Two –stage wavelet analysis assessment of dependencies in time series of disease incidence

Nina H. Fefferman, Jyotsna S. Jagai, Elena N. Naumova
Tufts University School of Medicine
Family Medicine and Community Health
136 Harrison Avenue, Boston, MA 02111 USA
e-mail: nina.fefferman@tufts.edu

Keywords: wavelet decomposition, time series, Fourier decomposition, loess smoothing, environmental epidemiology

Abstract

In epidemiology, techniques which examine periodicity in times series data can be used to understand weekly, biannual or seasonal patterns of disease. However, a simple understanding of periodicity is not sufficient to examine the possible influence of variation in incubation period, distributed sources of infection, and infection due to environmental factors, especially if these influences affect the rate of disease on various spatio-temporal scales. Wavelet analysis provides the ability to consider influences on various spatio-temporal scales. In order to examine the feasibility of using wavelets to assess dependencies over different spatio-temporal scales in a time series of disease incidence, we abstracted 10 years of daily records of ambient temperature and precipitation in addition to daily disease incidence data for Massachusetts for two enterically transmitted diseases. We eliminated periodic fluctuation in both seasonal and weekly case reporting using various techniques (Fourier transformation and "loess" smoothing) on each time series of disease data. These different methods were employed in order to examine the possible effect of removed periodicities on the variance of the data. We then performed a wavelet decomposition to examine the residuals from these analyses on a variety of temporal scales and examined the resulting correlations to the environmental data.

Introduction

Periodicity in time series data can be examined using a variety of techniques. In epidemiology, where a time series may consist of reported incidence of disease, these methods can be used to examine weekly oscillations, semi-annual cycles, or seasonal patterns of disease. Mathematical models of disease spread have shown that individual interaction rates can cause periodic epidemics in a population, even in the absence of external influences (Johansen, 1996). However, exposure to enteric pathogens in a population may operate within a far more complicated framework governed, not only by individual transmission, but also by variation in length of incubation period, distributed sources of infection, and variation due to environmental factors governing exposure. A simple understanding of periodicity may not be sensitive enough, especially if these influences affect the disease in a population on a variety of spatio-temporal scales.

Incidence of the diseases giardiasis and salmonellosis exhibit seasonal oscillations (Barwick et al. 2000; Furness et al. 2000; Lee et al. 2002; Naumova et al. 2000). Giardiasis is a parasitic infection, which has been shown to be a waterborne disease, and is associated with contaminated drinking and recreational water. Environmental conditions, such as daily precipitation are not only integral to seasonal variation but (at least in extreme cases such as flooding) can have profound effects on water purification plants (Curriero et al.2001). Swimming pools and other recreational water form common modes of transmission of waterborne pathogens during warmer months (Birkhead et al. 1989; CDC 2003). Salmonellosis is a bacterial infection and is typically

considered a foodborne disesase, which has been shown to be directly related to food 'spoilage' (Bean et al. 1996; Bentham and Langford 2001;Olsen et al. 2000). Therefore, increased incidence of salmonellosis is expected during the warmer months of the year and increased incidence can be predicted following moderate spikes in temperature over the average ambient temperature during any season. Environmental effects are direct causes of increased exposure however it is difficult to determine on what temporal scale these effects are acting. Annual oscillations could simply be the result of a confluence of individual transmissions coupled with a short-term immune memory, as described in Johansen (1996). If this were the case, one would expect to find no significant differences in disease rates regardless of environmental conditions across years. It would also be expected that disease rates to be independent of outside influences such as water supply utilized and types of food eaten.

In order to test these hypotheses, we examine the correlation of local variation of disease incidence, *with* and *without* the influence of global seasonal oscillation, with both temperature and precipitation fluctuations in cities that utilize different water supply systems. Should local environmental conditions have a direct affect on disease exposure risk, it would be expected that, after correcting for all periodic trends, there would remain correlation between reported incidence and variation in environment on a given temporal scale. It is expected that variation in ambient temperature will a higher correlation to the incidence of salmonellosis than to that of giardiasis; and likewise giardiasis will have a high correlation with precipitation especially in locales with vulnerable water supplies. The presence of such correlation would provide evidence that observed seasonal periodicity in the incidence of these diseases is regulated by seasonal effects other than direct environmental conditions.

