

A Bivariate Statistic for Temporal-Spatial Syndromic Surveillance

Marco Bonetti

Al Ozonoff

Laura Forsberg

Marcello Pagano

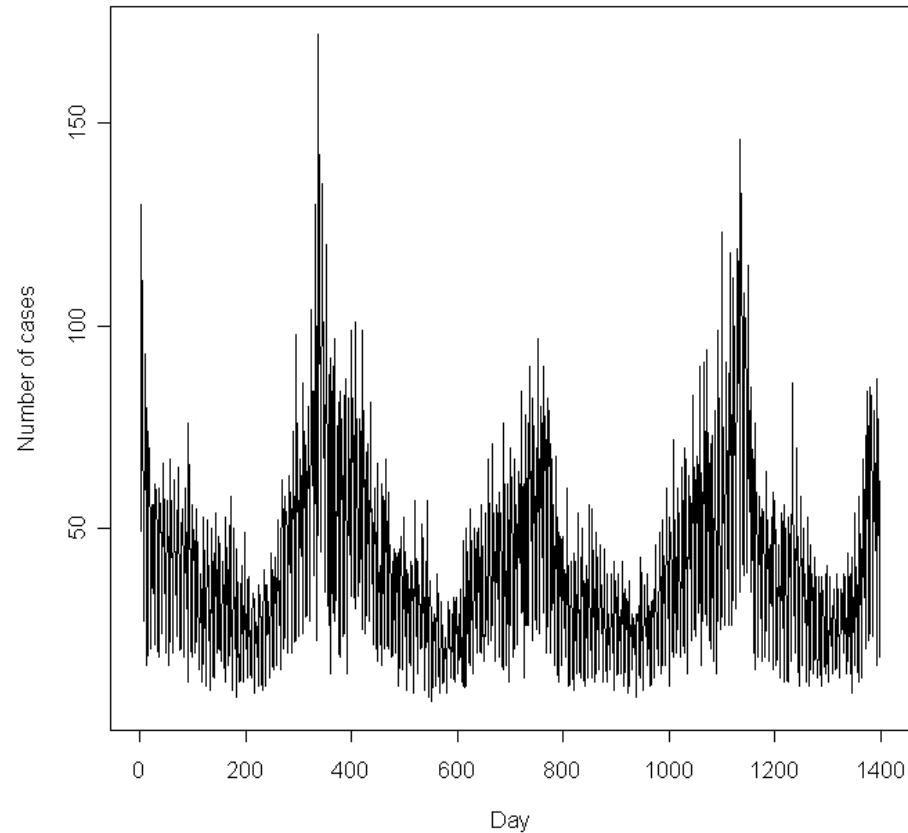
Department of Biostatistics
Harvard School of Public Health

Time series analysis of syndromic data

Traditional temporal surveillance using time series analysis

- Traditional temporal surveillance involves predicting case load (for example, number of “flu-like” cases or other syndrome presenting to a hospital ER) and sounding an alarm if the observed number of cases greatly exceeds the expected number.
- Typically models include a seasonal baseline which is sinusoidal, together with other known factors that influence case load such as day of week, holidays, etc.
- A more sophisticated approach is to use auto-regressive and/or moving average (ARMA) forecasting. Using the historical record, more accurate predictions can reduce the standard deviation of the observations, making a system more sensitive while keeping the false-positive rate fixed.

