Forecasting Chicken Pox Outbreaks using Digital Epidemiology

3/10/2016

Abstract

Public health surveillance systems are important for tracking disease dynamics. However not all diseases are reported, especially those with benign or mild symptoms. In recent years, social and real-time digital data sources have provided new means of studying disease transmission. Such affordable and accessible data have the potential to offer new insights into disease epidemiology at the national and international scales. I used the extensive information repository Google Trends, to examine the digital epidemiology of a common childhood disease in Australia, chicken pox, caused by varicella zoster virus (VZV), over an eleven-year period. I built a model to test whether Google Trends data could forecast recurrent seasonal outbreaks by estimating the magnitude and seasonal timing. I tested 8 different forecasting models, which are nested versions of each other, against a null cosine model that captured the general seasonality of chicken pox. I also included two models to 'fit', rather than 'forecast' the chicken pox data. The model that included the Google Trends data and a subsection of the parameters fit better than a 'full model' which included all parameters and the null model when examined by Akaike Information Criterion and a likelihood ratio test. These data and the methodological approaches provide a novel way to track, and forecast the global burden of childhood disease. I hope to exapnd this research into other childhood diseases for which surveillance is lacking.

Introduction

Childhood infectious diseases continue to be a major global problem, and surveillance is needed to inform strategies for the prevention and mitigation of disease transmission. Our ability to characterize the global picture of childhood diseases is limited, as detailed epidemiological data are generally nonexistent or inaccessible across much of the world. Available data suggest that recurrent outbreaks of acute infectious diseases peak within a relatively consistent, but disease-specific seasonal window, which differs geographically (1,2,3,4,5). Geographic variation in disease transmission is poorly understood, suggesting substantial knowledge gains from methods that can expand global epidemiological surveillance. Seasonal variations in host-pathogen interactions are common in nature (6). In humans, the immune system undergoes substantial seasonal changes in gene expression, which is inverted between European locations and Oceana (7). The regulation of seasonal changes in both disease incidence and immune defense is known to interact with environmental factors such as annual changes in day length, humidity and ambient temperature (8). Accordingly, quantification of global spatiotemporal patterns of disease incidence can help to disentangle environmental, demographic, and physiological drivers of infectious disease transmission. Furthermore, the recognition of the regional timing of outbreaks would establish the groundwork for anticipating clinical cases, and when applicable, initiating public health interventions.

Since childhood disease outbreaks are often explosive and short-lived (9), temporally rich (i.e., weekly or monthly) data are needed to understand their dynamics. Similarly, in order to establish the recurrent nature of outbreaks that occur at annual or multi-annual frequencies, long-term data are needed. Thus, ideal disease incidence data have both high temporal resolution and breadth (i.e., frequent observations over many years). Over the past decade, the internet has become a significant health resource for the general public and health professionals (10,11). Internet query platforms, such as Google Trends, have provided powerful and accessible resources for identifying outbreaks and for implementing intervention strategies (12,13,14). Research on infectious disease information seeking behaviour has demonstrated that internet queries can complement traditional surveillance by providing a rapid and efficient means of obtaining large epidemiological datasets (13,15,16,17,18). For example, epidemiological information contained within Google Trends has been used in the study of rotavirus, norovirus, and influenza (14,15,17,18). These tools offer substantial promise for the global monitoring of diseases in countries that lack clinical surveillance but have sufficient internet coverage to allow for surveillance via digital epidemiology.

Acquisition of Google Trends Data

Google Trends data were used to assess patterns of information seeking behavior over long time periods, from January 2004 to July 2015. To evaluate childhood disease information seeking behaviour, we obtained country-specific data from Google (19). Google Trends represent the relative number of searches for a specific key word (e.g. "chicken pox") standardized within each country such that the values range from 0 to 100. A search volume of 0 is assigned, by Google Trends, to weeks/months with a minimal amount of searches.

In order to relate Google Trends data to the dynamics of chicken pox (or other diseases of interest), care must be taken to select appropriate search terms. Chicken pox is the classical manifestation of disease, and therefore, language-specific queries of ‘’chicken pox’‘are a straightforward choice for data-mining. In contrast, infections with generic symptoms, such as fever and diarrhea, could arise from many other diseases, making it difficult to identify appropriate queries. In either case, search terms vary subtly from country to country. For instance, in the U.S.’‘chickenpox’‘is typically written as a single word, whereas in the U.K. and Australia, people refer to’‘chicken pox” as two words. Here I examined data from Australia, where the data were subset within the range that included consecutive weeks with > 0 search volume. Chicken pox data from Australia were digitized from (20), and age structure data were digitized from the United Nations (21).

Forecasting Outbreaks using Google Trends

To determine whether the information seeking behaviour observed in Google data, T, was able to forecast chicken pox outbreak magnitude and timing in Australia, I built and fitted multiple statistical models to forecast chicken pox case data. I evaluated the epidemiological information contained in Google Trends by comparing the Google Trends model with a seasonal null model that did not incorporate Google data. The null model lacked information seeking in the force of infection parameter, which we defined as the monthly per capita rate at which children age 0–14 years are infected. In order to estimate the number of symptomatic VZV infections per month, we multiplied the force of infection with an estimate of the population aged 0–14 years (21). All models were fitted to the case data from a VZV-vaccinated population (Australia), which exhibited reduced seasonality. To estimate the number of symptomatic VZV infections each month, I_t, I used Google Trends data from the previous two months, T_{t-1} and T_{t-2}, where t is time in monthly time steps. The full chicken pox process model tracked the force of infection, λ_t,

$$\lambda_t = \left[\beta_1 \cos\left(\frac{2\pi}{12}(t+\omega)\right) \mathrm{T}_{t-1}^{\alpha} + \beta_2 |\mathrm{T}_{t-1} - \mathrm{T}_{t-2}| + \beta_3\right] \epsilon_t. \tag{1}$$

The model also contained environmental stochasticity, ϵ_t, which was drawn from a gamma distribution with a mean of 1 and variance θ. I estimated 7 parameters for the full model: the mean and the phase of the seasonality (β_1 and ω), parameters scaling the Google Trends data (α and β_2), the baseline force of infection (β_3), the process noise dispersion parameter (θ), and the reporting dispersion parameter (τ) of a normal distribution, with a mean of 1, from which case reports were drawn. The parameters were estimated using maximum likelihood by iterated particle filtering (MIF) in the R-package pomp (22,23). We forecasted each model starting with 10000 parameter combinations generated from a sobol design, and replicated through pomp four times, with interatively smaller random walk standard deviations.

The process model (Eqn. 1) contained environmental stochasticity, ϵ_t, which was drawn from a gamma distribution with a mean of 1 and variance θ. In order to estimate the number of symptomatic VZV infections per month, I multiplied the force of infection, λ, with an estimate of the population aged 0–14 years (21), C,

$$I_t = \lambda_t \mathrm{C}. \tag{2}$$

I modeled the observation process, which represents the number of cases actually reported, to account for stochasticity in the reporting of symptomatic VZV infections. Case reports were drawn from a normal distribution with a mean report rate, $\rho = 1$, and dispersion parameter (τ) which was estimated.

$$\text{chickenpox}_t \sim \mathcal{N}(\rho I_t, \tau I_t). \tag{3}$$

I evaluated the epidemiological information contained in Google Trends by comparing the Google Trends model with a seasonal null model where the force of infection did not incorporate Google Trends data. The null model force of infection was modeled as:

$$\lambda_t = \left[\beta_1 \cos\left(\frac{2\pi}{12}(t+\omega)\right) + \beta_3\right] \epsilon_t. \tag{4}$$

In addition to the full model, I tested nested variations of the full model (Eqn. 1), including; a model without the cosine function;

$$\lambda_t = \left[\beta_1(\mathrm{T}_{t-1}^{\alpha}) + \beta_2|\mathrm{T}_{t-1} - \mathrm{T}_{t-2}| + \beta_3\right] \epsilon_t. \tag{5}$$

a model without the cosine function or the β_2 parameter;

$$\lambda_t = \left[\beta_1(\mathrm{T}_{t-1}^{\alpha}) + \beta_3\right] \epsilon_t. \tag{6}$$

a model without the cosine function or the α parameter;

$$\lambda_t = [\beta_1(\mathrm{T}_{t-1}) + \beta_2|\mathrm{T}_{t-1} - \mathrm{T}_{t-2}| + \beta_3] \epsilon_t. \tag{7}$$

a model without the cosine function, α, or the β_2 parameters;

$$\lambda_t = [\beta_1(\mathrm{T}_{t-1}) + \beta_3] \epsilon_t. \tag{8}$$

a model without the α parameter;

$$\lambda_t = \left[\beta_1 \cos\left(\frac{2\pi}{12}(t+\omega)\right) \mathrm{T}_{t-1} + \beta_2|\mathrm{T}_{t-1} - \mathrm{T}_{t-2}| + \beta_3\right] \epsilon_t. \tag{9}$$

a model without the β_2 parameter;

$$\lambda_t = \left[\beta_1 \cos\left(\frac{2\pi}{12}(t+\omega)\right) \mathrm{T}_{t-1}^{\alpha} + \beta_3\right] \epsilon_t. \tag{10}$$

and a model without the α or β_2 parameters;

$$\lambda_t = \left[\beta_1 \cos\left(\frac{2\pi}{12}(t+\omega)\right) \mathrm{T}_{t-1} + \beta_3\right] \epsilon_t. \tag{11}$$

In addition to the forecasting models, I also wrote two models to fit the Google Trends data to chicken pox data, without forecasting.

$$\lambda_t = \left[\beta_1(\mathrm{T}_t) + \beta_3\right]\epsilon_t. \tag{12}$$

$$\lambda_t = \left[\beta_1(\mathrm{T}_t^{\alpha}) + \beta_3\right]\epsilon_t. \tag{13}$$

