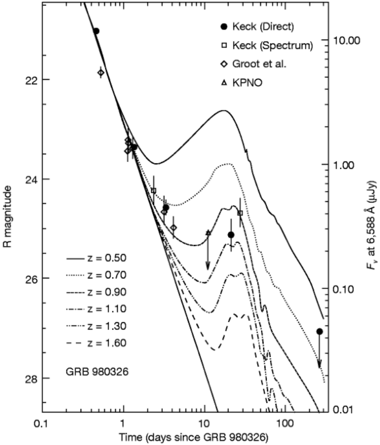
Data Analysis in COVID Days

Kipp Cannon and Catherine Beauchemin

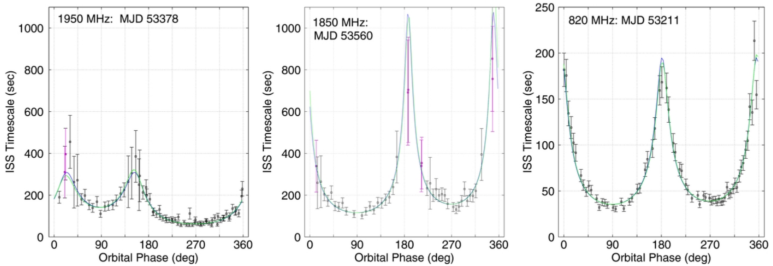
August 17, 2020

- We study light curves of supernovae to learn about their origins.



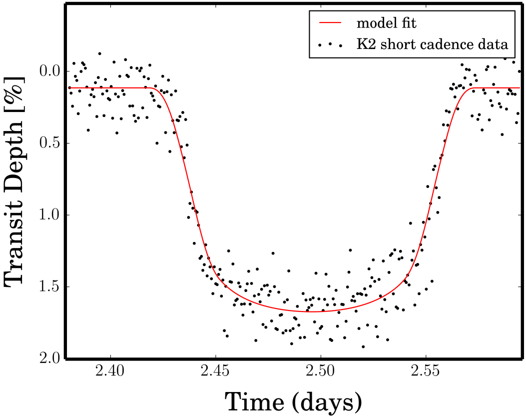
From Nature 401, 453–456, (1999).

- We study varying structure in radio pulses from pulsars to learn about the mechanisms of the source and of the interstellar medium.



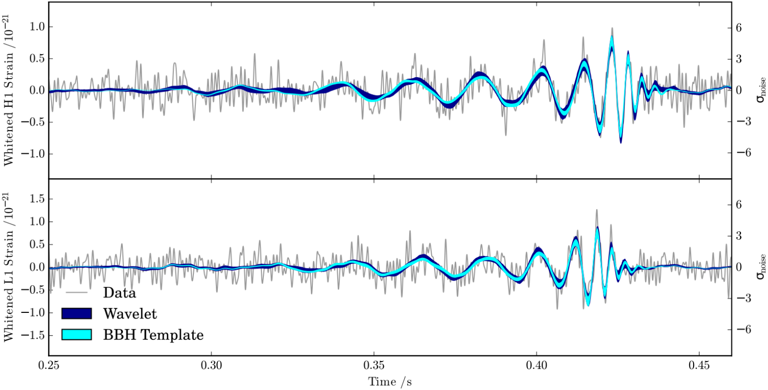
From Ap.J. 787(2), 161, (2014)

- We study the time-dependent brightness of stars to find and learn about planets.



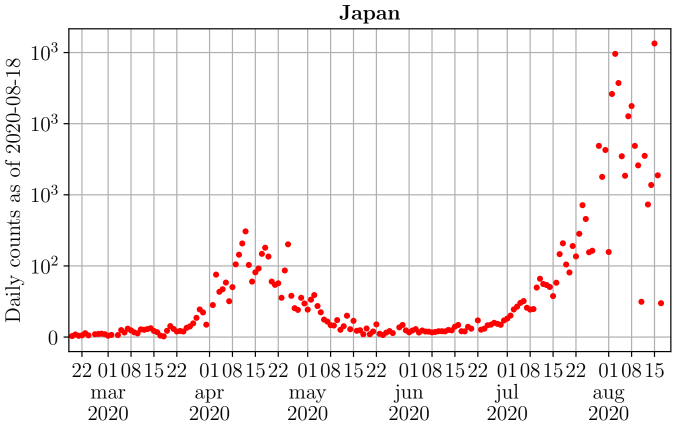
NASA Ames/T. Barclay

- We study the output of laser interferometers to find gravitational waves.



