A SIMULATION OF THE U.S. INFLUENZA OUTBREAK IN 2009-2010 USING A PATCH SIR MODEL BASED ON AIRPORT TRANSPORTATION DATA

D. I. WALLACE, M. CHEN
Dartmouth College
Hanover, NH, 03755, USA

We simulate the progression of the novel H1N1 virus outbreak in Spring 2009 using a patch SIR model for six U.S. regions. A sensitivity analysis shows that the timing of peak prevalence is greatly affected by epidemiological parameters but little affected by migration rates. The method presented here has the advantage of using publicly available transportation data for only a few large regions and epidemiological parameters that may be estimated from the first few cases and the households in which they occur. Comparison with data shows the method to have predictive value for timing of epidemics.

Keywords: SIR models, patch models, regular graphs, epidemiology, influenza, H1N1

1. Introduction

In March of 2009 a new strain of H1N1 influenza was detected in Southern California [1]. Within months cases were reported in distant sites across the U.S. Some, but not all, cases were traced to foreign travel. A compelling statistical argument has been made for linking regional outbreaks with frequency of inbound flights from Mexico [2]. However, large regional airline hubs serve all parts of the U.S. as well, so disease may well have been transmitted from within the U.S. once it arrived. This hypothesis is the one modeled and analyzed here.

The classic model of an infectious disease is given by three coupled ordinary differential equations for susceptible, infectious and recovered individuals (SIR models)[3]. This model is appropriate for large homogeneous well-mixed populations. A recent survey article points out the limitations of the SIR model in capturing spacial dynamics, pointing out that agent based models represent these dynamics but have their own computational

limitations [4]. Patch models make it possible to investigate spacial phenomena at an intermediate level, by representing regions of a certain size (the patches) as homogeneously mixed while linking well separated regions via migration. In either case, the accuracy of SIR models depends completely on the accuracy of the data used to estimate initial prevalence, transmission, and recovery rates for any given disease.

Global airline networks have been studied in an effort to model large scale disease propagation. One study, of the worldwide air network of 3,880 airports, indicates that properties of the statistical distribution of travel are thought to influence rates of disease propagation [5]. It would be useful to know to what extent a data from a few major airports capture the dynamics of disease transmission and, in particular, the timing and intensity of disease prevalence. In this study we use publicly available airport usage data to construct a patch SIR model for six U.S. regions with major airports and compare the results of the model with publicly available data from the 2009 influenza outbreak.

2. Methods

Models for diseases that spread from city to city consider populations that are separated into patches, each of which is described by an SIR model, with migration between patches given as a linear term. Equations 1-3 below give the general form for the patch SIR model used here.

$$S_i' = (a-b)(S_i)(1-S_i/k_i) - B\frac{S_i I_i}{(S_i+I_i+R_i)} - \sum_{j\sim i} m_{(i,j)}S_i + \sum_{j\sim i} m_{(j,i)}S_j \tag{1}$$

$$I_i' = B\frac{S_i I_i}{(S_i+I_i+R_i)} - (v+b+d)I_i - \sum_{j\sim i} m_{(i,j)}I_i + \sum_{j\sim i} m_{(j,i)}I_j \tag{2}$$

$$R_i' = vI_i - bR_i - \sum_{j\sim i} m_{(i,j)}R_i + \sum_{j\sim i} m_{(j,i)}R_j \tag{3}$$

Here the growth rate of the susceptible population, as a function of S_i only, is given by a logistic term. b is the natural death rate, d is the death rate of the disease only, B is the transmission coefficient, $m_{(i,j)}$ is the relative rate of travel from vertex i to vertex j, assumed to be equal for susceptible, infectious and recovered populations at vertex i. Sums denoted

by $j \sim i$ indicate vertices adjacent to the ith vertex. Note that making birth a function of only susceptible individuals is not the most general form that could be taken, but it is reasonable for short term diseases. In a previous paper [6] we show that stability of the disease free equilibrium for symmetric versions of such models depends only on local parameters, *i. e.* not migration rates.