Results

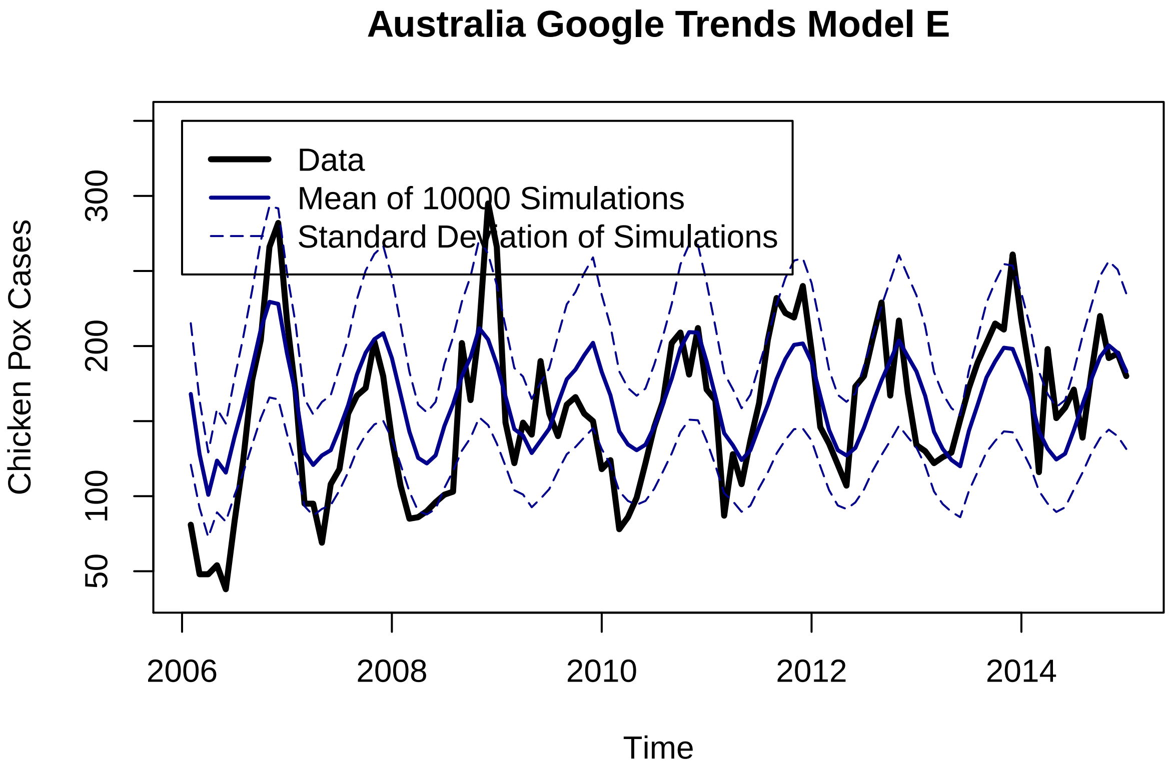
Models that included the cosine function (Eqns. $1, 9, 10, 11$) fit about 20 likelihood units better (Table 1), indicating the need for the inclusion of the cosine function. The cosine function is important in forecasting because without a seasonal function, the model would incorrectly forecast the next time step at all peaks and troughs. By including the cosine function, the models were able to correctly estimate the downturn after a peak, and upturn after a trough. Overall, model F (Eqn. 10) fit the best, despite estimating fewer parameters than the other models. It used one more parameter than the null model, yet fit 14 Log-likelihood units better. AIC and likelihood ratio tests were based off of this model.

The Google Trends model fit the case data and preformed better than the null model in Australia, as the null model AIC was $>$ 28 units above Google Trends model AIC. Since both models were seasonally forced, all models that included the cosine function captured the seasonal timing of outbreaks. However, the Google Trends model was able to predict the interannual variation in outbreak size (Fig~X), while the null model could not (Fig Y).

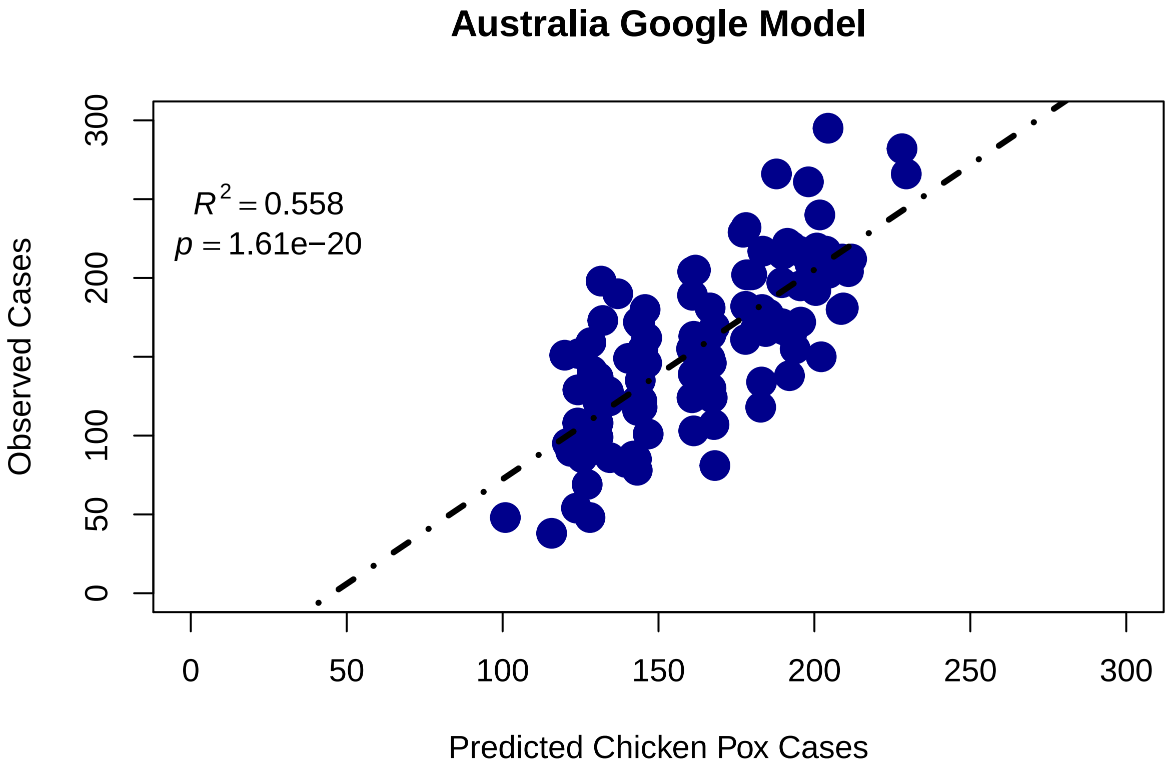
Equation #	Model	Model Structure	LogLik	Est # Params	AIC	Δ AIC
Eqn 1	A	Full	-565.43	7	1144.9	24.0
Eqn 4	H	Null	**-569.47**	6	**1150.9**	30.0
Eqn 5	B	No Cos	-584.96	6	1181.9	61.0
Eqn 6	C	No Cos, β_2	-585.02	5	1180.0	59.1
Eqn 7	D	No Cos, α	-586.08	5	1182.2	61.3
Eqn 8	I	No Cos, β_2, α	-585.63	4	1179.3	58.4
Eqn 9	E	No α	**-554.47**	6	**1120.9**	0.0
Eqn 10	F	No β_2	-563.35	6	1138.7	17.8
Eqn 11	G	No α, β_2	-558.32	5	1128.0	7.1
Eqn 12	M	No Forecast, β_2, α	-584.42	4	NA	NA
Eqn 13	N	No Forecast, β_2	-583.98	5	NA	NA

From these results, I simulated the Maximum-Likelihood parameter set for each the Null model (Eqn. 4) and the best fit Google Trends model (Model E, Eqn. 9) 10000 times to elucidate the improvement Google Trends adds to the model fit. I have included my code on how I simulated the model below (not shown - uncomment if you want to run it).

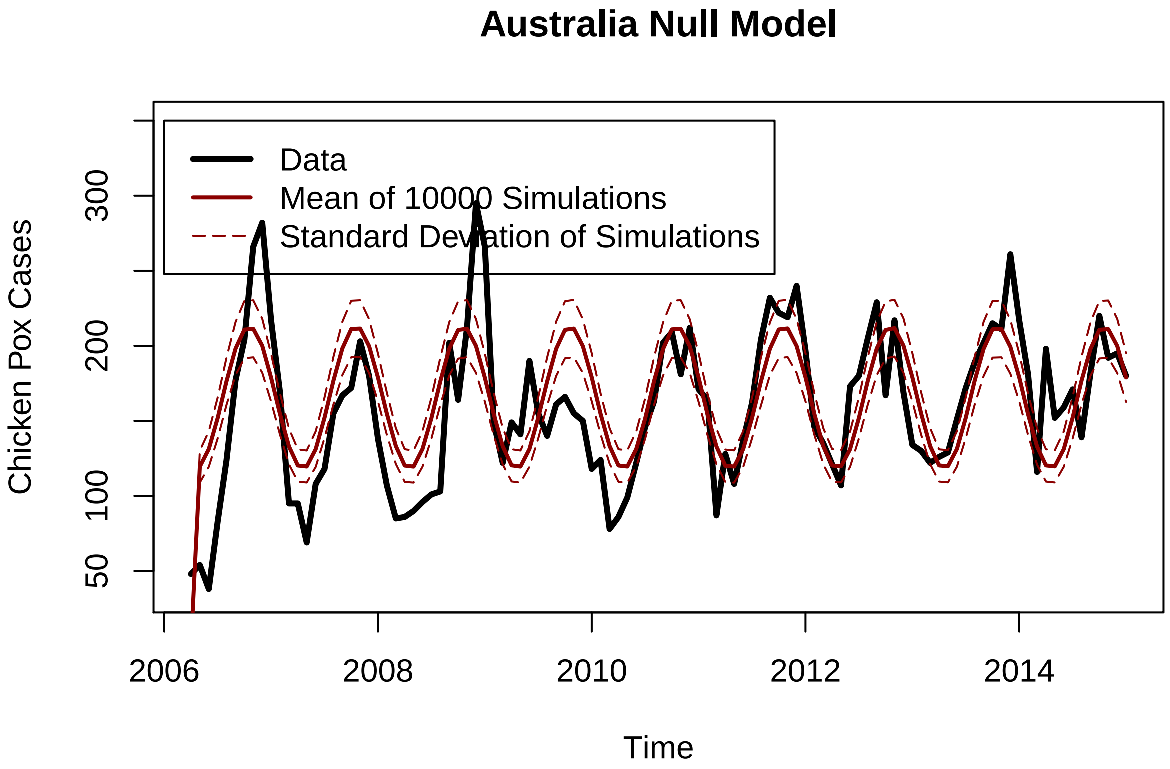
I then loaded the simulations and plotted it against the data, showing the means and standard deviations at each month.



Additionally, I examined the best forecasting Google Model fit vs the data and included the R-squared and p-values.

