Network effects of air travel on influenza and pneumonia

Ksenia Lepikhina, Aparajithan Venkateswaran, Josh Mellin

CSCI 5352, Fall 2019
University of Colorado Boulder

Abstract

Influenza and pneumonia are closely related diseases that are global in scale. Influenza (the flu) is a common cause of pneumonia. Most cases of influenza do not lead to pneumonia, however, when they do, they are severe, and often terminal. Here, we investigate the spread of influenza via air travel in the United States from 2009 – 2019. In particular, we build an agent-based network SIRS model that incorporates air travel at the state level to estimate the number of deaths due to influenza and pneumonia. Our model also accounts for seasonality in the probability of spreading influenza by incorporating a forcing sinusoidal function. Finally, using our model, we identify states that we believe play an important role in spreading influenza through air travel. We conclude that air travel does not explain a significant proportion of influenza spread. While the number of deaths (from simulations with air travel) is correlated with the true data, air travel itself is not a great predictor. After including seasonality, the estimates including air travel appear to estimate the data better than the estimate without air travel. Using our devised spreading centrality measure and the intuitive out-degree measure, we conclude that California, Illinois, Georgia and Texas are important states in spreading influenza via air travel.

1 Introduction

Influenza and pneumonia are closely related diseases that are global in scale. Influenza (the flu) is a common cause of pneumonia. Most cases of influenza do not lead to pneumonia, however, when they do, they are severe, and often terminal. In 2016, the flu and pneumonia, together, were the eighth leading cause of death in the U.S. Influenza is responsible for approximately 250,000 to 500,000 deaths around the world annually; in the United States alone, about 50,000 adults die from pneumonia [1].

In this paper, we hypothesize that air travel has a significant effect on the dissemination of the virus. In 1980, Knight claimed that the influenza virus can survive for up to an hour in the air in enclosed environments such as airplanes [2]. In 1982, Bean et al. concluded that the virus can survive for more than eight hours on hard surfaces such as stainless steel and plastic, and up to five minutes on hands after transfer from other surfaces [3].

In this report, we investigate the spread of influenza via air travel in the United States from 2009 – 2018. In particular, we build a network SIRS model that incorporates air travel

at the state level to estimate the number of deaths due to influenza and pneumonia. Our model also accounts for seasonality in the probability of spreading influenza. Finally, using our model, we identify states that we believe play an important role in spreading influenza through air travel.
The rest of this report is organized as follows: Section 2 will discuss some previous work on network SIRS models; Section 3 will present the data that we used; Section 4 will discuss our model in detail; Section 5 will discuss our results; and Section 6 will conclude our analysis and provide scope for future work.

2 Previous Work

Brownstein et al. assessed, with empirical data, the role of airline volume on the yearly inter-regional spread of influenza in the United States [4]. They characterized seasonality with band-pass filtering and showed that domestic and international air travel predicted the spread and mortality of influenza respectively. Kenah et al. developed an epidemic percolation network model to analyze the stochastic SIR model [5]. The results of Chan et al. showed that seasonal influenza strains originate in different countries [6]. This lends further evidence that air travel, or travel in general, plays an important role in global spread of influenza.
In our paper, we use data similar to [4], but the influenza mortality is from years 2009 – 2018. Our model is loosely inspired by the epidemic percolation network model, but for discrete time steps. Using this, we intend to simulate a monthly spread, and determine whether air travel truly affects the spread. Additionally, we identify states that play an important role in the spread of influenza.

3 Data

The data for this paper comes from three different sources. The first dataset, compiled by the Center for Disease Control and Prevention, consists of the number of deaths due to influenza and pneumonia by state for every week from the last quarter of 2009 to the first quarter of 2019 [7]. We use the US census data to determine the state level population from each year from 2009 – 2019 [8, 9]. The third dataset is a 10% sample of domestic airline travel information since the second quarter of 2007 collected by the U.S. Department of Transportation [10]. We construct a dynamic network with states as nodes and airline routes as edges. The number of deaths due to influenza and pneumonia will be a property of time and node (state). In essence, our data allows us to study the dynamics of the infectious disease on a dynamic network.

4 Methods

In this section, we describe the construction of our model with and without seasonality. Additionally, we discuss parameter estimation and centrality measures.

Constructing the networks In our directed network, the nodes represent each state and the edge weights represent the number of travellers. Because the air traffic data is a 10% sample, we multiply each edge weight in the data by 10. Further, to decrease temporal granularity, we assume a uniform travel distribution and artificially split the quarterly travel data into 3 individual months. This leads to 152 networks, one for each month.