Examining different temporal scales for possible influence of environmental fluctuation allows for the possibility that the most influential scale of effect is the seasonal periodicity itself for which the correction is used. In order to ensure that the elimination of annual oscillation is not masking a more subtle effect, two smoothing methods were used and the residuals of each were analyzed for evidence of local correlations. In addition, correlation analyses *without* first correcting for seasonal oscillation was also considered. Therefore, the two stages in the analysis are 1) wavelet decomposition and 2) correlations between wavelet crystal coefficients. This paper is considering the application of this two-stage wavelet analysis within the context of the two particular diseases giardiasis and salmonellosis, however, this method has broader applicability to any system where global periodicity may have a confounding effect in the determination of the impact of local conditions.

Methods

We abstracted 4,444 records of laboratory confirmed cases, without personal identifiers, for giardiasis and salmonellosis from the Massachusetts Department of Public Health (MADPH) Surveillance database. Data were collected for the cities of Boston, Lowell and Worcester over a ten-year period (3652 days) from January 1, 1993 through December 31, 2002. Time series of daily counts of reported laboratory-confirmed cases of the two diseases were created using the date of disease onset.

Daily measurements of maximum ambient temperature and precipitation were abstracted from the National Climatic Data Center (NCDC) Summary of the Day database (EarthInfo Inc, Boulder, CO). Boston, Lowell and Worcester each have local weather stations that covered the time period of interest. The meteorological data for Boston was complete; Lowell and Worcester were both missing small percent of data points for a few months during the 10-year study period. A daily average of available measurements for these months, were used to replace missing data.

Underlying seasonal periodic fluctuation in the daily reported incidence of each disease, in each of the three locations, was removed using two methods, Fourier transformation (Wavelet module, S+ Software, Version 6.1) and Loess smoothing. We then performed wavelet decomposition (Wavelet module, S+ Software, Version 6.1) on the residuals from these smoothing methods and the original time series itself, without any correction for seasonal oscillation. Wavelet decomposition was defined using:

$$x(t) = \sum_{k=-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} \left(d_{k,n} \psi_{k,n}(t) \right) \right) \text{ with the basis function } \psi_{k,n}(t) \qquad d_{k,n}(t) = \int_{-\infty}^{\infty} \overline{\psi_{k,n}(t)} x(t) dt$$

Smoothing and wavelet decomposition were also performed on the environmental characteristics, temperature and precipitation. A V-Spline2 wavelet was used since it provided a symmetric, periodic function, which is consistent with the seasonal trends. In order to examine the sensitivity of disease incidence to environmental factors, we examined the rank-correlation (Spearman) of the wavelet crystals for each disease with the wavelet crystals for the temperature and precipitation for the same locations. We also examined correlation among the wavelet coefficients between the diseases and environmental characteristics in each location. The correlations were calculated for each pair Boston – Worcester (BW), Boston – Lowell (BL) and Worcester – Lowell (WL), between each wavelet crystal for non-transformed data as well as for each type of transformation, Fourier and Loess. Spearman correlations and autocorrelations for wavelet coefficients were calculated for both outcome and both exposure variables. In order to examine the shape of relations between coefficients for each wavelet in greater details we applied a non-parametric smoothing procedure.

Results

The results of wavelet decompositions of the original time series for each disease in each of the three locations and the 9-wavelet crystals are shown in Figure 1. The large values in coefficients for the crystals 1 to 4 coincide with an increase in reported diseases. The mark in Figure 1C refers to a documented outbreak of giardiasis in Worcester in the summer of 1995 (Naumova et al, 2000). Multiple spikes in the wavelet coefficients for the time period of this outbreak are observed for crystals 1 to 4.