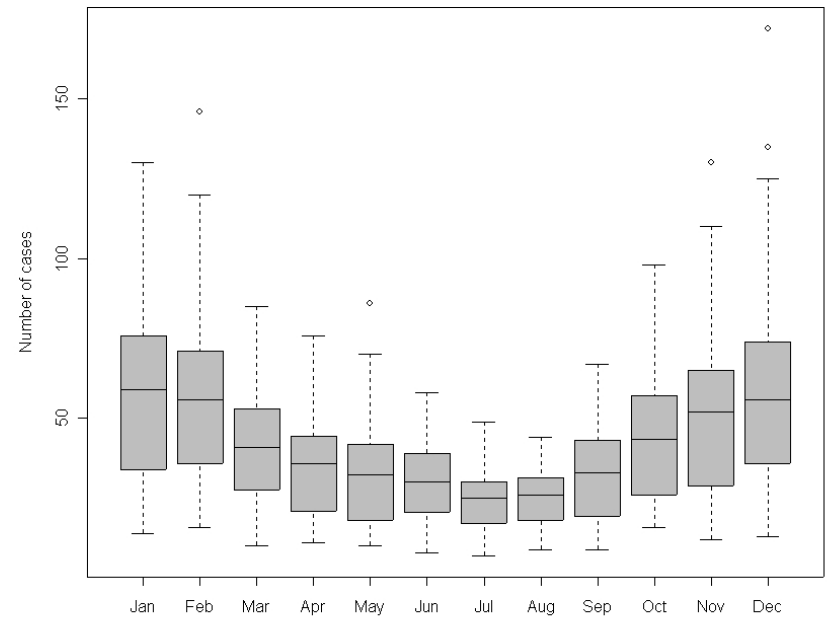
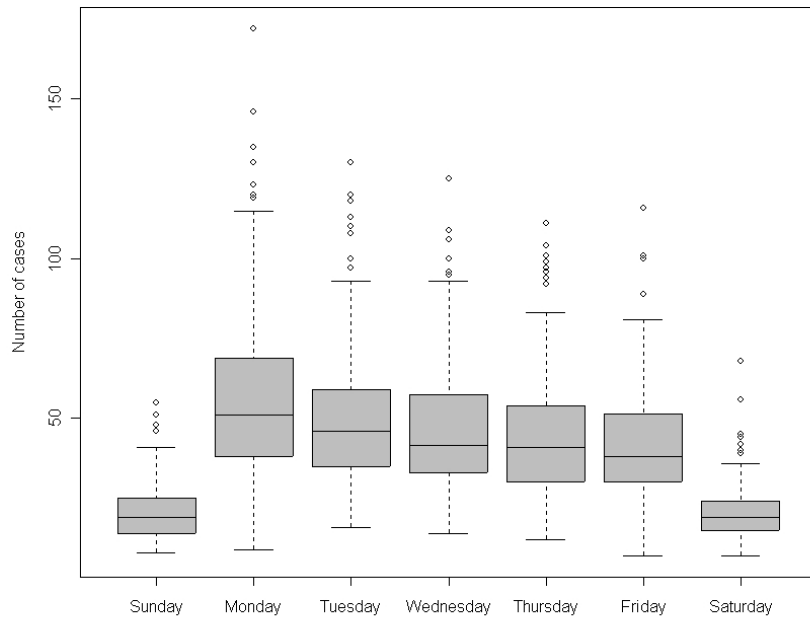
Eastern Massachusetts HMO Data



Raw data of daily case counts for upper respiratory infection within an Eastern MA HMO.

Number of Cases by Month and Day of Week

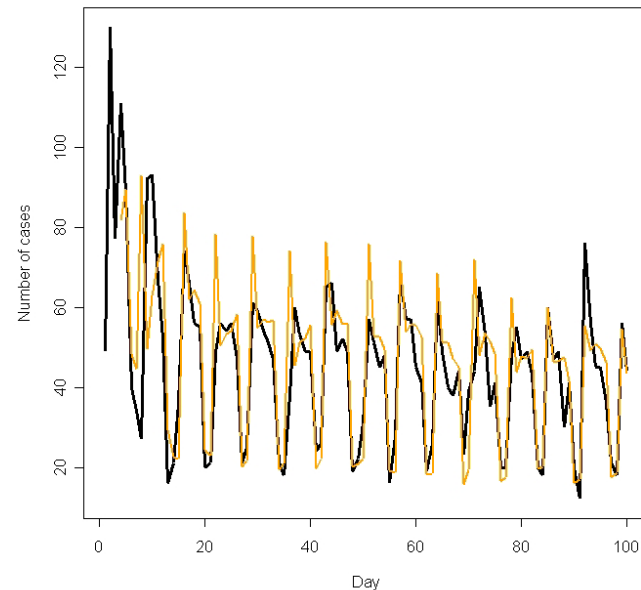
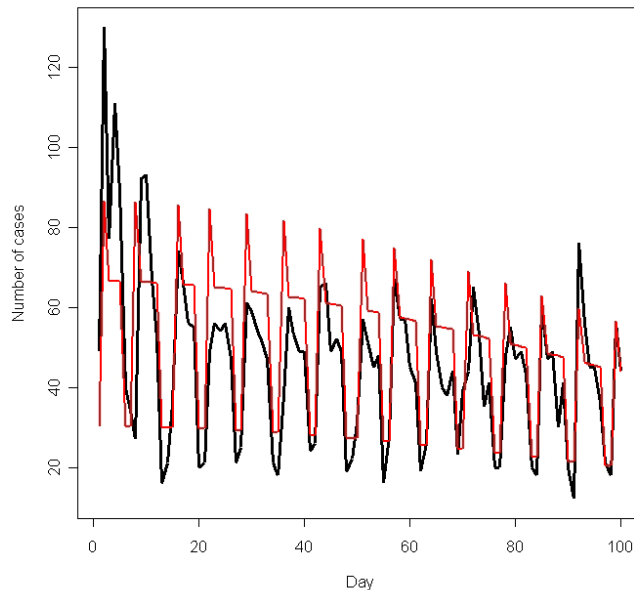
Eastern Massachusetts HMO