Model for spreading influenza We build an agent-based SIRS model on each network. Each state has agents (people) who are either infected, uninfected, or dead. During each time step i.e., each month, we attempt to spread a random infection to the uninfected population of each state with probability p_0. Then people travel according to our network with infected and uninfected people travelling at the same rate. The infected travellers at the destination attempt to transmit the infection to a fixed fraction of healthy people with probability p_t. Then, at the end of this cycle, people either recover, die, or continue to stay infected with probabilities p_r, p_d, p_i. We repeat this cycle, without the air travel component, two more times because a flu lasts for 1-2 weeks and each month has 4-5 weeks [11]. At the end of each year, we repopulate the state with our census data, keeping track of deaths due to influenza. Each of these infections, transmissions, and recoveries reduce to simple Binomial simulations. In state k, at time t, let $N_t^{(k)}$ be the total population, $I_t^{(k)}$ be the number of people already infected, and $y_t^{(k)}$ be the number of infected people who travelled to state k. Let $n_t^{(k)}$ be the number of people infected at random, $n_{y,t}^{(k)}$ be the number of people infected due to transmission, $r_t^{(k)}, d_t^{(k)}, i_t^{(k)}$ be the number of people who successfully recovered, the number who died, and the number who continue to have the infection respectively. These numbers are simulated by,

$$n_t^{(k)} \sim Binomial(N_t^{(k)} - I_t^{(k)}, p_0), \tag{1}$$

$$n_{y,t}^{(k)} \sim Binomial(N_t^{(k)} - I_t^{(k)} - n_t^{(k)}, 1 - (1-p_t)^{I_t^{(k)} + n_t^{(k)} + y_t^{(k)}}), \tag{2}$$

$$r_t^{(k)}, d_t^{(k)}, i_t^{(k)} \sim Multinomial(p_r, p_d, p_i), \tag{3}$$

where the probability in Equation 2 comes from noticing that person x is infected via transmission if at least one infected person successfully transmits the infection. Further, when we are evaluating the flu cycle without air travel, we set $y_t^{(k)} = 0$. And these values are used to update $N_{t+1}^{(k)}, I_{t+1}^{(k)}$ for the next time step.

Measuring performance Our model estimates the number of deaths. We can measure the average monthly error (normalized) using the following quantity,

$$\text{error} = \frac{1}{\#\text{months}} \cdot \frac{||\text{deaths}_{\text{true}} - \text{deaths}_{\text{est}}||_F}{||\text{deaths}_{\text{true}}||_F}, \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm.

Estimating parameters Our model consists of 5 parameters p_0, p_t, p_r, p_d, p_i. But we can reduce one of the parameters by noting that $p_r + p_d + p_i = 1$. One possible method for estimating these parameters is to exhaustively search $[0, 1]^4$ to find the the parameters that

minimize the error defined in Equation 4. This is computationally infeasible. Instead, we propose a second method — performing a random walk. We start the walk at a random estimate (or an educated guess). With this estimate, we calculate the error and compare it to an error tolerance that is chosen beforehand. Until we reach this threshold, we add a small amount of Gaussian random noise to the previous estimates. We accept these new parameters if the new error is smaller. Otherwise, the proposed parameters are discarded and the previous parameters are used in the next iteration. In this manner, we locally search for the parameters in a greedy fashion.
To make computations easier, we fix the probability of dying from influenza to be:

$$p_d = \frac{\sum_{\text{year}=2010}^{2018}(\text{deaths}_{\text{year}}) \cdot 0.125}{\sum_{\text{year}=2010}^{2018} \text{population}_{\text{year}}}$$

which was derived from a combination of the CDC death data from 2010 – 2018 as well as WebMD's statistics in "What Are Your Odds of Getting the Flu?" [12]. Now, we need to estimate only p_0, p_t, p_i which are the probability of a random infection, the probability of transmitting influenza, and the probability that an infected person continues to stay infected at the end of the flu cycle.

Modeling seasonality To model the seasonality of influenza, we allow the probability of transmitting (and contracting at random) to vary in a sinusoidal fashion. The motivation behind this is that flu seasons are periodic in nature, and sinusoidal functions are the simplest and most versatile of periodic functions. Throughout the flu season, the probability of infection/transmission should increase, especially around the peak of the flu season. In the middle of the flu season (typically December – February) people are more susceptible to the flu due to colder and drier weather leading to weaker immune systems (meaning more infections) [12].
In our monthly simulations, we allowed our probabilities to vary temporally as

$$p\left[a\cos\left(\frac{\pi}{6}t + t_0\right) + (1-a)\right]$$

where $0 \leq t \leq 12$ and $0 \leq a \leq 0.5$. Here, a represents how much we want to stretch the variations vertically, and t_0 represents the phase shift.

Control simulations In order to determine whether or not air travel actually affects the spread of influenza, we compare the local spread of influenza to the national spread by modeling the infection without and with air travel. For this study, we use the monthly infection spread where we assume the travel occurs once a month. Within that month, individuals are infected, the majority recover, and some die. As mentioned previously, our model includes a local infection and an infection that was spread via air travel. This control model excludes the air travel component.
Further, all of our simulations are performed with *artificially* split data – we assume a uniform air travel distribution when we split quarterly air travel data into monthly air travel data. To ensure that our simulations are not skewed by performing analysis on the month-level granularity, we repeat all of our methods with the quarterly air travel data.