Six major U.S. areas connected by large airports (CA, CO, TX, GA, IL, NY) were used as a basis for parametrizing the model in equations 1-3. For this simulation, the growth rate was set to $(a-b)(1-S_i/k_i) - dS_i$. The birth rate, a, and death rate, b, were taken as the U.S. average[7,8] and k_i was chosen to give an equilibrium population at the disease free equilibrium corresponding to the actual population of each region[9]. The transmission and death rates (B and d respectively) from this particular strain can be estimated from commonly known statistics reported in the newspaper as 1.5 new cases per three days [10], which correlate roughly with those in the literature. We used .5 as the transmission rate, estimated as a reproduction rate of 1.5 spread over a three day duration. Actual likelihood of transmission was reported to vary between 1 (for households with two members) and approximately .2 (for households with six members) [11]. A review of multiple swine flu studies puts the reproduction number for the 2009 outbreak between 1.4 and 1.6 new infections per infected individual, but does not say over how many days this occurs [12]. A World Health Organization study gives a reproduction rate of 1.58 and confirms the same three day period based on data from the early outbreak in Mexico [2]. The recovery rate can be estimated from data available at the Center for Disease Control and corresponds to a disease of 14 days duration [13].

The graph between patches was the complete graph on six vertices, with migration rates estimated from airport data[14]. To model migration rates we assumed that the traffic to and from destinations other than the six modeled was negligible.

Migration rates for each state included were two forms: migration from state 1 to state 2, *e. g.* ($m_{CA,NY}$) and total migration away from a specific state *e. g.* (m_{CA}).

Let q_{NY} be the total average daily traffic through New York airports (and define q similarly for all six regions). Let Σ_{NY} be the sum of average daily traffic through the five remaining airports in the model (and define Σ similarly for all six regions).

The calculations for the migration rates from one state to the other are

given by Equation 4.

$$m_{CA,NY} = q_{CA}q_{NY}/\Sigma_{CA} \tag{4}$$

This expression weights traffic from CA to NY according to the amount of traffic in NY. Notice that summing all the outgoing traffic from CA returns the total traffic in CA. The total migration away from a specific state is just the sum of the outgoing rates to all other states. The expression for New York is given by Equation 5.

$$m_{NY} = m_{NY,CO} + m_{NY,CA} + m_{NY,TX} + m_{NY,GA} + m_{NY,IL} \tag{5}$$

A simple model was created to test the migration rates with no disease present to verify that populations of the various regions remained approximately accurate at equilibrium. The model was then run with disease introduced in California and Texas. A second version of the model was run in which birth rates depended on the entire population of the region rather than on just susceptible individuals, in order to see if this adjustment to the model made any significant difference to the results. (It did not.)

The simulation was then compared with the patient data from clinics in each area (Regions 2,4,5,6,8,9) that report to the Center for Disease Control [15]. The regions for which FluView data was used included regions 2 (New York, New Jersey), 4 (Kentucky, Tennessee, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi), 5 (Minnesota, Wisconsin, Michigan, Illinois, Indiana, Ohio), 6 (New Mexico, Texas, Oklahoma, Louisiana, Arkansas), 8 (Utah, Colorado, Wyoming, Montana, South Dakota, North Dakota), and 9 (California, Nevada, Arizona), which correspond to NY, GA, IL, TX, CO, and CA, respectively. The regions used in this study did not include just the city in which the airport was located, but the surrounding area as well. Population data corresponding with the regions around each airport was taken from a 2010 report on tuberculosis incidence [16]. One assumption of the model was that the heavily populated regions near accounted for most of the disease prevalence.

A sensitivity analysis was conducted on parameters and initial conditions with respect to the error observed between peak case load (for each state and for the US total) in the model versus data. In a subsequent simulation, transmission rates were modified for each state for a best match with data (for the first version of the model). Table 1 shows parameters used for the original model. Not shown are initial conditions for infected individuals (2 in California, 1 in Texas, 0 elsewhere) or recovered individuals (0 for all regions).

