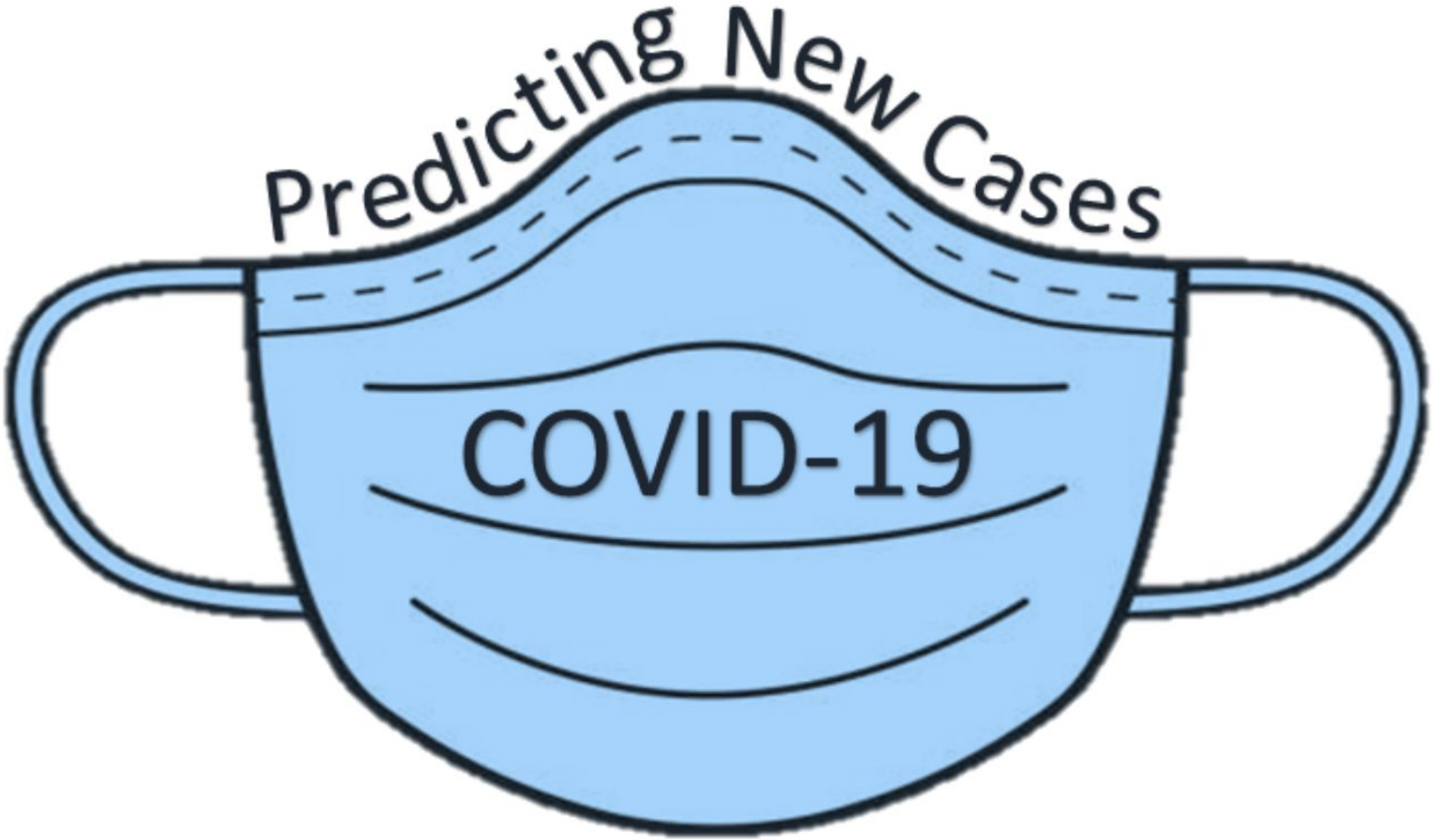
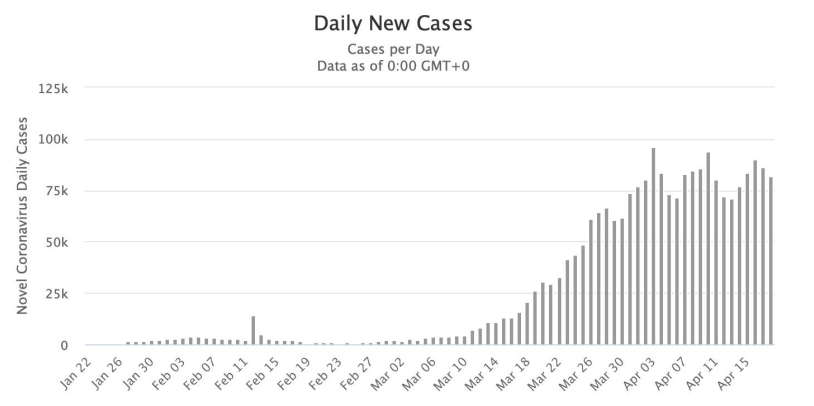


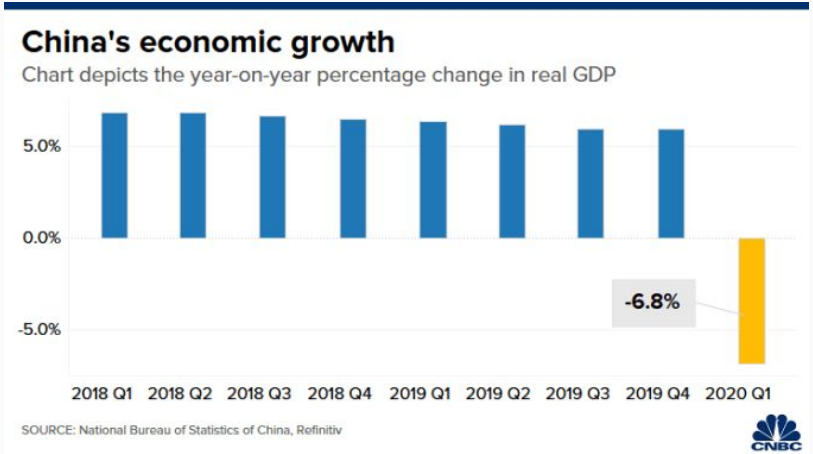
Group 1: Emily Damone, Taylor Krajewski, Alex Quinter, Kushal Shah, Euphy Wu, Jonathan Zhang

