Detecting changes in dispersion in COVID-19 incidence time series using a negative binomial model

Rachael Aber, Yanming Di, Ben Dalziel

Introduction

Evaluating infectious disease dynamics using only course-scale changes in incidence over time may fail to account for important parts of the social-biological context of transmission and control. For instance, the impact of non-pharmateutical interventions (NPIs; e.g., mask mandates) is often assessed in terms of changes in *mean* incidence, yet mask mandates are implemented and adhered to heterogeneously, so their effects may be context-dependent, leading to geographic and temporal variation in impact for the same policy. Similarly, the impacts of events/gatherings (such as national holidays) may result in changes in epidemic trajectories that differ based on context, and these changes may not be elucidated by simply examining aggregate mean changes in the US or even mean changes per county.

Implementing and evaluating targeted control strategies has typically required detailed data on individual-level variation in transmission rates, which are often unavailable. We present a method to identify shifts in population-level dispersion in incidence, so the impact of events or NPIs on superspreading dynamics can be assessed using freely available population-level data.

Metrics of population-level *variability* may be overlooked and useful ways to understand

epidemic dynamics and control. Population-level variability can be thought of as dispersion of data around the time-averaged trajectory. The ability to estimate changing rates of *dispersion* in epidemic time series would be a step towards a more complete view and predictive understanding of NPIs and gatherings at the individual as well as the population level. This is because individual-level heterogeneity in transmission scales up to affect population-level dynamics (Lloyd-Smith 2005), so variability in epidemic trajectories at the population level may provide information about individual-level variability in the transmission process.

From a modern theoretical ecology perspective, investigating beyond the first moment of a process has also been identified as important: ecological experiments are typically geared towards assessing the impacts of the mean strength of causal processes, however the variance about mean effects have been mostly ignored as a driver in biological assemblages, but may be as important as the mean (Benedetti-Cecchi 2003). Similarly, inference based on variability in epidemic time series' are emerging: for instance, Graham et al. (2019) use the mean and interannual coefficient of variation of measles incidence to construct a metric indicative of where a location may be on the path to elimination of the pathogen. Sun et al. (2021) found a combination of individual-based and population-based strategies was required for SARS-CoV-2 control, further highlighting the importance of considering population-level variability and its relationship to individual-level variability. Analyzing variability in epidemic dynamics in terms of bursts of incidence is also important for planning surge capacity (Wallinga 2018).

Causes of overdispersed disease incidence

Overdispersed disease incidence is suggestive of the underlying biological processes of superspreading (overdispersed *individual reproductive number*), demographic/environmental stochasticity affecting cases, or changes in propagation of the pathogen at the population level (i.e., population effective reproduction number changing in time). Note that some kinds of time dependence in the rate can cause autocorrelation, as can contagion (if it occurs outside of set periods), and heterogeneity (if an omitted variable is correlated in time) (Barron

1992). Also, demographic structure (e.g., age structure) has the potential to affect temporal autocorrelation in transmission rate - the effects of age structure can be captured by a model that includes an infection rate that varies over time (Earn et al. 1998).

Naively, the negative binomial distribution might accurately model a time series if there is a changing process mean: for example, if the mean of a Poisson distribution itself follows a gamma distribution, the resulting distribution is negative binomial (Cook 2009). Negative binomial regression (in contrast to Poisson regression) can account for unobserved heterogeneity, time dependence in the rate of a process and contagion that all lead to overdispersion (Barron 1992). Improved understanding of the causes of overdispersion in incidence may provide insight into interactions between host population structure and contagion processes, contributing to general predictive understanding in a range of systems, and understanding the limits of predictability.Previously, efforts to quantify the processes underlying overdispersed incidence have included the study of underlying offspring distributions. Also, comparison of simulated and observed incidence time series has been used to estimate the role of environmental and demographic stochasticity in measles: across community sizes, demographic stochasticity of measles becomes more important in small human populations, where dynamics can't be described as well simply by contact rate and birth rate (Grenfell et al. 2002). The final process potentially underlying overdispersed incidence, varying population effective reproduction number, has been studied extensively in investigations into seasonal forcing of population-level effective reproduction number.