Table 1. Parameters for the model

Notation	Value	Units
$k_{NY}, S_{NY}(0)$	19151072	individuals in NY area
$k_{CO}, S_{CO}(0)$	2599235	individuals in CO area
$k_{CA}, S_{CA}(0)$,	17293449	individuals in CA area
$k_{TX}, S_{TX}(0)$	12564909	individuals in TX area
$k_{GA}, S_{GA}(0)$	5540092	individuals in GA area
$k_{IL}, S_{IL}(0)$	9622245	individuals in IL area
$m_{NY,CO}$	$5.93 * 10^{-5}$	daily migration from NY to CO as a percent of regional population
$m_{NY,CA}$	0.000133771	daily migration from NY to CA
$m_{NY,TX}$	0.00012474	daily migration from NY to TX
$m_{NY,GA}$	0.000104113	daily migration from NY to GA
$m_{NY,IL}$	9.67E-05	daily migration from NY to IL
$m_{CO,NY}$	$5.30 * 10^{-5}$	daily migration from CO to NY
$m_{CO,CA}$	$5.86 * 10^{-5}$	daily migration from CO to CA
$m_{CO,TX}$	$5.46 * 10^{-5}$	daily migration from CO to TX
$m_{CO,GA}$	$4.56 * 10^{-5}$	daily migration from CO to GA
$m_{CO,IL}$	4.23E-05	daily migration from CO to IL
$m_{CA,NY}$	0.000137123	daily migration from CA to NY
$m_{CA,CO}$	$6.72 * 10^{-5}$	daily migration from CA to CO
$m_{CA,TX}$	0.000141255	daily migration from CA to TX
$m_{CA,GA}$	0.000117897	daily migration from CA to GA
$m_{CA,IL}$	0.000109452	daily migration from CA to IL
$m_{TX,NY}$	0.000125624	daily migration from TX to NY
$m_{TX,CO}$	$6.16 * 10^{-5}$	daily migration from TX to CO
$m_{TX,CA}$	0.000138777	daily migration from TX to CA
$m_{TX,GA}$	0.00010801	daily migration from TX to GA
$m_{TX,IL}$	0.000100273	daily migration from TX to IL
$m_{GA,NY}$	0.000100813	daily migration from GA to NY
$m_{GA,CO}$	$4.94 * 10^{-5}5$	daily migration from GA to CO
$m_{GA,CA}$	0.000111369	daily migration from GA to CA
$m_{GA,TX}$	0.00010385	daily migration from GA to TX
$m_{GA,IL}$	$8.05 * 10^{-5}$	daily migration from GA to IL
$m_{IL,NY}$	$9.23 * 10^{-5}$	daily migration from IL to NY
$m_{IL,CO}$	$4.52 * 10^{-5}$	daily migration from IL to CO
$m_{IL,CA}$	0.000101972	daily migration from IL to CA
$m_{IL,TX}$	$9.51 * 10^{-5}$	daily migration from IL to TX
$m_{IL,GA}$	$7.94 * 10^{-5}$	daily migration from IL to GA
m_{NY}	0.000518612	daily migration away from NY
m_{CO}	0.000254107	daily migration away from CO
m_{CA}	0.000572914	daily migration away from CA
m_{TX}	0.000534237	daily migration away from TX
m_{GA}	0.000445897	daily migration away from GA
m_{IL}	0.000413958	daily migration away from IL
a	0.0135	natural birth rate percent per day
b	0.008036	natural death rate percent per day
B	0.5	transmission rate individuals per day
d	0.01043575	death rate due to disease percent per day
v	0.07142857	recovery rate percent per day

3. Results

Figure 1 shows the infected populations in all cities and the total. The first two peaks are for the California and Texas regions, while subsequent peaks are for all other regions.

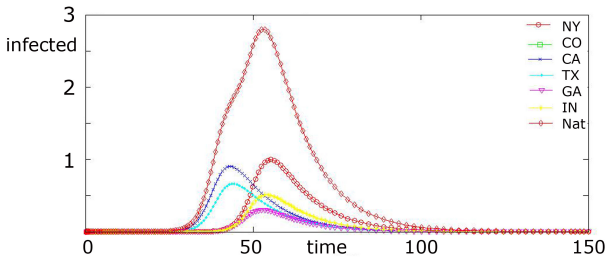

Figure 1. Model output for default parameters. Left scale is to be multiplied by 10^7.

3.1. *Disease prevalence*

The Center for Disease Control reports a total of 43,677 laboratory-confirmed cases of influenza (H1N1) 2009 in the United States through July 2009 [17]. They estimate the under reporting of influenza as 79 actual infections per reported case. Their statistical model estimates 1.8 million to 5.7 million cases occurred. Our model predicts 3.8 million cases, well within this range. However our model is only for heavily populated regions around the six major airports, so it seems like the model overestimates the number of cases. When broken down by region and reported cases we see consistently large overestimates of the peak prevalence, as in Table 2, which gives model calculations and data comparison for constant transmission rate of .5.

