COACTION PUBLISHING

Clustering of U.S. cities based on mortality from influenza and pneumonia

Eric Foster[1,2], Joseph Cavanaugh[1,2] and Philip Polgreen[1,2]

[1]University of Iowa, Iowa City, IA, USA; [2]Computational Epidemiology Research Group, Iowa City, IA, USA

Objective
To cluster cities in the United States based on their levels of mortality from influenza and pneumonia.

Introduction
Influenza is a major cause of mortality. In developed countries, mortality is at its highest during winter months, not only as a result of deaths from influenza and pneumonia but also as a result of deaths attributed to other diseases (e.g., cardiovascular disease). Understandably, much of the surveillance of influenza follows predefined geographic regions (e.g., census regions or state boundaries). However, the spread of influenza and its resulting mortality does not respect such boundaries.

Methods
Data on influenza and pneumonia mortality were collected from 97 cities over 11 years (1996 through 2007), as reported in the MMWR (1). We used a novel method of computing the pairwise distance between two time series based on the Mahalanobis Distance derived from the time-series state-space-modeling framework. Mahalanobis Distance is a scale invariant form of Euclidean Distance that also takes correlations of the data set into account. This is an extension of a previously devised Kullback-Leibler Information-based time-series-clustering discrepancy measure (2). All pairwise distances between cities were then used in a clustering procedure known as QT_Clust (3). This procedure was initially developed for the clustering of high dimensional genomic data. However, QT_Clust may be applied to many time-series-data sets where the trajectory rather than the process of a time series is of interest. A measure of cluster size and within-cluster distance is used to compare how geographically based influenza surveillance performs as opposed to nongeographically based surveillance.

Results
The average within-cluster distance for the nine census regions is 5205 units. Ignoring geography, we found that our nine largest clusters held 85 of the cities (87.6% of the total cities observed) and maintained an average within-cluster distance of 4918 units. This amounted to a 5.5% reduction in the within-cluster distance. The largest of these clusters held 33 cities from all but one census region and had a within-cluster distance of 4295 units, meaning that its within-cluster distance was 17.5% smaller than that of the mean within-cluster distance of the nine census regions, the largest of which only held 17 cities.

Conclusions
It is natural to think of geographic proximity as an indicator of how likely a city or region's pattern of influenza mortality mirrors that of another region. However, we hypothesize that the relatively high level of travel within the country will affect the pattern of mortality such that cities across the nation may resemble one another more closely than cities within a predefined geographic region. Our approach involved the creation of a discrepancy measure specifically designed for time series data and the application of a clustering routine that seeks to create high quality clusters (rather than high-inclusion clusters). The largest cluster held roughly a third of the observed cities and yet still had a low within-cluster distance when compared to the geographic census regions. This result suggests that many cities observe similar influenza and pneumonia mortality patterns despite varying geographical locations.
There are several limitations to this study. First, while our discrepancy works well in the presence of missing data, a preponderance of consecutive missing time points can negatively affect performance. We determined that 25 cities out of the original 122 cities reported too sporadically for analysis. Furthermore, our clustering technique depends upon the arbitrary selection of a 'quality criterion' that is very data driven. High quality clusters can be obtained, but this often leads to a large number of clusters. Conversely, a small number of clusters can be obtained by lowering the quality criterion. Future work will determine if we can use this time-series-clustering approach to find repeatable clusters that may or may not suggest changes to current geographic boundaries in an effort to coordinate future influenza surveillance activities.

Keywords
Time series; clustering; influenza

References

1. Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report. http://www.cdc.gov/mmwr.
2. Bengtsson T, Cavanaugh JE. State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. Environmetrics. 2008;19:103–21.
3. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Cold Spring Harbor Lab Press. 1999;9:1106–15.

*Eric Foster
E-mail: eric-foster@uiowa.edu

Citation: Emerging Health Threats Journal 2011, **4**: 11077 - DOI: 10.3402/ehtj.v4i0.11077