Testing Modeling Assumptions in the West Africa Ebola Outbreak

Keith Burghardt[1,*], Christopher Verzijl[2], Junming Huang[3,4], Matthew Ingram[5], Binyang Song[6], and Marie-Pierre Hasne[7]

[1]Department of Physics, University of Maryland, College Park, 20742, USA
[2]ABN AMRO Bank N.V., Amsterdam, Netherlands
[3]Compleχ Lab, Web Sciences Center, University of Electronic Science and Technology of China, 611731, P. R. China
[4]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, 02115, USA
[5]Department of Political Science and Center for Social and Demographic Analysis, University at Albany, State University of New York, Albany, 12222, USA
[6]SUTD-MIT International Design Centre & Engineering Product Development Pillar, Singapore University of Technology and Design, Singapore, 487372, Singapore
[7]Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland, 97239, USA
[*]keith@terpmail.umd.edu

ABSTRACT

The Ebola virus in West Africa has infected almost 30,000 and killed over 11,000 people. Recent models of Ebola Virus Disease (EVD) have often made assumptions about how the disease spreads, such as uniform transmissibility and homogeneous mixing within a population. In this paper, we test whether these assumptions are necessarily correct, and offer simple solutions that may improve disease model accuracy. First, we use data and models of West African migration to show that EVD does not homogeneously mix, but spreads in a much more predictable manner. Next, we estimate the initial growth rate of EVD within country administrative divisions and find that it significantly decreases with population density. Finally, we test whether EVD strains have uniform transmissibility through a novel statistical test, and find that certain strains appear more often than expected by chance.

Introduction

Since 2013, the Ebola virus outbreak in West Africa has become the largest such outbreak known. The epidemic first emerged in December 2013 in southern Guinea, but as of 11 May 2016, there have been about 28,600 cases of EVD in Guinea, Liberia, and Sierra Leone, and isolated cases in the Italy, Mali, Nigeria, Senegal, Spain, the United Kingdom, and the United States. 11,300 of these cases were fatal[1] and as high as these numbers are, they may be under-estimates, due to the poor quality of current data[2]. The goal of this paper is to better understand the spread of EVD, and test the assumptions of leading EVD models.

Individuals have often been assumed to homogenously mix with each other in many recent EVD models[3–7], but we show that, by applying recent work on the migration of diseases[8], homogeneous mixing is an especially bad approximation for EVD. We find that human migration patterns help predict where and when EVD originated and will appear, which would not be possible with a homogeneous mixing hypothesis. We also find evidence that the spread of EVD is much slower than other recent diseases, such as H1N1 and SARS[8], which may have helped health workers control the disease more effectively than they otherwise would have.

Furthermore, against our expectations, we find that the initial growth rate of EVD can decrease significantly with population density, possibly because higher population density areas are correlated with other attributes, such as better healthcare. A previous model[9], in contrast, found that higher density areas should contribute to a faster rate of disease spread. Our work suggest that location-specific initial growth rates better model EVD, although the underlying reason for this heterogeneity should be a topic of future research.

Finally, we create novel metrics for the relative transmissibility of EVD strains, which are robust to sparse sampling. These metrics add to previous work on EVD in Sierra Leone[10], and provide a novel understanding of EVD strains in Guinea. We find that the relative transmissibility of strains, as measured from these metrics, is not uniform, therefore, treating EVD as a single disease may be an inappropriate assumption[3–7].

These results, when taken together, suggest unexpectedly simple ways to improve EVD modeling. In the Discussion

section, we will explain how a meta-population model can potentially aid in our understanding of disease spread and growth. Furthermore, incorporating disease strain dynamics into this model could help us better predict which strains will become dominant in the future, which may improve vaccination strategies.

Results

Models of the West Africa Ebola outbreak have often assumed that the disease spreads via homogeneous mixing[3–7]. We find, however, that this assumption may not accurately model EVD when the disease first arrives in a given area. We will first discuss how the arrival time of EVD within a country or administrative area follows a predictable pattern due to the underlying migration model, in contrast to the mixing hypothesis. Next, we model the cumulative number of individuals infected in administration divisions at the first or second level in Guinea, Liberia, and Sierra Leone to estimate the initial growth rate of EVD. We find this growth rate varies significantly, and appears to decrease with the population density within the administrative division. Finally, we introduce models of how EVD disease strains spread to rule out uniform strain transmissibility.

How Does Ebola Spread?

Homogeneous mixing models assume that healthy individuals can get sick regardless of where they are, even when they are hundreds of miles from the origin of the infection. If this is true, then the disease should be quickly seen in all susceptible areas almost simultaneously. Although this approximation may be reasonable at short distances, there has to be a lengthscale when this would break down because, in the years since Ebola first emerged, no more than a handful of countries have become infected (Fig. 1).

Alternatively, one might assume that EVD spreads spatially. There is a significant positive correlation (Spearman $\rho = 0.26$, $p < 0.05$, $n = 56$ at the spatial resolution of administration divisions, 0.81, $p < 0.05$, $n = 8$ at the country resolution) between the arrival time of EVD in a administration divisions at the first or second level and the distance between division centroids to the outbreak origin, Guéckédou, Guinea. Furthermore, this assumption has been applied successfully to model EVD in Liberia[2].