As a part of exploratory analysis, Spearman correlations for coefficients of wavelet were calculated for both outcome and both exposure variables. The correlations were calculated for each pair Boston – Worcester (BW), Boston – Lowell (BL) and Worcester – Lowell (WL), between each wavelet crystal for non-transformed data as well as for each type of transformation, Fourier and Loess. Figure 2 shows an overall high correlation for temperature (Panel A) and precipitation (Panel B) for all nine crystals and all location pairs. The main exception is seen in the untransformed data in BL (Red, Solid) and WL (Green, Solid) pairs at low aggregation level (crystals 1 and 2). The diseases show higher correlations at the higher wavelet crystals, larger levels of aggregation. For giardiasis, the correlation is low until wavelet crystal level 4, which corresponds to 16 days. For salmonellosis, low correlation was observed for the crystals with aggregation up to 64 days.

The correlation between the wavelet decomposition for each disease and the environmental characteristics were considered. Wavelet crystal levels 1 to 5 displayed only sporadic significant correlations (Table 1). However, higher levels of aggregation displayed more significant correlations. Table 2 shows the correlations between each disease and the environmental parameters for crystals 6 to 9. After Fourier smoothing a significant relationship is seen between giardiasis and precipitation at wavelet crystal levels 8 and 9 in Lowell; and between samonellosis and temperature at wavelet crystal level 6 in all three locations. For untransformed and loess-smoothed data, significant correlation was observed between giardiasis and temperature at wavelet crystal level 7 in Boston; between salmonella and temperature at wavelet crystal 8 in Boston and Worcester. The correlation values seen for loess smoothed data and untransformed data were similar.

