through time. STL is based on loess *(3)*, which can model a numeric response as a function of any explanatory variables. After the STL modeling of the counts, we will add patient address and produce a time-space modeling using both STL and more general loess methods.

Methods: Daily counts in our exploratory study are from 14 hospitals in the Indiana system *(4)* for 2-2.5 years. We found that the square root transformation of the counts stabilized the dependence of the mean on the standard deviation and made remainder distributions much closer to normal. STL has model parameters that determine the smoothing window for each component. The parameters were chosen through the extensive use of data visualization.

Results: IS components show seasonal influenza including some double peaks within a season. IA components show a mild rise for some hospitals indicating growth in the number of patients. DW components for most hospitals have a peak on Mon., a drop through about Fri. then a rise back to a Sun. peak about as big as Mon. No significant effects appeared in the remainders, which can be modeled as independent, identically-distributed, normal random variables.

Conclusions: The use of loess and the square root transformation appear to be very promising, compared with other approaches, for modeling chief-complaint daily counts. Loess has already been used for the single task of estimating a yearly seasonal effect *(5)*. Square roots result in models that are considerably more parsimonious. Loess provides flexibility for accommodating patterns in the data, and it has very fast computational algorithms based on k-d trees. The next step of integrating spatial variables using general loess methods appears warranted.

References: *(1)* H.S. Burkom. Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance. *Johns Hopkins APL Tech. Digest*, 24, 2003. *(2)* R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J. Official Statist.*, 6:3–73, 1990. *(3)* W.. Cleveland, S.J. Devlin. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Amer. Statist. Assoc.*, 83:596–610, 1988. *(4)* S.J. Grannis, P.G. Biondich, B.W. Mamlin, G. Wilson, L. Jones, J.M. Overhage. How Disease Surveillance Systems Can Serve as Practical Building Blocks for a Health Information Infrastructure: the Indiana Experience. *AMIA Annual Symp. Proc.*, 286–290, 2005. *(5)* U.G. Dafni, S. Tsiodras, D. Panagiotakos, K. Gkolfinopoulou, G. Kouvatseas, Z. Tsourti, G. Saroglou. Algorithm for Statistical Detection of Peaks – Syndromic Surveillance System for the Athens 2004 Olympic Games. *Morbidity and Mortality Wkly. Rept.*, 53:86-94, 2004.