Using county-level SARS-CoV-2 time series from New York state, we investigate dispersion in case count time series and its relationship to a hypothesized changepoint. Specifically, we estimate and compare dispersion on either side of a putative changepoint in variability (Thanksgiving 2020), in order to test the hypothesis that dispersion changes significantly. It was recently found that k_t, the time-varying transmission heterogeneity for COVID-19, decreased over time and was

significantly associated with interventions to slow spread in Hong Kong (Adam et al. 2022). Since k_t underlies incidence overdispersion, the method developed here will allow practitioners in epidemiology to explore whether local events result in a change in dispersion of cases, potentially modulated by changes in k_t. Key unanswered questions remain in developing methodology to recover incidence dispersion estimates and using these estimates to test the effect of gatherings and NPIs.

Data

For this project, we use county-level COVID-19 incidence data. Source: USAFacts. The data is weekly case counts for each US county.

Approximately one month of data on either side of November 26th, 2020 was used in the analysis. This step was done to include a set "wave" of incidence in the model fit to each side of the hypothesized change. However, we acknowledge that there is a trade-off between including fewer waves and increasing the power of the proposed test by including more time series data.

We additionally use the full dataset to "scan" for variability breakpoints using a likelihood ratio test statistic.

In both the full data and the subset of data, rows that were all zeroes were omitted.

Methods

Modeling the proccess mean

We propose a generalized linear model (GLM) approach to quantify incidence dispersion in segments of a time series, using a single parameter, θ, while separately accounting for population size in the model. Here we account for unobserved heterogeneity, and

time dependence in rate - potentially due to contagion - (Barron 1992) by specifying an overdispersed conditional distribution.

Issues arise when modeling the entire time series as negative binomial - namely, that autocorrelation remains unaccounted for (Barron 1992). A recently proposed negative binomial regression model for time series of counts also accommodates serial dependence (Davis and Wu 2009). The approach presented here is similar, but we use a linear predictor, η, that includes a natural spline in time to account for autocorrelation in case counts. Natural splines are cubic splines which are linear outside of the boundary knots (Perperoglou et al. 2019). The degrees of freedom for the spline were chosen to be three, as there appeared to be no more than one knot in a time series of 30 days. In addition, increasing degrees of freedom above three results in overfitting to the time series and thus potentially insignificant changes in dispersion are more readily detected.

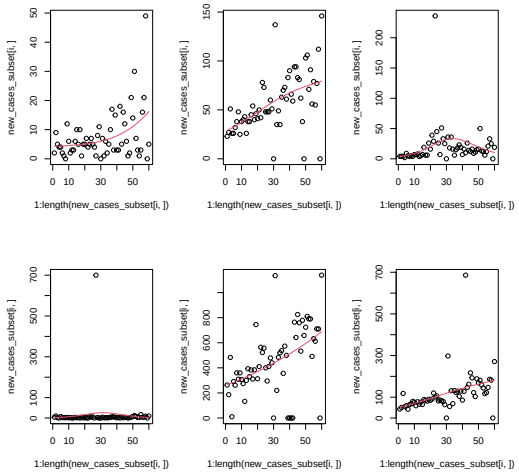

Figure 1: Fitted Curves With Three Spline Degrees of Freedom

$$g(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i)$$

$$log(E[Y_i]/n_i) = \sum_{m}^{M} \beta_m h_m(t_i)$$

,where n_i is the population size, Y_i is incidence, and t_i is the time point.

In this case, we use a log link function:

$$log(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i)$$

$$log(E[Y_i]) - log(n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i)$$

$$log(E[Y_i]) = \beta_1 (h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) + log(n_i)$$

IRLS

Iteratively reweighted least squares (IRLS) is used to obtain parameter estimates for the model. Briefly, IRLS works by computing an adjusted dependent variable, computing a weight matrix, and then calculating (using weighted least squares) coefficients estimates of the model used to calculate the linear predictor as well. This procedure is implemented via the NBPSeq R package (https://CRAN.R-project.org/package=NBPSeq) and from Di et al. (2011).