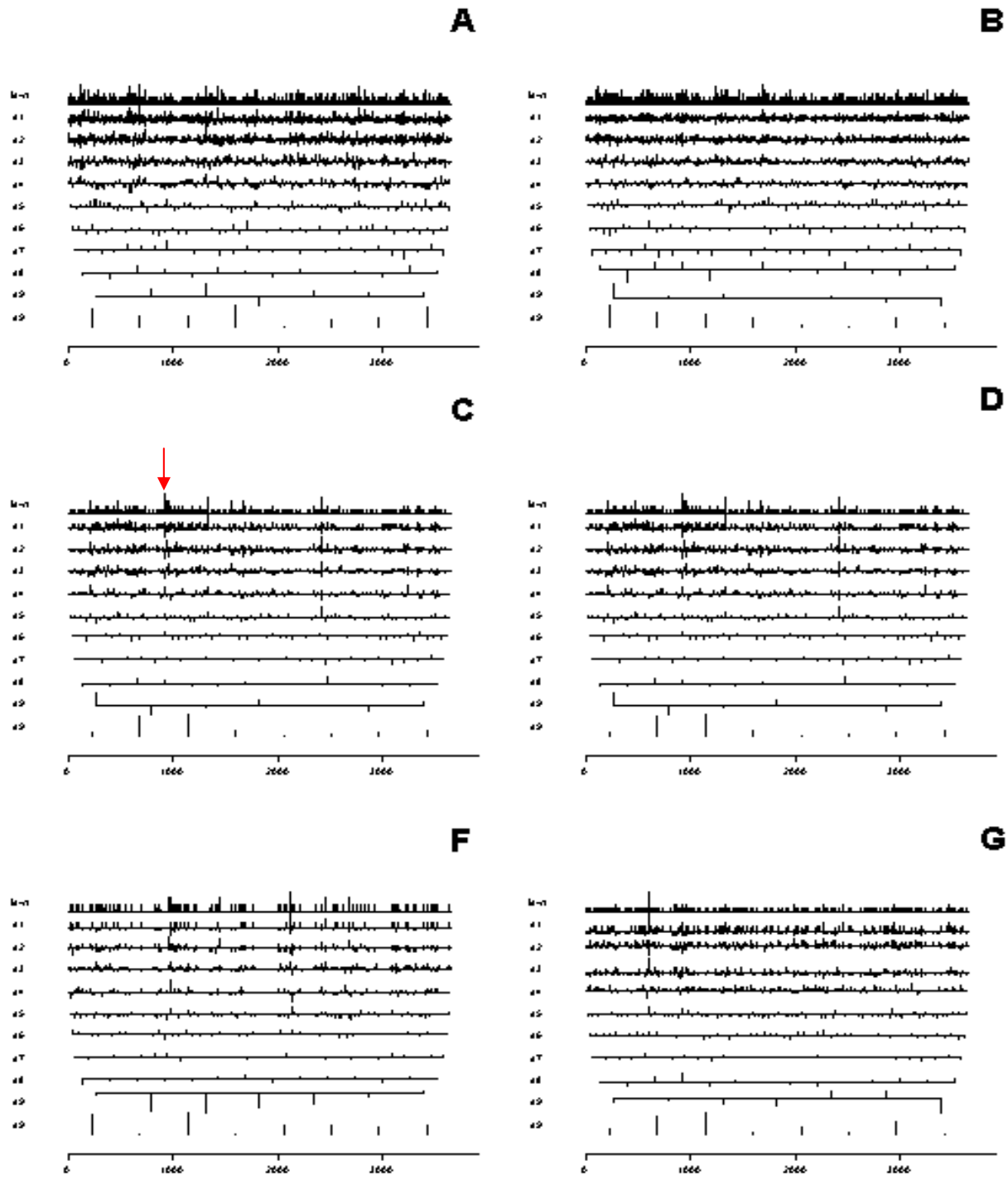

Figure 1. Discrete Wavelet Transforms (DWT) for untransformed data. A. Giardiasis in Boston, B. Salmonellosis in Boston, C. Giardiasis in Worcester, D. Salmonellosis in Worcester, E. Giardiasis in Lowell, F. Salmonellosis in Lowell. The mark in Panel C refers to a documented outbreak of giardiasis in Worcester in the summer of 1995.

