TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Analysis of daily corona data: a cautionary tale

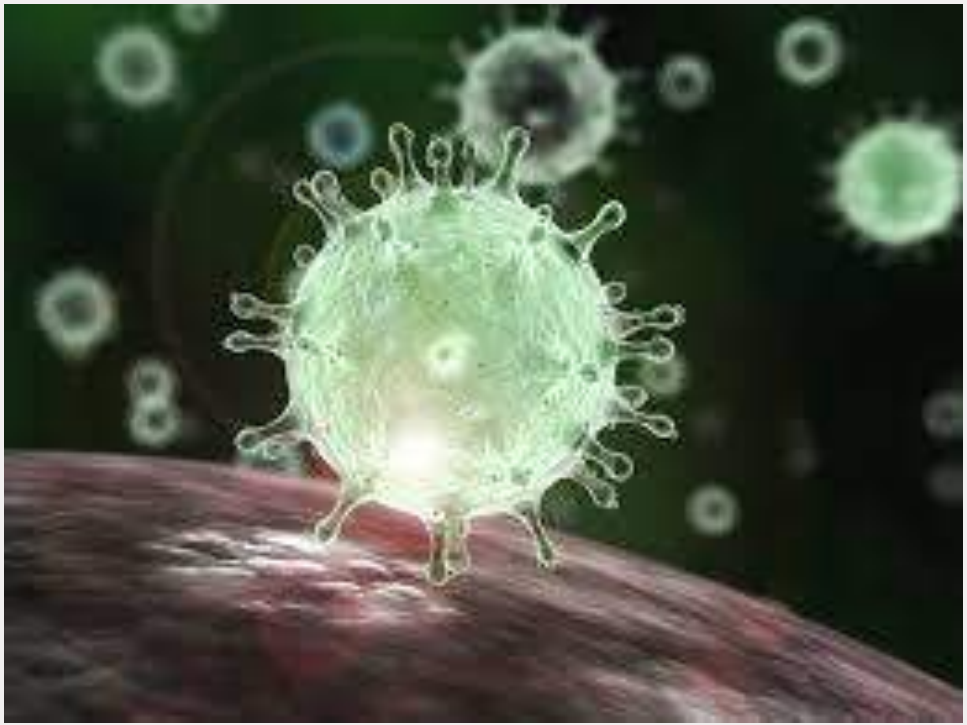
March 8, 2021

Edwin van den Heuvel

Mathematics & Computer Science

Content

1. Background Empirical Research
    a. Epidemiology
    b. Representative Sampling
    c. Sampling procedures corona
2. Epidemic Disease Models
    a. Differential equations
    b. Different analysis approaches
    c. Comparisons of curves
    d. Predictions of infections and deaths
    e. Prediction hospitalization capacity
    f. Generalized logistic curves
3. Governmental Interventions
    a. Discrete SEIR model
    b. Goodness-of-fit
    c. Daily effective contact-rate profile
4. Data science during a pandemic



TU/e

Background Empirical Research

Epidemiology

- Epidemiology studies the occurrence, distribution, and determinants of disease and health in individuals and (sub)populations

<u>Four focus areas</u>

- *<u>Diagnosis:</u>* determining and detecting disease
- *<u>Etiology:</u>* why has this person the disease and why now
- *<u>Prevention:</u>* how to reduce the risk of disease
- *<u>Prognosis:</u>* what is (changing) the disease progression

TU/e

Background Empirical Research

Epidemiology

- Proper data collection is essential:
  1. *Systematic reviews:* combining multiple studies
  2. *Randomized controlled trials:* human experiments
  3. *Cohort studies:* a representative group of participants is followed over time
  4. *Case-control studies:* controls are collected to match the group of disease cases
  5. *Cross-sectional studies:* sample of participants at one moment in time
  6. *Case reports:* individuals are described
- Scientific data collection principles
  - Representative sampling or probability sampling
  - Randomization of treatment allocation

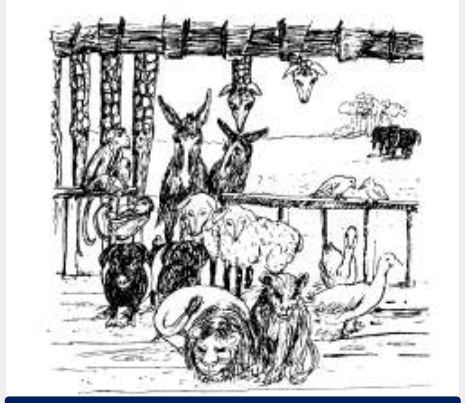
Background Empirical Research

Representative sampling

- Around the 1900's, discussions on sampling from populations initiate
  - Anders Nicolai Kaier – Director Norwegian Central Bureau of Statistics – Introduced the term 'representative method'
- Different meanings for samples exists[1]

