Regional Compartmental Epidemiological Modeling of Influenza in the United States with Wavelet Analysis

Anna Carmack

Biological Sciences
Applied and Computational Mathematics and Statistics
University of Notre Dame

Dr. Benjamin Ridenhour

April 23, 2014
Abstract

Each year, between 5% and 20% of United States residents contract seasonal influenza. Most epidemiological models depend on influenza-like illness (ILI) reports, collected weekly by the CDC from outpatient practices, resulting in noisy and incomplete counts for influenza across the country. To obtain a more accurate picture of seasonal influenza, wavelet analysis was employed to filter the available ILI data for each region of the United States and find the wavelet deviation best correlating with regional positive lab tests (an indicator of a true presence of influenza). That wavelet deviation was then fit to an SEIR compartmental model using a Markov chain Monte Carlo selection method to determine epidemiological parameters such as basic reproduction number ($R_0$) and generation time ($t_{\text{gen}}$) for each region during the influenza seasons. The average basic reproduction number was found to be 1.379 (1.34, 1.39), and the average generation time was found to be 2.563 days (2.45, 2.68). Using the nine US Census regions over ten years, the SEIR model fit to the wavelet analysis output estimated 26.30% (23.83%, 28.76%) of ILI regional reports to be true cases of influenza. Region and influenza season (year) were found to have little to no significance on influenza epidemic parameters or ILI proportion estimates, suggesting similar influenza dynamics across the United States.
1 Introduction

Seasonal influenza continues to affect a large portion of the United States population (estimated at 5-20%[1]), affecting health, productivity, and the economy. The economic burden of annual influenza epidemics has been estimated to be as high as $87.1 billion[2]. Though a vaccine is made available every year, usage varies from 35.8% to 69.3% (depending on the age group)[3], and only certain strains are included. Additionally, more than 200,000 are hospitalized in the US for influenza-related illness each year[4], and annual mortality counts for influenza-associated deaths range from 3,349 to 48,614[5].

However, these analyses are often completed retrospectively, examining all sources of data and attempting to find the true incidents of influenza, influenza-associated hospitalizations, and influenza-associated deaths. Suspected cases are reported each week to the Centers for Disease Control (CDC), but it is difficult to determine the actual cause of each incident. Without such knowledge, epidemics are over- and under-estimated, public health campaigns are mis-managed, and influenza continues to propagate around the United States.

Currently, the CDC uses Robert Serfling’s seasonality-based periodic regression model to analyze mortality as an epidemic determinant[6]. 122 cities across the United States contribute to the Mortality Reporting System, providing death counts and causes attributed to those deaths. When the percentage of deaths attributed to pneumonia and influenza (P&I) cross the 95% confidence interval of the periodic model (adjusted with a time-series trend) seasonal influenza is said to have crossed the “epidemic threshold.” This determination, along with virological surveillance, outpatient surveillance, hospitalization surveillance, and geographical reports combine to estimate the prevalence of influenza, attempting to ignore the noise of inaccurate reporting, incomplete reporting, and lack of reporting. Other proposals for influenza monitoring include Google search histories[7], Twitter content[8], and a variety of mathematical models[9][10][11].

Wavelet analysis provides an alternative simpler method to eliminate the noise in influenza reports and extract usable data. Influenza already naturally follows a wave pattern because of its seasonality (increasing in the winter months). Using the discrete wavelet transform of the data (finding the signal–actual influenza–through the noise–poor reporting), wavelet analysis breaks down the data into its wavelet coefficients. Correlation tests with influenza indicators (positive laboratory tests, for example) can then be used to determine the set of coefficients most representative of true influenza.

These methods can then be applied to regional data across the United States to determine if influenza epidemiological parameters differ by region. Differences could be caused by population density, geographical features, climate, and social interaction patterns. This knowledge could better inform regional and national influenza prevention and responses.
Additionally, aggregating the different regional estimates from this data could provide an accurate picture of influenza, including rather than obscuring the regional trends that make up the national influenza dynamics.

2 Methods

2.1 Influenza Data

The data for this work was taken from a variety of different sources across the United States. The original data, influenza-like illness (ILI) reports, was taken from the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). An influenza-like illness is defined as a fever of \(100^\circ F(37.8^\circ C)\) with a cough and/or sore throat (and without a known cause other than influenza)\[12\]. If all requirements are met, the case is reported to ILINet, consisting of over 2,900 providers across all 50 states. Approximately 1,800 of those providers report data regularly, including the total number of patients seen that week and the number of those patients presenting with an ILI.