Introduction

Introduction

- What is COVID-19?
  - Caused by severe acute respiratory syndrome coronavirus 2
  - First identified in early December 2019 in Wuhan, China
  - Transmitted by close contact and small droplets
  - Flu-like symptoms or asymptomatic
  - Diagnosis: rRT-PC
- Why do we Care?
  - Has infected over 2,000,000 people and 213 countries, areas or territories
  - Social and economic impact
  - High incidence rate
- Many experts have made excellent models and predictions that inspired us to create our own.





Project Goal

To predict the number of new cases of COVID-19 a country can expect on a chosen day given that country's previous number of recorded cases, global health security index, percentage of the population over 65 years old, and percentage of the population living in an urban setting.

Data

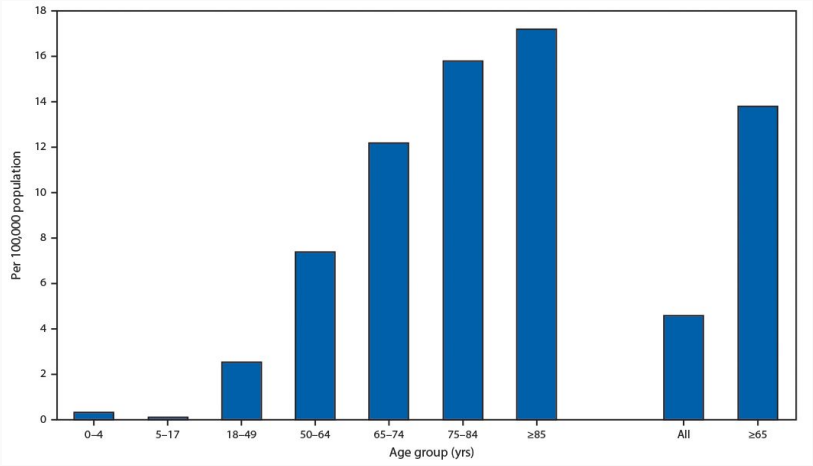
Hopkins COVID-19 Data

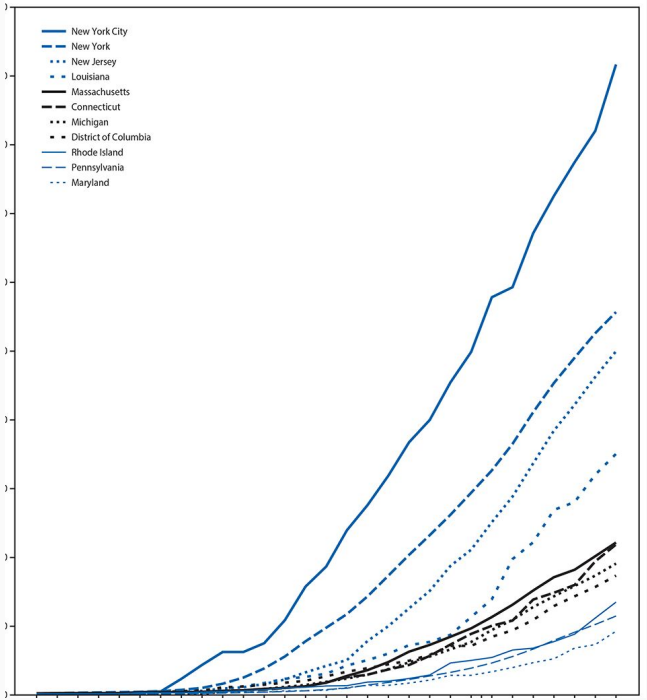
- COVID-19 Dataset
  - From the CSSE at Johns Hopkins University
  - Includes:
    - Country/Region
    - Province/State
    - The number of confirmed cases by day
    - The number of Deaths by day
    - The number of recovered cases by day

- Modification on the dataset
  - Create a variable for new cases by day for each country
  - Relabel the date at which a country passes 50 total cases as “Day 0”
  - Eliminate any country that had less than 50 total cases on Day 0
  - Eliminate any country with less than 5 observations in Poisson GLMM

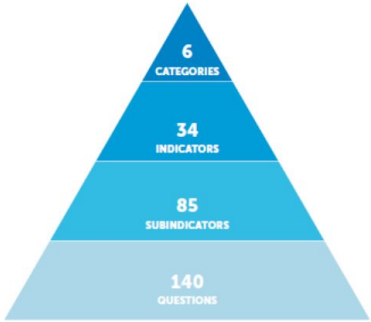
World Bank Data

- Population ages 65 and above (% of total population) by country
  - Older adults face higher COVID-19 risk
  - Highest COVID-19 associated hospitalization rate
- Urban Population (% of total population) by country
  - Population density might play a significant role in the acceleration of transmission





Global Health Security Index



PREVENT

1. PREVENTION
Prevention of the emergence or release of pathogens

DETECT

2. DETECTION AND REPORTING
Early detection and reporting for epidemics of potential international concern

RESPOND

3. RAPID RESPONSE
Rapid response to and mitigation of the spread of an epidemic

HEALTH

4. HEALTH SYSTEM
Sufficient and robust health system to treat the sick and protect health workers

NORMS

5. COMPLIANCE WITH INTERNATIONAL NORMS
Commitments to improving national capacity, financing plans to address gaps, and adhering to global norms

RISK

6. RISK ENVIRONMENT
Overall risk environment and country vulnerability to biological threats

Preview of Finalized Data

Below is a preview of what our final data set looks like. Each country has a row for each day of recorded cases from baseline (day of 50 cases) until April 3, 2020.