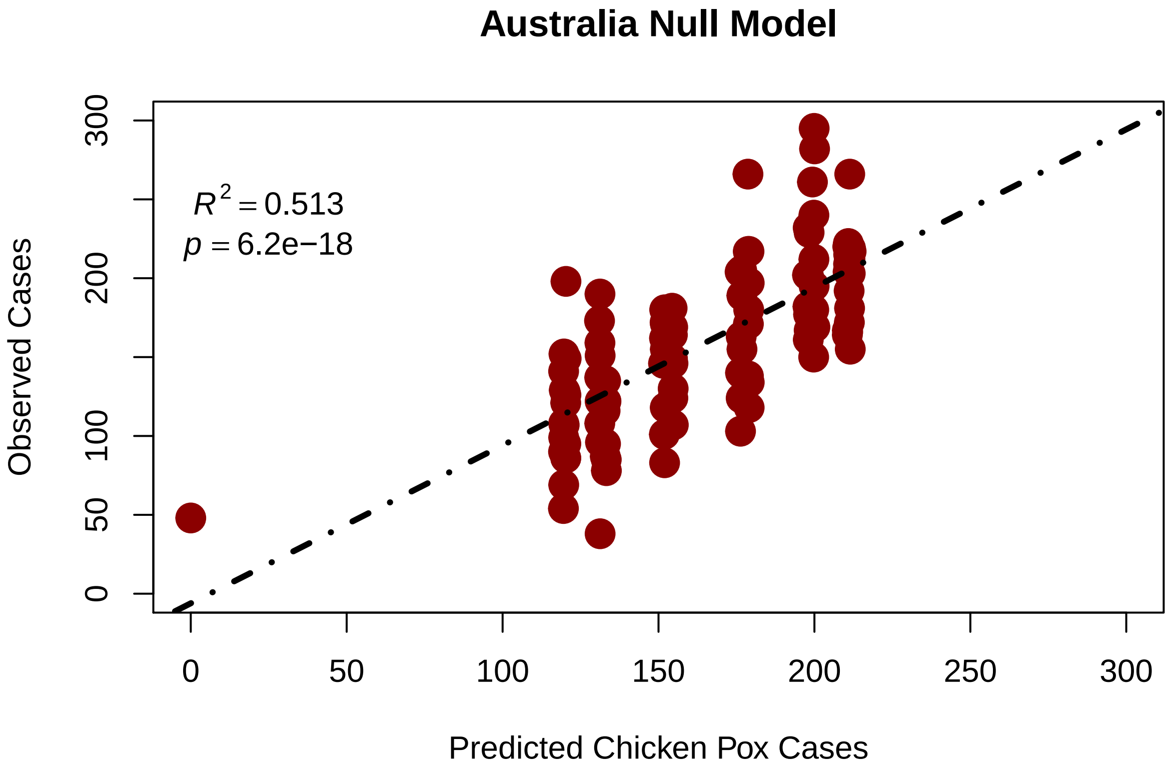


I did the same for the null model, showing the mean and standard deviations at each point.



The Google Trends model had a much larger standard deviation, allowing for the model (with the standard deviations) to capture most of the troughs and peaks throughout the time series. The null model had a smaller standard deviation, with more peaks and troughs outside these values.

I also examined the forecast fit for the null model (means) vs the data and included the R-squared and p-values.



When comparing the R-squared values for the null model and the best fit Google model, the Google Trends data was able to explain around 5 percent of the data. While not huge, it is significant (Table 1). To get a better idea of how Google Trends was better able to explain the interannual variation in chicken pox cases, I examined the peak and trough month of each year. I found the peak and trough month for each year, pulled out the number of cases in that month and created density distributions of the Google Trends and null models for each peak and trough for each year.

```
 col2rgb("darkred", alpha=TRUE)
```

```
##       [,1]
## red    139
## green    0
## blue     0
## alpha  255
```

```
redtrans <- rgb(139, 0, 0, 127, maxColorValue=255)


 col2rgb("darkblue", alpha=TRUE)
```

```
##       [,1]
## red      0
## green    0
## blue   139
## alpha  255
```

```
bluetrans <- rgb(0, 0, 139, 127, maxColorValue=255)
```

The results show that peaks in cases are typical near the end of the year (Oct/Nov, while a few years had peaks in other months).

```
month
```

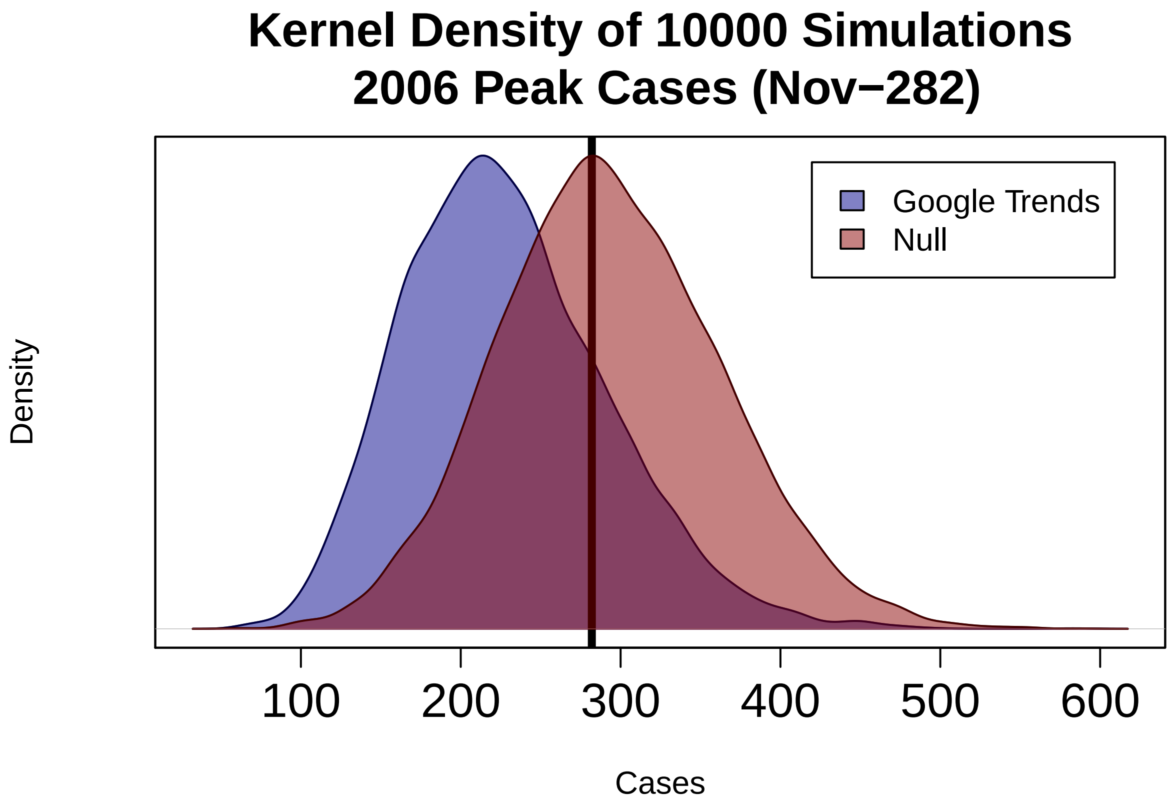
```
## [1] 11 10 11  5 11 11  8 11  9
```

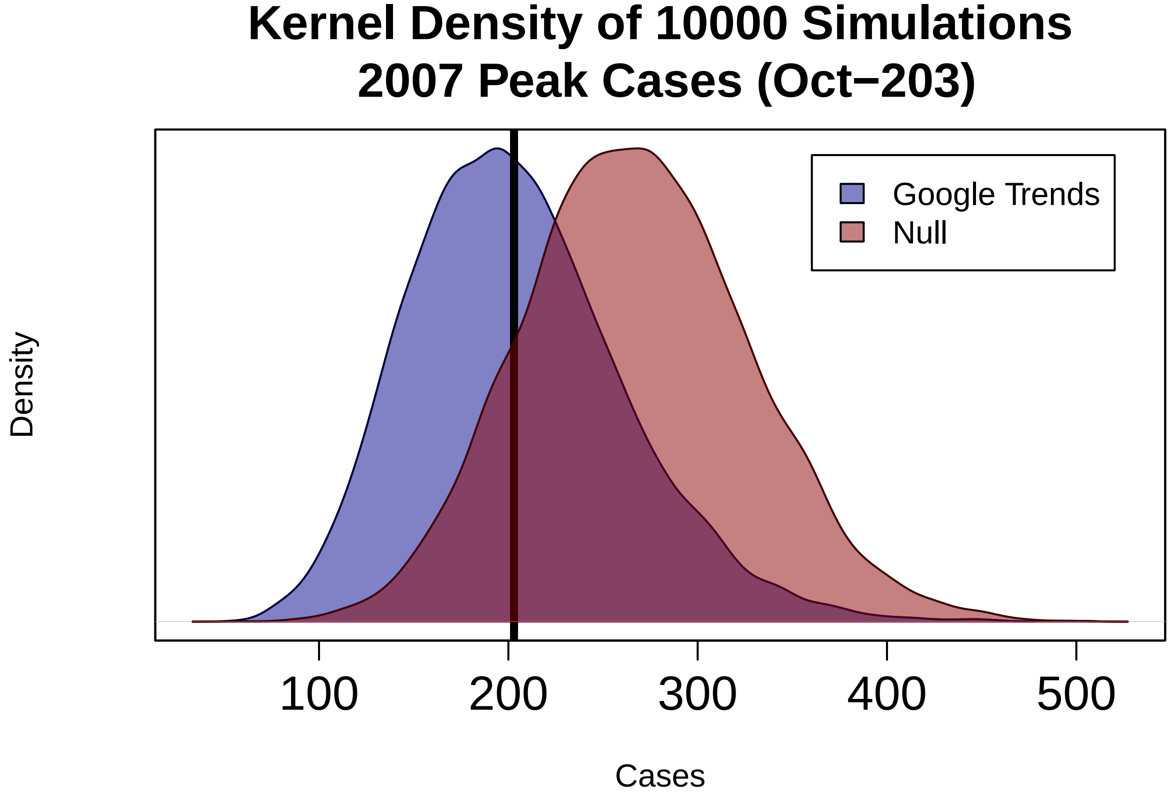
```
value
```

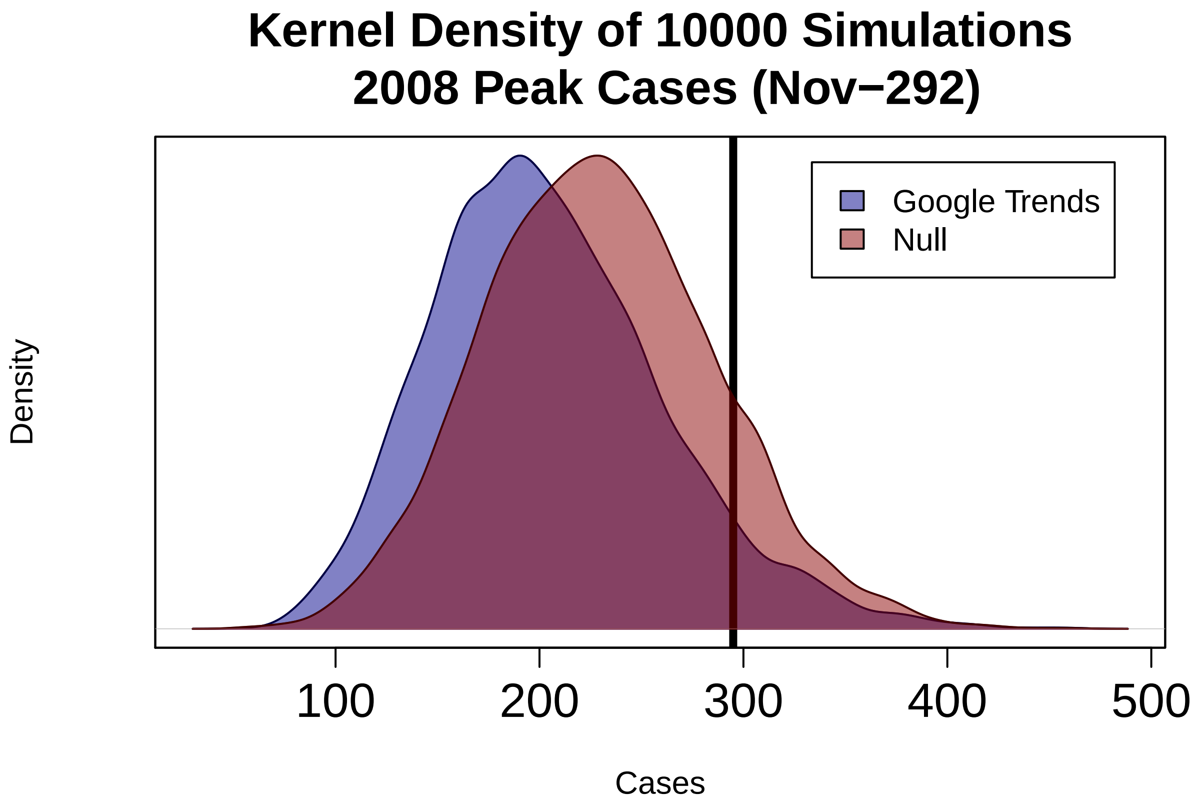
```
## [1] 282 203 295 190 212 240 229 261 220
```