Miniature

Typical or Ideal



Coverage (Arc of Noah)

Good estimation

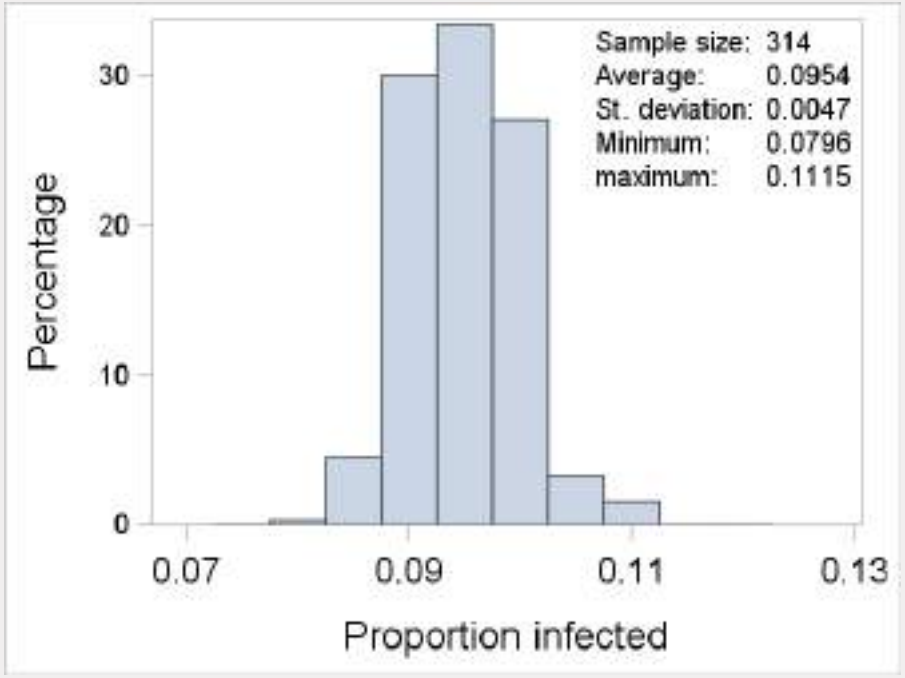
[1] Kruskal W, Mosteller F, Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939, *International Statistical Review*, 1980, **48**(2):169-195.

Background Empirical Research

Representative sampling

- Jerzy Neyman – Polish statistician – introduces probability sampling[1]
  - Quantifies bias
  - Develops confidence intervals
- Example:
  - Population is six schools
  - Percentage of infected children

School	# children	#infected	Proportion
1	590	40	0.068
2	280	50	0.179
3	900	30	0.033
4	440	30	0.069
5	360	70	0.194
6	570	80	0.140
Total	3140	300	0.0955



[1] Neyman J, On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, 1934, **97**(4):558-625

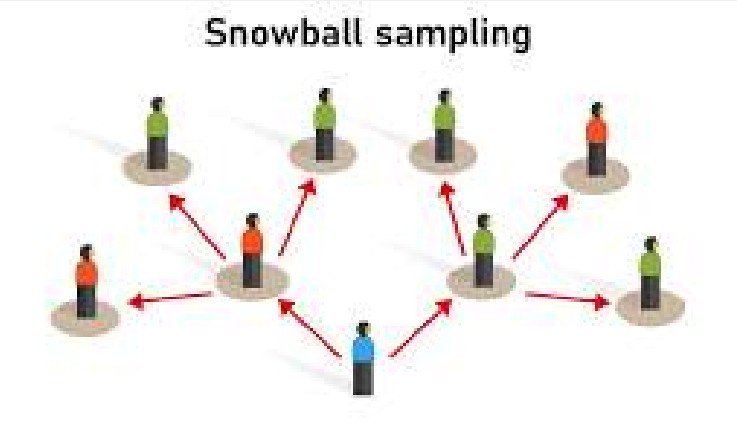
TU/e

Background Empirical Research

Sampling procedures corona

Purposive sample

- First wave: contact research



- Wave 2: test streets for volunteers and symptomatic people

Approach China

- Contact sampling
  - Wuhan: ≥ 1800 teams of epidemiologists
  - Percentage infected from contact: 1% - 5%

Date	Location	Contacts	Tested	Infected
17-02-2020	Shenzhen City	2842	2240	3.1%
17-02-2020	Sichuan Province	25493	23178	0.9%
20-02-2020	Guangdong Province	9939	7765	4.8%

- Examples fever clinics
  - Wuhan until 2nd week of January: 0/250 tests
  - Guangdong 01/01 – 14/01: 1/15000 tests
  - Hospital Beijing: 28/01 – 13/02: 0/1910 tests

TU/e

Background Empirical Research

Sampling procedures corona[1]