United States Data					
Day	**Total Cases**	**New Cases**	**GHS Index**	**% 65+**	**% Urban Setting**
0	51	36	83.5	15.80765	82.256
1	51	0	83.5	15.80765	82.256
2	57	6	83.5	15.80765	82.256
3	58	1	83.5	15.80765	82.256
4	60	2	83.5	15.80765	82.256
5	68	8	83.5	15.80765	82.256

Motivation

Motivation

Total Cases:

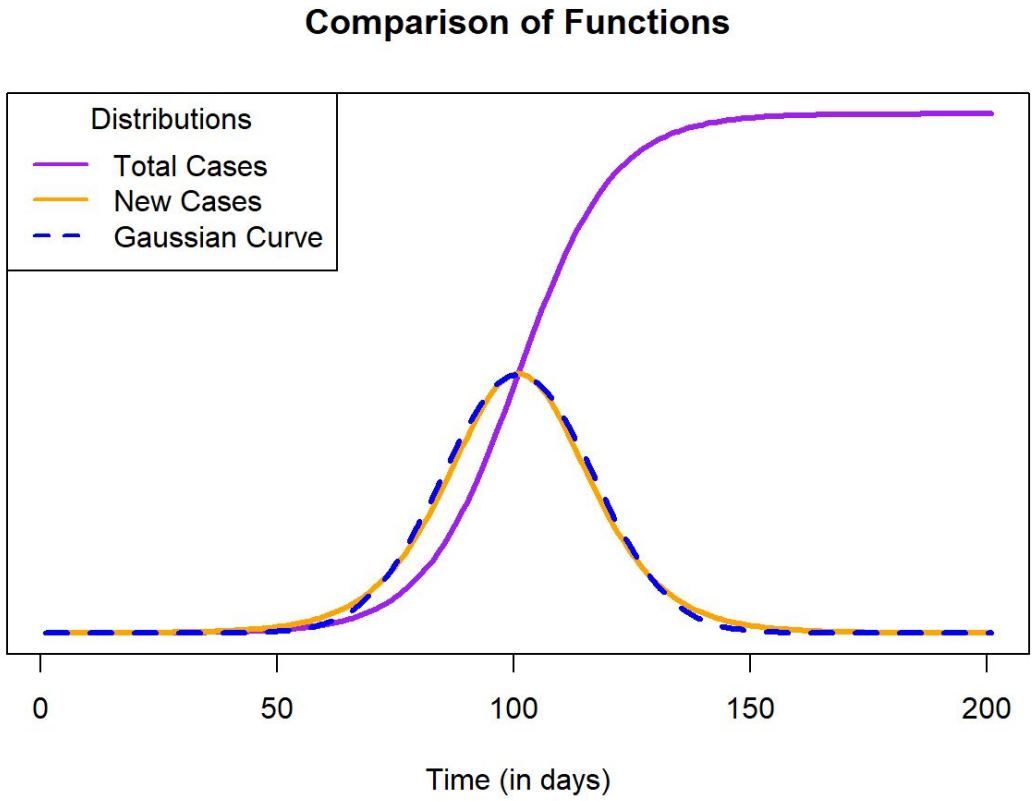
$$C(t) = \frac{KC_0}{C_0 + (K - C_0)e^{-rt}}$$

New Cases:

$$\frac{d}{dt} = \frac{rkC_0e^{rt}(K - C_0)}{\left((K - C_0) + C_0e^{rt}\right)^2}$$

Gaussian Curve:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t^2)}$$



Poisson GLMM

Poisson GLMM

y_{ij} = number of new cases in country i on day j

$$y_{ij} \sim Poisson(\lambda_{ij})$$

- $x_{ij} = (1, t_{ij}, t_{ij}^2, GHS, AgeGEQ65, UrbanPop)$
- Assuming an intercept, we have $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$

$$log(\lambda_{ij}) = x_{ij}\beta + \gamma_{i1} + \gamma_{i2}t_{ij} = x_{ij}\beta + z_{ij}\gamma_i$$

- γ_i is a vector of unobserved country-level random effects
- We assume $\begin{pmatrix} \gamma_{i1} \\ \gamma_{i2} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \right)$
  - γ_{i1} is the deviation of the i[th] country from the baseline number of new cases, which we have set to 50.
  - γ_{i2} is the deviation of the i[th] country from the average effect of the covariates in our model on number of new cases

Poisson GLMM

Let n be the total number of counties and n_i is the observed days of data in each country

Likelihood:

$$L(\beta, G|y, \gamma) = \prod_{i=1}^{n} \left(\prod_{j=1}^{n_i} f(y_{ij}|\lambda_{ij}) \right) \phi(\gamma_i|0, G) = \prod_{i=1}^{n} \int \left[\left(\prod_{j=1}^{n_i} f(y_{ij}|\lambda_{ij}) \right) \phi(\gamma_i|0, G) \right] d\gamma_i$$

and log-Likelihood:

$$l(\beta, G|y, \gamma) = \sum_{i=1}^{n} log \left[\int \left[\left(\prod_{j=1}^{n_i} f(y_{ij}|\lambda_{ij}) \right) \phi(\gamma_i|0, G) \right] d\gamma_i \right]$$

- Poisson PMF $f(y_{ij}|\lambda_{ij})$ with mean λ_{ij}
- Bivariate Normal PDF $\phi(\gamma_i|0, G)$ where $G = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$

Poisson GLMM - Estimation

Expectation of Complete Data log-Likelihood:

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= E[l_c(\theta)|\theta^{(t)}] \\
&= E\left[\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i} log(f(y_{ij}\lambda_{ij}^{(t)})) + log(\phi(\gamma_i|0, G^{(t)}))\right]\right] \\
&= \sum_{i=1}^{n}\left[\int\left[\sum_{j=1}^{n_i} log(f(y_{ij}|\lambda_{ij}^{(t)})) + log(\phi(\gamma_i|0, G^{(t)}))\right] f(\gamma_i|\beta^{(t)}, G^{(t)})d\gamma_i\right], \text{ where } \lambda_{ij}^{(t)} = e^{x_{ij}\beta^{(t)}+z_{ij}\gamma_i}
\end{aligned}
$$

If we could sample from the posterior distribution of γ_i we could approximate this integral through a MC approach

$$
Q(\theta|\theta^{(t)}) = \frac{1}{M}\sum_{i=1}^{n}\left[\sum_{k=1}^{M}\left[\sum_{j=1}^{n_i} log(f(y_{ij}|\lambda_{ijk}^{(t)})) + log(\phi(\Gamma_{ik}|0, G^{(t)}))\right]\right]
$$

where Γ_{ik} is one of the *M* samples from the i^{th} country, and $\lambda_{ijk}^{(t)} = e^{x_{ij}\beta^{(t)}+z_{ij}\Gamma_{ij}}$

Poisson GLMM - Estimation

To sample from this posterior, we tried three different approaches:

1. Independence Random Walk
2. Metropolis-within-Gibbs Random Walk
3. Adaptive Metropolis-within-Gibbs Random Walk

Metropolis-within-Gibbs Random Walk was the most suitable of these 3 approaches.

Poisson GLMM - Maximization

In the Maximization step, we maximize the Q-function with respect to β and G.

Based on our Q-function, $Q(\theta|\theta^{(t)}) = \frac{1}{M}\sum_{i=1}^{n}\left[\sum_{k=1}^{M}\left[\sum_{j=1}^{n_i} log(f(y_{ij}|\lambda_{ijk}^{(t)})) + log(\phi(\Gamma_{ik}|0, G^{(t)}))\right]\right]$,

the maximization process of β follows the maximization of parameters in a Poisson GLM with a few tweaks:

1. Each sample must be weighted $\frac{1}{M}$
2. An offset for the random effects $z_i\gamma_i$

And the maximization of G is the sample variance/covariance of the MC.

Poisson GLMM - Prediction

- We want to make predictions for a specific country. Consider the expectation of the posterior distribution as an estimator:

$$\begin{aligned}\hat{\gamma}_i &= E[\gamma_i | y_i, \hat{\beta}, \hat{G}] \\ &= \int \gamma_i f(\gamma_i | y_i, \hat{\beta}, \hat{G}) d\gamma_i\end{aligned}$$

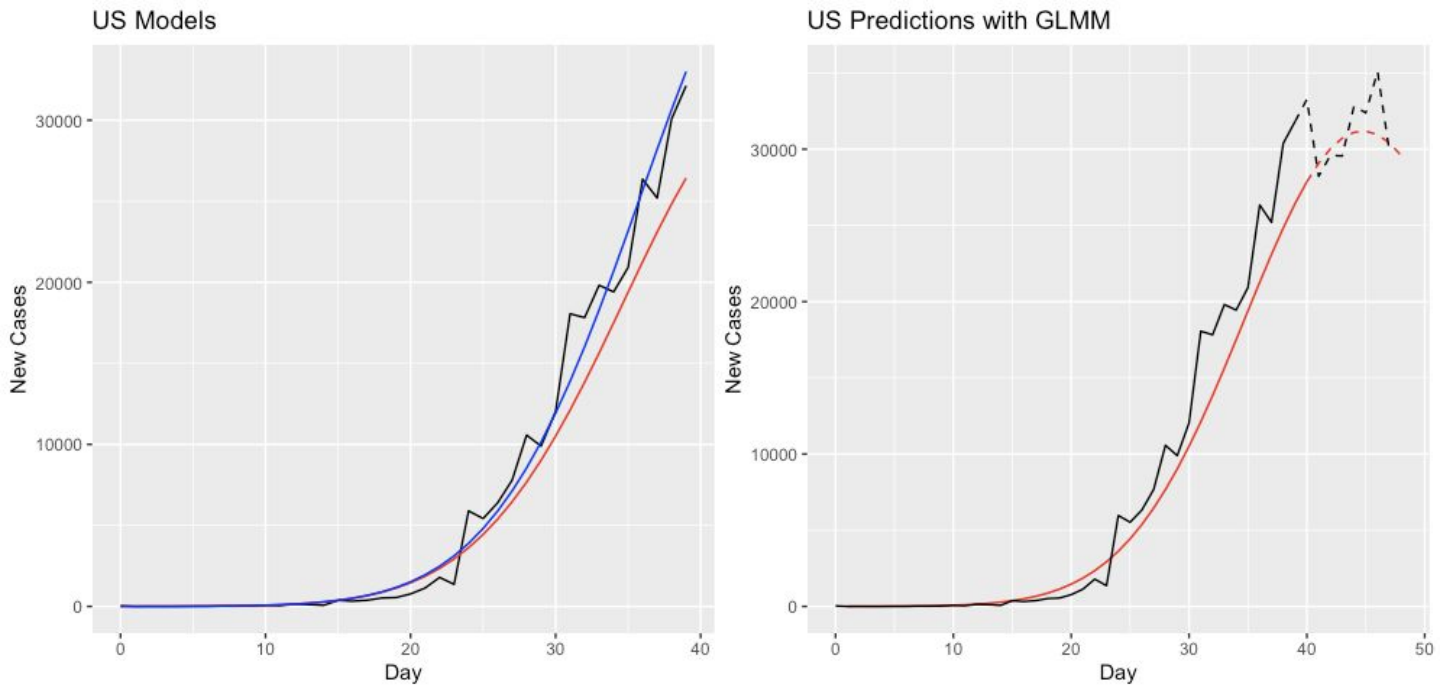
- But we already have a sample from this posterior; the Markov chains from the EM algorithm.
- Taking the mean of these chains should give us a prediction of the random effects for a specific country

$$E[\gamma_i | y_i, \hat{\beta}, \hat{G}] = \frac{1}{M} \sum_{k=1}^{M} \Gamma_{ik} \qquad E[y_{ij} | \gamma_i] = \frac{1}{1 + e^{-(x_{ij}\hat{\beta} + z_{ij}\hat{\gamma}_i)}}$$