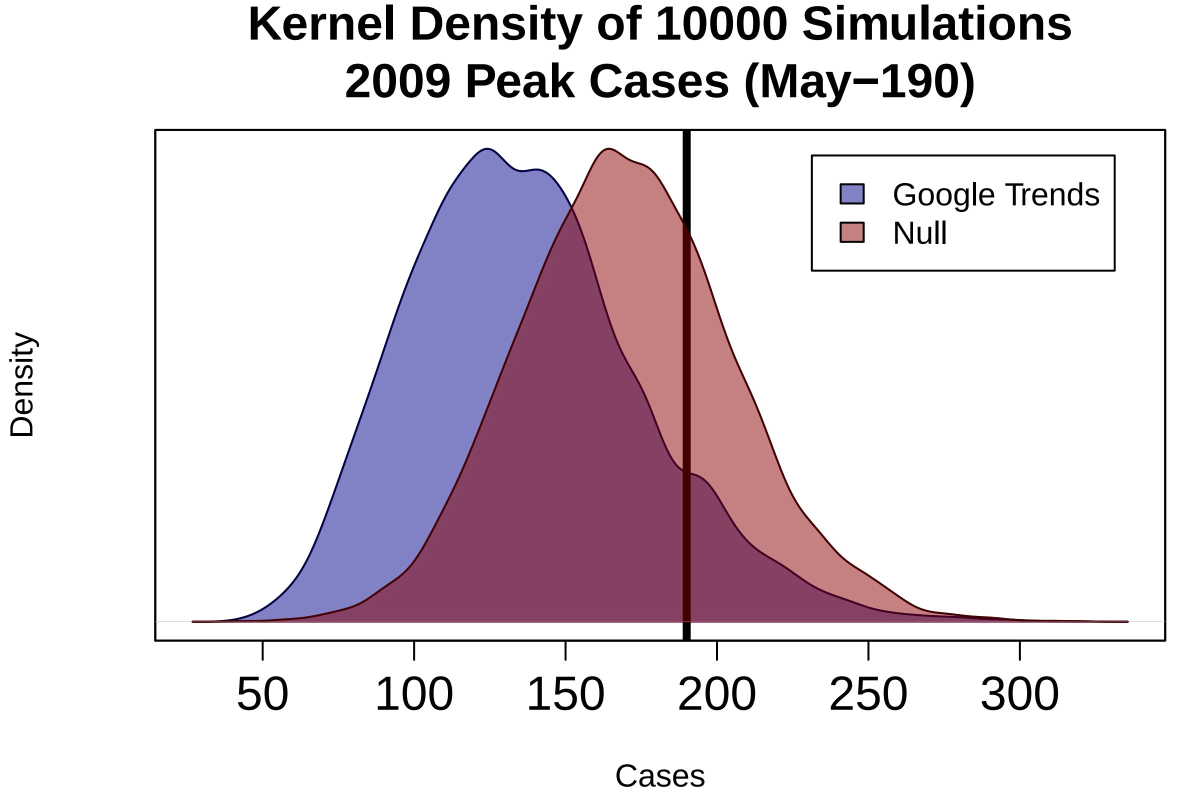
I then picked out each of the 10000 simulations for the two models at each of those months, assigning each it's own vector.

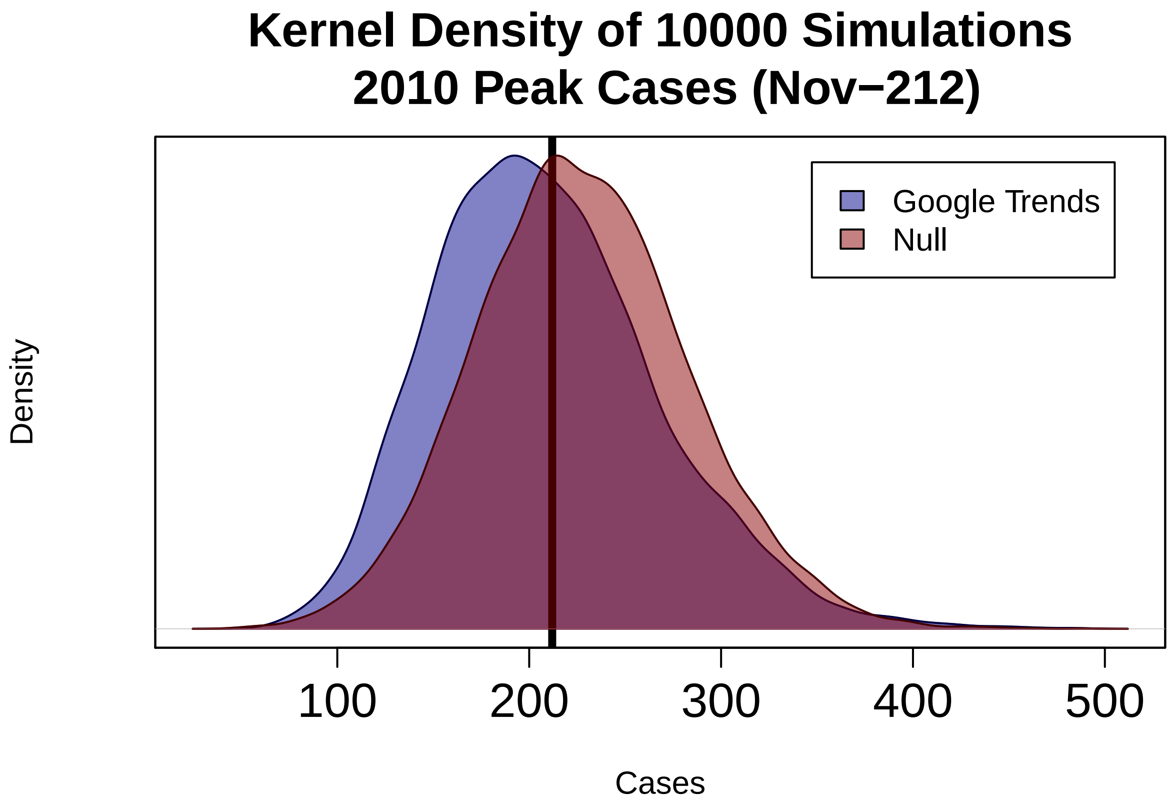
From that, I created density distributions for each model at the peak month for each year. This is a 3x3 matrix in pdf output, but had to make each year it's own figure to make it fit into an .rmd file.

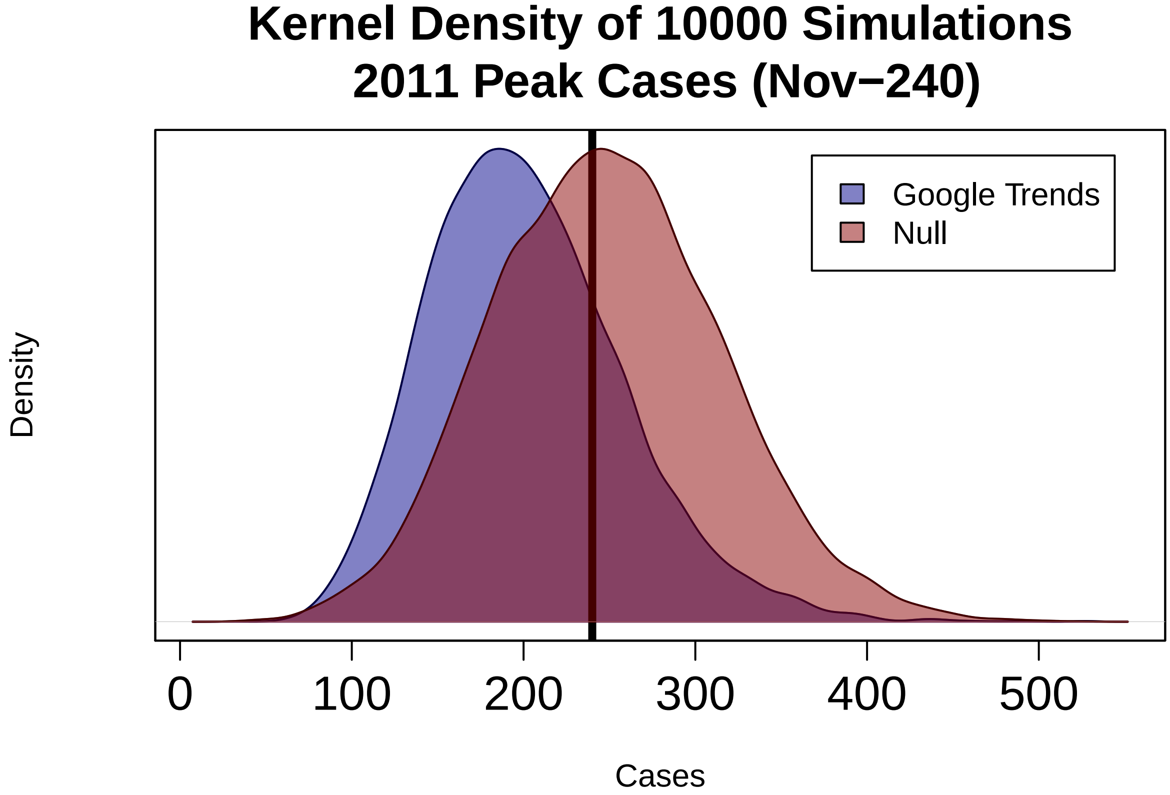

Kernel Density of 10000 Simulations
2006 Peak Cases (Nov−282)
Google Trends
Null
Density
100
200
300
400
500
600
Cases


Kernel Density of 10000 Simulations
2007 Peak Cases (Oct−203)
Google Trends
Null
Density
100
200
300
400
500
Cases