Additionally, providers can send samples to state laboratories to confirm a diagnosis of influenza. All state public health laboratories report those results to the World Health Organization (WHO) Collecting Laboratories or the National Respiratory and Enteric Virus Surveillance System (NREVSS). Laboratory results are then collected by the Centers for Disease Control (CDC) and released a week later. The 122 Cities Mortality Reporting System weekly reports were also used, specifically pneumonia and influenza (P&I) related deaths and total mortality counts (all causes).

Population data was taken from the United States Census Bureau - U.S. Department of Commerce. 2003 through 2008 populations were estimated using 2008 releases\[13\]. 2009 population was estimated using 2009 releases\[14\]. 2010 through 2013 populations were estimated using 2013 releases\[15\]. Regions were determined by US Census Divisions: New England (NE), Middle Atlantic (MA), East South Central (ESC), West South Central (WSC), South Atlantic (SA), East North Central (ENC), West North Central (WNC), Mountains (MTN), and Pacific (PAC).

This model only incorporates data within the standard influenza season, week 40 in the fall through week 20 in the spring of the following year, barring the 2008-2009 season and 2009-2010 season to account for the H1N1 pandemic. The 2008-2009 influenza season time-series was ended early to avoid the inclusion of H1N1 pandemic cases, and the 2009-2010 influenza season started earlier than typical.
2.2 Wavelet Analysis

The wavethresh package in R[16] was employed to decompose the available ILI data using a non-decimated transform. The cthresh command repeatedly divided the data into a smooth curve and the removed coefficient details, yielding 9 resolutions \((f_J(x)[17])\), where \(J = 0, 1, ..., 8\) and \(c_{J,k}\) identifies the father wavelet coefficients. At resolution \(J\), the wavelet decomposition is given by

\[
\phi_{J,k}(x) = \begin{cases} 
2^{J/2} & x \in [2^{-J}k, 2^{-J}(k + 1)], \\
0 & \text{otherwise.}
\end{cases}
\]

\[
f_J(x) = \sum_{k=0}^{2^J - 1} c_{J,k} \phi_{J,k}(x)
\]

where \(f_J(x)\) is the wavelet approximation of \(f(x)\) including resolutions \(0 - J\). With the wavelet decomposition, 9 \(f_J(x)\) wavelets were constructed in R using the nullevels command, selecting for the inclusion of each finer resolution level \((J)\). Each of these wavelet approximations was then removed from the original ILI reports \((f)\), creating 9 possible wavelet deviations \((\delta_J(f))\) of varying detail; any negative deviations were set to 0. Thus, the final equation for the wavelet deviation is given by

\[
\delta_J(f) = \text{argmax}(0, f - f_J(x))
\]

The same procedure was applied to data from the 122 Cities Mortality Reporting System, using both pneumonia and influenza (P&I) related deaths and total mortality counts (all causes) to confirm the appropriate wavelet deviation was selected.

2.3 Correlation

Each wavelet deviation \((\delta_J(f))\) was then compared to the percentage of lab tests confirmed positive in that region over the course of the 10 influenza seasons. Wavelet deviations were ranked by correlation with the lab results, and the ranks were averaged across years to determine the set of coefficients that best described all 10 influenza seasons and 9 regions.

2.4 SEIR Fit

An SEIR (Susceptible, Exposed, Infected, Recovered) compartmental model was used to mimic traditional influenza dynamics. In such models, individuals are tracked as they move from class to class, dependent on their interactions with infected individuals and the characteristics of their current classification.
\[
\frac{dS}{dt} = -\beta SI \\
\frac{dE}{dt} = \beta SI - \epsilon E \\
\frac{dI}{dt} = \epsilon E - \gamma I \\
\frac{dR}{dt} = \gamma I
\]

Individuals leave the susceptible class at rate $\beta$ after contact with an infected individual, assuming mass action kinetics in the system, and move into the exposed class. From the exposed class, individuals then progress to the infected class at rate $\epsilon$ and subsequently to the recovered class at rate $\gamma$.