Table 2. Peak prevalence results for constant transmission rate

Region	Peak Value (Model)	Peak Value (Data)	79 * Peak Value (Data)
NY	9,981,985	275	21,725
CO	1,442,394	2,134	168,586
CA	9,042,460	1,192	94,168
TX	6,614,867	652	51,508
GA	3,010,005	815	64,385
IL	5,106,028	1,763	139,277
National	28,000,000	10,050	793,950

However the model may just produce a distorted distribution. We can also compare the total number of cases in each region to the integrated infected compartment. If we assume a 14 day duration of disease and use the first fifteen weeks of data for each region to estimate total cases during this period we have the comparison described in Table 3, which gives model calculations and data comparison for constant transmission rate of .5. Total cases for the first 15 weeks of outbreak in each region are multiplied by 79 as recommended. Integrated infected compartments for each region are divided by a 14 day average duration of illness, consistent with the recovery rate. Our model predicts values about 15 times as large as the statistically adjusted data for the total number of cases and most of the individual regions.

Table 3. Total prevalence results compared

Region	79 * Total cases (Data)	Integrated Value (Model)/14
NY	236,052	14,975,307
CO	103,964	2,192,060
CA	474,869	13,758,223
TX	278,870	10,087,391
GA	514,053	4,551,770
IL	643,455	7,691,961
National	3,077,919	53,256,713

3.2. *Timing of peak prevalence*

The number of days until the model reached peak prevalence was compared to the peak in the corresponding data set for all regions and the total. For the default transmission rate of .5 errors were within a couple of weeks

for most regions. The exceptionally poor matches were for California and Texas. The data for these states had two separate peaks– an early one and a late one. The model placed a peak between the two. Adjusting the transmission rate for a single state could often bring the model and data into agreement, as is shown for the Colorado region in Figure 2.

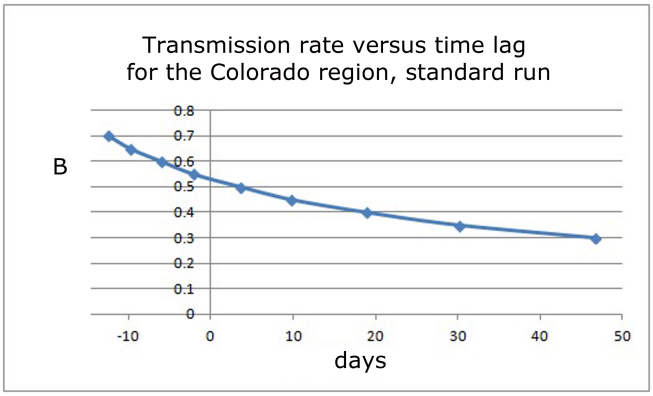

Figure 2. Time lag (peak of model - peak of data) versus transmission rates for Colorado

Figure 3 shows the results of altering transmission rates one at a time for each region. In this figure, California's two peaks are plotted separately, and these show the largest disagreement with the model, which produces a peak in the middle. Only the first peak of the Texas data was used for comparison.

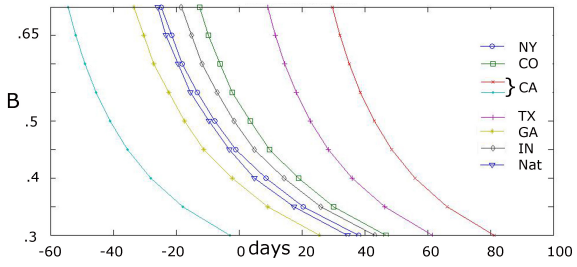

Figure 3. Time lag versus transmission rates for all regions

3.3. *Sensitivity of parameters*

A sensitivity analysis was carried out on all parameters. Specifically, we were interested in the error between the time to peak prevalence in the model and the data, for each region. Figure 4 shows, for Colorado, the result of varying each parameter up and down 10%. The result in Figure 4 is typical of all the regions and also the total. Disease related parameters have, in general, a much larger effect than migration rates, with the transmission rate having the largest effect of all. Table 4 summarizes the results for all regions with parameters listed in descending order of effect on time lag between data peak and model peak.