GW150914 signal reconstruction

- Here is a time series:



- What can we learn about the process that created this data?

- The ingredients:
  - A model or hypothesis about the nature of the underlying process.
  - A model for the statistical properties of the noise.

- The underlying process is a communicable infectious disease.
- New cases arise from pair-wise human interaction, not exposure to an environmental agent (like asbestos, radiation).
- Human interactions are complicated. There are patterns (co-workers, families, classmates, fellow commuters), but at the individual person-to-person level there is also a great deal of randomness.
- Not everyone can catch a disease: e.g., some people might be immune from previous exposure.
- Not everyone can transmit a disease: e.g., hospitalization isolates the infected.
- Try a simple model: **each sick person makes some number of other people sick**.

- The number of new sick people each sick person creates is called the "reproductive number", R.
- This model's behaviour is very simple, the number of sick people is given by

$$N(t) = N_0 R^{\frac{t - t_0}{t_{\text{infectious}}}} \tag{1}$$

where

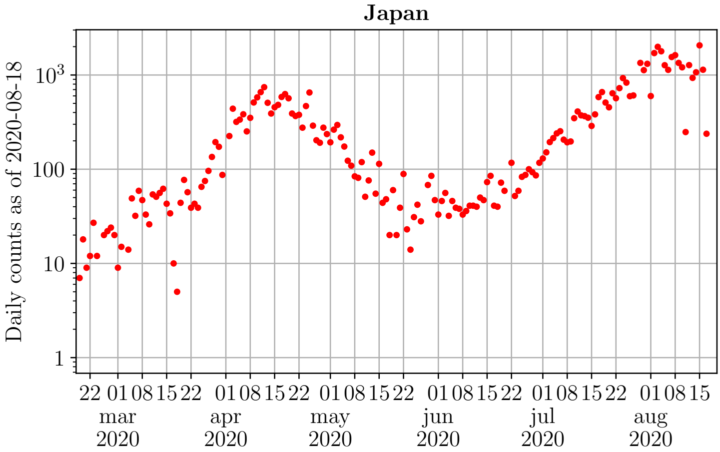
  - $N(t)$ is the number of cases at time t,
  - N_0 is the initial number of sick people at time t_0,
  - and a sick person is infectious for a period $t_{\text{infectious}}$.
- If $R > 1$ the number of sick grows exponentially, if $R < 1$ the number of sick decays exponentially, but it's always exponential.
- Taking the logarithm of both sides:

$$\ln N = \left[\frac{\ln R}{t_{\text{infectious}}}\right] t + \left[\ln N_0 - \frac{\ln R}{t_{\text{infectious}}} t_0\right] \tag{2}$$

which is, of course, the description of a straight line.

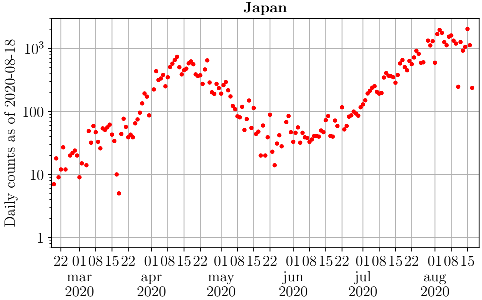
$$\ln N = \left[\frac{\ln R}{t_{\text{infectious}}}\right] t + \left[\ln N_0 - \frac{\ln R}{t_{\text{infectious}}} t_0\right]$$

- If we plot $\ln N$ (or N on a log scale) vs t we expect to see a straight line.
- Changing public health measures, changing patterns of behaviour (staying home from school, from work, avoiding restaurants and bars, using alcohol sanitizer), should change R over time, therefore we expect the slope to change over time.
- But always exponential: N grows or decays, but always exponentially.
- Let's check.



- Same data, log scale.

- What can we learn?
- The data does appear to be a sequence of straight line segments.
  - Our simple model appears to be sufficient to explain the data piece-wise.