Estimation Results - glmer vs MCEM

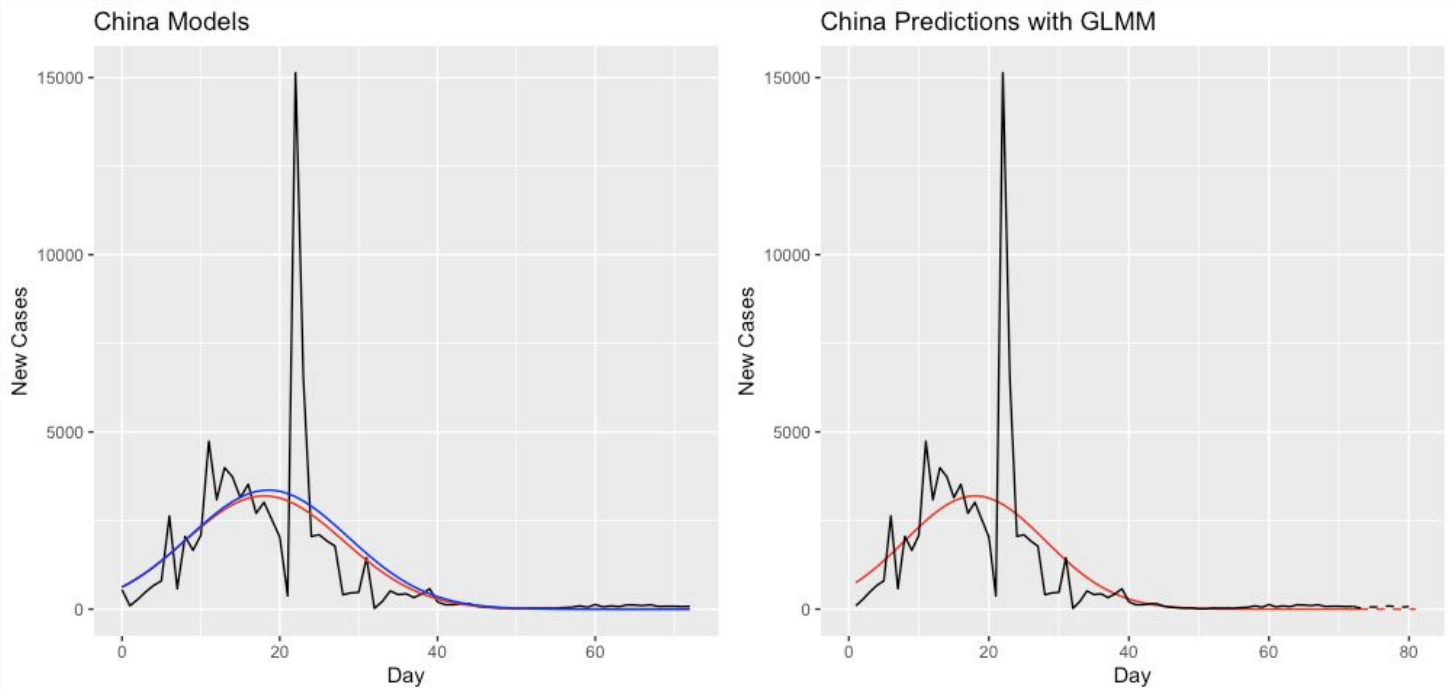
Parameter	MCEM	glmer
β_0	0.844	0.814
β_1	0.202	0.192
β_2	−0.005	-0.0049
β_3	0.028	0.033
β_4	0.010	0.014
β_5	−0.001	0.0026
γ_{11}	1.922	1.944
γ_{12}	-0.068	-0.010
γ_{22}	0.010	0.010

Poisson GLMM Results - US



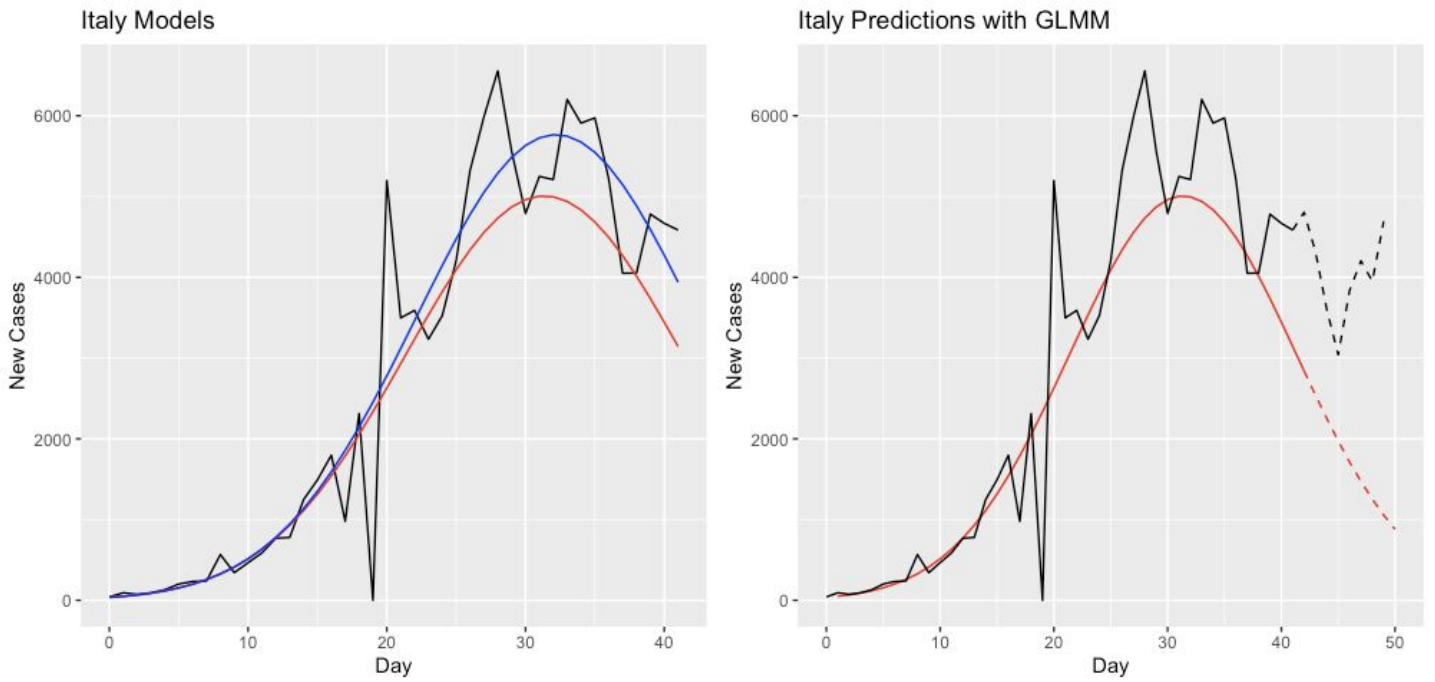
	MCEM	glmer
Random Effect Predictions	(-2.923, 0.246)	(-3.350, 0.259)