Inclusion of an offset term in the model

Regardless of whether the data is modeled as negative binomial or Poisson around the process mean (and regardless of whether quasi-likelihood methods are used), it's crucial to incorporate an offset term - sometimes called an exposure variable - in order to directly model counts (here, COVID-19 cases) per unit of observation (here, per individual). Essentially, instead of focusing on the infection rate, our model allows the focus to be on per person infection rate. In other words, an offset term was added to the model to account for case counts resulting from different population sizes - an offset for population size means that incidence dispersion properties may be assessed *while accounting for a population effect*. Note that if the offset model were perfect, the estimated coefficient on log population would be exactly one.

Identifying an appropriate model of the conditional distribution

To identify an appropriate conditional distribution, the value of Poisson model residual deviance (D) divided by its degrees of freedom (df) was used to verify that overdispersion is present under this model and found that only a few dispersion statistics were less than or equal to one. Since residual deviance of a model is expected to be asymptotically distributed as χ^2_{df}, where the degrees of freedom are the number of observations minus the number of estimated parameters, an indication that a model may be underestimating dispersion (that the data is inconsistent with the model) is that D/df » 1. The proportion of dispersion statistics that are greater than one in US counties over the Thanksgiving period is 0.9907526.

Directly comparing between quasi-Poisson and Poisson models via deviance residuals isn't possible because the deviance residuals are identical. Therefore, to evaluate whether the negative binomial conditional distribution is needed (opposed to a Poisson/quasi-Poisson conditional distribution with the same model of process mean), inspection of mean-variance relationships using a diagnostic plot is possible (Ver Hoef and Boveng 2007). In the quasi-Poisson model (called the linear negative binomial specification by Barron (1992)), the variance is a linear function of the mean, with a slope of one representing the Poisson special case. By contrast, the negative binomial model may more appropriately represent the data when the variance has approximately a quadratic relationship with the mean. Since the true mean and variance of the data set is unknown, one could approximate these quantities with estimates $\hat{\mu}$ by and $\widehat{\sigma^2}$, respectively. These quantities should be estimated using both the Poisson and the negative binomial model. Variance can be estimated by averaging squared residuals in each fitted mean category. The number of fitted mean categories (bins) should be chosen such that there are an adequate number of observations informing each variance estimate, and also an adequate number of variance estimates to visualize whether a trend is present. Visualization of individual squared residuals is also possible, but these may not be reliable estimates of the variance.

We use a diagnostic plot approach in which we rely on expected value of the residuals to infer "p" using the slope of a regression of log(Y) on log(X), where X is residualˆ2 - Y.

$$E[X] = \phi * \mu^p$$

$$log(E(X)) \sim log(\phi) + p * log(E(Y))$$

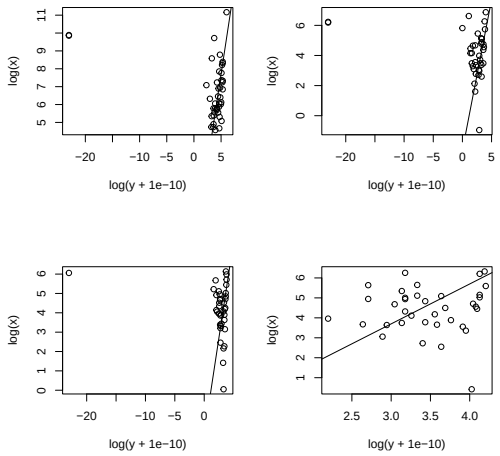

Figure 2: Diagnostic Plot

We use a second diagnostic plot, squared residuals versus squared fitted values to look for a linear trend in the Florida counties with a significant LRT test result for the Thanksgiving hypothesis. The constant dispersion model (NB2) model is appropriate if we observe an approximately linear trend.