We find, however, that migration in West Africa is more complex than either of these assumptions[11]. Intuitively, diseases should spread quickly between administrative divisions or countries with significant travel between them than between isolated areas. Therefore, it is reasonable to rescale distances such that "closer" areas have greater travel between them, following work by Brockman and Helbing[8]. We find that rescaling distances using migration patterns helps us better understand how quickly EVD spreads, and estimate where the outbreak started.

To our surprise, we find that the correlation between the arrival time and effective distance for EVD (Spearman $\rho = 0.95$, $p < 10^{-3}$, $n = 8$) was consistantly higher than the correlation between the arrival time and geodesic distance, see Figs. 1 & 2. A high Pearson correlation (0.96, $p < 10^{-3}$), and agreement with the normality assumption ($p > 0.05$, using the Kolmogorov-Smirnov normality test), also suggests a linear relationship between the arrival time and effective distance. Taking the slope of the plot gives us an effective velocity of spread, which we find to be 0.015days^{-1}, which is much slower than for previous diseases (≈ 0.1 days^{-1})[8]. An intuitive explanation for our finding is that lower overall migration in West Africa reduces the speed at which EVD spreads compared to other diseases.

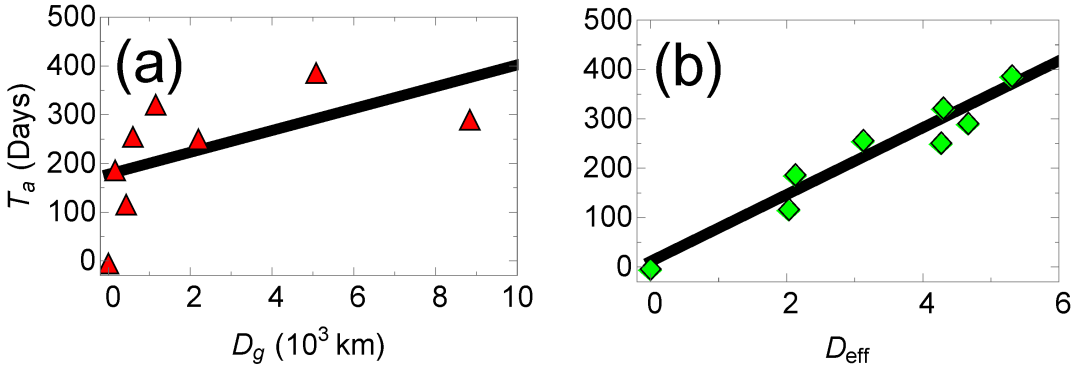

Figure 1. The arrival time, T_a, of EVD in a country versus (a) the great-arc length distance, D_g and (b) the migration-based effective distance, D_{eff} from the disease's point of origin (Guinea). The migration network used to construct the effective distance comes from census microdata[11]. The linear relationship between arrival time and effective distance suggests that there is a constant effective velocity of disease spread, in agreement with previous work on other diseases[8].

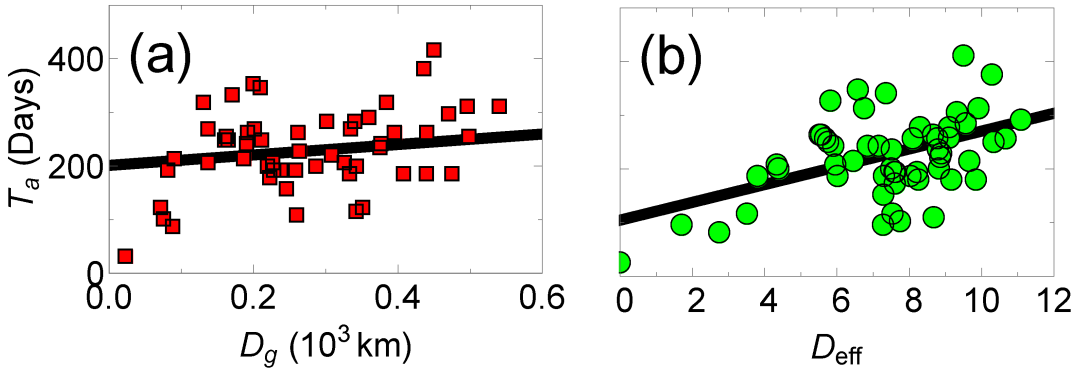

Figure 2. The arrival time, T_a, of patients with EVD in administration divisions at the first or second level within Guinea, Mali, Liberia, Sierra Leone, or Nigeria, versus (a) the great-arc length distance, D_{g}, and (b) the migration-based effective distance, D_{eff}, from the disease's point of origin (Guéckédou, Guinea). The migration network used to construct the effective distance comes from a radiation migration model[12] (similar results were found using the gravity migration models in[11]).