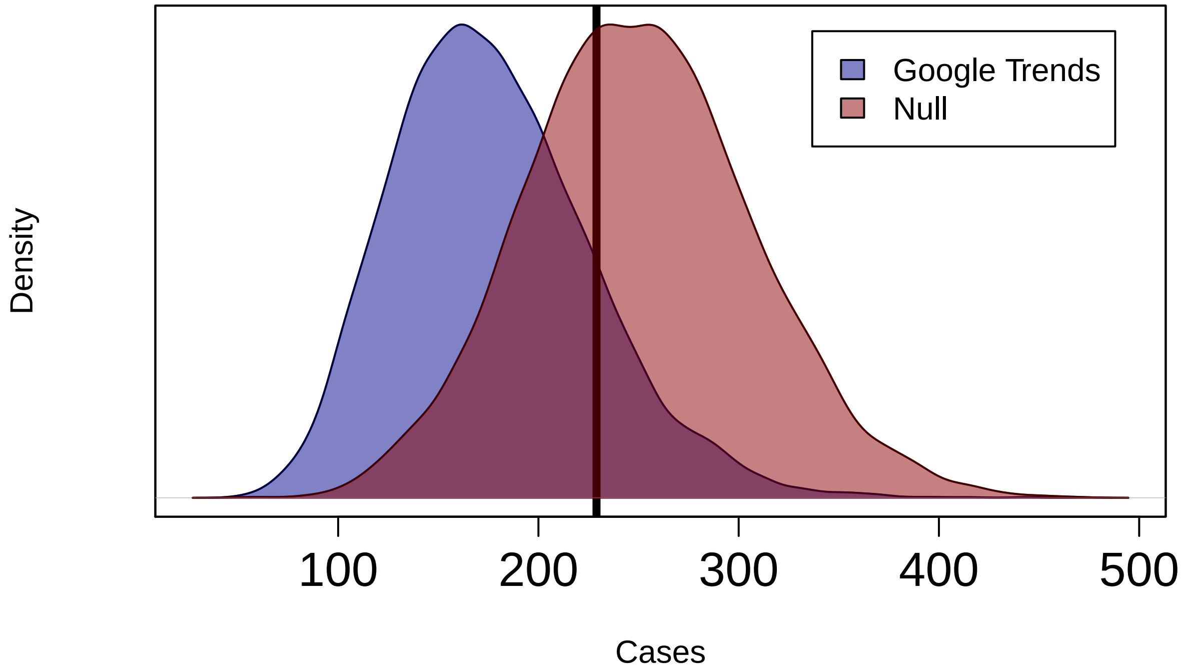

Kernel Density of 10000 Simulations
2008 Peak Cases (Nov−292)
Density
Google Trends
Null
100
200
300
400
500
Cases


Kernel Density of 10000 Simulations
2009 Peak Cases (May−190)
Density
Google Trends
Null
50
100
150
200
250
300
Cases

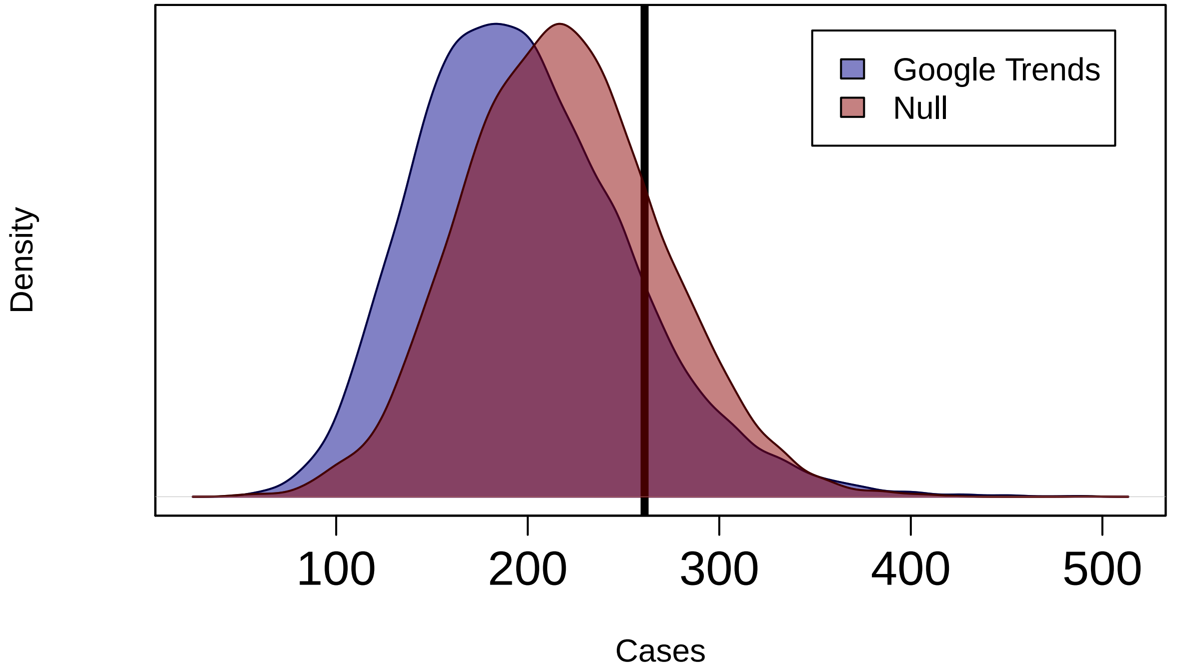

Kernel Density of 10000 Simulations
2010 Peak Cases (Nov−212)
Google Trends
Null
Density
100
200
300
400
500
Cases

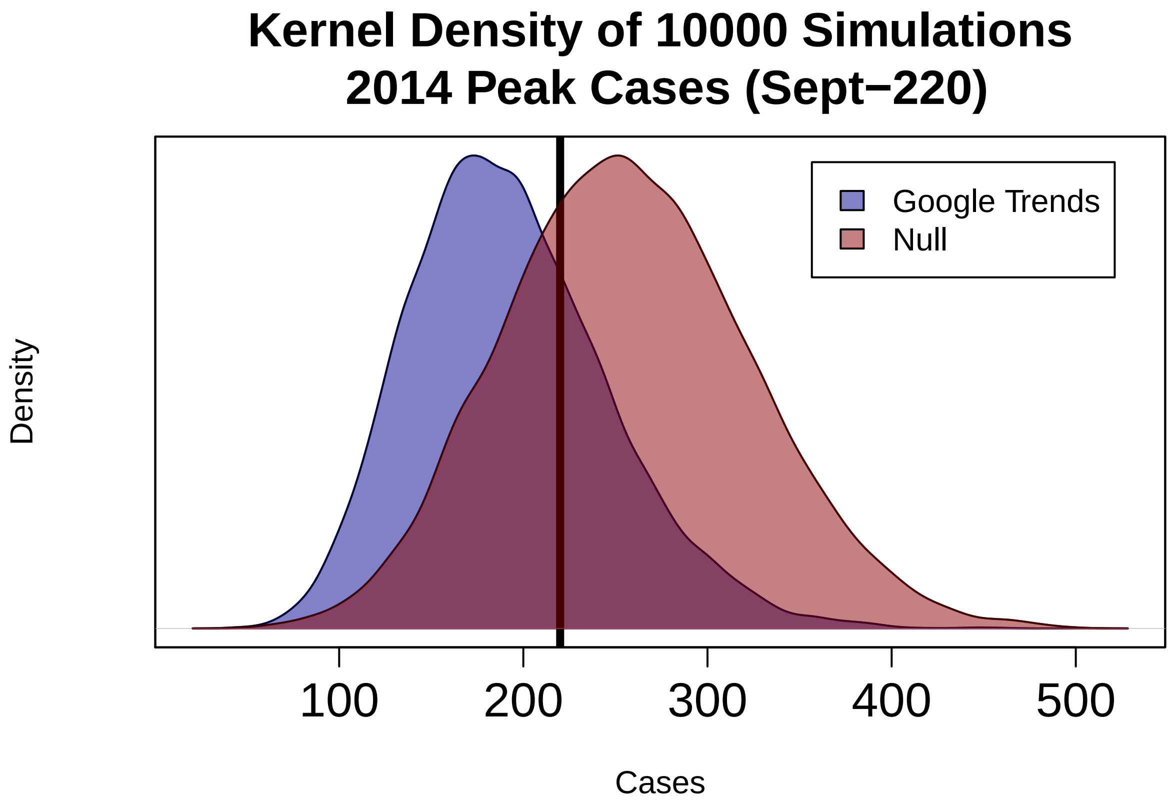

Kernel Density of 10000 Simulations
2011 Peak Cases (Nov−240)
Google Trends
Null
Density
0
100
200
300
400
500
Cases

Kernel Density of 10000 Simulations
2012 Peak Cases (Aug−229)



Kernel Density of 10000 Simulations
2013 Peak Cases (Nov−261)





These results are interesting in that the null model does a better job hitting the peaks in each year other than 2007. The null model performed very good at capturing the peak in 2006, 2010, 2011, and 2012.

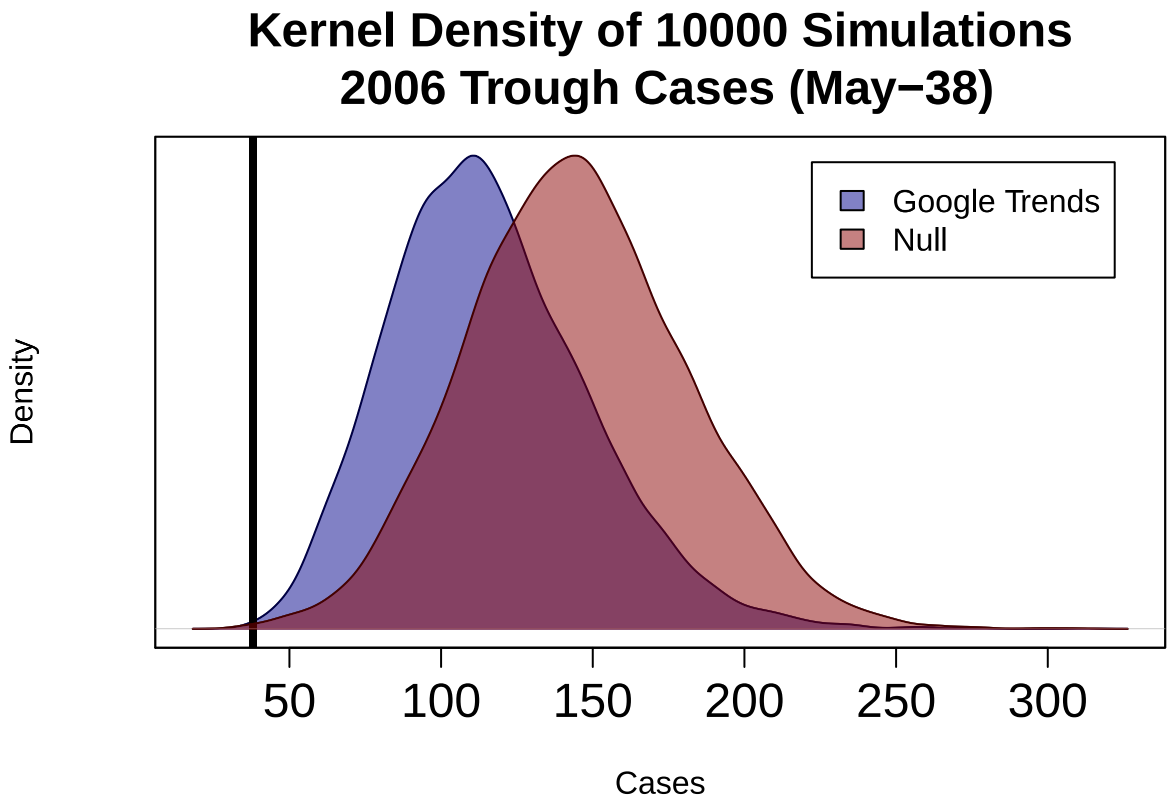
I did the same for the density distributions for each model at the trough month for each year. First I pulled out the months where the minimum number of cases occured each month, and how many cases there were.