These equations were then manipulated to extract the typical influenza parameters, such as $R_0$ and generation time. The basic reproductive number, $R_0$, relates $\beta$ and $\gamma$ ($R_0 = \frac{\beta}{\gamma}$), describing the number of new infections caused by each infected individual. An $R_0$ greater than 1 suggests the capability to cause an epidemic. The generation time, $t_{gen}$, indicates the number of days between onset of infectiousness for the original individual in the infected class and the onset of infectiousness for individuals one infects in the susceptible class, estimated at $\frac{1}{\epsilon} + \frac{1}{\gamma}$. A ratio term was added to mimic the dependency between the two individual parameters ($\epsilon$ and $\gamma$). Additionally, a scale term was used to account for the expected under-reporting of influenza.

Initial epidemic parameters ($R_0$, $t_{gen}$, the ratio term, and the scale term) were chosen based on a visual estimation of fit, and the model progressed as an adaptive proposal density[18] with a random walk (Markov chain Monte Carlo method)[19][20]) to find the parameters of best fit using the deSolve package in R[21]. A set of parameters would be accepted only if their aggregated estimate (the SEIR model has daily steps while the ILI reports and lab tests are weekly) multiplied by the population and incorporating the scale term had a residual sum of squares (in reference to the original wavelet deviation) less than that of the previous set of parameters, employing thinning in later generations to decrease autocorrelation. After 1000 generations of the random walk were accepted, the sampled parameters were extracted along with their distributions. Parameter sets with 4 normal distributions (suggesting an unbiased random walk) were chosen to model the epidemic.

This model was run for each region and influenza season included in the original data set (9 regions and 10 influenza seasons). Setting the initial conditions to the proportions equivalent to one exposed individual and the remaining individuals (average population over the two years included in that influenza season)
in the susceptible class, the model was run repeatedly until the accepted parameters’ distributions all followed normal distributions. This was done for each of the 90 region/influenza season ILI data sets using their wavelet deviations (See Appendix for sample posterior distributions of the epidemic parameters).

2.5 Statistical Analysis

Using the variance of the posterior distributions of the estimated epidemic parameters, the rma command from the metafor package in R created a linear regression model. The two categorical moderators, region and influenza season, were included to determine the significance of regional and seasonal influence on \( R_0 \), \( t_{gen} \), the ratio term, and the scale term. SEIR estimates were also compared to the original ILI reports, determining the proportions of ILI reports that can be classified as true cases of influenza. The lm command (stats in R) was used to determine regional and seasonal influence on the proportions.

3 Results

3.1 Data Manipulation

When applied to the 10 years of ILI reports, the wavelet analysis separated the seasonal data into 9 wavelet decompositions. Subtracting these wavelets from the original ILI data, each of the 9 wavelet deviations was compared to the regional positive lab tests (positive lab tests are indicative of a confirmed presence of influenza in the population, minimizing premature reports), and the wavelets were ranked according to their correlation to lab test results. The ILI report resulting wavelet deviation most reflective of positive lab tests was found to be the pattern remaining when the values in resolutions 0, 1, and 2 were removed from the original ILI reports. What remains were the coefficients representing resolutions 3-8, the equivalent of \( \delta_2(f) \). This wavelet deviation has been referred to as dev3 in the following results. An example can be seen in Figures 1 and 2, where the wavelet coefficients are seen for all resolutions in ENC, followed by the manipulation of ENC reports over the course of the ten influenza seasons.

The choice of the dev3 wavelet deviation was also confirmed by comparison of the weekly reported P&I-related deaths. Submitting the P&I data to wavelet decomposition, the dev3 wavelet deviation was again found to be most reflective of positive lab tests. The total mortality (all causes) reports were not reflected best in dev3, suggesting the trends seen in dev3 to be influenza-related and not simply seasonal disease trends.

3.2 SEIR Fit

For the 90 region/year data sets, \( R_0 \) values ranged from 1.18 to 2.23, with an average value of 1.37 (1.34, 1.39). The generation time \( (t_{gen}) \) values ranged from 2.00 to 4.79, with an average value of 2.57 (2.45, 2.68). Ratio values
Figure 1: Relative wavelet coefficients for ENC ILI reports at the baseline and 8 resolutions. The selected dev3 deviation removed the resulting wave patterns from resolutions 0, 1, and 2.

ranged from 0.37 to 0.65, with an average value of 0.50 (0.48, 0.52). Scale values ranged widely from 343.38 to 17873.78, with an average value of 2598.74 (1945.99, 3251.49). Distributions of accepted parameters can be found in the Appendix.

