Detecting changes in dispersion in COVID-19 case counts using a negative binomial model

JSM 2024

Rachael Aber, Yanming Di, Ben Dalziel

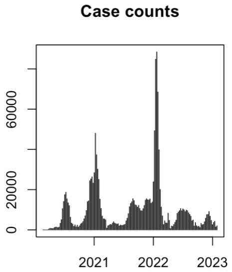
August 2, 2024

Oregon State University

Table of Contents

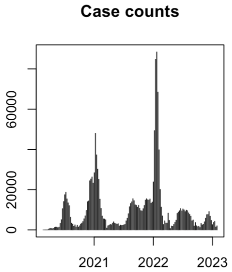
Table of Contents

Why study variability?



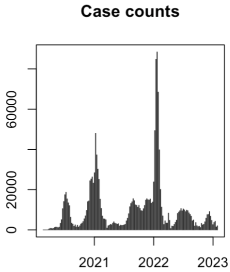
- Highly variable case count time series suggest transmission heterogeneity or changes in R

Why study variability?



- Highly variable case count time series suggest transmission heterogeneity or changes in R
- Metrics of variability are overlooked: "How is variability related to different phases of an epidemic?" [3]

Why study variability?



- Highly variable case count time series suggest transmission heterogeneity or changes in R
- Metrics of variability are overlooked: "How is variability related to different phases of an epidemic?"[3]
- Adam et al.[1] found that COVID-19 transmission heterogeneity decreased over time

Why study variability?

- Dispersion of a case count time series may be a useful metric–part of a framework that models variance flexibly

Why study variability?

- Dispersion of a case count time series may be a useful metric–part of a framework that models variance flexibly
- A 'mean crowding' parameter [5] was proposed: mean number per individual of other individuals in the same quadrat

Why study variability?

- Dispersion of a case count time series may be a useful metric–part of a framework that models variance flexibly
- A 'mean crowding' parameter [5] was proposed: mean number per individual of other individuals in the same quadrat
- Useful way to think about dispersion in case count time series, degree of dispersion is degree of clustering/crowding of cases (from the perspective)

Oregon State University

Why study variability?

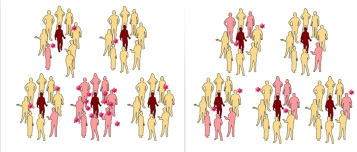
- Dispersion of a case count time series may be a useful metric–part of a framework that models variance flexibly
- A 'mean crowding' parameter [5] was proposed: mean number per individual of other individuals in the same quadrat
- Useful way to think about dispersion in case count time series, degree of dispersion is degree of clustering/crowding of cases (from the perspective)
- 

Table of Contents

Introduction to the method

- Let λ_t be epidemic intensity, I_t be incidence at time t, and θ_t be dispersion at time t

Introduction to the method

- Let λ_t be epidemic intensity, I_t be incidence at time t, and θ_t be dispersion at time t
- $$\mathrm{I_t = NB(\mu = \lambda_t, \theta_t = I_{t-1})}[4] \tag{1}$$

Introduction to the method

- Let λ_t be epidemic intensity, I_t be incidence at time t, and θ_t be dispersion at time t
- $$\mathrm{I_t = NB(\mu = \lambda_t, \theta_t = I_{t-1})}[4] \tag{1}$$
- Adjusted for population size using an offset in the model

Introduction to the method

- Let λ_t be epidemic intensity, I_t be incidence at time t, and θ_t be dispersion at time t
- $$\mathrm{I_t} = \mathrm{NB}(\mu = \lambda_\mathrm{t}, \theta_\mathrm{t} = \mathrm{I_{t-1}})[4] \tag{1}$$
- Adjusted for population size using an offset in the model
- Linear predictor includes a natural spline in time to account for autocorrelation in case counts (ns are)

$$log(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{2}$$

$$log(E[Y_i]) - log(n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{3}$$

$$log(E[Y_i]) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) + log(n_i) \tag{4}$$

Oregon State University

Introduction to the method

- Let λ_t be epidemic intensity, I_t be incidence at time t, and θ_t be dispersion at time t
- $$\mathrm{I_t} = \mathrm{NB}(\mu = \lambda_\mathrm{t}, \theta_\mathrm{t} = \mathrm{I_{t-1}})[4] \tag{1}$$
- Adjusted for population size using an offset in the model
- Linear predictor includes a natural spline in time to account for autocorrelation in case counts (ns are)

$$log(E[Y_i]/n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{2}$$

$$log(E[Y_i]) - log(n_i) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) \tag{3}$$

$$log(E[Y_i]) = \beta_1 h_1(t_i) + \beta_2 h_2(t_i) + \beta_3 h_3(t_i) + log(n_i) \tag{4}$$