- Purposive samples are considered scientifically unreliable
  - Symptomatic people are typically overrepresented
  - People who are concerned or need to travel are overrepresented
  - Health personnel are overrepresented
  - Lack of test capacity results into an underestimation of infected people
  - Sensitivity and specificity of tests should be known to be able to estimate rates
- **Conclusion: Numbers are unreliable**

Counter arguments:

- Purposive samples are sometimes better than probability sampling
  - Case-control studies
  - John Snow – cholera outbreak used purposive sampling (epidemiology)
  - Frequently used in market and opinion research
- Probability sampling fails:
  - Non-response is large – selection
  - Random sample may deviate strongly

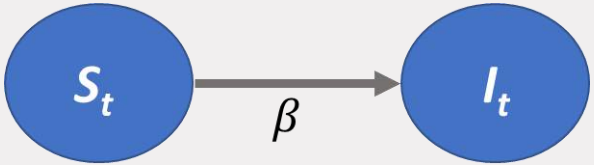
[1] Schnell R, Smid M. Methodological Problems and Solutions in Sampling for Epidemiological COVID-19 Research, *Survey Research Methods*, 2020 **14**(2):123-129).

Epidemic Disease Models

Differential Equations

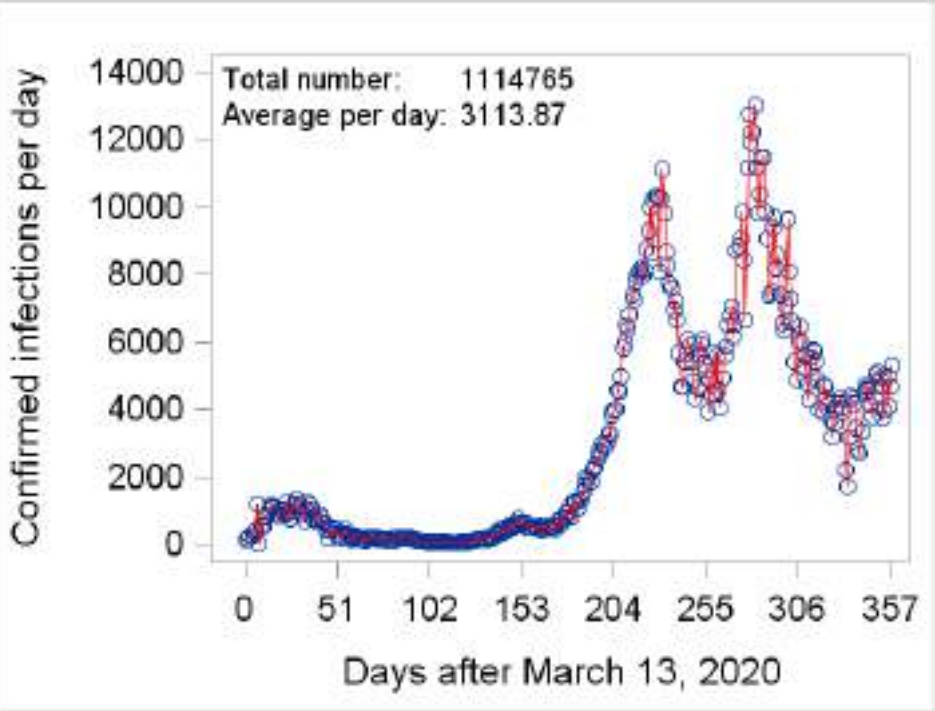
Susceptible-Infected (SI) model

- Compartmental model



  - S_t: number of susceptible individuals
  - I_t: number of infected individuals
  - β: daily number of effective contacts
  - $M = S_t + I_t$ total population
- Differential equation Verhulst:

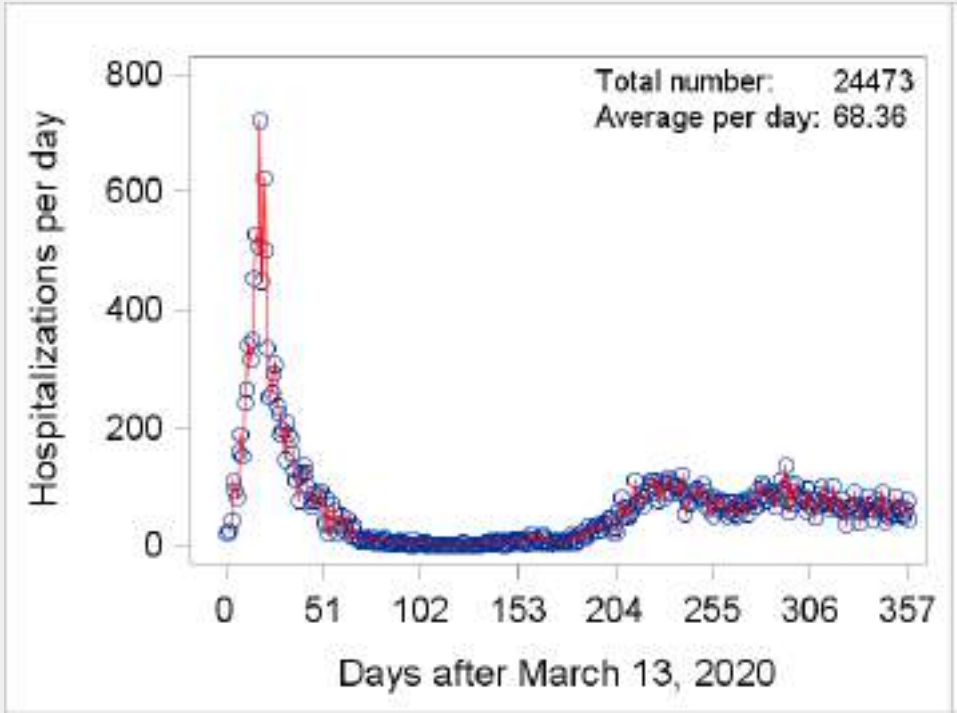
$$\frac{dI_t}{dt} = \beta \frac{I_t \cdot S_t}{M}$$