- From a line segment we can learn when the number of sick was 1 (or will be 1) but we cannot know R or $t_{\text{infectious}}$, only the combination $t_{\text{infectious}}^{-1} \ln R$.
- If we can identify dates when the slope changes we might learn something from those.
- We need to fit lines to the data: to minimize residuals, **we need to understand the noise**.

Noise Sources

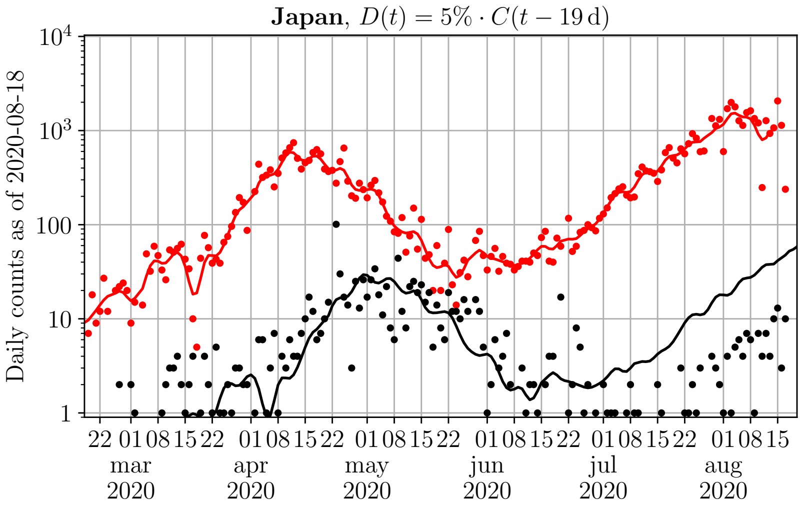
- Poisson shot noise:
  - When independent random events occur with some mean rate N, the actual number of them that are observed within a given interval is a Poisson-distributed random number. The standard deviation is $\sqrt{N}$.
- Reporting errors:
  - Health units lose data, then "fix" their mistake by reporting the cases later.
  - Governments interfere with data collection for the purpose of nationalistic propaganda.
- Periodic behaviour:
  - No case reporting on weekends.
  - People preferring to be tested on certain days of the week.
  - ...

Noise Sources

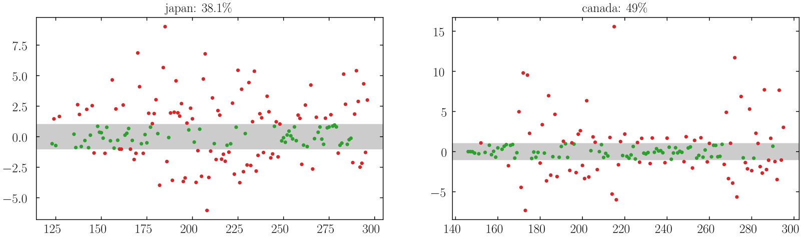
- Instead of generalizing the model to include the effects of periodicity in the process, we treat it as noise.
- An easy way to reduce it is to replace the data with a moving average. Our observations are

$$\text{data} = \text{exponential} + \text{noise} \tag{3}$$

- Because the noise-free function is assumed to be exponential, we use a moving time-symmetric (acausal) geometric mean: the underlying function is invariant under this transformation, while the noise is reduced.



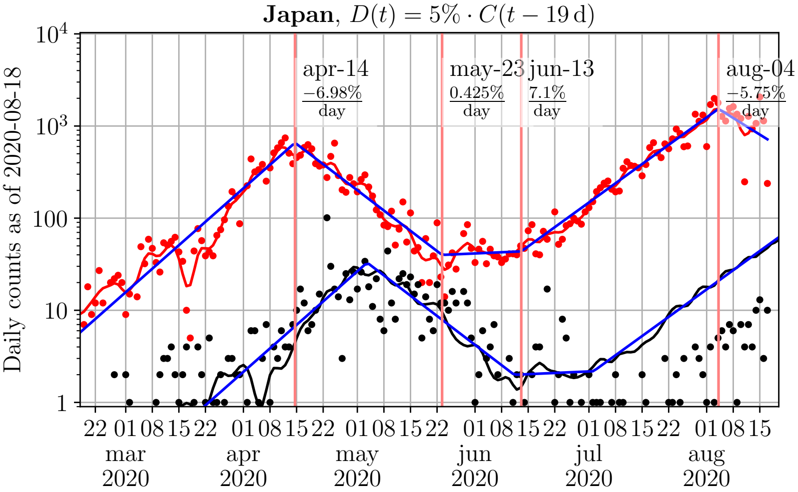
Japan, $D(t) = 5\% \cdot C(t - 19\,\mathrm{d})$
Daily counts as of 2020-08-18
10^4
10^3
100
10
1
22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15
mar 2020
apr 2020
may 2020
jun 2020
jul 2020
aug 2020