- Model fit on a rolling basis to each time series (one estimate)

Oregon State University

Negative binomial model

- $$f_t(I) = \binom{I+\theta-1}{I} \left(\frac{\mu}{\mu+\theta}\right)^I \left(\frac{\theta}{\mu+\theta}\right)^\theta \qquad (5)$$

Oregon State University

Negative binomial model

- $$f_t(I) = \binom{I+\theta-1}{I}\left(\frac{\mu}{\mu+\theta}\right)^I\left(\frac{\theta}{\mu+\theta}\right)^\theta \tag{5}$$

- $$E(I) = \mu \tag{6}$$
  $$Var(I) = \mu + \frac{\mu^2}{\theta} \tag{7}$$

Oregon State University

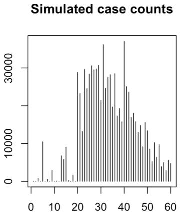
Negative binomial model

- $$f_t(I) = \binom{I+\theta-1}{I} \left(\frac{\mu}{\mu+\theta}\right)^I \left(\frac{\theta}{\mu+\theta}\right)^\theta \tag{5}$$

- $$E(I) = \mu \tag{6}$$
$$Var(I) = \mu + \frac{\mu^2}{\theta} \tag{7}$$

- LRT framework: at each time step along a time series, fit null and full(θ change) model, conduct LRT to produce p-value at each time point

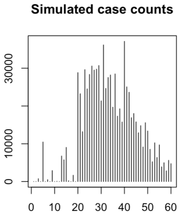
Simulated time series data set



- Validity/power simulations: Gaussian and uniform epidemic curves with an attack rate of 0.1

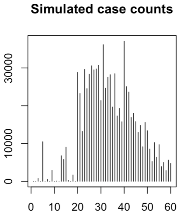
Oregon State University

Simulated time series data set



- Validity/power simulations: Gaussian and uniform epidemic curves with an attack rate of 0.1
- Varying magnitude of θ change, location of the change, underlying population size, and epidemic curve shape (allowed)

Simulated time series data set



- Validity/power simulations: Gaussian and uniform epidemic curves with an attack rate of 0.1
- Varying magnitude of θ change, location of the change, underlying population size, and epidemic curve shape (allowed)
- Epidemic curves over 60 time steps each were produced, and the likelihood-ratio test (LRT) procedure was applied to each

Application to empirical data

- Weekly case counts from US counties between 2020-01-04 and 2023-03-18

Application to empirical data

- Weekly case counts from US counties between 2020-01-04 and 2023-03-18
- Estimated θ_t for 154 time steps for 144 US counties (IRLS procedure implemented via the NBPSeq package[7] and from Di et al.[6])

Application to empirical data

- Weekly case counts from US counties between 2020-01-04 and 2023-03-18
- Estimated θ_t for 154 time steps for 144 US counties (IRLS procedure implemented via the NBPSeq package[7] and from Di et al.[6])
- We investigated large counties (largest three counties in each state)

Oregon State University

Table of Contents

Results: simulated data

- Negative binomial/LRT method is robust to differences in population size (for population sizes examined)

Oregon State University

Results: simulated data

- Negative binomial/LRT method is robust to differences in population size (for population sizes examined)
- Illustrated that an increase in θ is associated with decreased variability in simulated incidence time series (same relationship is observable in the empirical time series), with an increase in θ corresponding to a decrease in variability around the trend in incidence

Oregon State University

Results: simulated data

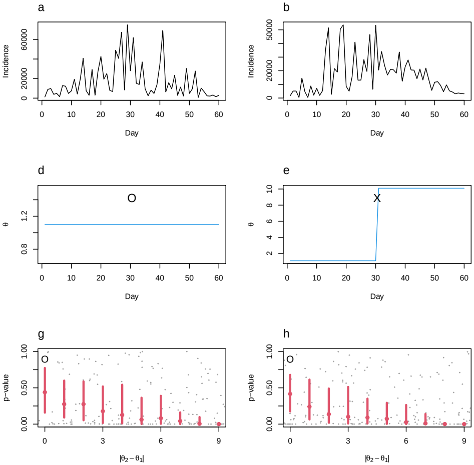


Figure: Detecting dispersion changes in incidence time series in populations of different sizes. A/B: Simulated incidence when dispersion is constant/changes. C/D: Constant/changing dispersion used in generation of above. E: Performance of LRT with simulated data that has different absolute differences in theta (horizontal axis of each pane) illustrates p-value distribution across