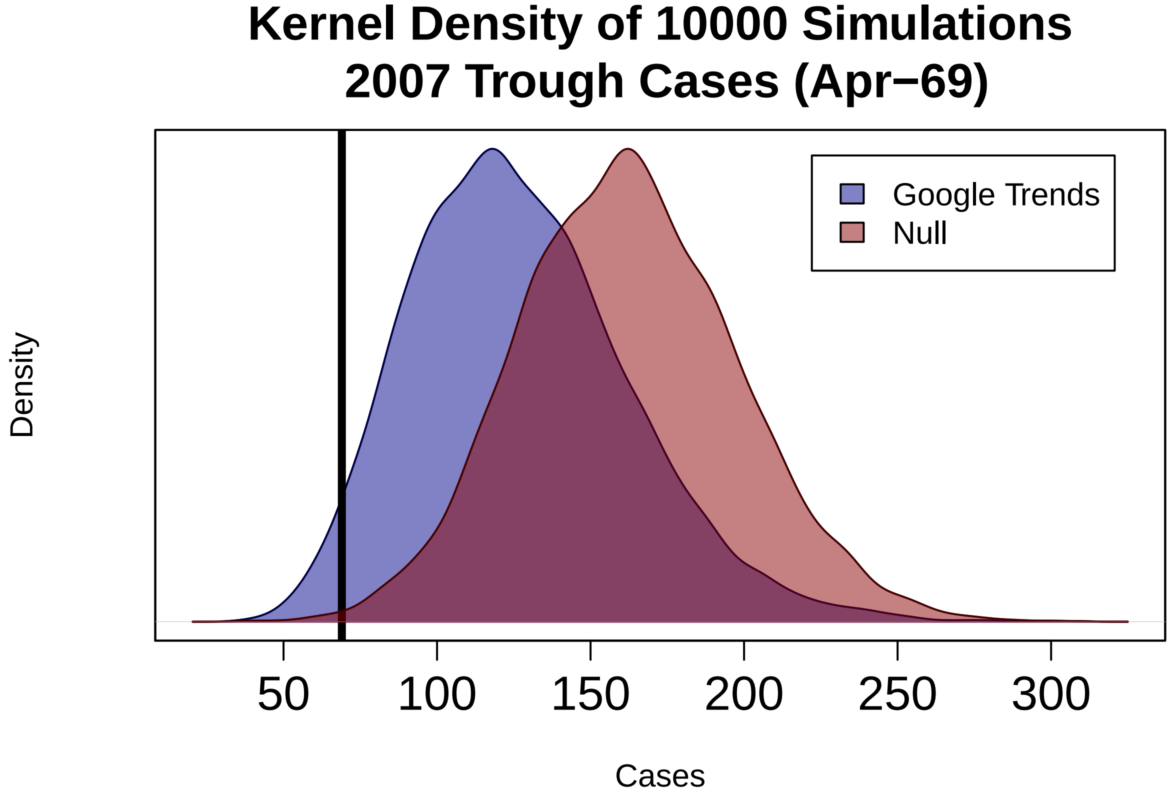
```
Nmonth
```

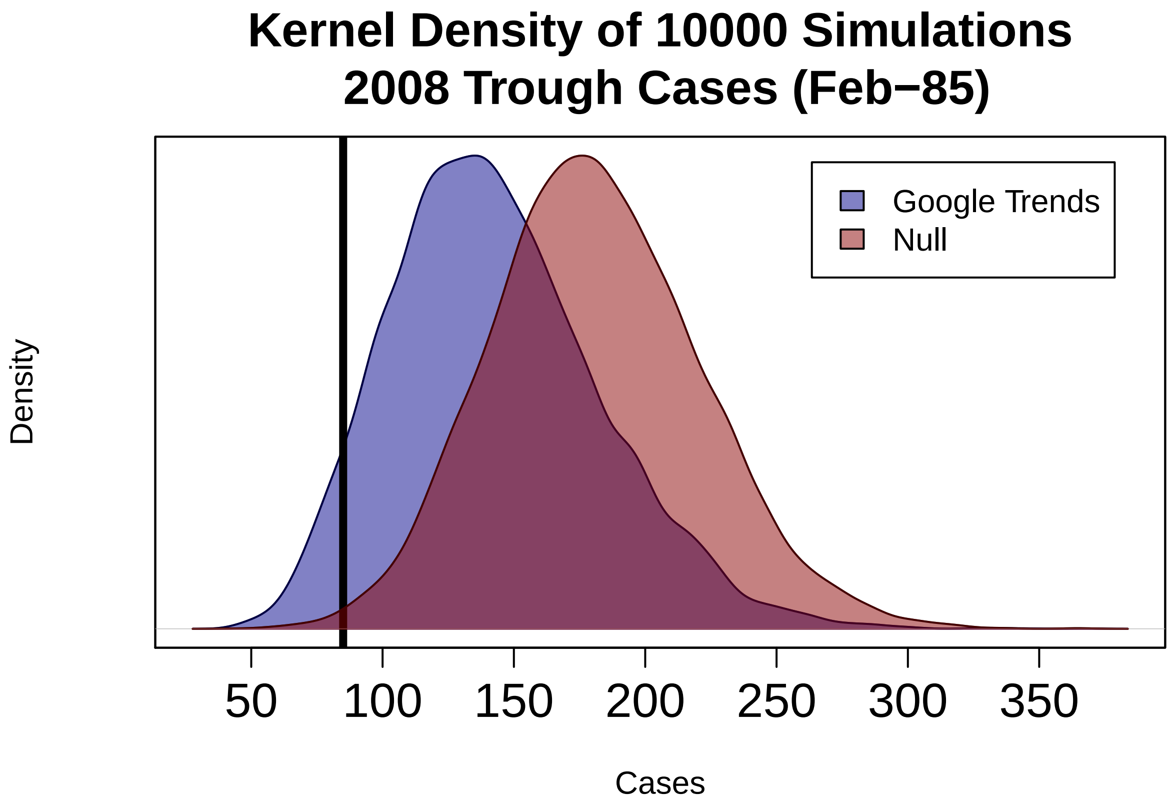
```
## [1]  5  4  2 12  2  2  4  2  2
```

```
Nvalue
```

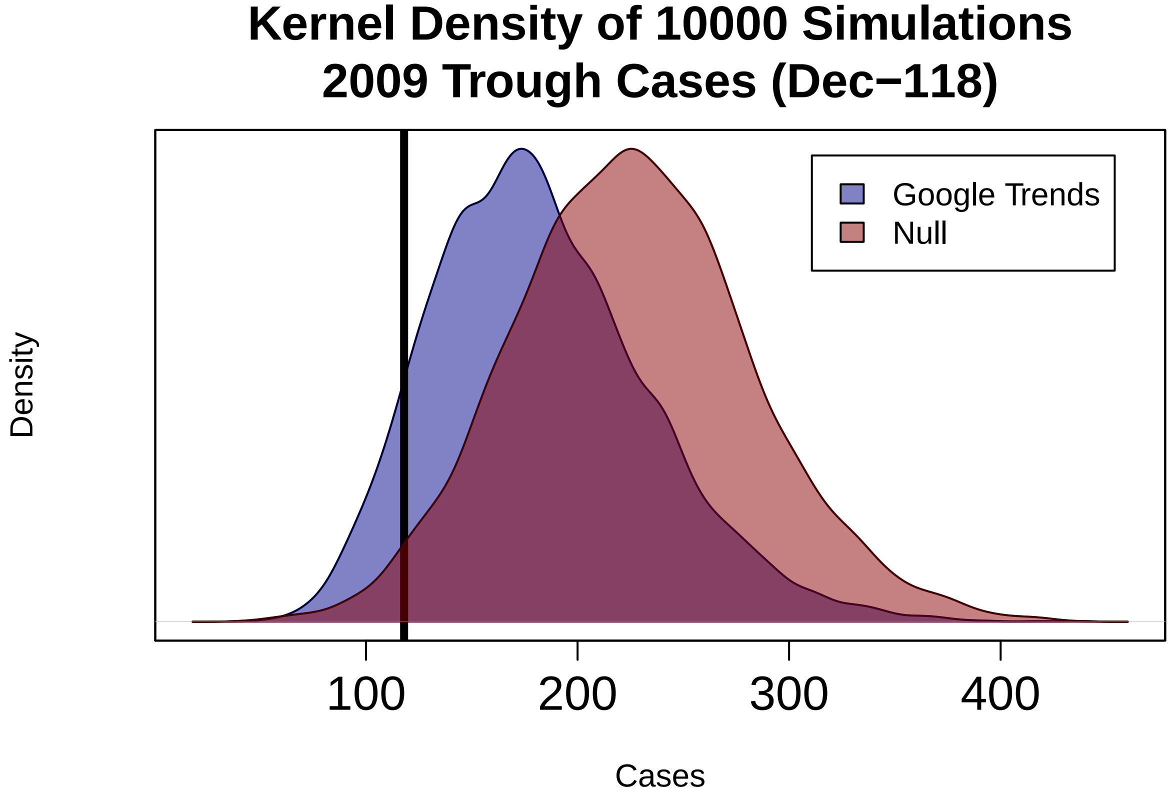
```
## [1]  38  69  85 118  78  87 107 122 116
```


Kernel Density of 10000 Simulations
2006 Trough Cases (May−38)
Google Trends
Null
Density
50
100
150
200
250
300
Cases