Centrality measures To identify important states in the spread of influenza, we measure the spreading centrality of each state. The spreading centrality is a measure of the importance of each state given that it was the initial seed of the infection. To compute the spreading centrality for each state k, we run the model by infecting only state k in each time step. Previously estimated parameters are used here. For this simulation, the probability of randomly contracting an infection is set to 0 for all other states. The value of a state i's spreading centrality is computed as the total number of deaths occurring at the end of the simulation. A running tally of the number of deaths is kept throughout the simulation in order to measure this. This process is repeated multiple times for each state in order to obtain a good average spreading centrality value.
We also compare these results from a naive measure of importance – out-degree. This metric was used because, intuitively, as more people travel out from a state, the more likely this state can spread a larger infection. Therefore, we would expect those states that have a higher out-degree to also record a higher spreading centrality score.

5 Results

In this section, we discuss our results from our model and centrality measures. We conclude this section by enumerating two flaws in our model.

5.1 Effect of air travel on deaths

In this section, we present the results from our model (on a monthly basis) from October 2009 to April 2019. We first discuss aggregate results across the US, and then we analyze four states in particular – Alabama, Alaska, California, and Colorado. Alaska is selected as it has one of the smallest populations in the US and is very remote from the other states. California is chosen because it has the largest population in the US. Alabama is chosen because our model severely over estimates the number of deaths. Finally, Colorado is selected because it's where we all live and we were interested in seeing how well our model performed in our state.
First, we present the number of deaths on a monthly basis without including seasonality. The results are shown in Figure 1. Since we do not force seasonality here, we do not effectively capture the periodic variations in the true data. Further, our estimates are very low. An important observation here is that the estimates with air travel are higher than the estimates without air travel. This makes sense because by allowing *long range links* with air travel, the infection can spread further and increase the number of deaths. A second important observation is that the simulation with air travel has some periodic variation. This can be attributed to the fact that these peaks correspond with summer vacations and winter holidays, both of which see increased air traffic.
Figure 2 shows the number of deaths on a monthly basis, after incorporating seasonality. Clearly, these results outperform the results in Figure 1. Both the red and blue curves are sinusoidal in nature, which is a reflection of the sinusoidal forcing we imposed on p_0 and p_t. Both the air travel and no air travel lines do not appear to be changing based on the year but rather appear to follow the same trend. However if zoomed in, we see that generally,

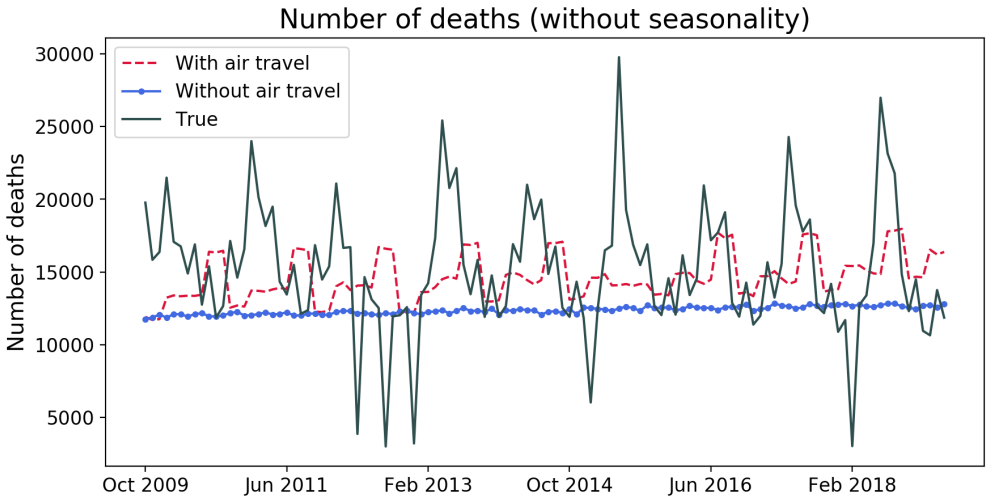

Figure 1: The number of deaths across US measured without seasonality. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

the number of deaths is subtly increasing over time. This model still does not capture the steep increases and sharp dips observed in the true data. This might indicate that we need a more sophisticated forcing function than sines and cosines.