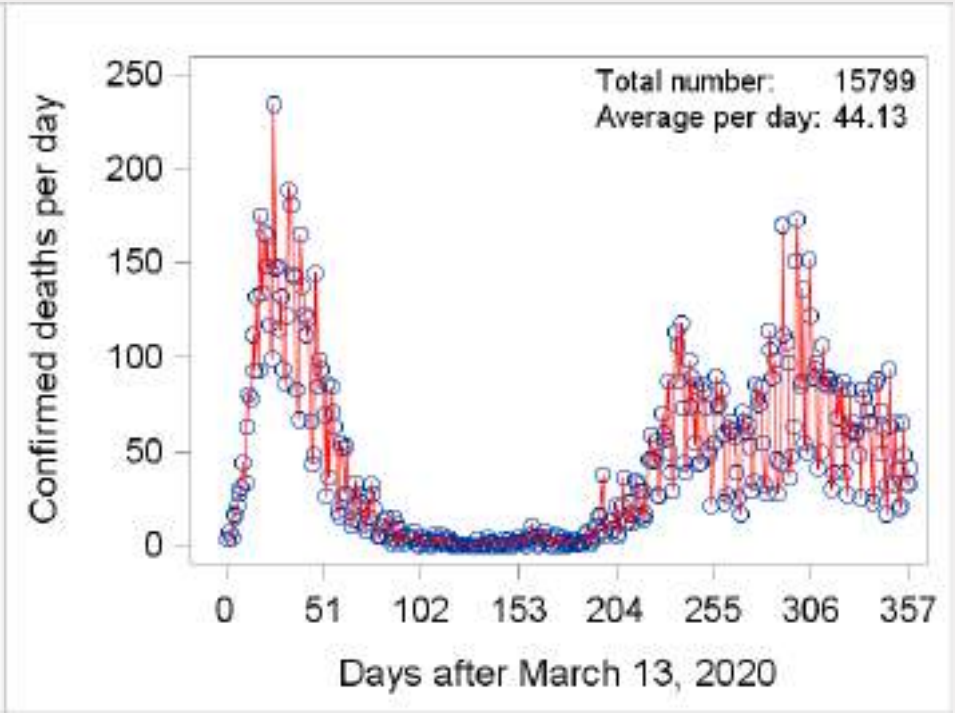


TU/e

Epidemic Disease Models

Differential Equations





TU/e

Epidemic Disease Models

Differential Equations

Solution differential equation:

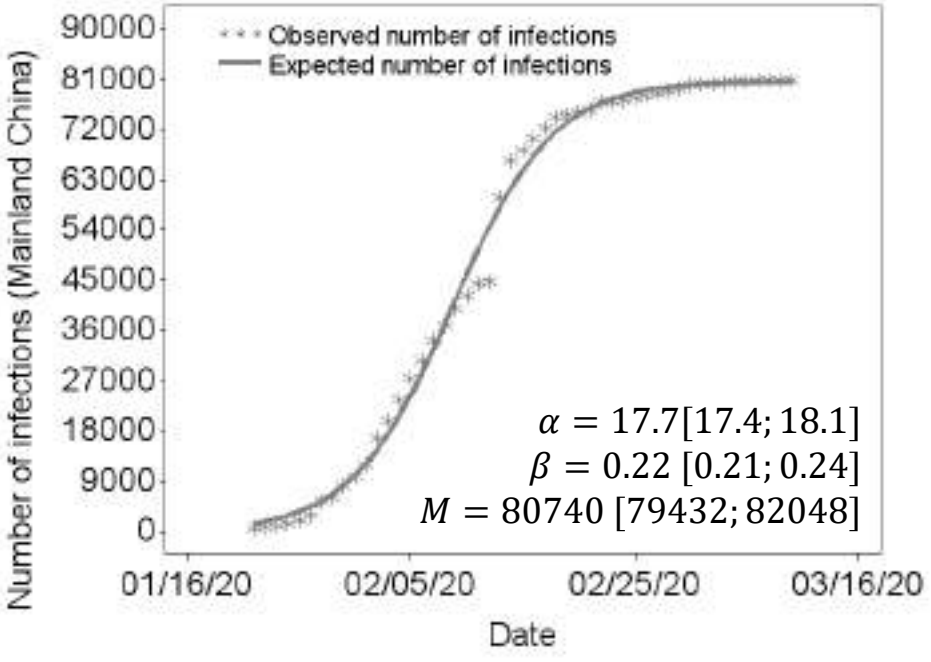
$$\mathbb{E}(I_t) = M/[1 + \exp(-\beta(t - \alpha))]$$