The reason for the lower correlations at the first or second administrative level is in part because we use migration models to determine the effective distance, and the disease spread for several months before it was detected[7]. Despite the lower quality data, however, we can still use it to determine the most likely origin of EVD, which we compare to the known origin, Guéckédou, Guinea. Quickly finding where a disease originated is important to help understand what caused it (e.g., what was the vector), and to predict where and when it will arrive, which can allow health workers to prepare[8]. Previously, it was found that the correlation between the arrival time and the effective distances from the disease origin is higher than correlations from areas where the disease did not originate[8]. We therefore ranked the correlation between the arrival times and effective distances from all the administrative divisions where Ebola was found between 2013 and 2016 (Fig. 3). We find that Guéckédou, Guinea has the highest correlation out of 56 infected administrative divisions, suggesting that, even with poor data, the region EVD originated can immediately be found.

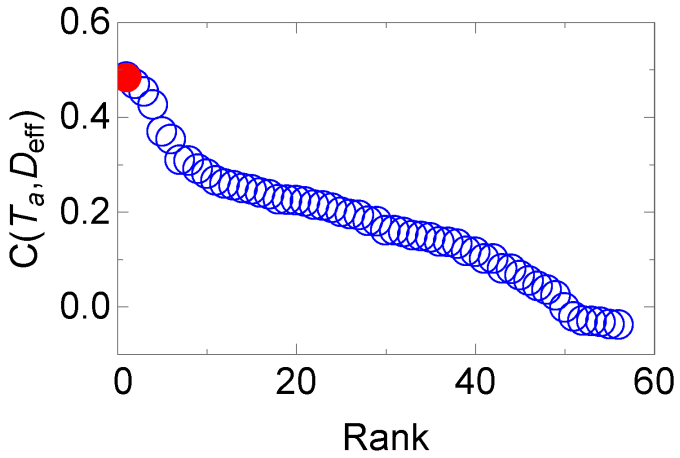

Figure 3. The (ranked) correlations between arrival time, T_a, and migration-based effective distance, D_{eff}, between administrative divisions, calculated via Eq. 5 in[8]. The migration network used to construct the effective distance comes from a radiation migration model[12] (similar results were found using the gravity migration models in[11]). In red is the true origin, which is found to have the highest correlation. Our work strengthens a previous finding[8], that the origin of a disease can be accurately estimated by correlating the arrival time to the effective distance.

In conclusion, we find strong evidence that a migration network can elucidate how quickly Ebola spreads, when and where it will arrive and where the infection began even with limited data. Furthermore, we see that alternative hypotheses for how EVD spreads, such homogeneous mixing, and nearest neighbor interactions, provide quantitatively poorer agreement with data.

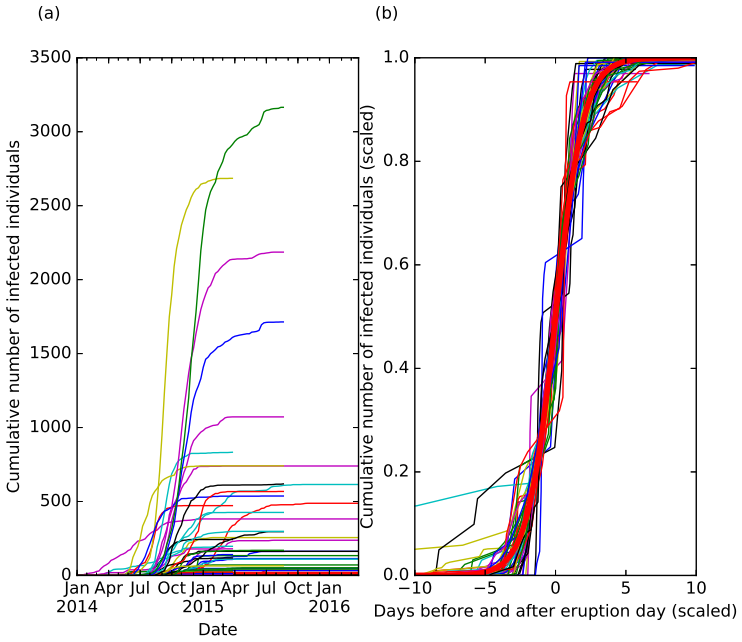

Figure 4. The cumulative number of infected individuals over time within administration divisions at the first or second level in Guinea, Sierra Leone and Liberia, from the patient database dataset[1]. It is clear that the rate and size of infections are heterogeneous, but when we fit the data to logistic functions, and renormalize the coefficients to $p_n \to 1, \tilde{t} = (t - t_0)q_n$ in (b), we see that the distributions collapse. A logistic function is plotted in red.

The Growth of EVD Across Administration Divisions

In this section, we show that the cumulative number of Ebola cases within administration divisions at the first or second level is well-approximated by a logistic function (Fig. 4a). We use this finding to estimate the initial growth rate of EVD for each infected administrative division, where, to our surprise, we find that the initial growth rate decreases with population density. The logistic function has been used to model past diseases[13], and is equivalent to the Susceptible-Infectious (SI) model when 100% of individuals are initially susceptible. To fit the SI model to our data, however, we would have to make a simplistic assumption that only a fraction p_n of individuals are susceptible, in order to explain why a small fraction of the population is ever infected. We do not claim this describes the actual dynamics of EVD, although we will explain later why the cumulative number of cases should approximately follow this distribution. For an administrative division n, the cumulative number of infected individuals over time, t, is simply:

$$i_n(t) = \frac{p_n}{1 + e^{-q_n(t - t_{0n})}}, \tag{1}$$

and the initial growth rate is q_n.