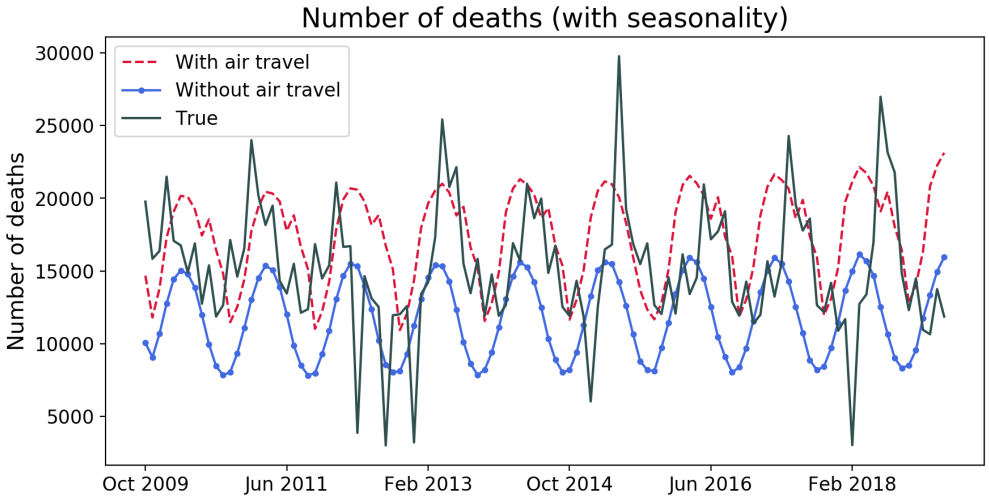

Figure 2: The number of deaths across US measured with seasonality. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

Our results so far have been on the aggregate level. We also looked at how well our model did for specific states to see if it was able to model the number of deaths in smaller or larger states better. The plots for the four chosen states are shown in Figures 4, 5 in Appendix A.

Based on these investigations, we observe that our model performs well for some states but not all. For example, Alabama and Colorado have similar populations ($\sim$ 5 million) and our model significantly overestimates the number of deaths in Alabama and does a reasonable job with Colorado. For a state with a small population such as Alaska, our model severely underestimates the number of deaths. For a state with a large population such as California, our model also underestimates but not significantly.
One of the reasons that Alabama may be severely overestimated is because in reality. Alabama is a fairly warm and humid state that may see fewer cases of influenza. We can also see that the true data for Alabama is fairly flat and does not change seasonally very much. Our seasonal forcing may not apply to warmer and more humid states. California may be well estimated because it is more populated, and there is a constant flux of people moving. Colorado may also fit the same boat besides the fact that Colorado experiences four seasons and that the temperature can vary significantly. Alaska is heavily underestimated and this is likely due to the cooler and drier temperatures than in other states. Alaska likely sees more cases of influenza and more deaths caused by the disease due to the climate.
We performed the same analysis on quarterly aggregated air traffic data and got similar results. These results are shown in Appendix B.

5.2 Centrality measures

Figure 3 shows the centralities as measured by the out-degree and the *spreading* centrality described in Section 4. Table 1 shows the top 10 states ranked by these centrality measures. As we can see there is a strong correlation between our intuitive model for key states and the results from our model. In fact the Pearson correlation coefficient between the two sets of results is 0.54. This lends evidence to our belief that air travel is correlated with the spread of influenza. We believe that California, Illinois, Florida and Texas are the key states in spreading influenza.

Rank	Out-degree	Spreading Centrality
1	California	Colorado
2	Florida	Illinois
3	Texas	Georgia
4	Georgia	Utah
5	Illinois	Texas
6	New York	Montana
7	Colorado	California
8	North Carolina	Idaho
9	Arizona	Nevada
10	Nevada	Minnesota

Table 1: Top 5 states ranked by out-degree and spreading centrality

Three interesting states that are negatively correlated are Florida, Montana and Utah. The large spreading centrality and small out-degree of Montana and Utah can be explained by observing the destination of these travellers. These travellers go to populated states such as