- I_t: confirmed number of infections
- M: expected maximum number
- β: growth rate
- α: turning point
- t: time determined in days

- Parameter α is implicit

$$\alpha = \beta^{-1}\log(M \cdot [I_0]^{-1} - 1)$$

  - I_0 number of confirmed infections at the start of data



TU/e

Epidemic Disease Models

Different analysis approaches

Analysis of accumulated data

- Non-linear regression with normal errors

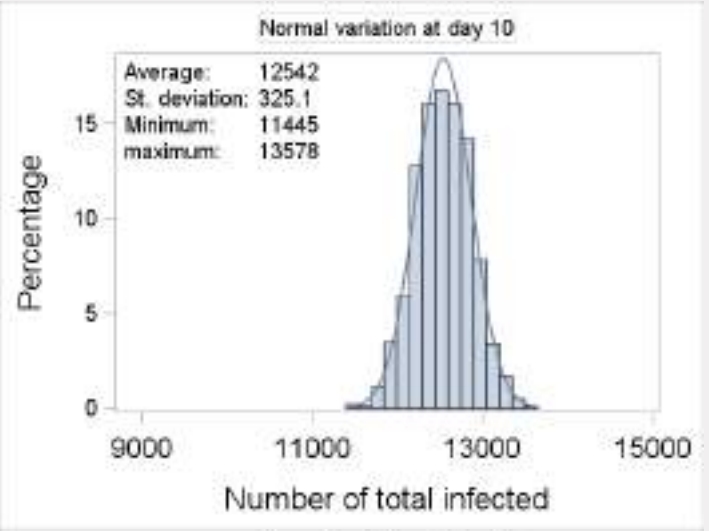
$$I_t = \frac{M}{1+\exp(-\beta(t-\alpha))} + e_t$$

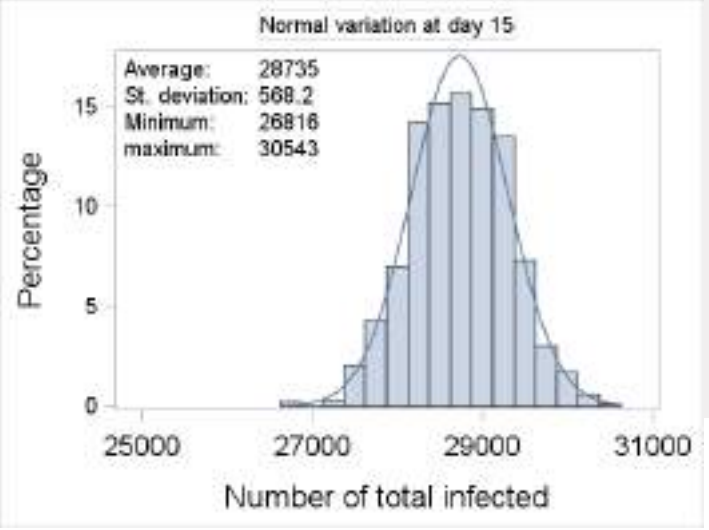
  - With $e_t \sim N(0, \sigma^2)$ i.i.d.
  - **All three** parameters are estimated
- Model adjustments:
  - Heteroscedastic error structures

$$\sigma_t^2 = \mathbb{V}(e_t) = \sigma^2 F_{\alpha,\beta}(t)[1 - F_{\alpha,\beta}(t)]$$