3.3 Meta-Analysis

From the SEIR models, 90 sets of parameters were extracted, along with their sample distributions of accepted values. Using the \texttt{rma} command in the R package \texttt{metafor}\cite{22}, a mixed regression was used to determine the significance of influenza season and region on each parameter, resulting in four sets of significance values, displayed in Table 1. The 2010-2011 influenza season and the ENC region were set as reference levels in dummy coding. For $R_0$, the 2003-2004 season, the 2008-2009 season, the 2009-2010 season, and the 2012-2013 season were found to be statistically significantly (p-value $\leq 0.05$) influencing the parameter. No regions had statistically significant influences. The $t_{\text{gen}}$ parameter had relatively similar significant influences in influenza season; however, the ESC
Figure 2: Wavelet manipulation for ENC over the 10 year period begins at the original ILI reports. The aggregated wavelet is then removed from the ILI reports, leaving resolutions 3–8 coefficients intact. Negative cases are then removed to yield the $\text{dev}_3$ wavelet deviation.

region was found to be statistically significantly influencing $t_{\text{gen}}$. The ratio parameter showed no significant influence for any factors, and the scale parameter was only significantly influenced by the MA region. The linear model’s intercept was found to be statistically significant for all parameters. While yearly trends have a slight influenza on influenza dynamics parameters, regions have little to no influence on those same parameters.

The regression models accounted for 61.51%, 28.65%, 4.47%, and 8.15% of heterogeneity ($R^2$) for $R_0$ values, $t_{\text{gen}}$ values, the ratio terms, and the scale terms, respectively.
Table 1: When linear regression models are constructed to determine the significance of influenza season and region (both categorical variables), very little statistically significant influence was observed, suggesting similar influenza dynamics across the United States every year.

<table>
<thead>
<tr>
<th>Moderator</th>
<th>$R_0$ p-value</th>
<th>$t_{gen}$ p-value</th>
<th>ratio p-value</th>
<th>scale p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0***</td>
<td>0***</td>
<td>1e-04***</td>
<td>0***</td>
</tr>
<tr>
<td>2003-2004</td>
<td>0***</td>
<td>0.0038**</td>
<td>0.4678</td>
<td>0.0964*</td>
</tr>
<tr>
<td>2004-2005</td>
<td>0.1733</td>
<td>0.119</td>
<td>0.9233</td>
<td>0.0933*</td>
</tr>
<tr>
<td>2005-2006</td>
<td>0.5275</td>
<td>0.946</td>
<td>0.3745</td>
<td>0.1123</td>
</tr>
<tr>
<td>2006-2007</td>
<td>0.3037</td>
<td>0.8844</td>
<td>0.7156</td>
<td>0.0515*</td>
</tr>
<tr>
<td>2007-2008</td>
<td>0.2077</td>
<td>0.8839</td>
<td>0.7462</td>
<td>0.782</td>
</tr>
<tr>
<td>2008-2009</td>
<td>5e-04***</td>
<td>0.0203*</td>
<td>0.6007</td>
<td>0.5642</td>
</tr>
<tr>
<td>2009-2010</td>
<td>0***</td>
<td>0.4989</td>
<td>0.6599</td>
<td>0.6755</td>
</tr>
<tr>
<td>2011-2012</td>
<td>0.2493</td>
<td>0.0999*</td>
<td>0.8283</td>
<td>0.2558</td>
</tr>
<tr>
<td>2012-2013</td>
<td>3e-04***</td>
<td>0.7037</td>
<td>0.9967</td>
<td>0.115</td>
</tr>
<tr>
<td>ESC</td>
<td>0.0511</td>
<td>0.0233*</td>
<td>0.1149</td>
<td>0.5151</td>
</tr>
<tr>
<td>MA</td>
<td>0.7649</td>
<td>0.2243</td>
<td>0.0825*</td>
<td>0.0068**</td>
</tr>
<tr>
<td>MTN</td>
<td>0.4987</td>
<td>0.2827</td>
<td>0.9705</td>
<td>0.1032</td>
</tr>
<tr>
<td>NE</td>
<td>0.2065</td>
<td>0.6805</td>
<td>0.2172</td>
<td>0.5706</td>
</tr>
<tr>
<td>PAC</td>
<td>0.6128</td>
<td>0.0615*</td>
<td>0.4547</td>
<td>0.1618</td>
</tr>
<tr>
<td>SA</td>
<td>0.3744</td>
<td>0.1676</td>
<td>0.6608</td>
<td>0.6032</td>
</tr>
<tr>
<td>WNC</td>
<td>0.2669</td>
<td>0.6435</td>
<td>0.519</td>
<td>0.3031</td>
</tr>
<tr>
<td>WSC</td>
<td>0.5107</td>
<td>0.9268</td>
<td>0.1962</td>
<td>0.8952</td>
</tr>
</tbody>
</table>