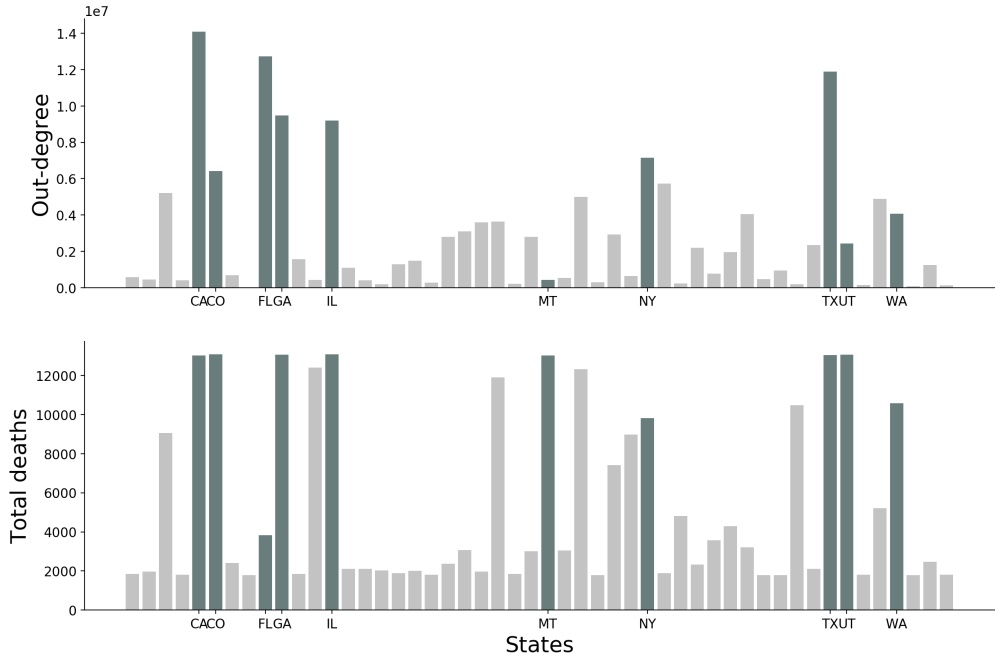

Figure 3: Centralities as measured by out-degree and our *spreading* centrality.

California, Colorado, Texas and Washington. Therefore, they have a greater chance to infect more people. Travellers from Florida on the other hand visit less populated states, which explains why its spreading centrality is small compared to the out-degree. The out-flow for these outliers are shown in Appendix C. Once we remove these interesting outliers, our correlation coefficient increases to 0.69.

Overall we can conclude that states with a higher out-degree generally also have a higher spreading centrality. It is interesting to note, however, that this is not the only driving force in a high spreading centrality score, as there are deeper interactions occurring that allow some smaller out-degree states to still get a high spreading centrality score.

5.3 Discussion

One issue that we can take note of in our analysis is that the CDC death data included the number of deaths during a given week between the fourth quarter of 2009 to the first quarter of 2019 from influenza *and* pneumonia. It's possible that when we extrapolate the probability of deaths from our data, we are creating a representation error because we are overestimating the number of infected people due to pneumonia and are likely miscalculating the number of infected since our calculation introduces some amount of rounding error. Representation error occurs when the model variables are not exactly the same thing as the real system.

Another potential flaw in our model is revealed when trying to fit parameters. We observed two different things when we were estimating parameters. Firstly, the error (as measured by Equation 4) would not drop below 0.56 no matter where we initialized the parameters. Secondly, there is no unique minima in this parameter space.

We conjecture that this is because, our system is over-specified for five parameters. In

other words, the five parameters are not sufficient for our model. This explains the first observation because there is a limit beyond which we cannot squeeze more information from these parameters. This also directly explains the second observation – there are multiple sets of parameters that reach the same limit in different ways. Further, this also explains the variability in our results across different states. By aiming to get a smaller error across all states, the model compromises the diversity across different states. We think that by having individual sets of parameters for each state, we can overcome this issue. On the other hand, then we begin to overfit the model.

6 Conclusion

Overall, we can conclude that air travel does not explain a significant proportion of influenza spread. While the number of deaths (from simulations with air travel) is correlated with the true data, air travel itself is not a great predictor. Without the inclusion of seasonality, our model appears to simply increase the number of deaths slowly over the course of the 10 years, without any fluctuations. After including seasonality, the estimates including air travel appear to estimate the data better than the estimate without air travel. Using the spreading centrality measure and the intuitive out-degree measure, we conclude that California, Illinois, Georgia and Texas are important states in spreading influenza via air travel.

An important misrepresentation in our model was generality. We assumed the parameters to be constant over time and space. Instead, we wish to have a separate set of parameters for each state thereby capturing the state-wise variation more accurately without compromising on the aggregate results. The tradeoff with this approach is the computational complexity and the potential for overfitting our model. To make the problem more tractable, we could implement Markov Chain Monte Carlo methods to estimate parameters more efficiently than random walks. MCMC would give us a distribution for our parameters. We could then look at the maximum a posteriori estimate or the posterior mean.

Another potential improvement can be having meta-population model within each state or county. Though the flu has multiple strains and a single individual can be infected multiple times within one season, on average, adults are infected twice per decade and children contract the disease every other year [11]. We could use this to more accurately model the spread of influenza. We could also decrease granularity to city or county level. In our model, we treated LAX and SFO as a single node. If we had more data that captured this granularity, it would make our model more accurate. In addition to this, we could incorporate ground transportation data. We conjecture that ground transportation will contribute more to spreading influenza as there are air travel restrictions for sick people.

Finally, we could also include "ghost" nodes for international airports. It's highly likely that some proportion of influenza is brought in internationally. Adding these ghost nodes could increase the number of infected people within the US which would then increase the number of deaths.

References

[1] "What is the connection between influenza and pneumonia?." https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/pneumonia/what-is-the-connection.html. Accessed: 2019-08-10.

[2] V. Knight, "Viruses as agents of airborne contagion," *Annals of the New York Academy of Sciences*, vol. 353, no. 1, pp. 147–156, 1980.