The dynamics are highly heterogeneous (Fig. 4a), therefore, it would seem un-intuitive for a single function to fit all the data. However, after fitting each cumulative distribution to the logistic model, and rescaling the variables: $p_n \to 1, \tilde{t} = (t - t_0)q_n$, we find that the distributions collapse (Fig. 4b).

Small variations in q_n very quickly become substantial variations in the infection size later on, therefore, we want to understand how q_n varies across administrative divisions. When plotting q_n versus population density (Fig. 5) with more than 20 infections total, we notice that, although q_n is ≈ 0.1 days^{-1}, which is similar to a previous EVD outbreak[14], q_n varies significantly across administrative divisions. This is in contrast to many previous models, which assume a global parameter can describe the growth of Ebola[3–7]. Furthermore, we find that q_n decreases significantly with population density (Spearman $\rho = -0.48$, $p < 10^{-3}$, $n = 44$), plausibly because better healthcare may exist in higher density areas. This contrasts with a previous model, which predicted a positive scaling relation between the mean growth rate and population of cities[9]. Therefore, not only is the growth rate of Ebola unexpectedly heterogeneous, but the dependence on population density may help us understand why this is the case.

What Makes the Data Collapse? In the SI model, 100% of individuals are eventually infected, therefore, to find agreement with data, our model has to assume a small proportion of individuals in each division are susceptible to the disease. This seems

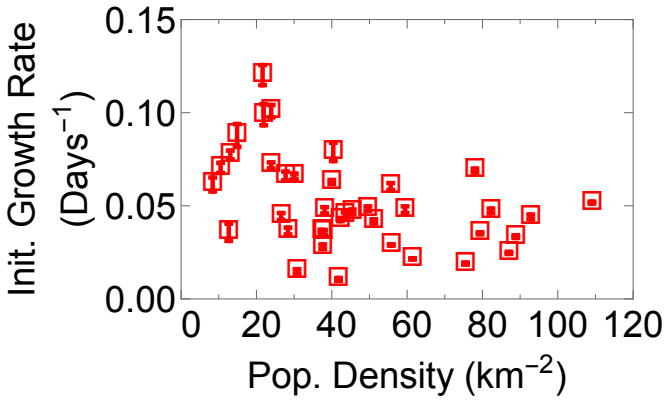

Figure 5. The initial growth rate, q_n, versus population density for the PSD dataset with more than 20 infections (outliers not visible in this plot are: Kissidougou, Conakry, and Coyah Guinea; Western Area Urban and Rural Sierra Leone; Montserrado Liberia), where error bars are standard deviations. We find that the initial growth rate drops significantly with population density (Spearman $\rho = -0.48$, $p < 10^{-3}$, $n = 44$).

implausible; more likely, all individuals are susceptible and, as they become aware of an infection, they reduce their interactions or otherwise reduce the overall disease transmissibility. To demonstrate this hypothesis, we created a more realistic, although simplistic, disease model, in which susceptible (s) individuals can become infected (i), but then recover or are removed[15]. The SIR model, like the SI model, significantly overshoots the cumulative number of cases in the absence of intervention. We therefore have the disease transmissibility decrease over time as a result of public health interventions, starting a fixed time after the outbreak starts. We call this model the Susceptible - Decreasingly Infectious - Recovered (SDIR) model.

The equations are:

$$\frac{ds_n}{dt} = -a(t)\, s_n i_n \tag{2}$$

$$\frac{di_n}{dt} = a(t)\, s_n i_n - b\, i_n \tag{3}$$

$$r_n = 1 - s_n - i_n \tag{4}$$

in normalized units, s_n the susceptible fraction of the population, i_n infected and r_n removed (recovery or death), for each administrative division.

With the recovery rate γ kept constant, we model intervention through exponential fall-off of the infection rate β with k, after some delay t_0:

$$a(t) = a\, e^{-k(t-t_0)}\ . \tag{5}$$

We see rough quantitative agreement with empirical data (e.g., Fig. 6), which suggests a similar model underlies the dynamics, in agreement with previous work[3–6].

Are Strains Uniformly Transmissible?

In this section, we use EVD genome sequences to determine what strains appear more often than expected by chance in Guinea and Sierra Leone. Our results suggest that EVD strains do not necessarily have uniform transmissibility. We are not aware of any previous models that take the strain of EVD into account, although a previous paper found certain EVD strains in Sierra Leone have different growth rates[10]. Unlike the previous paper, however, we can use our method to understand the transmissibility of strains in Guinea, where the sampling rate is otherwise too low.

