Exploring source-sink dynamics of pre-vaccination measles epidemics using a spectral formulation of Granger causality

Kate Behrman, Alexander Mikheyev and Erin Taylor

Abstract—Disease transmission can occur on many scales, from cell to cell or between individuals, even between distant large-scale communities. In each of these cases, however, an effective disease outbreak strategy requires containment of the infected entity. Here, we focus on large-scale disease transmission at the community scale and introduce a method for inferring the spatial network of disease spread. This network is useful in disease containment efforts because it can identify those communities that act as sources of infection as well as reveal communities that are at risk of becoming secondary and tertiary sites of an outbreak. The method uses a measure of Granger Causality to evaluate the relationships between multiple time series of reported disease cases and the result is a directed network indicating the causal links of disease transmission between cities. We illustrate the application of this method on a data set of measles cases throughout 60 cities in England and Wales and indicate possible extensions of the method to other ecological applications.

I. INTRODUCTION

INFECTIOUS diseases spread through direct contact from person-to-person. Many epidemiological models assume that all individuals have the same probability of becoming infected and infecting other individuals in the community. For many respiratory and sexually diseases this assumption does not hold leading to incorrect estimates of the reproductive number, a measure commonly used to determine the severity of an outbreak and advise public heath intervention (Meyers et al. 2005). The incorporation of the explicit pattern of contact to model infectious disease spread increases the accuracy of model predictions and assessment of containment strategies (Loyd and May 2001, Meyers et al. 2005, Meyers et al. 2006).

Spatial contact network epidemiology models the geographic locations of individuals and the explicit contact between them in space. In its simplest form, a spatial contact network is a graph of a community of individuals represented by nodes with contact between two individuals represented by a connecting edge. In a directed or semi-directed contact network the edges are not symmetric indicating that transmission is only possible in one direction.

For large-scale analysis of infectious diseases it is infeasible to model the interactions of all individuals. Instead, each node represents a city or discrete community to which the infection is likely to spread. Knowledge of the network structure is essential for accurate model predictions and containment, however the structure of the actual disease networks is often unknown. Therefore, the network structure is estimated from previous knowledge, transportation records, or individual based microsimulations (Colizza et al. 2005, Riley 2007). Because these techniques do not utilize data from the current disease outbreaks they attempt to model, they may not accurately represent the actual network structure.

In this work, we address the problem of inferring the large-scale spatial contact network of a disease and formulate a method for modeling the network from data collected about actual disease spread. Our method employs a measure called Granger Causality to evaluate data from a time series of reported disease cases. Granger causality is a statistical measure used to deduce the relationship between two simultaneous time series. It is based on the premise that a relationship is causal from time series one to two if including past measures of series one more accurately predicts time series two (Granger 1969). Granger Causality has been used infer the directed network structure from spatially explicit time series data (Kaminski et al. 2001) and has been widely applied and accepted in econometrics, engineering, and neurobiology. A closely related measure to granger causality is the directed transfer function (DFT). The DFT is a multivariate approach that uses the entire covariance structure to deduce all directed pair wise interactions in a multivariate dataset (Kaminski and Blinowska 1991). DFT along with Ganger Causality accurately reconstructs the network relations of simulated data and neurobiological systems (Kaminski et al. 2001).

We extend the use of the DFT, in the well-accepted granger causality framework, to construct the spatial contact network for a complex ecological system, namely disease transmission, and demonstrate how this method can be applied to inferring relationships in other ecological time series data sets. To illustrate this approach, we present a case study analyzing the measles outbreaks in 60 cities across England and Wales from 1944 to 1966.

II. METHODS

A. Granger Causality and Direct Transfer Function

Consider time series data from channels $X_1(t)$ and $X_2(t)$,
which can be modeled by the following autoregressive process, where $p$ is the number of lagged observations included in the model (the model order):

$$X_1(t) = \sum_{j=1}^{p} A_{11} X_1(t-j) + \sum_{j=1}^{p} A_{12} X_2(t-j) + E_1(t)$$

$$X_2(t) = \sum_{j=1}^{p} A_{21} X_1(t-j) + \sum_{j=1}^{p} A_{22} X_2(t-j) + E_2(t)$$

$X_1$ is said to Granger-cause $X_2$ if the coefficients of $A_{21}$ are jointly non-zero, and vice versa. This definition can be straightforwardly expanded to more than three time series in order to compute conditional Granger causality.