[3] B. Bean, B. Moore, B. Sterner, L. Peterson, D. Gerding, and H. Balfour Jr, "Survival of influenza viruses on environmental surfaces," *Journal of Infectious Diseases*, vol. 146, no. 1, pp. 47–51, 1982.

[4] J. S. Brownstein, C. J. Wolfe, and K. D. Mandl, "Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states," *PLoS medicine*, vol. 3, no. 10, p. e401, 2006.

[5] E. Kenah, D. L. Chao, L. Matrajt, M. E. Halloran, and I. M. Longini Jr, "The global transmission and control of influenza," *PloS one*, vol. 6, no. 5, p. e19515, 2011.

[6] J. Chan, A. Holmes, and R. Rabadan, "Network analysis of global influenza spread," *PLoS computational biology*, vol. 6, no. 11, p. e1001005, 2010.

[7] "Deaths from pneumonia and influenza (p & i) and all deaths, by state and region." https://data.cdc.gov/Health-Statistics/Deaths-from-Pneumonia-and-Influenza-P-I-and-all-de/pp7x-dyj2. Accessed: 2019-09-10.

[8] "State intercensal tables: 2000-2010." https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-state.html. Accessed: 2019-09-10.

[9] "State intercensal tables: 2010-2018." https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html. Accessed: 2019-09-10.

[10] "Airline origin and destination survey (db1b)." https://www.transtats.bts.gov/tables.asp?db_id=125&DB_Name=. Accessed: 2019-09-10.

[11] Z. Villines, "How long does the flu last? timeline and recovery," Apr 2019.

[12] C. DerSarkissian, "Flu statistics: What are your odds of getting the flu?," Nov 2017.

[13] L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon, "Networks and the epidemiology of infectious disease," *Interdisciplinary perspectives on infectious diseases*, vol. 2011, 2011.

[14] S. Berger, "These are the states with the longest and shortest commutes - how does yours stack up?," Feb 2018.

A Plots for four states using monthly data

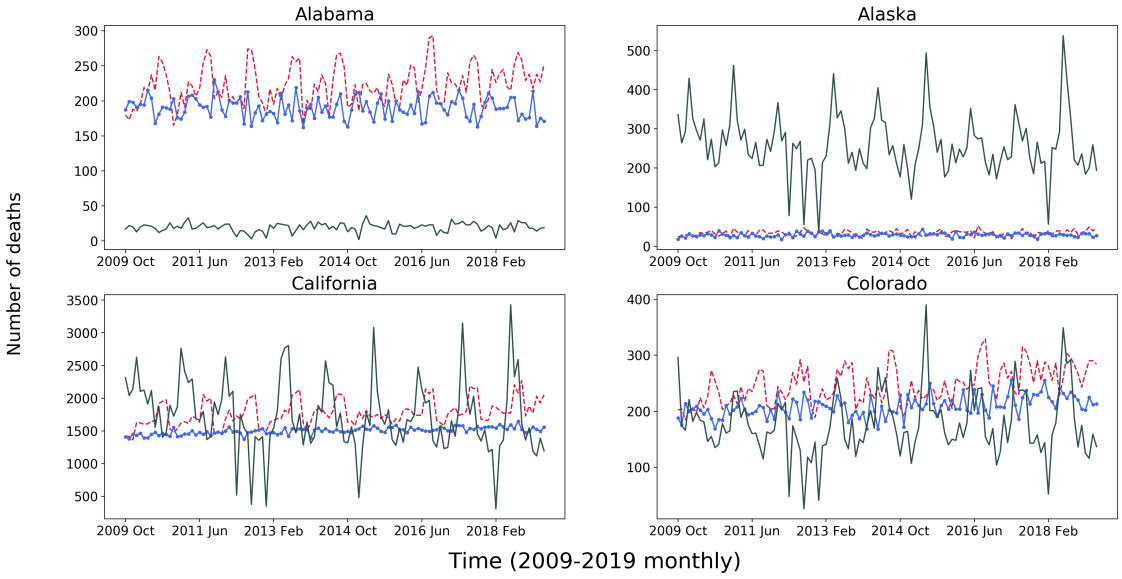

Figure 4: The number of deaths in Alabama, Alaska, California, Colorado measured without seasonality. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