Proposed hypothesis test

We tested the hypothesis that the incidence dispersion parameter changes at the Thanksgiving changepoint of interest: $\theta_1 \neq \theta_2$. Hypothesis tests were conducted to assess whether θ changes after Thanksgiving 2020. The Wald test was used, where the estimates of θ in different segments are considered independent. This approach is possible because we used IRLS estimates of θ - the standard error of θ was found utilizing the observed information matrix.

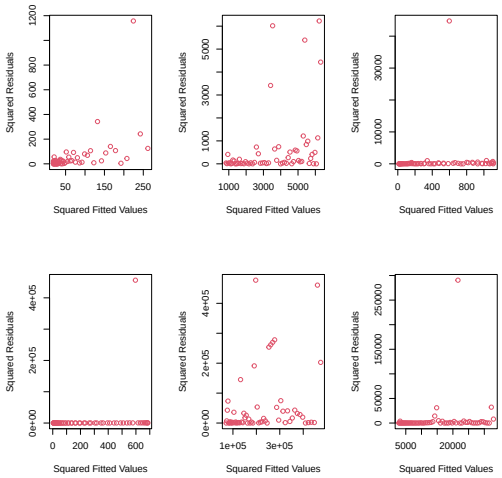

Figure 3: Diagnostic Plot 2

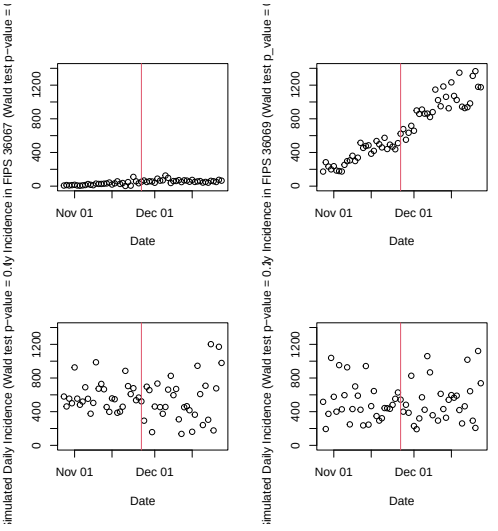

Figure 4: Empirical change in dispersion

$$Cov(\hat{\theta}_1, \hat{\theta}_2) = 0$$

The form of the Wald test statistic is:

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{SE(\hat{\theta}_1 - \hat{\theta}_2)}$$

$$= \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\widehat{Var(\hat{\theta}_1)} + \widehat{Var(\hat{\theta}_2)}}}$$

Results

Thanksgiving 2020: Wald test

For every county, the model was fit separately to either side of November 26th, 2020 in order to compute values of $\hat{\theta}_1$ and $\hat{\theta}_2$. A test of the two-sided alternative that θ before is different from θ after (different theta after Thanksgiving) was performed.

Simulation

Validity and power of the Wald test

To ensure that the hypothesis testing framework is valid, a simulation with known (equal) values of theta on either side of Thanksgiving 2020 was run in order to ensure that the distribution of resulting p-values from the test of the two-sided alternative is approximately uniform on the interval from zero to one. In both the validity and power simulations, Gaussian and uniform epidemic curves (see figure above which demonstrates correct behavior of the test in either case) are used. The area under these curves (final outbreak size) is set to be the population size divided by ten.

To implement these constraints on the integral in the Gaussian case:

we use a normalizing constant of (n/10) * sqrt($2pi\ \sigma^2$)

$$\int_0^\infty (n/10)/\sqrt{2\pi(\sigma^2))}e^{-(t-61)^2/(2\sigma^2)}dt = n/10$$

So, we have the desired Gaussian function:

$$\mu_t = (n/10)/\sqrt{2\pi\sigma^2}e^{-(t-61)^2/(2\sigma^2)}$$

So that integration will result in a final outbreak size of n/10.