3.4 ILI Proportions

When the estimated parameters were returned to the SEIR model with each of the 90 (region and influenza season) initial conditions of proportions equivalent to 1 exposed individual and the remainder of the population in the susceptible class, counts of true influenza cases could be approximated. The percentage of influenza cases captured under the SEIR model compared to those originally reported as ILI ranged from 3.77% (NE in the 2011-2012 influenza season) to 58.05% (WSC in the 2004-2005 season). A histogram of all proportions can be seen in Figure 3. The average proportion of ILI reports included in the SEIR estimate by region/influenza season was 26.30% (23.83%, 28.76%).

Aggregating the number of cases estimated over each season (summing all 9 regions) and comparing to the reported ILI counts, proportions ranged from 9.95% (2011-2012 season) to 37.89% (2007-2008 season). On average, 25.14% (18.93%, 31.35%) of ILI reports were included in the aggregated SEIR estimate of influenza each year. Overall influenza estimates ranged from 39,605 (2011-2012 season) to 225,057 (2009-2010). Within the 10-year period, the models estimated an average of 101,951 (41,728, 208,878) people become infected with influenza each year out of the average 410,922 (254,627, 567,218) ILI reports.
Figure 3: The proportion of ILI reports estimated to be true cases of influenza ranged from 3.77% to 58.05%, with an average proportion of 26.30% (23.83%, 28.76%). The distribution of those proportions fell on a relatively normal distribution.

The collection of SEIR estimates can then be compared to the 10-year ILI reports for each region (Figures 4 & 5). While all 90 estimates identified at least a small peak, some region/influenza season cases were more reflective of ILI reports than others. It was also noted that the Pacific region had a particularly low incidence when compared to the size of its population. A linear regression model was used to determine if regions or seasons were over-influencing the proportion of ILI reports estimated to represent true influenza cases. At a significance value of 0.05, only the 2007-2008 influenza season had a significant influence on the proportion.
### Flu Season SEIR Estimates ILI Reports Percentage

<table>
<thead>
<tr>
<th>Flu Season</th>
<th>SEIR Estimates</th>
<th>ILI Reports</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003-2004</td>
<td>62711</td>
<td>208728</td>
<td>30.04%</td>
</tr>
<tr>
<td>2004-2005</td>
<td>91624</td>
<td>248031</td>
<td>36.94%</td>
</tr>
<tr>
<td>2005-2006</td>
<td>53128</td>
<td>221525</td>
<td>23.98%</td>
</tr>
<tr>
<td>2006-2007</td>
<td>54624</td>
<td>248949</td>
<td>21.94%</td>
</tr>
<tr>
<td>2007-2008</td>
<td>150432</td>
<td>397063</td>
<td>37.89%</td>
</tr>
<tr>
<td>2008-2009</td>
<td>49042</td>
<td>319823</td>
<td>15.33%</td>
</tr>
<tr>
<td>2009-2010</td>
<td>225057</td>
<td>883925</td>
<td>25.46%</td>
</tr>
<tr>
<td>2010-2011</td>
<td>153153</td>
<td>556223</td>
<td>27.53%</td>
</tr>
<tr>
<td>2011-2012</td>
<td>39605</td>
<td>398158</td>
<td>9.95%</td>
</tr>
<tr>
<td>2012-2013</td>
<td>140135</td>
<td>626799</td>
<td>22.36%</td>
</tr>
</tbody>
</table>