  - With $F_{\alpha,\beta}(t) = [1 + \exp(-\beta(t-\alpha))]^{-1}$
  - Autoregressive error structure

$$\mathrm{CORR}(e_t, e_{t-1}) = \rho$$



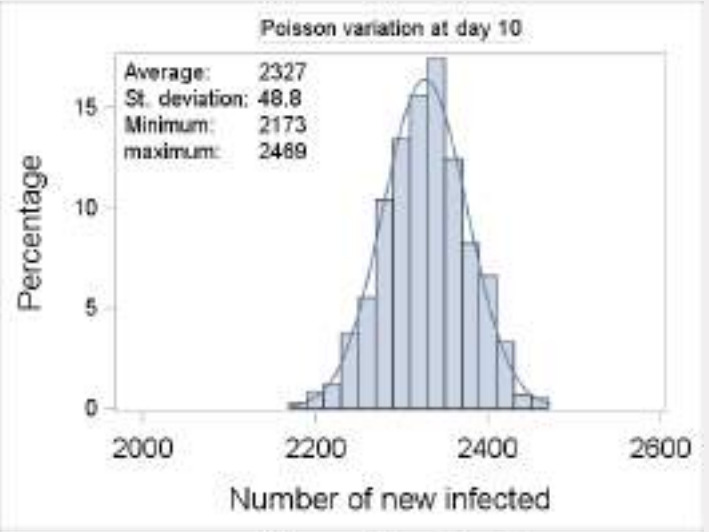


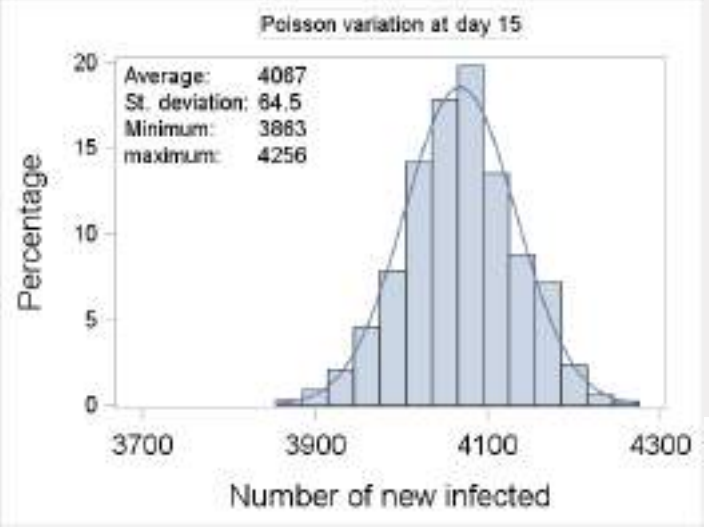
Epidemic Disease Models

Different analysis approaches

Analysis daily new infections

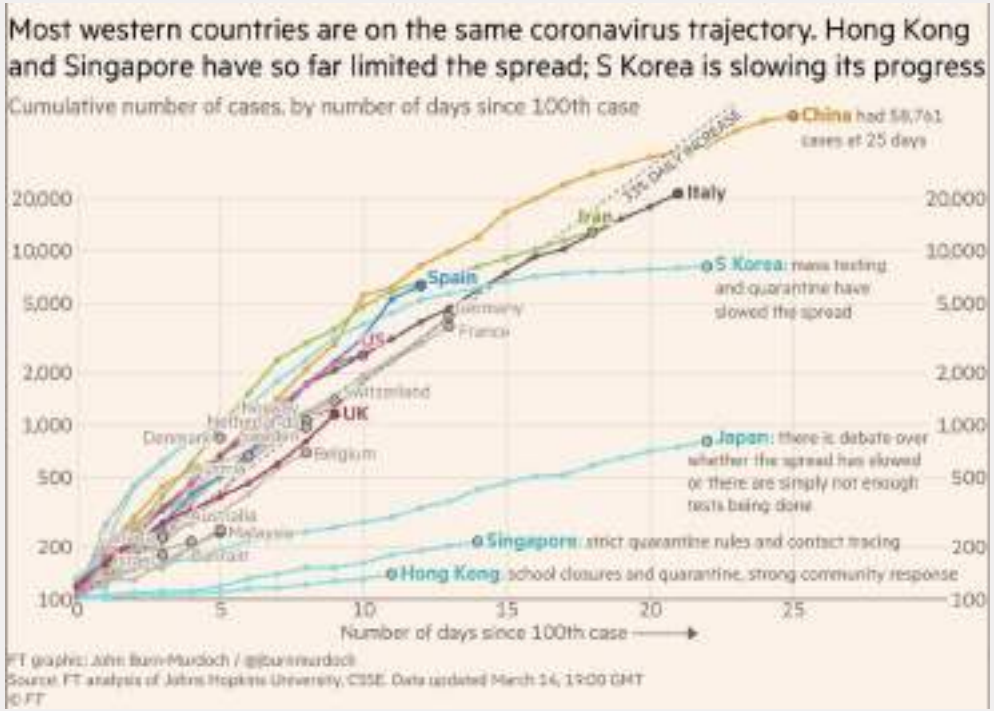
- Non-linear Poisson regression for daily new infections $\Delta I(t+1)$
  - $I(t+1) - I(t) | I(t) \sim \text{Poisson}(\lambda_t)$
  - $\lambda_t = \beta I(t)[1 - I(t)/M]$
  - Requires only estimates for β and M
  - Can be conducted with maximum likelihood estimation
- Assumes β is constant over time
- M is difficult to estimate when the full curve is not available