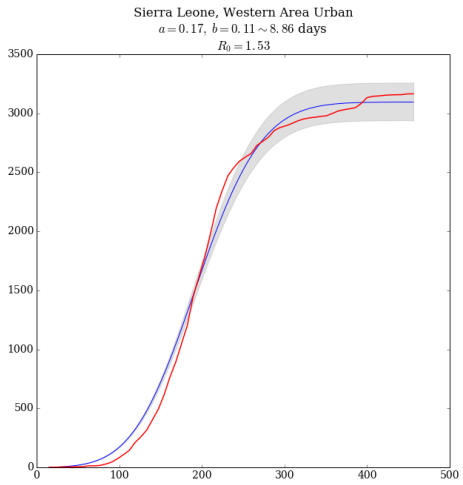

Figure 6. The fit of SDIR to the cumulative number of infected individuals in a typical county infected with EVD, where the blue line is the best fit and the shaded area represents standard errors. While the fits are poor compared to the logistic function, we see qualitative similarities to the logistic function.

Modeling Strain-Dependent Infection Probabilities

We use meta data from Ebola nucleotide sequences isolated from patients in Guinea[16] between April, 2014 and January, 2015, and Sierra Leone between late May 2014 and January, 2015[17–19], to determine when and where a strain of EVD was found, then use kernel-density estimation (KDE, see Methods) to estimate the spatial probability distribution of being infected with EVD strains up to some overall constant[20]:

$$P_E(\vec{x},t,h,\Delta t|s) = \frac{C}{n_s h}\sum_{i\in s} K\left(\frac{\vec{x}-\vec{x}_i}{h}\right) H\left(\Delta t - |t-t_i|\right). \tag{6}$$

Here, K represents a kernel with bandwidth h, around the position of each observed sequence $1 \leq i \leq n_s$, with position $\vec{x}$ at time t_i, C is an overall constant and H is the Heaviside step function. We therefore use a KDE for the sequence with a sliding time window of length Δt to represent old data becoming irrelevant as time moves forward. For the rest of the paper, our kernel is chosen to be a radially symmetric Gaussian[1]:

$$K\left(\frac{\vec{x}-\vec{x}_i}{h}\right) = K\left(\frac{\|\vec{x}-\vec{x}_i\|}{h}\right) = \frac{1}{h\sqrt{2\pi}}\mathrm{Exp}\left(\frac{\|\vec{x}-\vec{x}_i\|^2}{2h^2}\right). \tag{7}$$

By summing these probabilities, we can then find the relative probability of being infected with EVD:

$$P_E(\vec{x},t,h,\Delta t) = \frac{C}{nh}\sum_{i} K\left(\frac{\vec{x}-\vec{x}_i}{h}\right) H\left(\Delta t - |t-t_i|\right), \tag{8}$$

where $n = \sum_s n_s$. There are two free parameters: the kernel bandwidth h and sliding time window width Δt, but knowing the true infection pattern allows for these parameters to be estimated (see Fig. 7 for EVD in Guinea).

We apply the KDE to EVD in Guinea and Sierra Leone, and test whether the estimated probabilities of becoming infected correlate with the number of infected individuals within each time window. A high correlation would represent close agreement between the KDE and actual spatial probability, and would increase our trust in these findings. In Guinea, we find Spearman correlations up to 0.5 ($p < 0.01$, $n = 68$, see Fig. 7) which suggests that the KDE is in strong agreement with data. In Sierra Leone, however, these correlations are negligable. The Guinea dataset will therefore be the focus of this section.

From Fig. 7, we find that the model best correlates with Guinea's infection data if $h = 0.5$ and $\Delta t = 120$. Our findings, however, are robust to h and Δt. We let h vary between 0.1 and 1 degrees latitude and longitude, which represents length-scales of roughly 10 to 100 km, and Δt vary between 30 and 300 days.

[1] Initial results suggest this choice does not change our findings qualitatively.

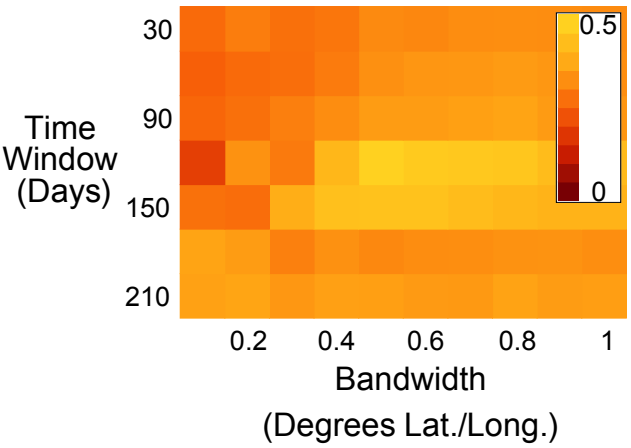

Figure 7. A heat map of the Spearman rank correlation between Eq. 8 and the number of individuals infected with EVD within a time Δt, as a function of the bandwidth h and the time window width, Δt in Eq. 8. The correlation between the model and data is highest (Spearman $\rho = 0.50$, $p < 0.01$, $n = 68$) when the bandwidth is 0.5 degrees and the time window width is 120 days.