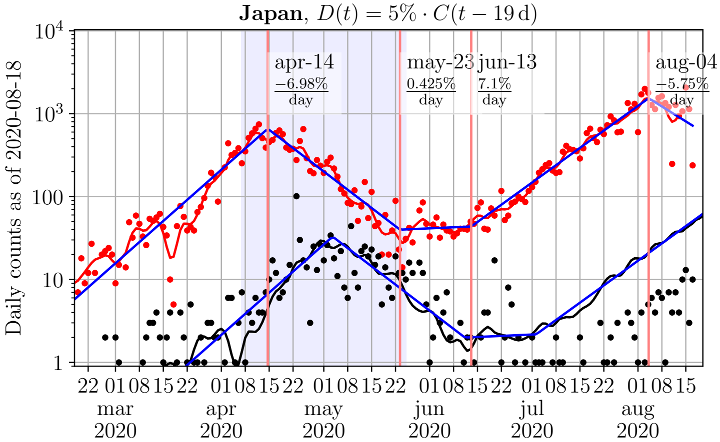
- Differences between observed case count, and 3-day moving Gaussian geometric mean for Japan and Canada.
- In both cases, the results appear to be independent random variables (no more correlations). The variance is time dependent, especially for Canada, and larger than expected for a Poisson process, but we can measure it and accommodate it.

Maximum Likelihood Fit

- We use the algorithm described by V. Muggeo "Estimating Regression Models with Unknown Break-Points", Statist. Med. 2003; 22:3055–3071 (DOI: 10.1002/sim.1545).
- Solves for the pieces-wise linear function that minimizes the weighted sum-of-square residuals.
- Requires the number of break points to be specified.
- To choose the number of segments, we use the Bayesian information criterion to select the model with the greatest support.


Japan, $D(t) = 5\% \cdot C(t - 19\,\mathrm{d})$
Daily counts as of 2020-08-18
10^4
10^3
100
10
1
apr-14
$\frac{-6.98\%}{\mathrm{day}}$
may-23
$\frac{0.425\%}{\mathrm{day}}$
jun-13
$\frac{7.1\%}{\mathrm{day}}$
aug-04
$\frac{-5.75\%}{\mathrm{day}}$
22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15 22 01 08 15
mar 2020
apr 2020
may 2020
jun 2020
jul 2020
aug 2020

Interpretation

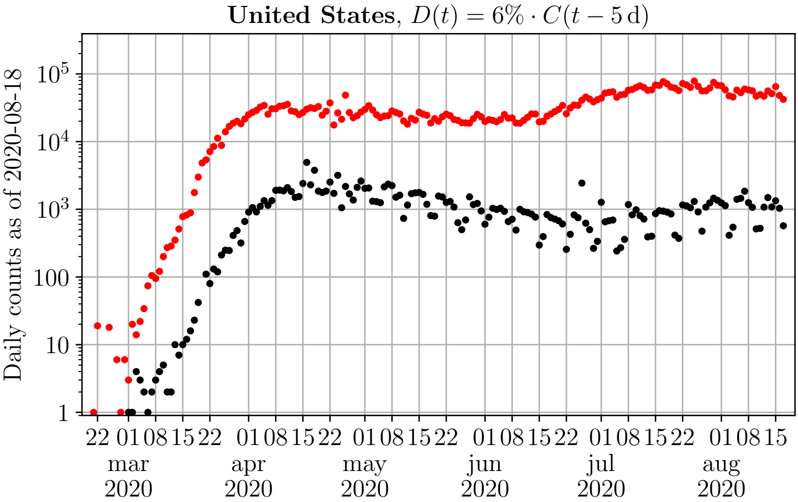


- The grey region is Tokyo's state of emergency.