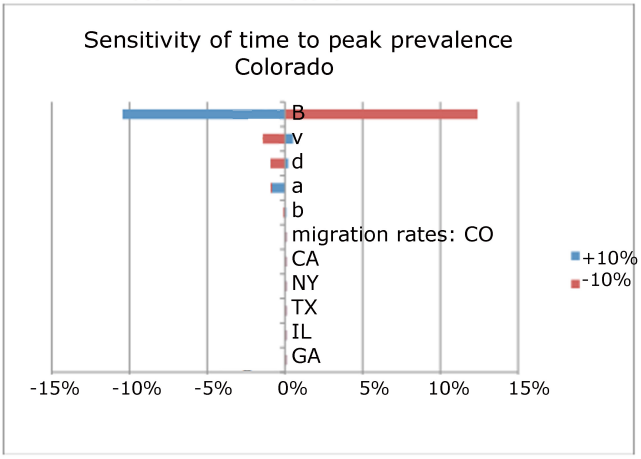

Figure 4. As each parameter is varied from its default value by +/- 10%, the time to peak prevalence varies. The transmission rate is seen to have by far the largest effect; migration rates have almost no effect.

Table 4. Sensitivity of error in peak time prediction

Region	Top 5	Negligible effect
NY	B, v, d, a, b	$m_{CO}, m_{CA}, m_{TX}, m_{NY}, m_{IL}, m_{GA}$
CO	B, v, d, a, b	$m_{CO}, m_{CA}, m_{NY}, m_{TX}, m_{IL}, m_{GA}$
CA	B, v, d, b, m_{CA}	$m_{NY}, m_{CO}, m_{TX}, m_{IL}, a, m_{GA}$
TX	B, v, d, b, m_{CA}	$m_{NY}, m_{CO}, m_{TX}, m_{IL}, a, m_{GA}$
GA	B, v, d, b, a	$m_{NY}, m_{CA}, m_{CO}, m_{TX}, m_{IL}, m_{GA}$
IL	B, v, d, b, a	$m_{NY}, m_{CA}, m_{CO}, m_{IL}, m_{TX}, m_{GA}$
NAT	B, v, m_{NY}, a, d	$b, m_{CO}, m_{CA}, m_{IL}, m_{TX}, m_{GA}$

4. Conclusions

4.1. *Estimates of disease prevalence*

The data we used gives the percent of those patients who, having presented with flu-like symptoms, tested positive for influenza[15]. Early cases were detected in California and Texas[18]. These were taken as initial conditions in the model, with disease incidence at all other locations taken to be zero. The FluView application has a disclaimer that explains that the data is from both the U.S. World Health Organization (WHO) Collaborating Laboratories and the National Respiratory and Enteric Virus Surveillance System (NREVSS). The disclaimer also mentions that the methods/testing practices varied from region to region. One last thing the disclaimer states is that WHO collaborating laboratories report on the influenza A subtypes of H1 or H3 while the majority of NREVSS laboratories do not report them.

There are approximately 145 participating laboratories in the U.S. How we chose to aggregate this data influences how our model compares to it. Perhaps if we had included more regions in the vicinity of each airport the local estimates would have been closer to the CDC estimate of 79 times the reported cases. The model prediction for the total is within the CDC estimate for the U.S. of 1.8 to 5.7 million cases [17] for April to July of 2009.

In addition, altering the transmission rate would also change the predicted prevalence. Our model does not take into account behavioral changes of individuals which may take place when an outbreak is known and publicized, as data for the resulting change in transmission rates was not available. These changes may also be modeled explicitly if relevant parameters are known [19]. Doing so should lower the estimated number of cases in each region.

The estimates of prevalence we obtain from the integral of infected individuals over the time period under consideration (and divided by duration of the disease) were consistently higher than CDC estimates, indicating that a patch SIR model would have to be adjusted in order to predict the disease burden in advance.

4.2. *Timing of peak prevalence*

With two notable exceptions the timing of peak prevalence in our model was within a week or two of peak number of cases reported in the data. With transmission rates within a range of .4 to .55 we have good agreement on timing with data from GA, NY, CO, IL and the whole U.S., as shown in Figure 3. Strangely, the simulations for states where the disease starts

(TX, CA) have the worst agreement with data. In both cases the data show double peaks. Texas has an early peak followed by a period of low case reports and then a second peak, which may be a second strain of virus. It is possible that, due to overland travel from Mexico, the original number of cases was under-reported in Texas, which could certainly alter the timing of the peak in that state. California also shows two peaks, but this is probably because the disease started in the San Diego region and later reached Northern California. We suspect that if we had treated the California region as two separate patches we would have had better agreement with the data. Nonetheless, the model produced results within a week or two of the peaks in the data set for most regions, indicating that patch models based on airport migration data could be a useful tool in predicting the timing of disease spread.