Measuring the Relative Transmissibility of Strains

To measure the relative transmissibility of each strain, we calculate the relative probability an individual will be infected by strain A, versus any other strain. We do not compare between strains, because it may be difficult to know if a virus belongs to one strain or another, and furthermore, many infections were not found to belong to a major claude. If Strain A is found more (less) often than chance, we would say that it is stronger (weaker) than other strains overall. This is quantified by the equation:

$$Q_s(i+1) = \frac{P_s(\vec{x}_{i+1}, t_{i+1}, h, \Delta t)p(s)}{P_E(\vec{x_{i+1}}, t_{i+1}, h, \Delta t)} - I(i+1 \in s), \tag{9}$$

where I is the indicator function that a new infection's strain is s or not. In words, this quantity measure how often an infection is seen comared to chance. We define the success, S, of a strain as:

$$S_s = \frac{\sum_{i=i_0}^{i_0+m-1} Q_s(i)}{m}, \tag{10}$$

where the sum is from the first time the strain is seen, until the last time, after m sequences are sampled. P-values can be determined by bootstrapping to find how often S_s could be at least that large by chance. We find that S_s is significantly different than chance, (Fig. 8), therefore, Ebola strains might not be uniformly transmissible.

Our method may be compared to the method used by Meyer, Elias, and Hohle[21], who compared the success of invasive meningococcal disease (IMD) strains. Common surveillance algorithms were found to be practical but these appear to be more useful for a chronic rather than an acute outbreak, such as EVD.

In Guinea, we find that SL1 has one of the most negative values of S_s (Fig. 8), where all values are statistically significant ($p < 0.05$, n varies with Δt). Even seemingly weaker strains, however, have a success metric which is more negative (meaning the strain is stronger) than expected by chance. This is likely because few strains are concurrent in time, and therefore, comparing values across strains is difficult.

Discussion

In conclusion, we find several factors not often accounted for, that may improve the accuracy of modeling EVD. First, EVD appears to spread with a constant effective velocity through a migration network, in disagreement with the homogeneous mixing hypothesis (Figs. 1 & 2) commonly used to model diseases like Ebola. By taking into account the role a migration network has in disease spread, we can predict where EVD will arrive with greater accuracy than before. This method can also accelerate the process of identifying the index case by determining which administrative division is the origin through arrival time-effective distance correlation maximization. Second, we find that the growth of EVD at the finer spatial resolutions can be well described by three scaling parameters, and the initial growth rate decreases with the population density, contrary to our intuition, and a recent model[9], which suggests that a population density-dependent EVD model may more accurately predict