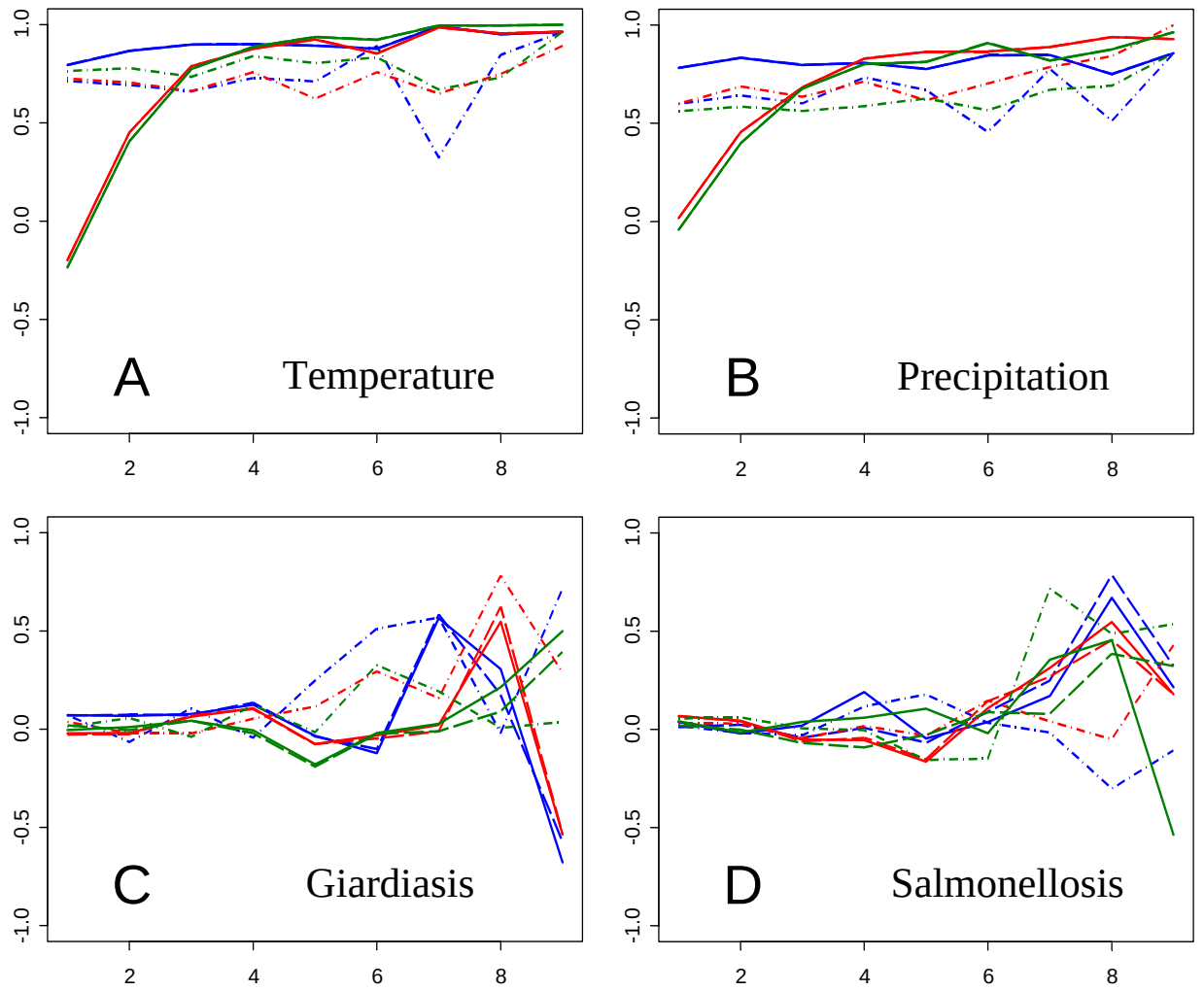

Figure 2. Spearman correlations between locations for transformed and untransformed data for each of 9 wavelet crystals. Untransformed data (Solid line), Loess Smoothed data (Dashed), Fourier Smoothed data (Dotted-Dashed). Correlation between Worcester and Lowell (WL, Green Line), Correlation between Boston and Lowell (BL, Red Line), Correlation between Boston and Worcester (BW, Blue).

Table 1. Significant correlations for wavelet crystals 1-5 between disease and environmental characteristics. Sample size varies for each crystal: W1 (n = 1826), W2 (n = 913), W3 (n = 456), W4 (n = 228), W5 (n = 114).

	Boston	**Worcester**	**Lowell**
	Giardia vs. Temperature		
Fourier	(w1) 0.056	(w1) 0.063	
	Giardia vs. Precipitation		
Fourier			(w3) 0.022
No Smoothing	(w5) -0.200		
	Salmonella vs. Temperature		
Loess			(w3) 0.116
No Smoothing	(w3) 0.116		
	Salmonella vs. Precipitation		
Fourier	(w1) 0.033	(w5) -0.191	(w2) 0.098
Loess	(w5) -0.189	(w5) -0.191	(w3) 0.101
No Smoothing			(w3) 0.101

Table 2. Correlations for various wavelet crystals between disease and environmental characteristics, those in bold are significant. Sample size varies for each crystal: W6 (n = 57), W7 (n = 29), W8 (n = 14), W9 (n = 7).

	Wavelet Level 6 (w6)			Wavelet Level 7 (w7)		
	Boston	**Worcester**	**Lowell**	**Boston**	**Worcester**	**Lowell**
	Giardia vs. Temperature					
Fourier	0.017	-0.024	0.225	0.351	0.160	-0.087
Loess	-0.144	0.108	0.068	**0.398**	0.242	0.035
No Smoothing	-0.157	0.123	0.070	**0.402**	0.266	0.029
	Giardia vs. Precipitation					
Fourier	-0.074	-0.234	-0.145	0.050	0.044	**0.483**
Loess	0.054	-0.028	0.098	-0.201	-0.132	0.306
No Smoothing	0.050	-0.014	0.110	-0.198	-0.063	0.317
	Salmonella vs. Temperature					
Fourier	**0.292**	**0.273**	**0.326**	0.047	0.436	-0.006
Loess	0.086	-0.041	0.144	0.249	0.239	0.342
No Smoothing	0.078	0.123	0.133	0.321	0.267	0.346
	Salmonella vs. Precipitation					
Fourier	**0.286**	0.017	-0.054	0.067	0.223	**0.385**
Loess	0.126	-0.061	-0.033	0.005	-0.088	-0.300
No Smoothing	0.127	-0.014	-0.052	-0.081	-0.064	-0.280