The daily and monthly case count box plots reveal that both mean and variance of the case counts exhibit both daily and seasonal effects. These must be accounted for in any time series modelling of the case count data.

Time Series Analysis

Eastern Massachusetts HMO



Standard modelling of case counts typically includes a sinusoidal curve to account for seasonal variation, and day-of-week indicators to account for weekly variation. The first 100 days of data are shown, with the red curve showing predicted values for the case counts. Although the model fits quite well in places, there are still portions of the graph where the fit is poor. The orange curve shows the same model, but with a third-order auto-regressive AR(3) component added. After fitting the seasonal model plus AR(3), the resulting residuals (observed minus expected case counts) are a closer approximation to a normal distribution and have a smaller standard deviation than those from the seasonal model alone.

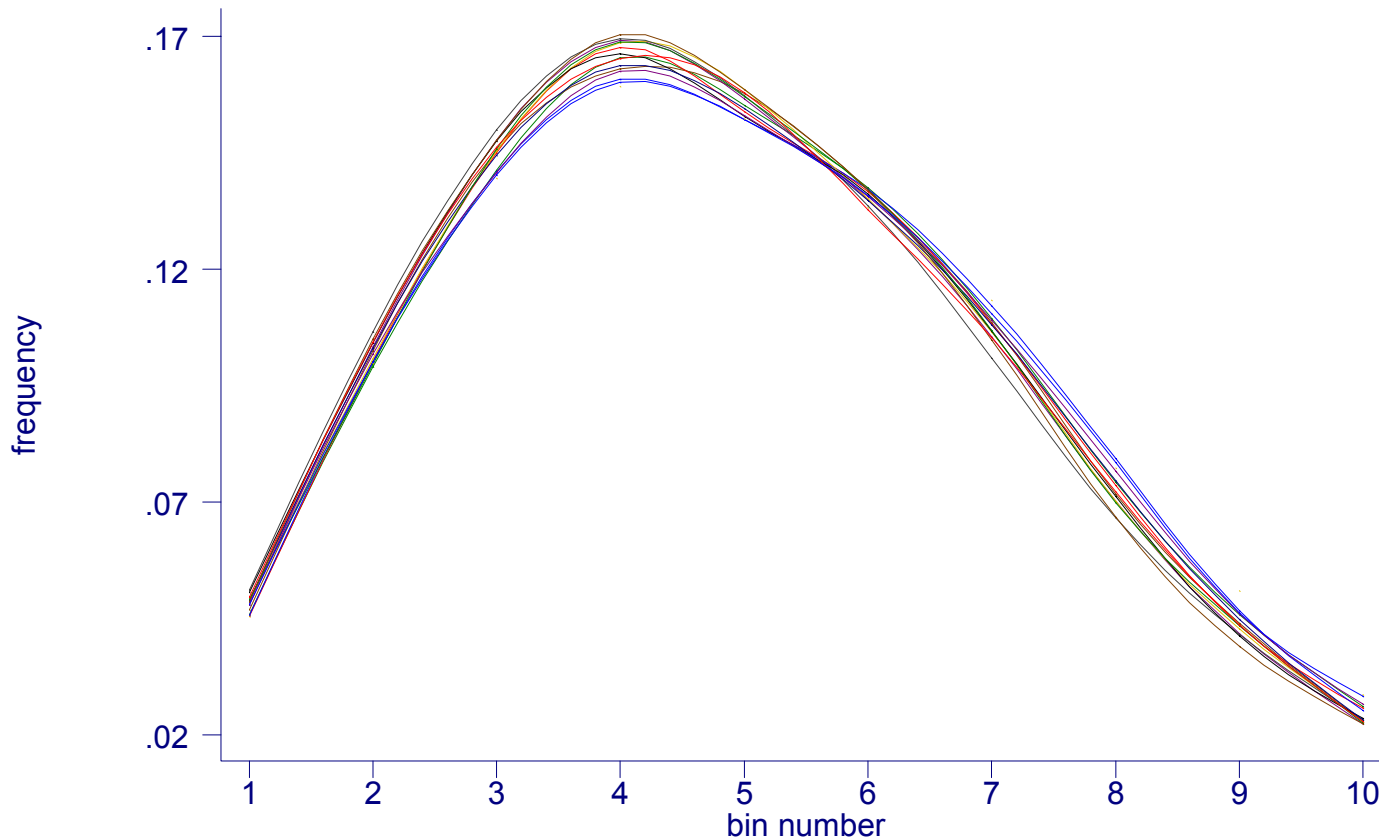
The M -statistic

A statistic designed for the detection of spatial aberrations in a population

- Given a collection of points (for example, addresses of patients with upper respiratory symptoms arriving at an ER), one possible summary of the spatial data is the distribution function of the Euclidean distances between points.
- For a particular syndrome and underlying population, this distribution of distances may exhibit stationarity. If so, we have an established distribution with which to compare emerging cases on a daily basis.
- The M -statistic is a test statistic based on this principle of comparing the observed distribution of distances to the expected distribution based on historical data.

Distribution of the Distances

Eastern Massachusetts HMO