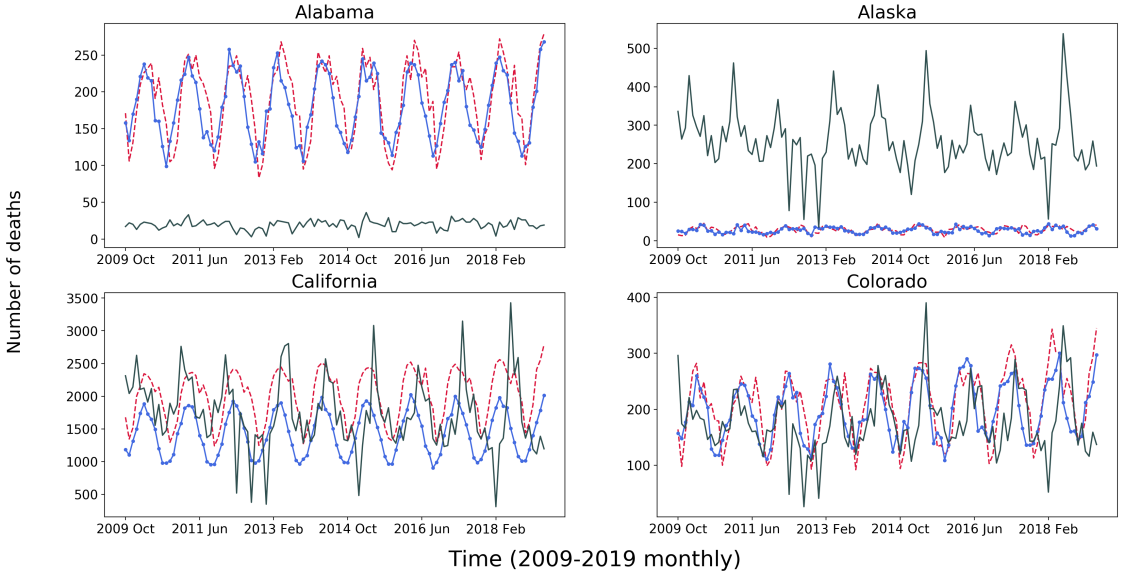

Figure 5: The number of deaths in Alabama, Alaska, California, Colorado measured with seasonality. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

B Plots using quarterly data

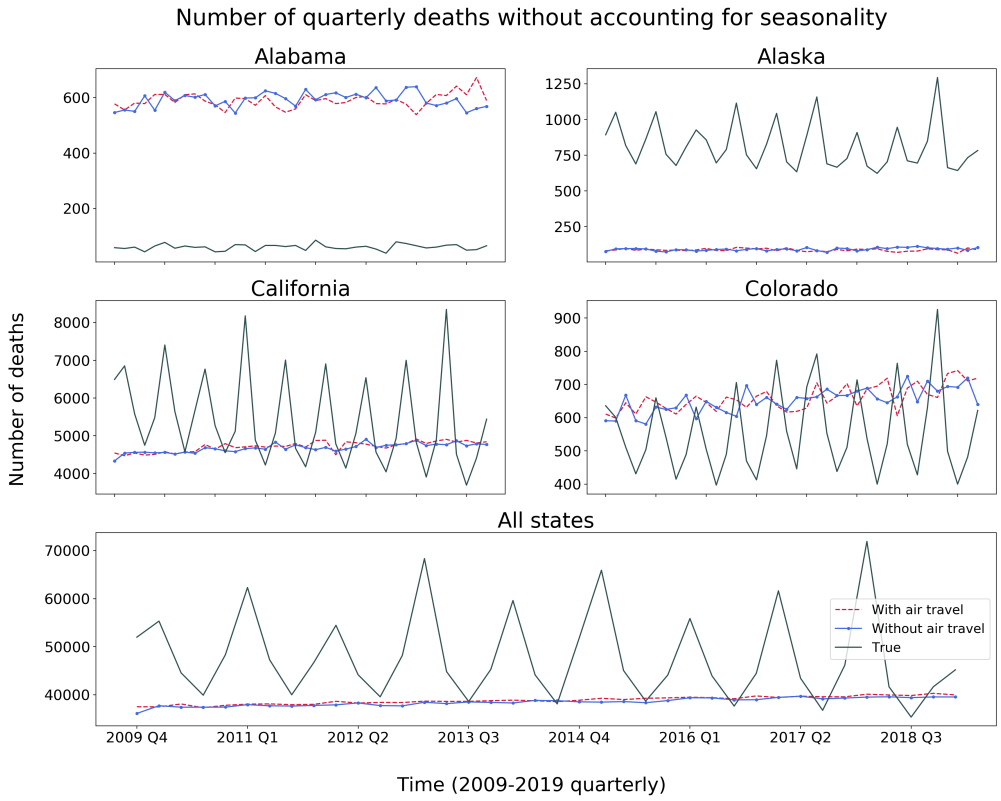

Figure 6: The number of deaths in Alabama, Alaska, California, Colorado, and across all states measured without seasonality. Each time step corresponds to a single quarter. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