Recently Kaminski et al. (2001) has shown that the directed transfer function (DTF) can be interpreted within the framework of traditional Granger causality, but in the frequency spectrum:

$$\theta_{21}^2(f) = |H_{21}(f)|^2 = |[A^{-1}(f)]_{21}|^2,$$

where the spectral matrix $A(f)$ is computed as $-\sum_{j=0}^{p} A(j)e^{-i2\pi j f}$ for the off-diagonal elements, and -1 on the diagonal.

### B. Determining Disease Transmission Networks

The DTF along with its equivalent measure, Granger Causality, has been used to evaluate causal relations in neural systems (Babolini et al., 2005 and Kaminski et al. 2001). In these studies, the data sets are time series of electroencephalographic recordings of the brain taken over several different channels and DTF analysis indicates which sections of the brain are causally related. We extended this approach to a measles time series data set, which contained measles cases reported every two weeks in 60 different cities across England and Wales from 1944 to 1966 (Bolker and Grenfell, 2005). The result of DTF analysis, in this case, should be a large-scale spatial transmission network indicating the causal influence between cities in regards to disease transmission.

We performed our analysis by first fitting a multivariate autoregressive model of order nine to the measles data set where each of the 60 cities was considered to be a separate data channel. We then calculated the DTF using the coefficients from the fitted MVAR model. The result was a measure of the influence of each city on the remaining 59 cities in the data set. Since DTF is calculated in the frequency domain, however, this influence is measured as a function of frequency, with the greatest frequency corresponding to the shortest sampling interval. In the case of the measles data set, the shortest time sampling interval was 2 weeks, equivalent to a frequency of 0.5. Thus, in order to obtain a single value for causal influence, we integrated DTF over the frequency range $[0, 0.5]$.

We used the empirical distribution technique introduced by Kaminski et al. (2001) in order to determine the significance of the DTF values calculated from our test data set. This required us to generate a surrogate data set by shuffling our time series from each channel in a random and independent fashion. We then fit an MVAR model to this surrogate data and calculated DTF measures from the model. We repeated this process on 1,000 surrogate data sets in order to obtain an empirical distribution of our causality measures, which was then used to estimate the behavior of the null model case. DTF values from our test data were considered significant if the probability of them arising in the null model distribution was less than 0.05 ($p < 0.05$).

Once a threshold for significance was determined, we generated the spatial network of causal disease transmission links. If the DTF from city $i$ to city $j$ was greater than the threshold of significance, a directed edge from city $i$ to city $j$ was added to the spatial transmission network.

### III. RESULTS

An example DTF calculated from our measles data set is shown in Figure 1. In this case the DTF is from city 27 (London) to city 4 (Blackburn). It can be seen that the influence of London on Blackburn peaks around week 30. Integrating the DTF over frequencies from 0-0.5 gives us a single measure of the influence of London on Blackburn, which can then be easily compared to other DTF values between the remaining cities.

If we assume that the transmission influence of a city is proportional to its population size (Rodriguez and Torres-Sorando, 2001), then larger cities should have greater DTF values. Figure 2 shows that this is indeed the case. The DTF measure increases as a function of city population size according to a power law. Thus, our measure of DTF is capturing relevant features in the measles data set.

The final spatial transmission network is shown in Figure 3. This network is laid out so that the location of the nodes...
corresponds to the geographical coordinates of the various cities in England and Wales. A blue line indicates that the edge is directed from the node with a larger vertex ID to the node with the smaller ID and a red line indicates that the edge is directed from the node with a smaller vertex ID to the node with the larger ID. Two cities are highlighted, London and Birmingham, which are two of the most highly populated cities in the data set. As expected, these nodes have a large number of connections indicating that they exert a high degree of influence on other cities. In addition, nodes that are very close to one another in the center of the network have a higher degree than those cities on the network periphery, indicating that the DTF measures may be picking up the influence of geographic proximity in disease transmission. Our future work will quantify this relationship between proximity and DTF measures while controlling for population size since this has been seen to have a significant impact on DTF values.