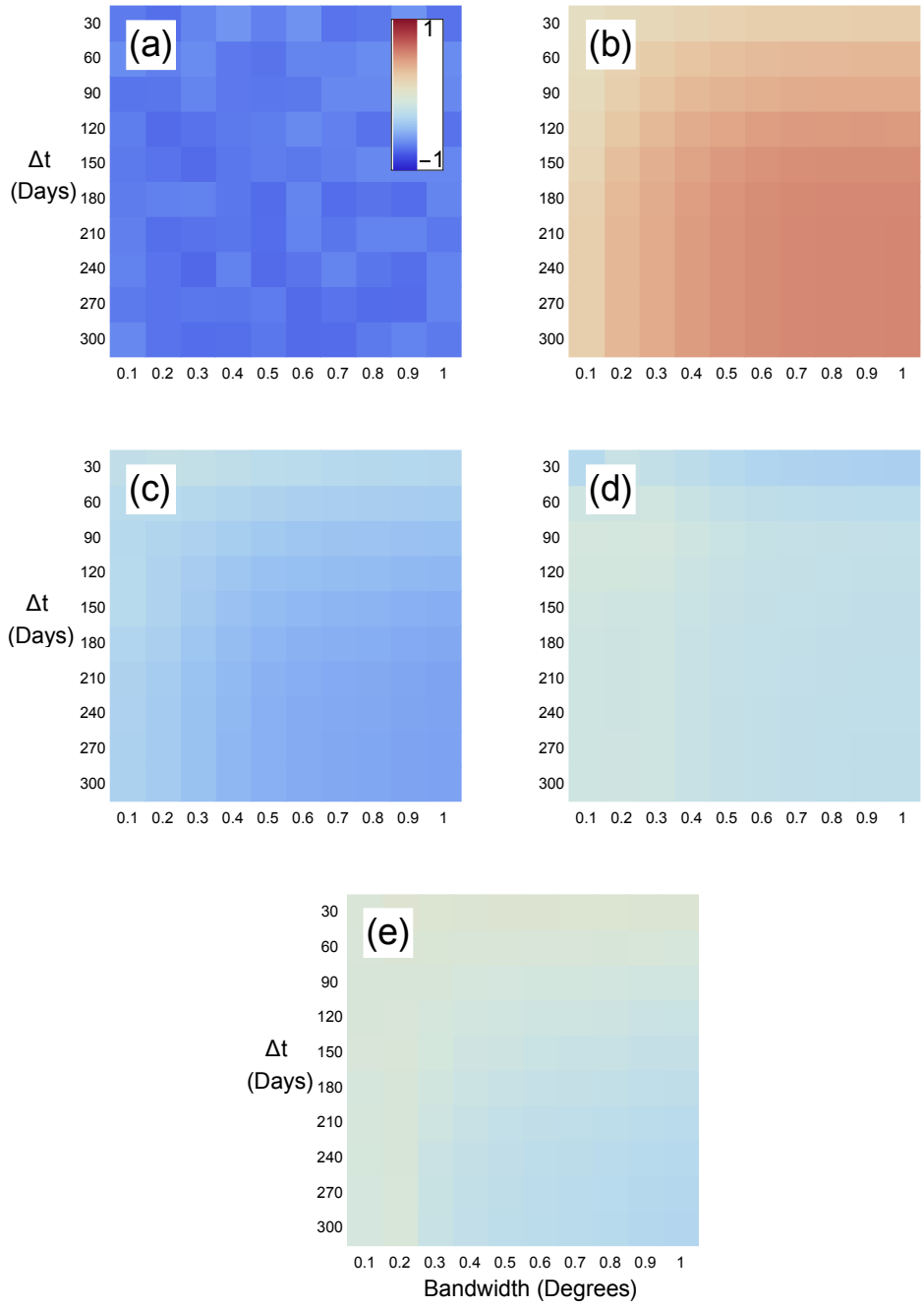

Figure 8. The success metric for strains in Guinea for (a) SL1, (b) GN1, (c) GN2, (d) GN3, (e) GN4, across various time window widths and bandwidths, where red values correspond to strains seen less often, and blue more often, than expected between the time the strain first appeared and last was seen. Sequence data comes from[16].

the spread of Ebola when the disease first arrives at an administrative division. One plausible explanation for this result is that higher population density areas recieve better healthcare than other areas, but more work is necessary to understand this behavior. Finally, we find a wide variation in the transmissibility of different strains of EVD, which suggests that modeling each disease strain, when this information is known, can improve the prediction task by reducing the heterogeneity of the data. In addition, our method may improve vaccination strategies if vaccines are made for particularly transmissible strains as well as the most common ones.

One way to take these factors into account would be through a meta-population model on top of a migration network with high spatial resolution. A meta-population model treats areas such as cities, districts, or countries as nodes on a network, with links connecting them to represent the flow of individuals from one area to another. Diseases within each area (node) are modeled with a compartmental model, under the assumption that the homogeneous mixing approximation is more accurate in smaller areas than for the entire network. Previous work has already found that a meta-population model[8], or spatio-temporal model[2], can accurately predict the spread of diseases. Finally, a strain-dependent transmissability may further improve model accuracy.

Our approach towards testing assumptions in disease modeling is not restricted to EVD, but can be applied toward other diseases of epidemiological concern. Future work is therefore necessary to test our methods on other diseases and check whether a meta-population model will better predict disease spread.

Methods

This section explains how data on the cumulative number of infected individuals at the first and second administration level (e.g., counties or districts) was gathered, how the migration network was constructed, and how we found and used strain data in Fig. 8.

Infection Data from Humanitarian Data Exchange and World Health Organization

For the 2014-2015 EVD epidemic, we are aware of two main data sources on the cumulative number of infections: World Health Organization (WHO) patient database and the WHO weekly situation reports[1]. The patient database data produces results similar to the Situation Reports, but cor consistancy, all plots in this paper use the patient database. We focused on data from December 2013 to January, 2016 for the major West African countries effected: Guinea, Liberia, Mali, Nigeria, Senegal, and Sierra Leone.

There was a significant amount of work parsing data from the patient database. Administration names in the data had multiple spellings, some administrative areas appeared individually but also aggregated with nearby areas, and finally spaces, accents and other characters appeared inconsistently. These were cleaned and harmonized. Although in the most effected countries, we have the cumulative number of cases at the second administration level, in three countries (Liberia, Mali, and Nigeria) we only have data at the first administrative level (region, not district).

Migration Data from Flowminder

Data on intra- and international migration in West Africa is taken from Flowminder[11]. For modeling migration at fine spatial resolutions, the Flowminder data contains three different data sets: (1) within countries (capturing exclusively intranational movement of people); (2) between countries (capturing exclusively international movement of people); and (3) within and between countries, capturing both intra- and international movement of people. All three data sets include the West African countries most affected by EVD as well as Benin, Cote d'Ivoire, Gambia, Ghana, Guinea Bissau, Senegal, and Togo, which had few, if any, EVD cases. In the third dataset, Flowminder collected census microdata from Public Use Microdata Series (IPUMS) on what country (or countries) an individual resided during the previous year.

To create a migration network, we first matched the Flowminder node coordinates to known district centroids (errors between centroids and node coordinates were $\pm$10km). Having matched nodes to administration names, we associated each node to the district-level arrival time of EVD, recording the date that the WHO patient database first records more a case in each administrative area. Four gravity model parameters were used to estimate the traffic between administration areas. Three were fit to migration using cell phone data in Cote d'Ivoir, Kenya, and Senegal, respectively, while the final one was fit to IPUMS data. We found that all produce similar fits, and work equally well to estimate the effective distances between areas. In addition, using population data from Geohive (next section), we created a radiation migration model, found to more accurately estimate migration patterns than gravity models[12], we therefore used this for all figures that use migration networks in this paper.