A comparison of the distribution of distances for upper respiratory syndrome, using several non-overlapping time periods. This demonstrates that the data exhibit the stationarity required by the M-statistic. This null distribution can be compared to the observed distribution on a daily basis to determine whether the emerging cases follow a different spatial pattern from that which we would expect.

In practice, we calculate the M-statistic in a two step process.

1. The null distribution of the distances is estimated empirically. This distribution is summarized by binning the distances into k bins. The variance-covariance matrix of this k -dimensional vector is estimated via resampling methods.
2. We calculate the M-statistic for a subpopulation, by comparing its empirical cdf to that calculated in step 1. This is done by calculating a Mahalanobis-like distance between the two vectors of bin counts. Let the observed counts in this data be denoted by \mathbf{o} , the expected counts be denoted by \mathbf{e} and the estimated variance-covariance matrix be denoted by \mathbf{S} . Then we calculate:

$$M = (\mathbf{o} - \mathbf{e})' \mathbf{S}^{-1} (\mathbf{o} - \mathbf{e})$$

We can determine significance by empirical methods.

Powers for the M Statistic

Eastern Massachusetts HMO

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 6	0.16	0.18	0.12
446 (center)	0.29	0.17	0.17
185	0.21	0.15	0.14
364	0.19	0.11	0.13
212 (edge)	0.09	0.05	0.07
Standard deviation	7.1	9.4	14.1

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 9	0.25	0.26	0.13
446 (center)	0.55	0.26	0.3
185	0.49	0.34	0.25
364	0.42	0.25	0.23
212 (edge)	0.14	0.07	0.1
Standard deviation	7.1	9.4	14.1

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 12	0.37	0.38	0.17
446 (center)	0.75	0.57	0.48
185	0.72	0.56	0.42
364	0.64	0.41	0.31
212 (edge)	0.17	0.12	0.13
Standard deviation	7.1	9.4	14.1

Powers for the HMO data with varying cluster sizes (6, 9, 12) and at varying locations. Clusters were placed at the center of census tracts (446, 185, 364, 212) throughout eastern Massachusetts.