Poisson GLMM Results - China



	MCEM	glmer
Random Effect Predictions	(4.200, -0.0216)	(4.009 -0.0104)

Poisson GLMM Results - Italy



	MCEM	glmer
Random Effect Predictions	(4.200, -0.0216)	(4.009 -0.0104)

Machine Learning

Machine Learning

Random Forest

Why we chose this method:

- Nonparametric
- Allows for feature importance estimates
- Ensemble of decision trees can generate improved predictions

Limitations at the outset:

- Ignores correlation when applied to full dataset
- Extrapolation (prediction at future time points) potentially biased towards average results

Prophet (Time-Series Modeling)

Why we chose this method:

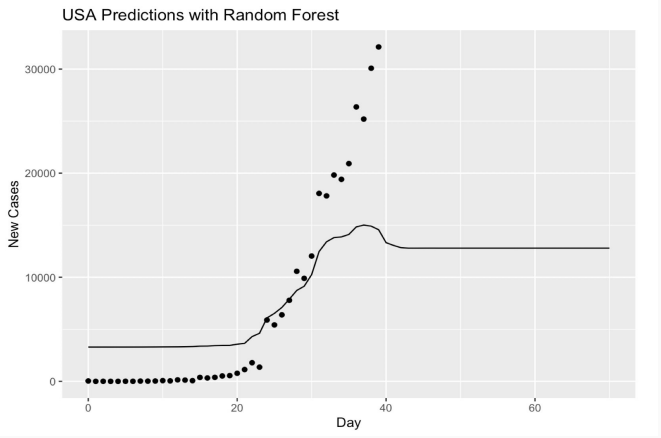
- Allows for forecasting of future cases
- Logistic parameterization

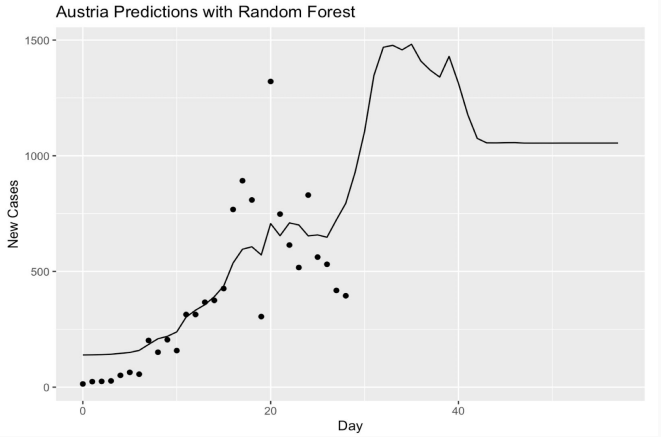
Limitations at the outset:

- Must be applied to each country individually
- Does not include covariates included in other models

Random Forest

- Unreliable fits to current data, resulting in unreasonable predictions
- Makes predictions for a country *not* based on its previous time points, but based on data from other “similar” countries at the same time point
- Decision tree-based structure results in extrapolations that average-out after a certain time point
- These results are **unsurprising** given the weaknesses of machine learning methods applied to longitudinal datasets (especially with unfinished time cycles), which will be further elaborated during the discussion of limitations





Prophet

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

y(t) *~ Total cases* ***g(t)****~ non-periodic changes/trend* ***s(t), h(t)*** *~seasonal and holiday trends*

With a logistic assumption, the non-periodic change is modeled as follows

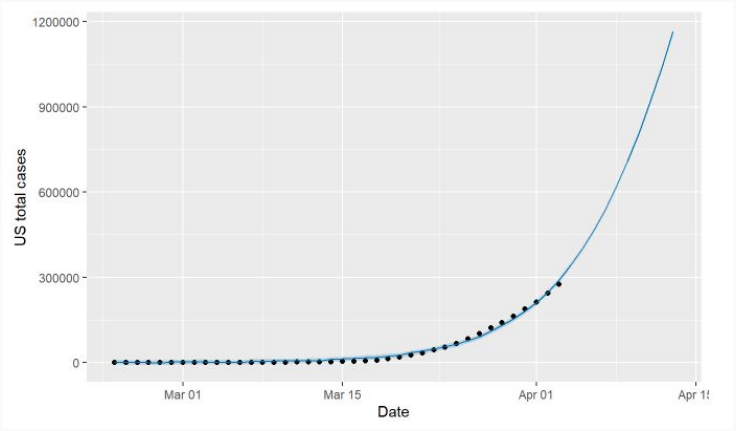
$$g(t) = \frac{C(t)}{1 + exp(-k(t - m))}$$