Population data from GeoHive

To determine the population density, and to estimate the migration network from the radiation migration model[12], we first find the population and area of each administrative division. First, we collected the area administrative divisions in West Africa from www.geohive.com. Next, we collected population in those districts from population census records. For districts providing

multiple census datasets, we collected from the latest report: Guinea, 2014; Liberia, 2008; Mali, 2009; Nigeria 2011; Senegal 2013; and Sierra Leone, 2004. Some population data is probably out-of-date and may lead to some quantitative inaccuracy. However, we believe that newer data will confirm our initial conclusions.

Temporal-Spatial Resolution of Sequences To determine the success of individual strains, we gathered supplementary data found in recent papers on EVD sequences in Sierra Leone and Guinea[16–19], which recorded the strain, time and location where EVD was sequenced. In the future we hope to make our own phylogenetic tree from the sequences, but for now we use the strain labels from the supplementary data itself.

A substantial fraction of sequences do not belong in any significant claude, and data from the Sierra Leone data shows that some sequences may belong to one strain or another depending on the sequences and phylogenetic tree method used. We therefore tested the success of each strain by comparing the likelihood of finding EVD with that strain versus all other sequences, using the KDE.

Although the values seen in Fig. 8 comes directly from Eq. 10, the p-values required a method that can tell us what the values of Eq. 10 would typically be, if strains were not more likely to be seen compared to chance. To determine this, we generate Bernoulli random variables (1 with probability p_i, and 0 with probability $1-p_i$) with

$$p_i = \frac{P_s(\vec{x}_{i+1}, t_{i+1}, h, \Delta t)}{P_{EVD}(\vec{x_{i+1}}, t_{i+1}, h, \Delta t)} \tag{11}$$

for each sequence $i+1$, determine the success measure S, and record whether S is greater than the empirical data. We bootstrapped using the $\{p_i\}$ values 10^4 times to obtain our p-values.

References

1. Organization, W. H. Ebola data and statistics.
2. Merler, S. *et al.* Spatiotemporal spread of the 2014 outbreak of ebola virus disease in liberia and the effectiveness of non-pharmaceutical interventions: a computational modeling analysis. *Lancet Infect. Dis.* **15**, 204–211 (2015).
3. Lewnard, J. A. *et al.* Dynamics and control of ebola virus transmission in montserrado, liberia: a mathematical modelling analysis. *The Lancet* **14**, 1189–1195 (2014).
4. Chowell, D., Castillo-Chavez, C., Krishna, S., Qiu, X. & Anderson, K. S. Modelling the effect of early detection of ebola. *The Lancet: Infectious Diseases* **15** (2015).
5. Camacho, A., Kucharski, A. J., Funk, S., Piot, P. & Edmunds, W. J. Potential for large outbreaks of ebola disease. *Epidemics* **9**, 70–78 (2014).
6. Pandey, A. *et al.* Strategies for containing ebola in west africa. *Science* **346**, 991–995 (2014).
7. Chowell, G. & Nishiura, H. Transmission dynamics and control of ebola virus disease (evd): a review. *BMC Medicine* **12** (2014).
8. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
9. Schlapfer, M. *et al.* The scaling of human interactions with city size. *J. R. Soc. Interface* (2014).
10. Luksza, M., Bedford, T. & Lassig, M. Epidemiological and evolutionary analysis of the 2014 ebola virus outbreak. *arXiv:1411.1722 [q-bio.PE]* (2014).
11. Wesolowski, A. *et al.* Commentary: Containing the ebola outbreak – the potential and challenge of mobile network data. *PLoS Currents Outbreaks* (2014).
12. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
13. Berger, R. D. Comparison of the gompertz and logistic equations to describe plant disease progress. *Phytopathology* **71**, 716–719 (1981).
14. Chowella, G., Hengartnerb, N., Castillo-Chaveza, C., Fenimorea, P. & Hyman, J. The basic reproductive number of ebola and the effects of public health measures: the cases of congo and uganda. *J. Theor. Biol.* **7**, 119–126 (2004).
15. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London* **115**, 700–721 (1927).

16. et al., M. W. C. Temporal and spatial analysis of the 2014-2015 ebola virus outbreak in west africa. *Nature* **524**, 97–101 (2015).

17. et al., S. K. G. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science* **12**, 1369–1372 (2014).

18. Tong, Y.-G. *et al.* Genetic diversity and evolutionary dynamics of ebola virus in sierra leone. *Nature* **524**, 93–96 (2015).

19. Park, D. J., Dudas, G. & et al., S. W. Ebola virus epidemiology, transmission and evolution during seven months in sierra leone. *Cell* **161**, 1516–1526 (2015).

20. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Chapman and Hall/CRC, 1986).

21. Meyer, S., Elias, J. & Hohle, M. A space-time conditional intensity model for invasive meningococcal disease occurence. *Biometrics* **68**, 607–616 (2012).

Acknowledgments

The authors would like to thank the Santa Fe Institute and St. John's College for providing us a productive working environment at the Complex System Summer School, and Samuel Scarpino for our many fruitful discussions there. Finally, K. B. would like to thank Michelle Girvan for her many insightful suggestions.

Author contributions statement

J.M., K.B. and C.V. conceived the models, C.V. and M.I. cleaned WHO Situation Reports and the patient database data on Ebola, all authors analyzed results, and all authors reviewed the manuscript.

Additional information

Competing financial interests The authors declare no competing financial interests.