Table 2: Aggregating the ILI reports and SEIR estimates across all 9 regions each year, 25.14% (18.93%, 31.35%) of ILI reports were predicted to be true cases of influenza. The overlap with the regional proportion confidence level (23.83%, 28.76%) supports the apparent homogeneity of influenza across the United States.
Figure 4: SEIR estimates can be compared to the original ILI reports, covering week 40 to week 20 of each influenza season on a standardized scale of cases per 100,000 residents.
Figure 5: SEIR estimates can be compared to the original ILI reports, covering week 40 to week 20 of each influenza season on a standardized scale of cases per 100,000 residents.
4 Discussion

Regardless of vaccination campaigns and public health awareness, seasonal influenza continues to have significant effects on the United States and countries around the world. One of the biggest challenges in management and prevention is determining accurate counts of influenza. While ILI reports are suggestive of the presence of influenza, they are by no means complete. The realization and quantitative estimation of the presence of influenza in a community comes from a compilation of a variety of factors: ILI counts, mortality reports, laboratory results, and regional reports. However, a simpler method would require less information and still yield an accurate prediction in a timely matter. Wavelet analysis aims to do just that. While not estimating total cases across the United States, this method clearly estimates the composition of ILI reports (true influenza or excess reports).

Positive laboratory test results for influenza were chosen as the baseline when determining the accurate wavelet deviation for ILI counts because of the ability to detect influenza presence. Though not all samples are sent to the laboratory, a positive lab test confirms the presence of influenza in the community. While the lack of positive lab tests does not, conversely, confirm the absence of influenza in the community, it suggests a less prominent presence or possible total absence. The figure below (Figure 6) compares the original ILI reports time series with the positive laboratory test time series. From these two time series, the wavelet deviation ($\delta_{J}^*_{3}$) was derived, the base of the SEIR model. By simplifying the necessary input to weekly ILI reports and laboratory results, wavelet analysis presents a simpler method to estimate the influenza presence across the United States.

As noted, the wavelet deviation ($\delta_{J}^*_{3}$) best reflecting the positive lab tests in the ILI reports was found to be the deviation when the coefficients of resolutions 0, 1, and 2 (the largest wave patterns) were removed, $\delta_{J}^*_{3}$. Though this could have been insignificant, the same wavelet deviation was determined to be reflective of positive lab tests in the P&I deaths (from the 122 Cities Mortality Report), which follows the same seasonal patterns of influenza. Additionally, total deaths (all causes) did not correlate with the same wavelet deviation, supporting the significance of the wavelet deviation $\delta_{J}^*_{3}$ and its inclusion of seasonal trends specific to influenza.

The SEIR compartmental model was then fit to the 33-week periods taken from each region’s wavelet deviation using a random walk method. On the regional level, $R_0$ values ranged from 1.18 to 2.23, with an average value of 1.37 (1.21, 1.67), similar to previous estimates of 1.3 for seasonal influenza in the United States[23]. Generation time values, however, were lower than previous estimates of 3.6 days[24], ranging from 2.00 days to 4.79 days, with an average value of 2.56 (2.00, 3.79) days. Though lower, generation time is more difficult to estimate when only weekly, rather than daily, reports are available, and,
Additionally, the reports are distorted because of the assumed under-reporting. With an estimated ratio of 0.5, approximately 1.3 days are spent in both the exposed and infected classes before recovery. The scale term shows the biggest range (343.38, 17873.78), accounting for the expected under-reporting.

When the parameters were compared in a mixed regression model, season (as a categorical variable) displayed little to no statistically significant (p-value ≤ 0.05) influence, and region (a categorical variable as well) had even less influence. While these results would be unexpected when considering the varying population distributions, unique social structures, and climates of each of the 9 regions included in this study, the apparent homogeneity of influenza dynamics across the United States simplifies public health efforts when tracking and reacting to influenza. Regional models will continue to be necessary to address local epidemic peaks and small-scale reactionary efforts (vaccination, education campaigns, etc.), but influenza can also apparently be effectively monitored on the large-scale national level. While this study did not investigate wavelet analysis of national ILI reports and laboratory results, future work could explore the similarities and differences seen with results from the national level.