	Wavelet Level 8 (w8)			Wavelet Level 9 (w9)		
	Boston	**Worcester**	**Lowell**	**Boston**	**Worcester**	**Lowell**
	Giardia vs. Temperature					
Fourier	0.284	**0.569**	0.125	0.750	0.357	0.607
Loess	0.442	**0.622**	0.248	0.178	0.107	-0.286
No Smoothing	0.473	**0.596**	0.200	0.178	0.071	-0.285
	Giardia vs. Precipitation					
Fourier	0.446	0.103	**0.631**	0.678	0.750	**0.785**
Loess	-0.204	0.354	0.055	-0.107	0.000	-0.321
No Smoothing	-0.235	0.354	0.165	-0.107	0.071	-0.321
	Salmonella vs. Temperature					
Fourier	0.446	0.402	0.020	0.321	0.714	0.750
Loess	**0.938**	**0.785**	0.459	**0.892**	0.428	0.071
No Smoothing	**0.925**	**0.596**	0.455	**0.892**	0.071	0.357
	Salmonella vs. Precipitation					
Fourier	0.393	0.213	-0.125	0.500	0.071	0.750
Loess	-0.367	0.116	-0.288	0.035	0.714	0.357
No Smoothing	-0.437	0.354	-0.266	0.036	0.071	0.071

Discussion

The examination of any environmental influence on disease counts involves countless levels of complexity. When examining water- or food-borne illnesses, for which the indirect influence of environmental conditions can have as great an impact on disease incidence as the direct effects, it is especially difficult to discover whether or not minor fluctuations play a significant role. On a larger scale these local fluctuations are themselves seasonally governed, making the task of isolating their potential affects, apart from annual oscillation, even more challenging. Any attempt to correct for well-understood patterns of oscillation runs the risk of obfuscating the impact of smaller-scale influences. In order to minimize this risk, we have chosen to employ two different mathematical techniques of correcting for periodic oscillation in the hopes that each would mask different possible influences; it is still possible that all of the global corrections employed share in common the elimination of relevant environmental fluctuations, hence we also analyzed the data without correcting for seasonal trends as well.

Both giardiasis and salmonellosis show higher correlations at the higher wavelet crystals (larger levels of aggregation). This implies a greater variation at smaller levels of aggregation and less variability with increased aggregation, which is exactly what could be expected from our understanding of seasonal variation of disease patterns in New England. It should be noted that the higher correlation for greater levels of aggregation (high crystal levels) can be thought of as revealing trends on a much greater temporal scale. For giardiasis, the correlation is low for wavelet crystal levels 1 through 4, which correspond to temporal scales up to 16 days. Since the incubation period for giardiasis is known to average 14 days, these are exactly the crystal levels we would expect to exhibit significant correlation to environmental conditions. For salmonellosis, low correlation was observed for the crystals with aggregation up to 64 days. The correlations seen for the loess smoothed data and the untransformed data were very similar (Table 2), suggesting loess smoothing did not sufficiently remove the seasonal patterns.

Since Lowell is located 30 miles north of Boston, which is in turn 44 miles east of Worcester, each experiences significantly different weather patterns. This can be seen in the low aggregation level (in crystals 1 and 2) in the untransformed data in the Boston-Lowell and Worcester-Lowell correlation pairs (Figure 2). The weather patterns of Boston and Worcester display high correlation throughout the wavelet crystals. The mark in Figure 1C denotes a documented outbreak of giardiasis in Worcester in the summer of 1995 (Naumova et al., 2000). Multiple spikes in the wavelet coefficients for the time period of this outbreak are observed for crystals 1 to 4. The figures show that potential outbreaks can be seen through the various wavelet crystals.