To ensure that the hypothesis test has adequate power, a simulation using unequal values of theta on either side of Thanksgiving was repeatedly run in order to ensure that the distribution of resulting p-values is heavily right-skewed. The resulting probabilities of rejection are tabulated below: the proportion of tests that reject is around 0.05 for those instances where the null is true (the probability of the normal variable having 0.05 of the probability density below is 0.05). Additionally, the proportion of tests that reject is high in instances where the alternative hypothesis is true. We also simulated data assuming that theta was a function of the mean of the process in order to assess the robustness of the test (Supplement).

Additionally, the relationship between the p-values resulting from the Wald test and the effect size is as expected.

Validity and power of the LRT

The tabulated Type I error rate and power, as well as the relationship between p-values and effect size is presented below for the likelihood ratio test approach.

	Type.I.Error.Rate	Power.at.3	Power.at.9
1	0.04	0.18	0.88

Figure 5: Type I Error Rate and Power - Wald Test

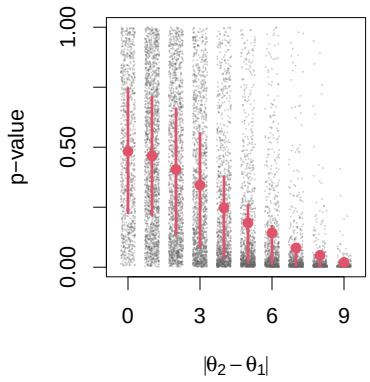

Figure 6: Wald Test p-values v. Theta Difference

Biogeography of Superspreading Dynamics

The spatial pattern of theta difference after Thanksgiving 2020 shows that there is spatial correlation in the degree of change. In particular, spatial clusters of increased dispersion occur on the East Coast and in the South.

Additionally, the likelihood ratio test statistic does not necessarily track with changes in mean incidence in a county, but rather captures dispersion changes. In some time windows, the Likelihood-Ratio statistic can't be computed due to zero variance: the variance of mu is a function of mu and dispersion, both of which are zero in a stretch of zeroes.

In visualizing a surface of the likelihood ratio test statistic and associated direction of the change in theta, it appears that