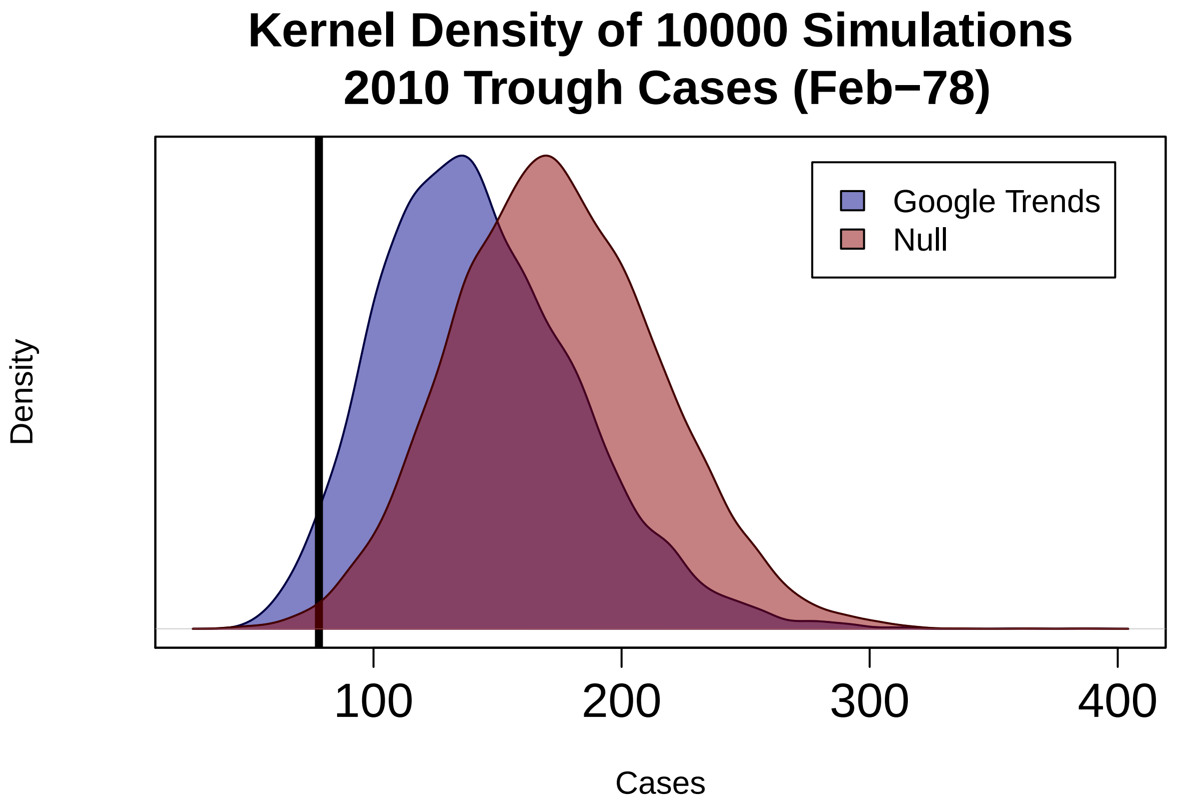

Kernel Density of 10000 Simulations
2007 Trough Cases (Apr−69)
Google Trends
Null
Density
50
100
150
200
250
300
Cases

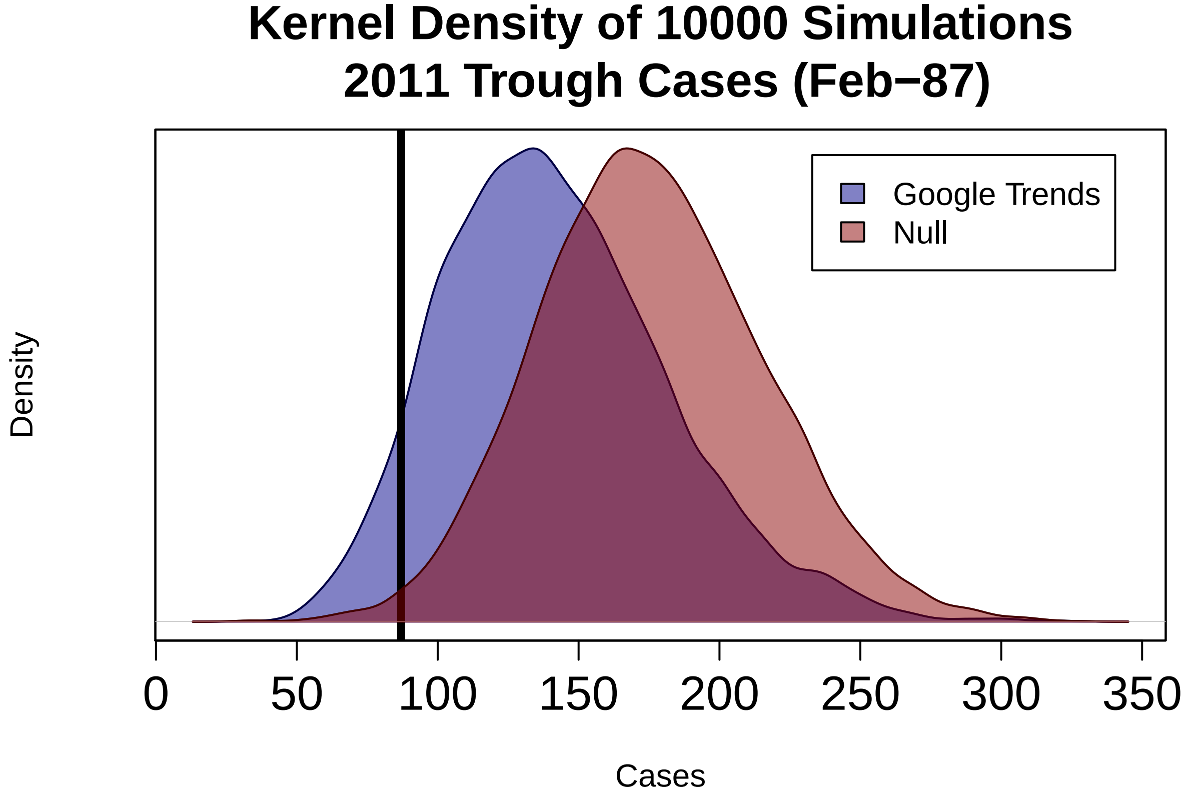


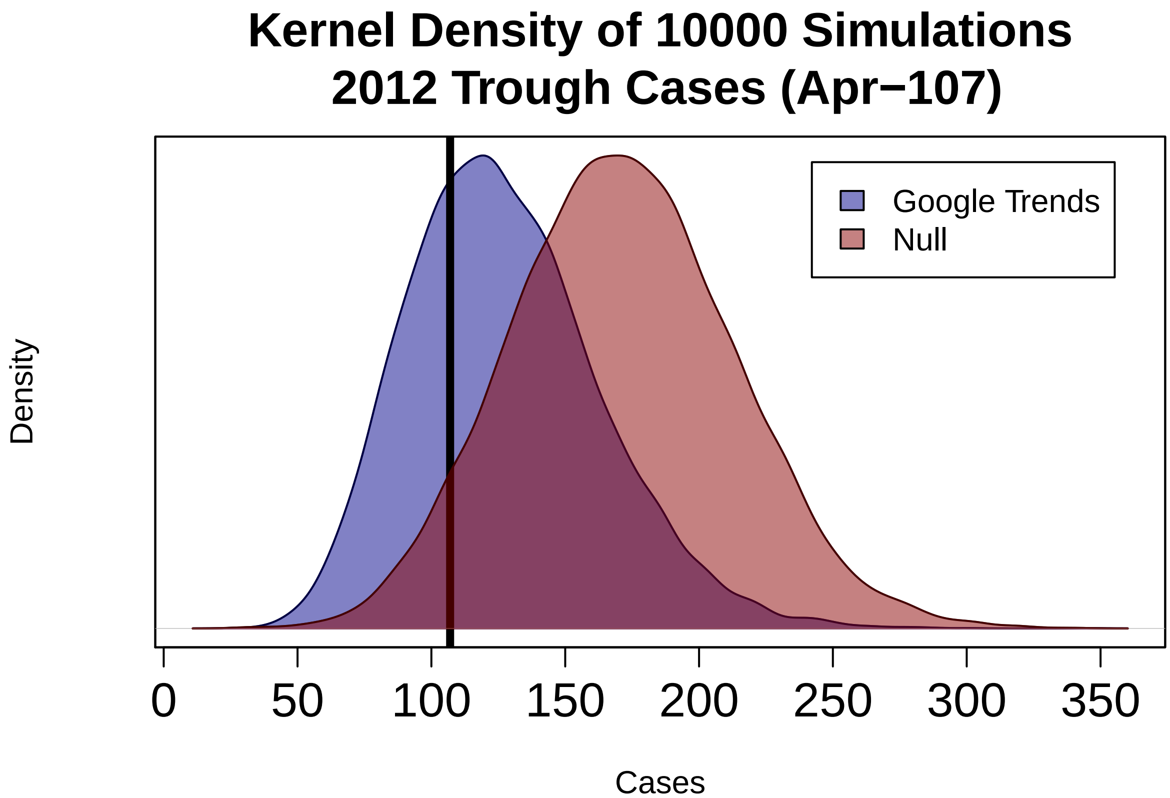
Kernel Density of 10000 Simulations
2008 Trough Cases (Feb−85)