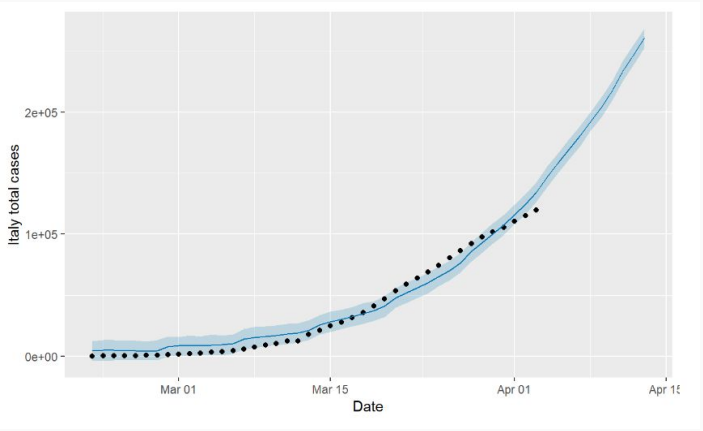
C(t) ~ function for the carrying capacity ***k(t)*** ~growth rate ***m*** ~ offset parameter

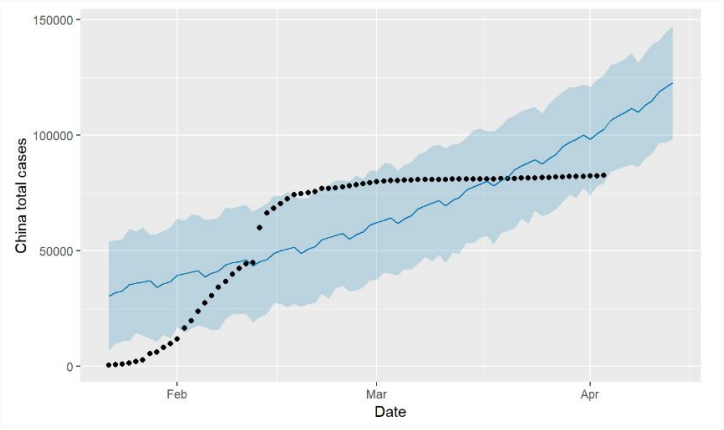
Prophet

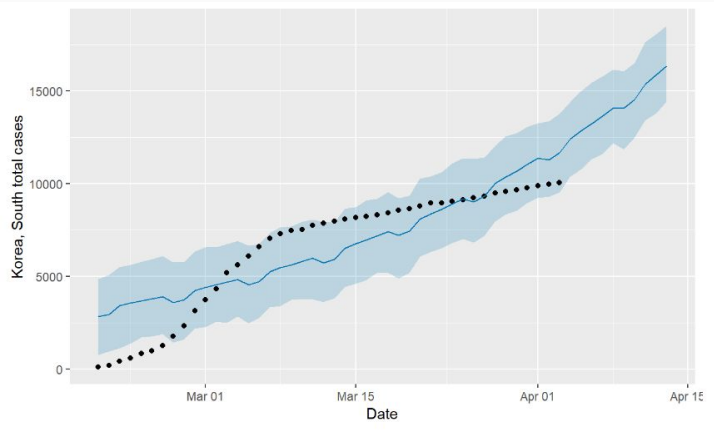
```
forecast <- function(country, data, numPred){
  dat <- data[data$Country.Region == country, ]
  prophet_dat <- as.data.frame(dat$date)
  prophet_dat$y <- as.numeric(dat$total_cases)
  prophet_dat$cap <- as.numeric(dat$TotalPop*.01)

  colnames(prophet_dat) <- c("ds", "y", "cap")

  now <- prophet(prophet_dat, growth= "logistic")
  future <- make_future_dataframe(now, periods=numPred)
  future$cap <- rep(prophet_dat[1,3], length(future$ds))

  forecast <- predict(now, future, )
  print(plot(now, forecast, plot_cap=F, uncertainty = T, ylabel = paste(country, "total cases")))
  return(list(now,forecast))
}
```

Prophet Continued









Discussion

Comparison of Poisson GLMM & Machine Learning Methods

Poisson GLMM vs. Random Forest

- Random forest limited due to vastly different country-specific disease curve shapes and spreads
- Parametric approach advantageous for COVID dataset
  - Maintains known disease curve shape
  - Future prediction requires extrapolation, which benefits from a fixed modeling structure