	Type.I.Error.Rate	Power.at.3	Power.at.9
1	0.08	0.2	0.83

Figure 7: Type I Error Rate and Power - LRT Test

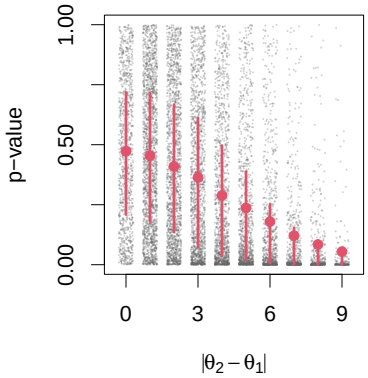

Figure 8: LRT p-values v. Theta Difference

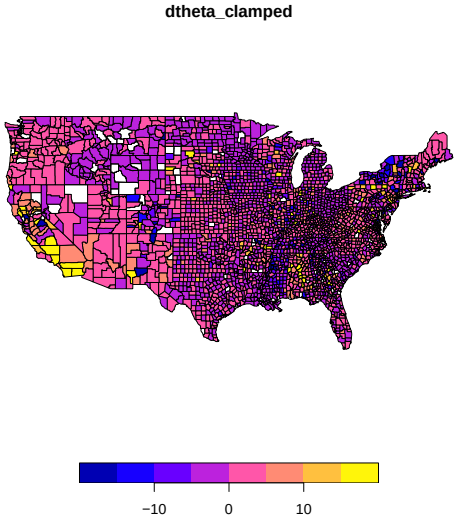

Figure 9: Mapping the Thanksgiving hypothesis across the US

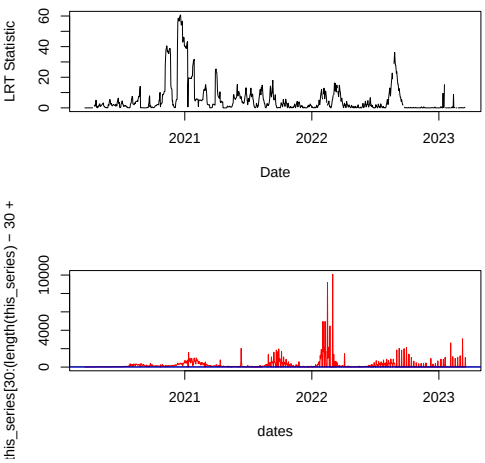

Figure 10: LRT Scan Figure

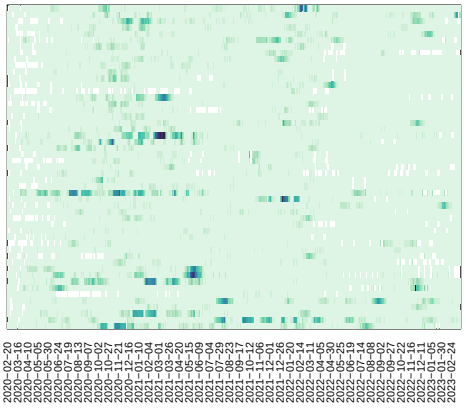

Figure 11: The LRT Surface

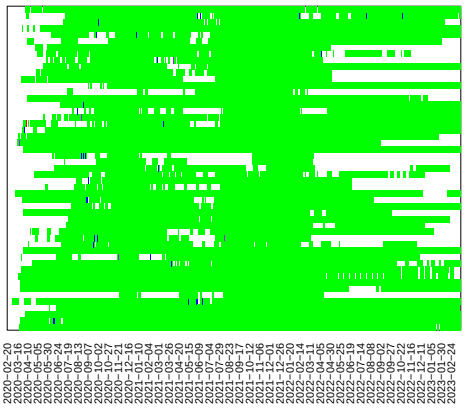

Figure 12: The Theta Difference Surface

References

Adam, D., Gostic, K., Tsang, T., Wu, P., Lim, W. W., Yeung, A., Wong, J., Lau, E., Du, Z., Chen, D., Ho, L.-M., Martín-Sánchez, M., Cauchemez, S., Cobey, S., Leung, G., & Cowling, B. (2022). Time-varying transmission heterogeneity of SARS and COVID-19 in Hong Kong. https://doi.org/10.21203/rs.3.rs-1407962/v1

Cook, J. D. (n.d.). Notes on the Negative Binomial Distribution. 5.

Davis, R. A., & Wu, R. (2009). A negative binomial model for time series of counts. Biometrika, 96(3), 735–749.

Graham, M., Winter, A. K., Ferrari, M., Grenfell, B., Moss, W. J., Azman, A. S., Metcalf, C. J. E., & Lessler, J. (2019). Measles and the canonical path to elimination. Science, 364(6440), 584–587. https://doi.org/10.1126/science.aau6299

Karlis, D., & Xekalaki, E. (2000). A Simulation Comparison of Several Procedures for Testing the Poisson Assumption. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 355–382. https://doi.org/10.1111/1467-9884.00240

Killick, R., & Eckley, I. A. (2014). changepoint: An R Package for Changepoint Analysis.

Journal of Statistical Software, 58, 1–19. https://doi.org/10.18637/jss.v058.i03

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066), Article 7066. https://doi.org/10.1038/nature04153

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. BMC Medical Research Methodology, 19(1), 46. https://doi.org/10.1186/s12874-019-0666-3

Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity. II. The Poisson distribution. Biometrika, 53(1), 183–190. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. (n.d.). https://doi.org/10.1126/science.abe2424

Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. (n.d.). https://doi.org/10.1126/science.abe2424

Venables W.N., Ripley B.D. (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Ver Hoef, J. M., & Boveng, P. L. (2007). QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA? Ecology, 88(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Wallinga, J. (2018). Metropolitan versus small-town influenza. Science. https://doi.org/10.1126/science.aav1003