Our goal has been to present the potential of implementing a two stage wavelet analysis in order to isolate the influence of environmental factors on disease incidence on a variety of temporal scales. As a result, many aspects of the implementation examined were simplified for clarity of presentation. It may therefore be the case that greater detail in each specific analysis would yield results more clearly supporting or rejecting the influence of various environmental factors on each scale. The more crucial of these aspects include the selection of an appropriate basis function for the wavelet. In this study, we have relied on the default basis for our v-spline2 wavelet, however this choice may well significantly affect the outcome of the wavelet decomposition, significantly influencing our correlation analyses. An appropriate window size for the loess smoothing process is crucial to the effective removal of underlying seasonal patterns. Varying window sizes prior to wavelet decomposition would allow for the selection of a window which would provide a cleaner correction of periodic oscillation. In performing the Fourier decompositions, we chose in this study to ignore boundary effects. The Fourier decomposition produces large values at the edges of the signal. In this analysis Spearman correlations were used to reduce their effect however, further research into the elimination of these boundary effects is necessary if we are to employ the resultant findings in any meaningful way. Lastly, an understanding of the sensitivity of this method to ranges of outliers (outbreaks) or locally increased variation for varying lengths of time would lend considerably greater credibility to any interpretation of its results.

A more fundamental consideration in the application of this two-stage process lies in the form of the data itself. Because the shape of seasonal oscillations can be dissimilar from those of localized fluctuations, the influence of the method used to correct for this periodicity may differentially affect the various fluctuations within a single smaller scale. While it has been postulated that the incidence of some diseases may well be fractal in nature, reporting of these diseases is almost certainly not. In weighing the potential benefit of correcting for

confounding periodic variation, the level of self-similarity should be considered and the possible implications of this differential effect should be noted in any interpretation of decomposed residuals.

While no method can be said in certainty to remove the risk of masking influential effects, the process of wavelet analysis applied to environmental data *without* seasonal oscillations can provide greater insight into their direct affects on disease incidence. Wavelets are a powerful tool which can be used to determine potential outbreaks in surveillance data for waterborne and food-borne enteric infections.

References

Barwick RS. Levy DA. Craun GF. Beach MJ. Calderon RL. 2000. Surveillance for waterborne-disease outbreaks--United States, 1997-1998. *MMWR. CDC Surveill Summ. 49(4):1-21.*

Bean NH. Goulding JS. Lao C. Angulo FJ. 1996. Surveillance for foodborne-disease outbreaks--United States, 1988-1992. *MMWR. CDC Surveill Summ.45(5):1-66.*

Bentham G. Langford IH. 2001. Environmental temperatures and the incidence of food poisoning in England and Wales. *Int J Biometeorol. 45(1):22-6.*

Birkhead G. Vogt RL. 1989. Epidemiologic surveillance for endemic Giardia lamblia infection in Vermont. The roles of waterborne and person-to-person transmission. *Am J Epidemiol. 129(4):762-8.*

CDC, Centers for Disease Control and Prevention. 2003. Surveillance data from Swimming Pool Inspections Selected States and Counties, United States, May-September 2002. *MMWR 52 (22) 513-6.*

Curriero FC. Patz JA. Rose JB. Lele S. 2001. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994. *Am J Public Health. 91(8):1194-9.*

Furness BW. Beach MJ. Roberts JM. 2000. Giardiasis surveillance--United States, 1992-1997. *MMWR. CDC Surveill Summ. 49(7):1-13.*

Johansen A. 1996. A Simple Model of Recurrent Epidemics. *J Theor Biol.* 178(1):45-51

Lee SH. Levy DA. Craun GF. Beach MJ. Calderon RL. 2002. Surveillance for waterborne-disease outbreaks--United States, 1999-2000. *MMWR. CDC Surveill Summ. 51(8):1-47.*

Naumova EN. Chen JT. Griffiths JK. Matyas BT. Estes-Smargiassi SA. Morris RD. 2000. Use of passive surveillance data to study temporal and spatial variation in the incidence of giardiasis and cryptosporidiosis. *Public Health Rep. 115(5):436-47.*

Olsen SJ. MacKinnon LC. Goulding JS. Bean NH. Slutsker L. 2000. Surveillance for foodborne-disease outbreaks--United States, 1993-1997. *MMWR. CDC Surveill Summ. 49(1):1-62.*