Comparisons of the M Statistic to Other Cluster Detection Methods

	M Statistic		Spatial Scan Statistic		MEET	
	.05 level	.01 level	.05 level	.01 level	.05 level	.01 level
<i>6rural01</i>	0.816	0.653	0.998	0.992	0.196	0.057
<i>6rural02</i>	0.753	0.546	0.991	0.986	0.221	0.072
<i>6rural04</i>	0.428	0.194	0.973	0.946	0.229	0.064
<i>6rural08</i>	0.293	0.094	0.971	0.937	0.213	0.055
<i>6rural16</i>	0.204	0.053	0.969	0.936	0.229	0.062
<i>6mixed01</i>	0.885	0.759	0.936	0.871	0.925	0.833
<i>6mixed02</i>	0.853	0.704	0.939	0.871	0.896	0.771
<i>6mixed04</i>	0.767	0.578	0.937	0.873	0.838	0.654
<i>6mixed08</i>	0.692	0.472	0.941	0.876	0.817	0.599
<i>6mixed16</i>	0.602	0.372	0.949	0.886	0.832	0.602
<i>6urban01</i>	0.907	0.805	0.922	0.818	0.941	0.870
<i>6urban02</i>	0.859	0.722	0.903	0.823	0.920	0.830
<i>6urban04</i>	0.905	0.799	0.892	0.794	0.961	0.902
<i>6urban08</i>	0.855	0.705	0.913	0.824	0.983	0.951
<i>6urban16</i>	0.738	0.705	0.926	0.836	0.986	0.950
<i>6rumix1</i>	0.992	0.974	1.000	0.999	0.964	0.910
<i>6rumix2</i>	0.986	0.954	0.999	0.997	0.952	0.871
<i>6rumix4</i>	0.934	0.814	0.997	0.987	0.930	0.793
<i>6rumix8</i>	0.862	0.689	0.996	0.986	0.931	0.772
<i>6rumix16</i>	0.774	0.535	0.996	0.982	0.941	0.804
<i>6two01</i>	0.994	0.980	1.000	0.998	0.970	0.923
<i>6two02</i>	0.987	0.949	0.999	0.996	0.962	0.895
<i>6two04</i>	0.975	0.924	0.992	0.974	0.971	0.912
<i>6two08</i>	0.939	0.830	0.991	0.968	0.977	0.936
<i>6two16</i>	0.835	0.634	0.987	0.947	0.975	0.915
<i>6urbmix1</i>	0.997	0.990	0.987	0.950	0.998	0.995
<i>6urbmix2</i>	0.990	0.967	0.984	0.950	0.995	0.984
<i>6urbmix4</i>	0.986	0.954	0.966	0.901	0.991	0.969
<i>6urbmix8</i>	0.954	0.876	0.954	0.871	0.990	0.960
<i>6urbmix16</i>	0.873	0.696	0.935	0.811	0.984	0.935
<i>6three01</i>	1.000	1.000	1.000	0.999	0.999	0.997
<i>6three02</i>	0.999	0.997	1.000	0.999	0.998	0.992
<i>6three04</i>	0.996	0.984	0.996	0.981	0.994	0.973
<i>6three08</i>	0.981	0.932	0.992	0.964	0.989	0.952
<i>6three16</i>	0.908	0.755	0.977	0.916	0.983	0.918

Powers for the M Statistic, Spatial Scan Statistic and Maximized Excess Events Statistic to detect “hot spot” spatial clusters of disease. Powers are calculated from the benchmark data set, designed to compare cluster detection methods (Kulldorff, Tango, Park, 2003). Gray shaded boxes are for powers of the M statistic and those that are equivalent to the M. Powers that are less than and greater than M, are indicated by yellow and blue shading, respectively.

Temporal-spatial surveillance

Using a bivariate statistic to combine temporal and spatial data

- Time series modelling yields a statistic N based on residuals (observed minus expected case counts). Residuals are approximately $N(0, \sigma_N^2)$, where standard deviation σ_N can be estimated from the data.
- The spatial statistic M is asymptotically independent of N . The log-transformed statistic $\log(M)$ is approximately normal with standard deviation σ_M estimated from the data.
- The problem is now to construct an appropriate rejection region for the bivariate statistic $(\log(M), N)$ with a controlled false positive rate.

- One approach: Exploit the approximate bivariate normality of the statistics.
- With a prespecified alternative, we can train the statistic using a quadratic classification rule for bivariate normal populations. We construct a quadratic form that distinguishes values of $(\log(M), N)$ under the null from values under the alternative, while keeping control of the false positive rate.
- The rule is given by a quadratic form:

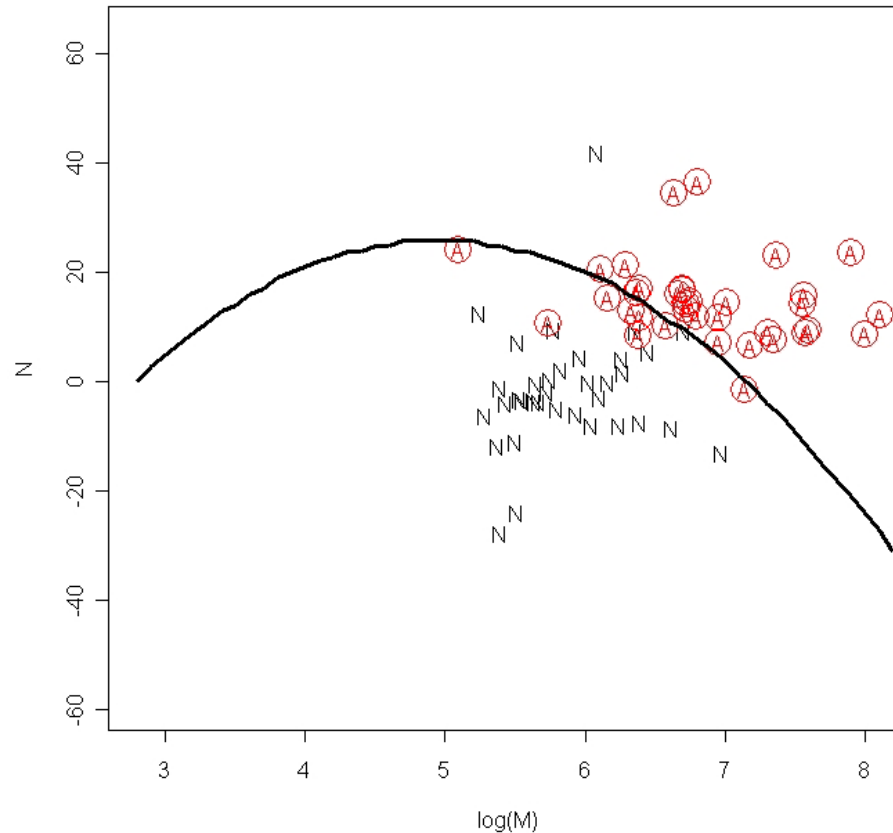
$$Q(x, y) = \alpha_1 x^2 + \beta_1 y^2 + \alpha_2 x + \beta_2 y$$

where

$$\alpha_1 = \frac{1}{2} \left(\frac{1}{\sigma_M} - \frac{1}{\sigma_{M^*}} \right), \quad \beta_1 = \frac{1}{2} \left(\frac{1}{\sigma_N} - \frac{1}{\sigma_{N^*}} \right)$$

$$\alpha_2 = \frac{\mu_M}{\sigma_M} - \frac{\mu_{M^*}}{\sigma_{M^*}}, \quad \beta_2 = \frac{\mu_N}{\sigma_N} - \frac{\mu_{N^*}}{\sigma_{N^*}}$$

Quadratic Classification



Plot of the bivariate statistic (simulated data), showing data points from the null (black N) and alternative (circled red A) populations. The quadratic discriminant curve is set to keep the false-positive rate at a pre-specified level. Points above the curve are classified to the alternative population; below the curve, to the null.

Powers for the Bivariate Statistic

Eastern Massachusetts HMO

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 6	0.16	0.18	0.12
446 (center)	0.47	0.28	0.14
185	0.40	0.26	0.09
364	0.36	0.21	0.08
212 (edge)	0.21	0.13	0.07
Standard deviation	7.1	9.4	14.1

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 9	0.25	0.26	0.13
446 (center)	0.74	0.57	0.28
185	0.70	0.54	0.24
364	0.64	0.42	0.20
212 (edge)	0.37	0.23	0.10
Standard deviation	7.1	9.4	14.1

Number of cases added	Holiday Indicator		
	Wknds/ Hols	Wkdays	Day after
N + 12	0.37	0.38	0.17
446 (center)	0.90	0.79	0.54
185	0.88	0.77	0.48
364	0.85	0.65	0.35
212 (edge)	0.59	0.35	0.16
Standard deviation	7.1	9.4	14.1

Powers for the Bivariate statistic when utilized on the Eastern Massachusetts HMO data. Note that by comparison to these powers are greater than those obtained by using the M statistic or time series analysis alone.

Conclusions

- Separate statistical methods for temporal or spatial analysis can be combined into a bivariate statistic, and this approach might be applied effectively in a syndromic surveillance setting.
- Any gains in power utilizing the bivariate statistic argue for the importance of using a temporal-spatial approach to syndromic surveillance whenever possible.
- Further research into the statistical methodology behind temporal-spatial surveillance should be considered an essential component of ongoing syndromic surveillance efforts.