The timing also suggests that only a few cases arriving from abroad have the potential to spread through migration within the country. It is unnecessary to suppose that every outbreak was initiated through travel from Mexico. Regions assumed to acquire the disease by migration from Texas or California (or each other) had peak prevalence substantially later than those two states, both in the model and in the data. Had infected visitors from abroad been the source of infection it is unlikely that the timing would be so consistent.

4.3. *General utility of this method*

Although the transportation data is only an average, our sensitivity analysis shows that small variations in it do not have a noticeable effect on timing of peak prevalence. Of all the parameters in this model, the transmission rate has the largest effect. Unfortunately, this parameter can be difficult to estimate and can vary regionally and over time as people increase their awareness of a possible epidemic. How well a patch SIR model will predict the timing of an epidemic is highly dependent on getting an accurate estimate of the transmission rate and other disease-related parameters. However, even a rough estimate of the timing of an epidemic, in advance, is very useful. The method presented here has the advantage of using publicly available transportation data for only a few large regions and epidemiological parameters that may be estimated from the first few cases and the households in which they occur. Comparison with data shows the method to have predictive value.

References

1. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team, *N Engl J Med* **360** 2605-2615 (2009)
2. C. Fraser *et al*, *Science* **324** 1557 (2009)
3. R. M. Anderson and R. M. May, *Nature* **280**, 361 (1979).
4. M. J. Keeling and L. Danon, Mathematical modelling of infectious diseases *British Medical Bulletin***92** (2009)
5. V. Colizza, A. Barrat, M. Barthelémy, and A. Vespignani *PNAS* **103 7** 20152020 (2006)
6. Wallace, Hu-Wang, Chen *Symmetries and Groups in Contemporary Physics (Nankai Series in Pure, Applied Mathematics and Theoretical Physics)*, World Scientific, in press (2013)
7. Center for Disease Control and Prevention (Atlanta 2009), [Data File]. Retrieved May 3, 2012 from`http://www.cdc.gov/nchs/births.htm`
8. Center for Disease Control and Prevention (Atlanta 2009), [Data File]. Retrieved May 3, 2012 from `http://www.cdc.gov/nchs/deaths.htm`
9. Center for Disease Control and Prevention (Atlanta 2010). Retrieved May 3, 2012 from `http://www.cdc.gov/tb/statistics/reports/2010/pdf/report2010.pdf`
10. A. Gardner (ABC News, New York 2009). Retrieved May 3, 2012 from `http://abcnews.go.com/Health/Healthday/story?id=8439860`
11. S. Cauchemez, *et al*, *N Engl J Med* **361** 2619-2627 (2009)
12. B. J. Coburn, B. G. Wagner and S. Blower, *BMC Medicine* **7 30** (2009)
13. Center for Disease Control and Prevention (Atlanta 2010) Retrieved May 3, 2012 from `http://www.cdc.gov/h1n1flu/qa.htm`
14. Airports Council International, North America (Montreal 2009), [Data File] Retrieved May 3, 2012 from `http://aci-na.org/content/airport-traffic-reports`
15. Center for Disease Control and Prevention (Atlanta, 2010). [Data Set]. Retrieved Oct. 3, 2012 from `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`
16. Center for Disease Control and Prevention (Atlanta 2010).[Data Set]. Retrieved Oct. 3, 2012 from the Division of Tuberculosis Eliminations online ordering system at `http://www.cdc.gov/tb/`.
17. C. Reed, F. J. Angulo, D. L. Swerdlow, M. Lipsitch, M. I. Meltzer, D. Jernigan, *et al*, *Emerg Infect Dis* (2009) Retrieved Jul. 31, 2013 from `http://wwwnc.cdc.gov/eid/article/15/12/09-1413.htm`
18. Center for Disease Control and Prevention (Atlanta 2009). *MMWR* **58 15**, 400-402. Retrieved Oct. 10, 2012 from `http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5815a5.htm`
19. S. Funk, Marcel Salathé and V. A. A. Jansen, *J. R. Soc. Interface* **7** 12471256 (2010)