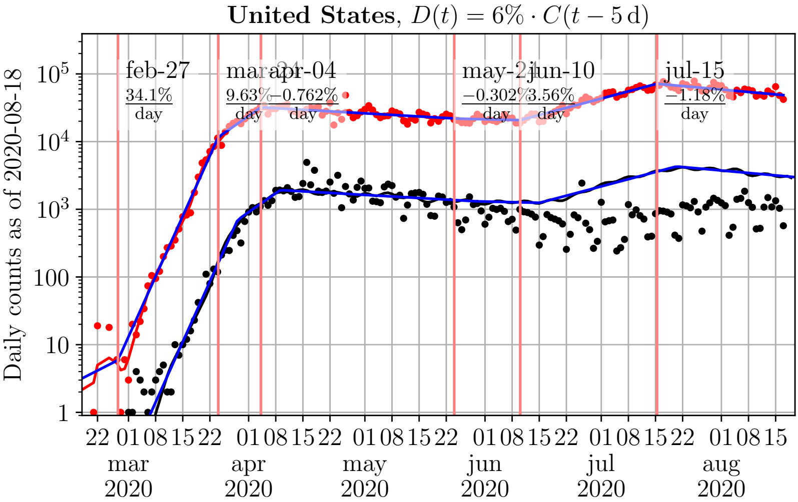
Summary

- There is little evidence that anything other than declaring a state of emergency is an effective intervention in Japan.
- The case counts grew until the state of emergency, they fell throughout, and immediately began growing again when it was lifted.
- If you remember the news announcing a "party rental room cluster", or a "Kabukicho bar cluster", and blaming rising cases in Tokyo in July on these, in fact the evidence does not support the hypothesis of a series of large impulse source events: the data are consistent with our assumption that each sick person makes some number of other people sick, uniformly. Everyone is equally responsible for the spread.
- There was certainly no change in July, the growth had started a month before then.

Let's do Another: USA



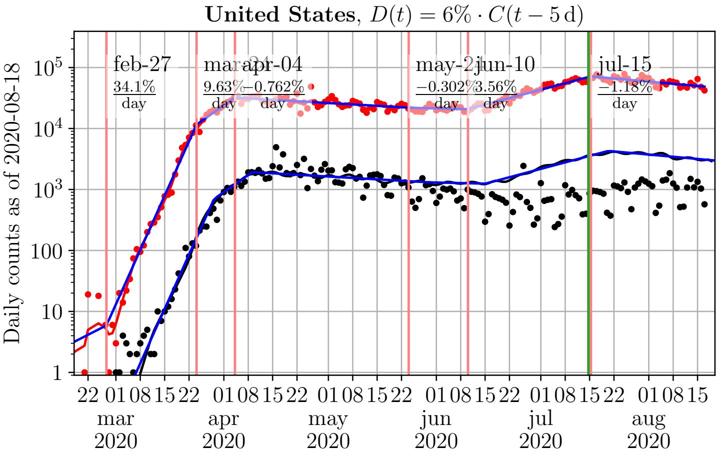
Let's do Another: USA



Interpretation

- USA public health units originally reported data to the Centres for Disease Control (CDC), and independent government agency.
- Public health units were ordered to cease sending data to the CDC, and submit it only to the Department of Health and Human Services (HHS).
- HHS is directly under the control of the White House via the Secretary of Health and Human Services.
- The reporting procedure changed on **July 15**.

Interpretation



- Hmm …

Remarks

- What I have shown is a kinematic model: a description of the observed data, not a model of the system that produced it.
- We can construct a dynamical model. The standard form are "SEIR" models for the main states they assume exist:
  - **S**usceptible
  - **E**clipsing (infected but not infectious)
  - **I**nfectious
  - **R**esolved or removed (cured and now immune, or dead)
- Couplings and delays are defined, they can be linear or bilinear or non-linear.
- Allow you to infer other information, like how many undetected infected people are in the population, or how many unreported deaths are occurring.