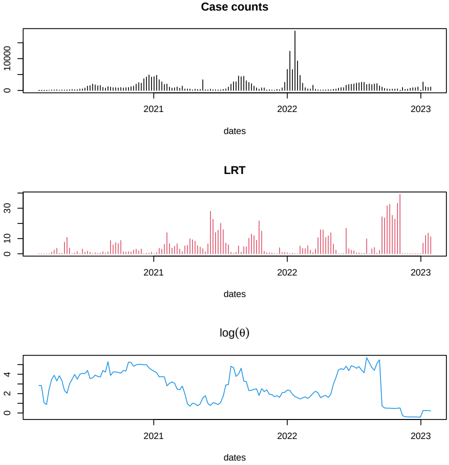


Figure: Method applied to case counts between 2020-01-04 and 2023-03-18 for Jefferson County, AL. A: Case counts . B: LRT statistic C: Log dispersion parameter.

Results: empirical data

- Highly overdispersed incidence observed more frequently later in time series (consistent) (Most dispersed category in Fig reaches)

Results: empirical data

- Highly overdispersed incidence observed more frequently later in time series (consistent) (Most dispersed category in Fig reaches)
- Evidence for a change in θ was observed across many counties (evidenced by concentration of low p-values around peak incidence)

Results: empirical data

- Highly overdispersed incidence observed more frequently later in time series (consistent) (Most dispersed category in Fig reaches)
- Evidence for a change in θ was observed across many counties (evidenced by concentration of low p-values around peak incidence)
- High dispersion may indicate less diffuse epidemics that are potentially more subject to climate forcing[2], or increased locally experienced mean density [5]

Results: empirical data

- Highly overdispersed incidence observed more frequently later in time series (consistent) (Most dispersed category in Fig reaches)
- Evidence for a change in θ was observed across many counties (evidenced by concentration of low p-values around peak incidence)
- High dispersion may indicate less diffuse epidemics that are potentially more subject to climate forcing[2], or increased locally experienced mean density [5]
- Raising variance relative to mean implies spatiotemporal "crowding" of cases (i.e. localized surges) which may necessitate more surge capacity in hospitals and testing centers

Oregon State University

Results: empirical data

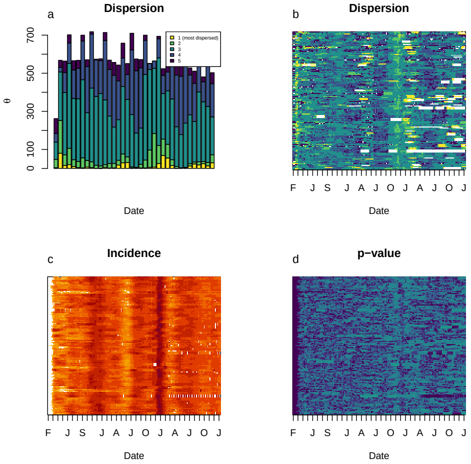


Figure: Incidence and dispersion in large counties in the US. A: Binned log of the dispersion parameter. B: Log of the dispersion parameter for each of the large counties (y-axis). C: Log incidence for each of the large counties (y-axis). D: LRT p-values for each of the large counties (y-axis).

Oregon State University

Table of Contents

Concluding remarks

- LRT procedure performs well for relevant range of population sizes

Concluding remarks

- LRT procedure performs well for relevant range of population sizes
- Dispersion parameter and LRT statistic don't simply reflect changes in process mean

Concluding remarks

- LRT procedure performs well for relevant range of population sizes
- Dispersion parameter and LRT statistic don't simply reflect changes in process mean
- Dispersion is high at unexpected times (near peak incidence)(changes)

Oregon State University

Concluding remarks

- LRT procedure performs well for relevant range of population sizes
- Dispersion parameter and LRT statistic don't simply reflect changes in process mean
- Dispersion is high at unexpected times (near peak incidence)(changes)
- Methods that use time series are crucial (due to); timing/allocation of public health resources can be achieved (with, pop less effective)

Table of Contents

[1] Dillon Adam et al. *Time-varying transmission heterogeneity of SARS and COVID-19 in Hong Kong*. Tech. rep. ISSN: 2693-5015 Type: article. Mar. 2022. DOI: 10.21203/rs.3.rs-1407962/v1. URL: https://www.researchsquare.com/article/rs-1407962/v1 (visited on 03/28/2022).

[2] Benjamin D. Dalziel et al. "Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities". In: *Science* 362.6410 (Oct. 2018). Publisher: American Association for the Advancement of Science, pp. 75–79. DOI: 10.1126/science.aat6030. URL: https://www.science.org/doi/10.1126/science.aat6030 (visited on 10/05/2022).

[3] Matthew Graham et al. "Measles and the canonical path to elimination". en. In: *Science* 364.6440 (May 2019), pp. 584–587. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau6299. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aau6299 (visited on 08/02/2021).

Thanks!