IV. DISCUSSION

As illustrated by our example of disease transmission networks, Granger causality may be a useful technique for uncovering relationships in complex ecological time series data sets. For example, spatial and temporal dynamics of the measles epidemic in the current example result from the interactions between numerous partially isolated populations. In ecological literature, such interacting populations are known as a metapopulation and many species exhibit metapopulation dynamics. As shown here, Granger causality may be able to find source and sink populations in such a structure. This would be useful from a basic biological perspective, as well as from the point of view of conservation biology -- when establishing conservation priorities, the sources should be preserved at the expense of the sinks.

Granger causality may also help inferring which species regulate others in a food web. For instance, a simple Lotka-Volterra type interaction where a predator and prey species reciprocally regulate each other’s populations may give rise to complex population dynamics, such as periodic oscillations. However, periodic oscillations may also result from fundamentally different types of interactions, such as one species tracking the cyclical variations of the other. This application would be similar to the original purpose of Granger causality, namely to track correlations between financial instruments, so it should also work for the analogous case of interspecific interactions. Even more promisingly, conditional Granger causality may be used to infer population regulation in multi-species food webs, which are capable of even more complex behavior.

As with any statistical measure, it is desirable to test the statistical significance of Granger causality against some plausible random model. In the case of a two-system interaction (e.g., A and B), we may be interested in comparing the Granger causality measure of A\( \rightarrow \)B with that of A\( \rightarrow \)A. We would conclude that Granger causality is significant if our measure for A\( \rightarrow \)B is greater than that of A\( \rightarrow \)A. The test can be accomplished through the use of surrogate data techniques, generating a number of autoregressive time series using just data from system B and computing a distribution of Granger causality measures. In the conditional Granger causality case with an additional system C (e.g., A\( \rightarrow \)B\( \rightarrow \)C), the autoregressive model should be computed using data from B and C. Obviously, the quality of the test will depend greatly on how well the data set is described by a linear autoregressive model, and deviations from the model assumptions may produce erroneous results.

Although Granger causality is a powerful technique for the analysis of time series data, it is correlative. Thus, despite its name, it cannot infer causality, particularly in the case where some variables are unobserved. In the worst case, Granger causality may suggest causality between variables that do no
directly interact, for instance if they are both caused by an unmeasured variable. Subject to these caveats, Granger causality may be able to provide valuable insights into systems where the number of interacting components is well understood through other ecological observations. For example, in a fully sampled system of interacting populations, such as animals living in a set of meadows deep in a forest, Granger causality measures may be reflective of actual causality. However, when using Granger causality, it is always important to specify the interacting components of a system and to explicitly justify the exclusion of others.

There are a number of useful extensions to the traditional Granger causality measure, which are not extensively explored in this report but should prove to be useful in the future. First, Granger causality can be re-formulated in the frequency spectrum, an approach first developed for the analysis of neural spike data. This approach permits the identification of frequencies at which the interaction occurs, which may provide additional insight into the functional nature of the relationship under study. Second, non-parametric Granger causality based on Fourier transform or wavelets can potentially overcome the limitations of autoregressive models, although more work needs to be done on understanding their properties and developing suitable tests of statistical significance.

ACKNOWLEDGMENT

The authors would like to thank the Santa Fe Institute for providing a stimulating and mind-opening environment at their annual Complex System Summer School where this work was engendered. They would also like to thank their advisors Mark Kirkpatrick, Timothy Keitt, Ulrich Mueller, William Ogle and Robert Holt for their insights and support.

REFERENCES


Kate D. Behrman is a Ph.D. candidate at The University of Texas in the Section of Integrative Biology. She currently works under the advisement of Dr. Mark Kirkpatrick and D. Timothy Keitt.

Alexander Mikheyev has recently graduated with a Ph.D in Ecology, Evolution and Behavior from the University of Texas. He is currently employed as a Principal Investigator at the Okinawa Institute of Science and Technology.

Erin R. Taylor is a Ph.D. candidate at The University of Florida in the Electrical and Computer Engineering department. She currently works under the advisement of Dr. William Ogle in the Biomedical Engineering department researching genetic networks and biological modeling.