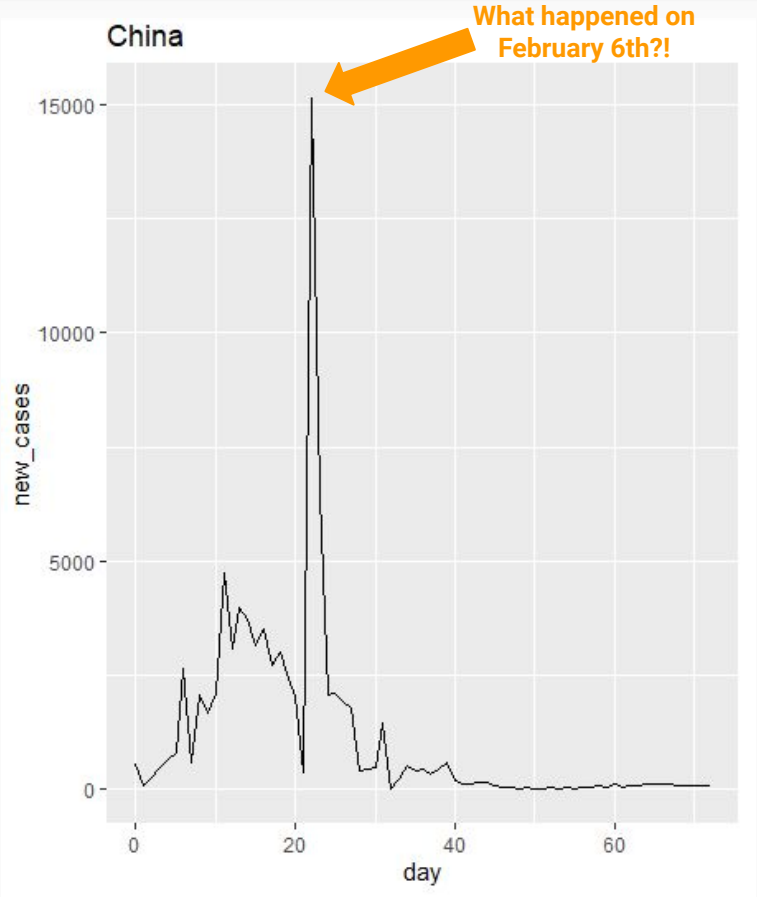
Poisson GLMM vs. Prophet

- On countries that have more data available, Prophet can have good predictions, but is limited to these countries only
- Prophet works better for seasonal and weekly trends, which is not represented in COVID data so far
- Poisson GLMM better for worldwide predictions because it can model using all of the data

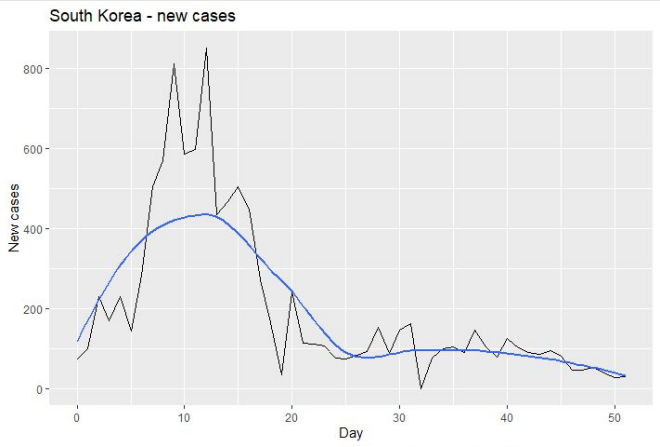
Limitations

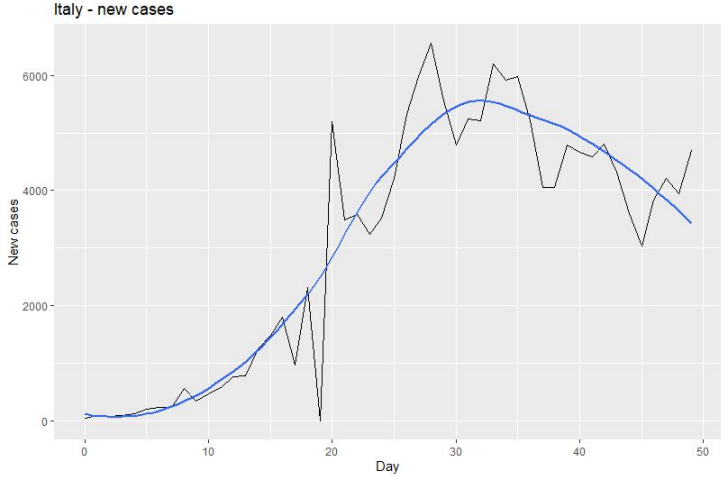
Situational Limitations

- Due to the rare occurrence of pandemics, such as COVID-19, it is very difficult to base a pandemic model off of previously modeled pandemics due to socio-cultural changes between events.
- Lack of data on movement restrictions that have a great impact on spread of disease
- Lack of widespread testing and possibility of infected individuals lacking symptoms make it impossible to know how many cases of COVID-19 there truly are
- Countries not accurately reporting data, lag time between cases confirmed and reported



Limitations





Statistical/Computing Limitations

- MCMC approaches have become somewhat of a 'gold standard,' but with complications
    - Writing a likelihood in closed form
    - Data augmentation that involves imputation of more unknowns than feasible to handle
    - Difficulty implementing the model in real time due to high level of computational effort.
- More recent data has shown countries with a bell curve-type ascent but linear descent (e.g. Italy) and flatter peaks (e.g. USA)

Questions?

References

References

“Coronavirus.” World Health Organization, World Health Organization, www.who.int/health-topics/coronavirus#tab=tab_1.

"Coronavirus disease 2019 (COVID-19)—Symptoms and causes". *Mayo Clinic*. Retrieved 14 April 2020.

China says its economy shrank by 6.8% in the first quarter as the country battled coronavirus

De Angelis, Daniela et al. “Four key challenges in infectious disease modelling using data from multiple sources.” Epidemics vol. 10 (2015): 83-7. doi:10.1016/j.epidem.2014.09.004

Geographic Differences in COVID-19 Cases, Deaths, and Incidence — United States, February 12–April 7, 2020

Google Cloud Platform, Google, console.cloud.google.com/marketplace/details/johnshopkins/covid19_jhu_global_cases.

Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1–30, 2020

"Interim Guidelines for Collecting, Handling, and Testing Clinical Specimens from Persons for Coronavirus Disease 2019 (COVID-19)". *Centers for Disease Control and Prevention (CDC)*. 11 February 2020.

Junling Ma, “Estimating epidemic exponential growth rate and basic reproduction number.” Infectious Disease Modelling, Volume 5 (2020): 129-141. doi: 10.1016/j.idm.2019.12.009. eCollection 2020.

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c. Journal of biomedical informatics, 89, 56-67.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

“The Global Health Security Index.” GHS Index, www.ghsindex.org/.

Walters, Caroline E et al. “Modelling the global spread of diseases: A review of current practice and capability.” Epidemics vol. 25 (2018): 1-8. doi:10.1016/j.epidem.2018.05.007

“World Bank Open Data.” Data, 7 Apr. 2020, data.worldbank.org/.

“World Wide Daily New Cases” www.worldometers.info