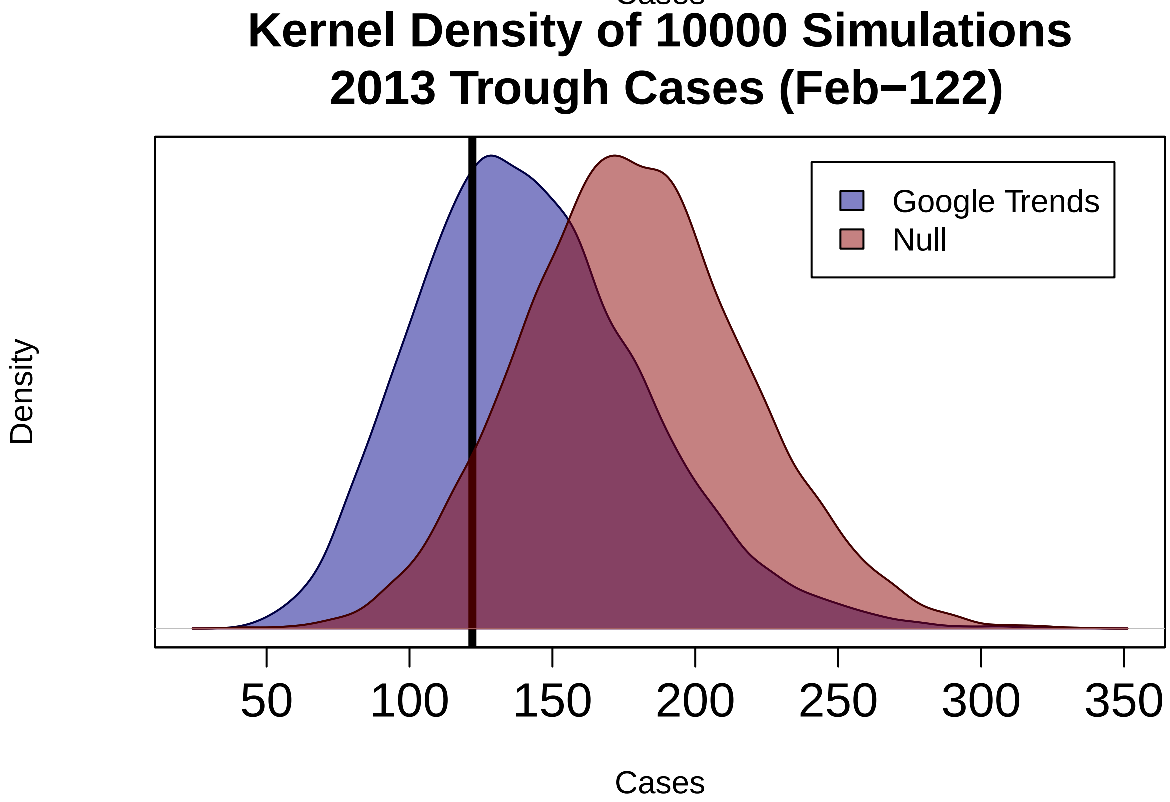


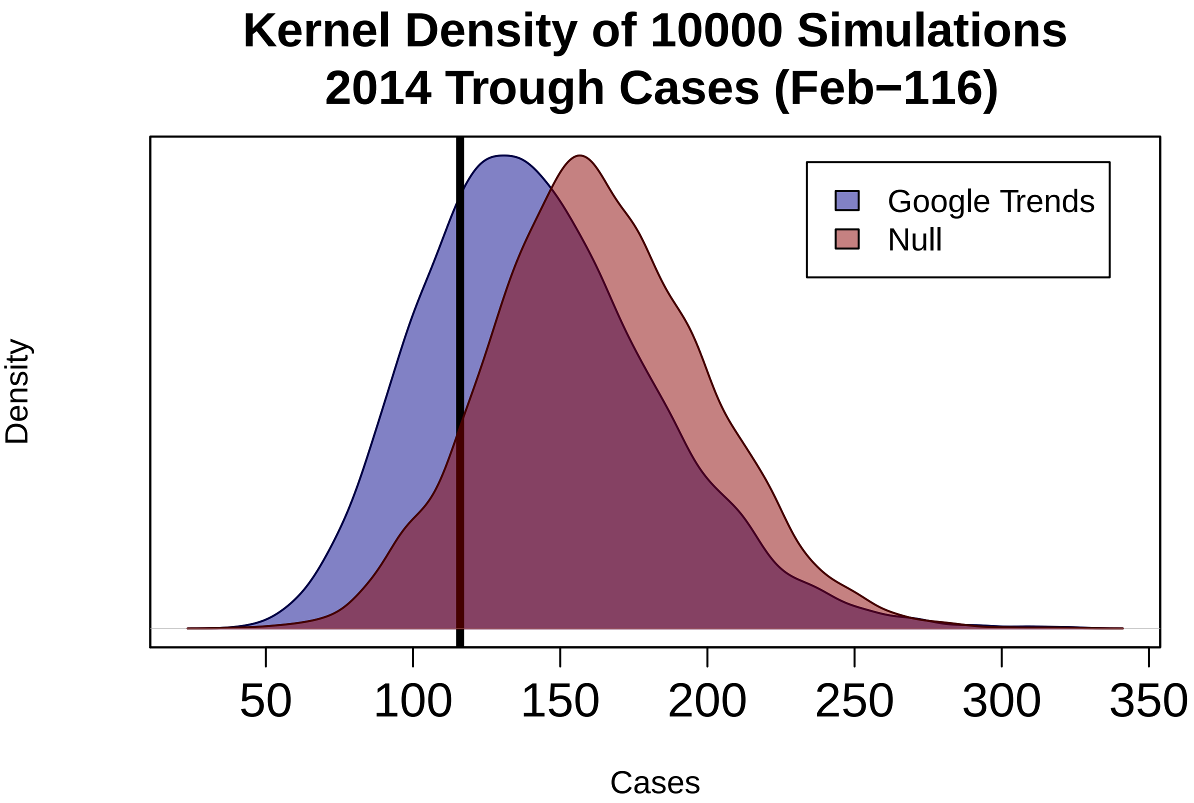
Kernel Density of 10000 Simulations
2009 Trough Cases (Dec−118)


Kernel Density of 10000 Simulations
2010 Trough Cases (Feb−78)
Google Trends
Null
Density
100
200
300
400
Cases


Kernel Density of 10000 Simulations
2011 Trough Cases (Feb−87)
Google Trends
Null
Density
0
50
100
150
200
250
300
350
Cases


Kernel Density of 10000 Simulations
2012 Trough Cases (Apr−107)
Google Trends
Null
Density
0
50
100
150
200
250
300
350
Cases


Kernel Density of 10000 Simulations
2013 Trough Cases (Feb−122)
Google Trends
Null
Density
50
100
150
200
250
300
350
Cases



This figure best explains why the Google Trends model is a better fit to chicken pox data than the null model. While the Google Trends model best captured the actual peak in 2012, 2013, and 2014, it's density distribution was always closer to the actual cases than the null model. The trough in 2006 is hard to characterize as the model is trying to also estimate initial conditions, which could explain why neither the Google Trends model or the null model were able to accurately forecast the number of cases here in May, 2006. Of the models tested that included Google Trends data, model E (Eqn. 9) was best able to forecast chicken pox incidence. It performed better than the null model that captured the mean seasonality of chicken pox incidence. Interestingly, the null model was better able to capture chicken pox peak months, but performed poorly in capturing the troughs each year.

Finally, it may come as a surprise that the two models 'fitting' chicken pox data (models M and N), rather than forecasting, performed similar to the forecasting models that did not include the cosine function. This may be due to the fairly poor correlation between Google Trends data in Australia to the actual case data (due to vaccination). If I chose a different country that does not vaccinate, such as Thailand, I would expect the model fits to be better.

Conclusion

By taking advantage of freely available, real-time, internet search query data, we were able to validate information seeking behaviour as an appropriate proxy for otherwise cryptic chicken pox outbreaks and use those data to forecast outbreaks one month in advance. This modeling approach, which incorporated Google Trends, was able to better forecast outbreaks than models that ignored Google Trends. These results suggest that information seeking can be used for rapid forecasting, when the reporting of clinical cases are unavailable or too slow.

Studies of disease transmission at the global level, and the success of interventions, are limited by data availability. Disease surveillance is a major obstacle in the global effort to improve public health, and is made difficult by underreporting, language barriers, the logistics of data acquisition, and the time required for data curation. I demonstrated that seasonal variation in information seeking reflected disease dynamics, and as such, was able to reveal global patterns of outbreaks and their mitigation via immunization efforts. Thus, digital epidemiology is an easily accessible tool that can be used to complement traditional disease surveillance, and in certain instances, may be the only readily available data source for studying seasonal transmission of non-notifiable diseases. I focused on chicken pox and its dynamics to demonstrate the strength of digital epidemiology for studying childhood diseases at the population level, because VZV is endemic worldwide and the global landscape of VZV vaccination is rapidly changing. Unfortunately, there is still a geographic imbalance of data sources: the vast majority of digital epidemiology data are derived from temperate regions with high internet coverage. However, because many childhood diseases remain non-notifiable throughout the developing world, digital epidemiology provides a valuable approach for identifying recurrent outbreaks when clinical data are lacking. It remains an open challenge to extend the reach of digital epidemiology to study other benign and malignant diseases with under-reported outbreaks and to identify spatiotemporal patterns, where knowledge about the drivers of disease dynamics are most urgently needed.

Citations

1 London, W. P. & Yorke, J. A. Recurrent Outbreaks of Measles, Chickenpox and Mumps. I. Seasonal Variation in Contact Rates. American Journal of Epidemiology, 1973, 98, 453-468

2 Metcalf, C. J. E.; Bjørnstad, O. N.; Grenfell, B. T. & Andreasen, V. Seasonality and Comparative Dynamics of Six Childhood Infections in Pre-vaccination Copenhagen. Proceedings of the Royal Society B: Biological Sciences, 2009, 276, 4111-4118

3 van Panhuis, W. G.; Grefenstette, J.; Jung, S. Y.; Chok, N. S.; Cross, A.; Eng, H.; Lee, B. Y.; Zadorozhny, V.; Brown, S.; Cummings, D. & Burke, D. S. Contagious diseases in the United States from 1888 to the present New England Journal of Medicine, 2013, 369(22), 2152-2158

4 Altizer, S.; Dobson, A.; Hosseini, P.; Hudson, P.; Pascual, M. & Rohani, P. Seasonality and the Dynamics of Infectious Diseases. Ecology letters, 2006, 9, 467-84

5 Grassly, N. & Fraser, C. Seasonal infectious disease epidemiology Proceedings of the Royal Society B: Biological Sciences, 2006, 273, 2541-50

6 Martinez-Bakker, M. & Helm, B. The influence of biological rhythms on host–parasite interactions Trends in ecology & evolution, Elsevier, 2015

7 Dopico, X. C.; Evangelou, M.; Ferreira, R. C.; Guo, H.; Pekalski, M. L.; Smyth, D. J.; Cooper, N.; Burren, O. S.; Fulford, A. J.; Hennig, B. J. & others Widespread seasonal gene expression reveals annual differences in human immunity and physiology Nature communications, Nature Publishing Group, 2015, 6

8 Stevenson, T. J. & Prendergast, B. J. Photoperiodic time measurement and seasonal immunological plasticity Frontiers in neuroendocrinology, Elsevier, 2015, 37, 76-88

9 Keeling, M. & Rohani, P. Modeling infectious diseases in humans and animals Princeton University Press, 2008

10 Higgins, O.; Sixsmith, J.; Barry, M. & Domegan, C. A literature review on health information seeking behaviour on the web: a health consumer and health professional perspective ECDC Technical Report, Stockholm, 2011

11 Brownstein, J.; Freifeld, C. & Madoff, L. Digital disease detection—harnessing the Web for public health surveillance New England Journal of Medicine, 2009, 360(21), 2153-2157

12 Bryden, J.; Funk, S. & Jansen, V. A. Word usage mirrors community structure in the online social network Twitter EPJ Data Science, 2013, 1, 1-9

13 Salathé, M.; Bengtsson, L.; Bodnar, T. J.; Brewer, D. D.; Brownstein, J. S.; Buckee, C.; Campbell, E. M.; Cattuto, C.; Khandelwal, S.; Mabry, P. L. & Vespignani, A. Digital epidemiology PLoS Computational Biology, 2012, 8, 1-5

14 Hulth, A.; Rydevik, G.; Linde, A. & Montgomery, J. Web queries as a source for syndromic surveillance. PloS one, 2009, 4(2), e4378

15 Shaman, J. & Karspeck, A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109, 20425-30

16 Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S. & Brilliant, L. Detecting influenza epidemics using search engine query data. Nature, Nature Publishing Group, 2009, 457, 1012-1014

17 Desai, R.; Lopman, B. A.; Shimshoni, Y.; Harris, J. P.; Patel, M. M. & Parashar, U. Use of internet search data to monitor impact of rotavirus vaccination in the United States. Clinical Infectious Diseases, 2012, cis121

18 Desai, R.; Hall, A.; Lopman, B.; Shimshoni, Y.; Rennick, M.; Efron, N.; Matias, Y.; Patel, M. & Parashar, U. Norovirus disease surveillance using Google internet query share data. Clinical Infectious Diseases, 2012, 55(8), e75-e78

19 Google Google Trends. https://www.google.com/trends/. 2015

20 Australian-Government National Notifiable DIseases Surveillance System. http://www9.health.gov.au/cda/source/rpt1sela.cfm. Accessed May 1, 2015 2015

21 UN http://esa.un.org/unpd/wpp/ Accessed June 18, 2015 2015

22 King, A. A.; Nguyen, D. & Ionides, E. L. Statistical Inference for Partially Observed Markov Processes via the R Package pomp Journal of Statistical Software, 2015, In Press

23 King, A. A.; Ionides, E. L.; Bretó, C. M.; Ellner, S. P.; Ferrari, M. J.; Kendall, B. E.; Lavine, M.; Nguyen, D.; Reuman, D. C.; Wearing, H. & Wood, S. N. pomp: Statistical Inference for Partially Observed Markov Processes, 2015