Those same parameters, taken from the SEIR fit to the dev3 wavelet deviations, were fed back into the initial conditions (one individual categorized as exposed at week 20 and the remainder of the region’s individuals susceptible, proportionally). At the region/influenza season level, 26.30% (23.83%, 28.76%) of ILI reports were estimated to be true cases of influenza. Aggregating across the nine regions each year, 25.14% (18.93%, 31.35%) of ILI reports were estimated to be true cases, or an average of 101,951 (41,728, 208,878) cases out of the average 410,922 (254,627, 567,218) annual reports. The linear regression model found no region to have a significant influence on that proportion, echoing the similarity between the region/influenza season level of analysis and the nationwide aggregate estimate. This number is predictably lower than estimates of the overall incidence of influenza in the United States each year (between 5% and 20%); these case estimates are only proportions of the reported suspected cases (suggesting a need for improvement upon reporting techniques, determining standards more reflective of a virologically confirmed case of influenza). There are many cases that are never reported and are not estimated using this method. However, there is great potential for the analysis of ILI reports and laboratory results.

A scale term was included to account for under-reporting, but its full potential was not employed in this analysis. While this work focused on the composition of ILI reports, the calculated scale terms could theoretically be removed when the estimated parameters are fed back into the SEIR model. This would yield estimates of influenza in the total US population, rather than in the suspected cases (ILI reports) as done here.

Additionally, it was noted in the ILI reports and laboratory results that the
2009 H1N1 pandemic was clearly out of cycle with the regular seasonal influenza pattern across the 10 year reports, especially in the MA region. The wavelet deviation also contained the sudden peak, but such data was excluded from this analysis because of the timing following the regular 2008-2009 seasonal influenza and the pandemic nature of the strain. Instead, the 2008-2009 season was cut short when modeling to ignore the climbing ILI reports, and the 2009-2010 season was extended at the beginning to accommodate the earlier presence of seasonal influenza following the H1N1 pandemic. While an SEIR model could, predictably, be fit to the pandemic peak, the parameters found were ignored when determining average values and distribution to prevent the undue influence of pandemic dynamics.

This work only begins to explore the potential uses of wavelet analysis for influenza and other disease modeling, and potential future studies have been suggested. Most importantly, similar work needs to be done using the national level of data, including ILI reports and laboratory results. Regional influences on influenza dynamics appear to be minimal, and national estimates should, therefore, follow similar parameters and proportions. Our aggregate estimates can incorporate regional patterns, but it is possible such patterns are lost when all ILI reports are considered at once. However, this can not be confirmed without further application of wavelet analysis. Results could also be improved by considering the sub-typing of influenza laboratory results. All influenza type laboratory tests were grouped in this application of wavelet analysis and SEIR fit, but fitting separate wavelet deviations for different sub-type results would break down ILI reports into not only true influenza cases but also the sub-type of those cases. Additionally, this method was only employed for seasonal influenza. While pandemics are rarer, the presence of the 2009 H1N1 peak in the wavelet deviation suggests the possible employment of wavelet analysis when working with non-seasonal influenza or other non-cyclical disease patterns. Further explorations of wavelet analysis could also include the reliability of real-time use to determine incidence and quickly react to an unexpected or early peak. The dev3 wavelet deviation could adjust as the most recent ILI reports and laboratory results are released each week. While this application focused on seasonal influenza in the United States, wavelet analysis provides many possible areas of exploration in epidemiology and other time-series based data sets.
Figure 6: Comparison of the ILI reports and positive laboratory results in one year (NE 2007-2008) yields the dev3 wavelet deviation and the compartmental model fit.
References


Appendices

Figure 7: Example posterior distribution of accepted $R_0$ values in MA 2005-2006 SEIR fit
Figure 8: Example posterior distribution of accepted $t_{\text{gen}}$ values in MA 2005-2006 SEIR fit
Figure 9: Example posterior distribution of accepted ratio values in MA 2005-2006 SEIR fit
Figure 10: Example posterior distribution of accepted scale values in MA 2005-2006 SEIR fit
Figure 11: Distribution of estimated $R_0$ values for the 90 region/season influenza epidemics
Figure 12: Distribution of estimated $t_{gen}$ values for the 90 region/season influenza epidemics
Figure 13: Distribution of estimated ratio values for the 90 region/season influenza epidemics
Figure 14: Distribution of estimated scale values for the 90 region/season influenza epidemics