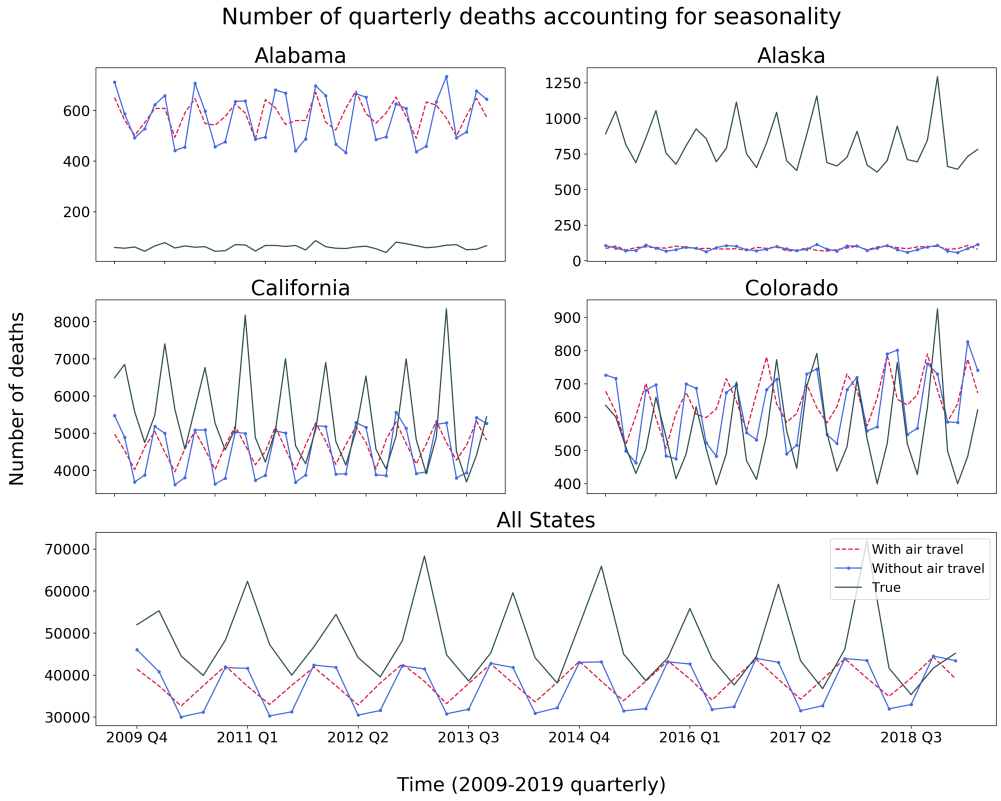

Figure 7: The number of deaths in Alabama, Alaska, California, Colorado, and across all states measured with seasonality. Each time step corresponds to a single quarter. The black line represents the real data, the red and blue lines represents the results from our model with and without accounting for air travel respectively. In this simulation, we do not include seasonality.

C Out-flow from Utah, Montana, Florida

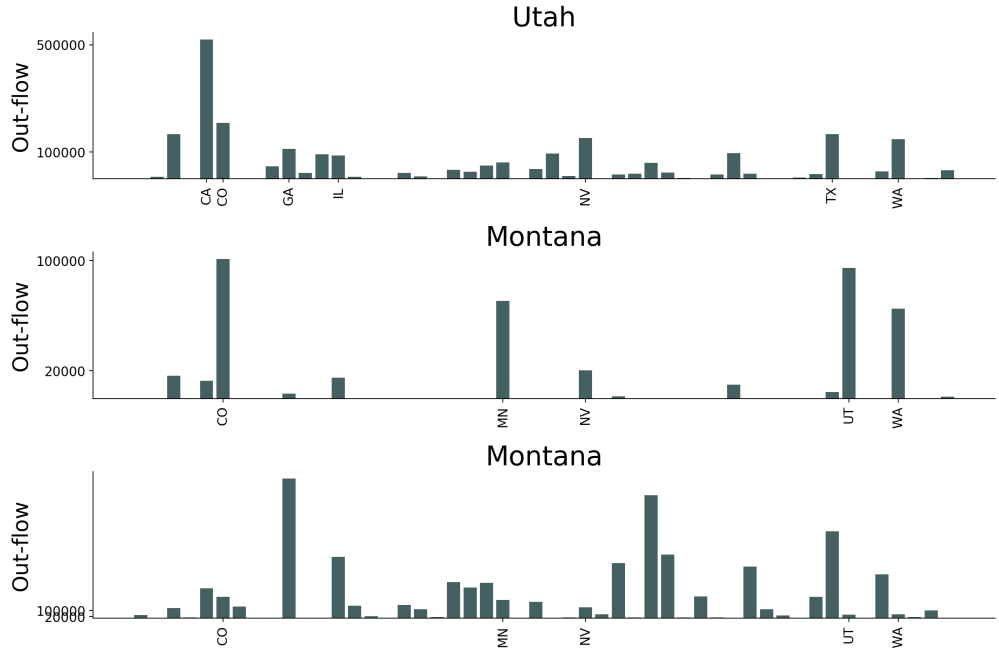

Figure 8: Average out-degree from Utah, Montana, and Florida.