Epidemic Disease Models

Comparison of curves (cumulative)



Exponential growth:

$$\log\left(\mathbb{E}\left(I(t)\right)\right) = \log(M) + \log\left(F_{\alpha,\beta}(t)\right)$$

- With $F_{\alpha,\beta}$ logistic distribution
- Not a linear function in t

- Differences in log scale are more difficult to see
- With limited data curves all start out similar
- Start at 100 events is arbitrary

TU/e

Epidemic Disease Models

Comparison of curves (cumulative)

Pairwise comparisons:

- Data up to March 25, 2020
- Starting point ≥ 100 infections
- Likelihood ratio test $H_0: \beta_{NL} = \beta_O$
  - Other parameters are country specific

Vergelijk	LRT	P-value	Better
Netherlands vs. Italy	17.3	<0.001	Italy
Netherlands vs. Belgium	5.17	0.023	Netherlands
Netherlands vs. Germany	31.9	<0.001	Netherlands
Netherlands vs. England	8.37	0.004	Netherlands
Netherlands vs. Sweden	2.01	0.156	NA
Netherlands vs. China	0.04	0.841	NA
Netherlands vs. South Korea	28.3	<0.001	Netherlands

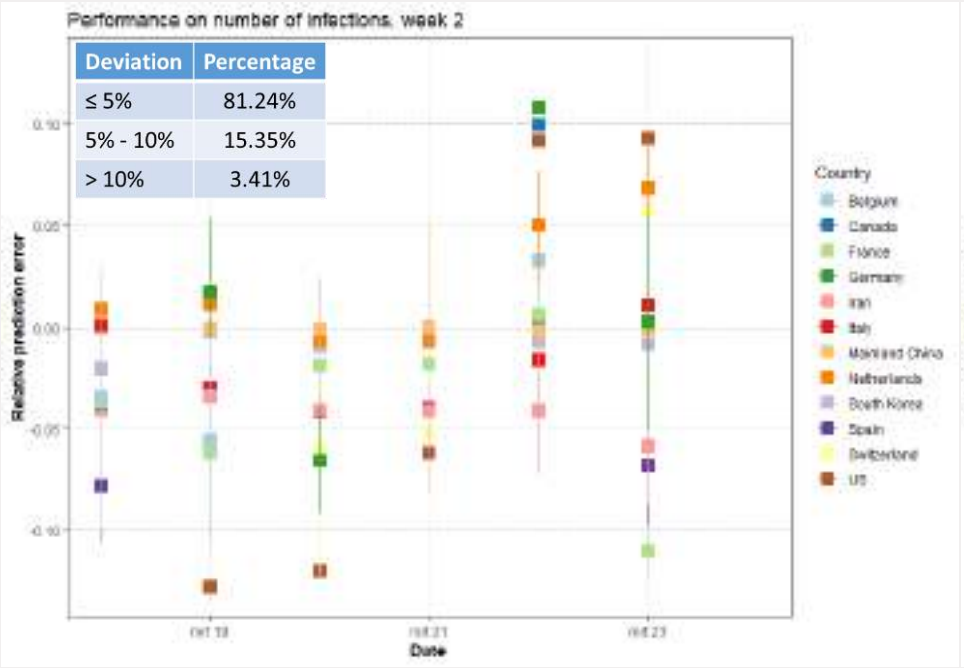
Comparisons problematic:

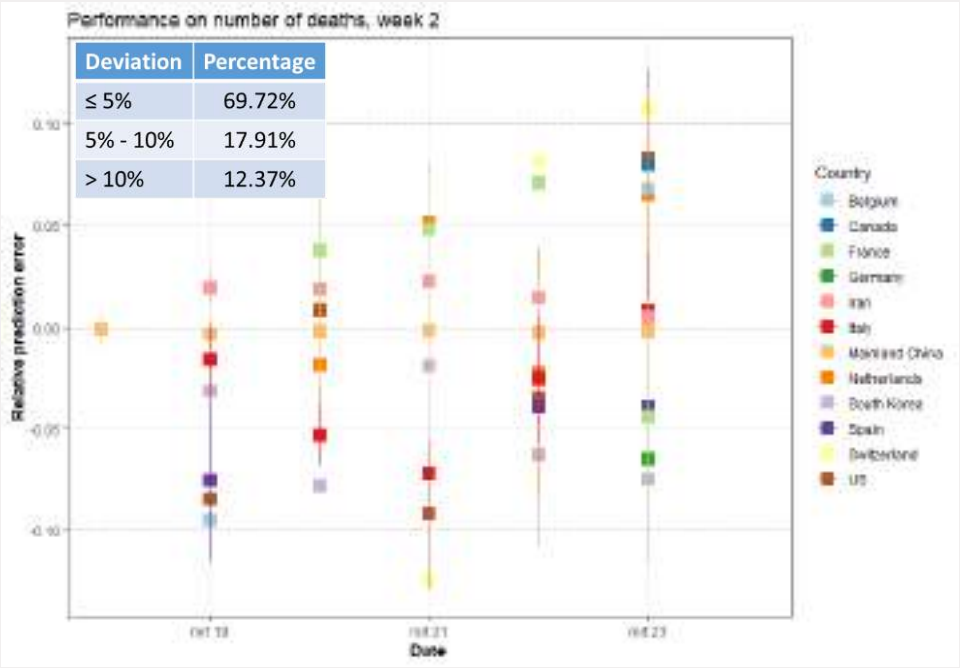
- Countries are not evolving parallel
  - Bias in parameter estimates
  - Influences of governmental measures affects growth rate
- Data related problems
  - Netherlands tested less than other countries like Italy and Spain
  - Test policy changes over time
- Starting point has strong influence
  - Starting at first death, then Netherlands vs. Sweden: p<0.001

TU/e

Epidemic Disease Models

Predictions of infections and deaths (cumulative analysis)



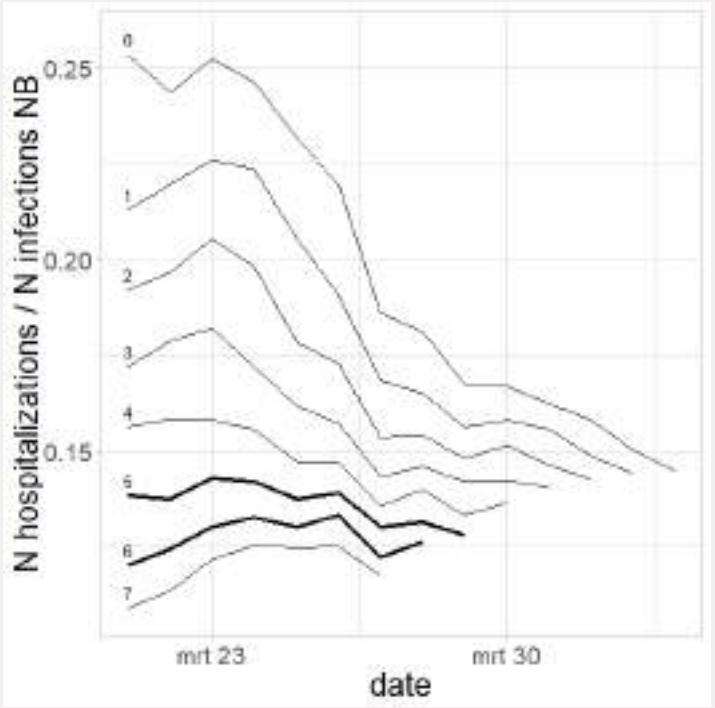


Verhulst logistic growth models

Predictions of infections and deaths

Predictions hospitalizations:

- Ratio of number of hospitalizations and number of infections (see Figure)
  - Calculated at different time lags
  - Calculated for North Brabant
- Ratio is almost constant at 5 or 6 days
  - At 5 days: 13.6%
  - At 6 days: 12.7%
- Our predictions were used to plan
  - Capacity for number of hospitalizations
  - Capacity for the number of ICU's



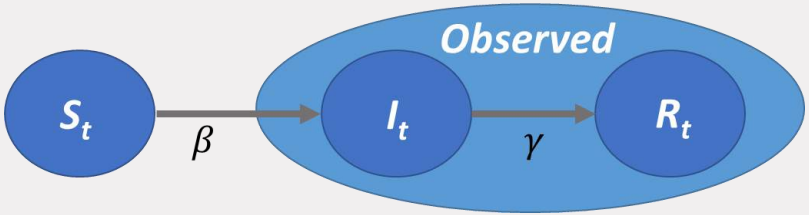
TU/e

Epidemic Disease Models

Generalized logistic curves

Data is more complicated:

- Suceptible-Infected-Removed model



- S_t: number of susceptible individuals
- I_t: number of infected individuals
- R_t: number of removed individuals
- $M = S_t + I_t + R_t$ total population
- β: daily number of effective contacts
- γ: daily removal rate

- Differential equations:

$$\frac{dI_t}{dt} = \beta \frac{I_t S_t}{M} - \gamma I_t$$

$$\frac{dR_t}{dt} = \gamma I_t$$

- We only observe $Y(t) = I_t + R_t$

$$\frac{dY(t)}{dt} = \beta(Y(t) - R_t)\left(1 - \frac{Y(t)}{M}\right)$$

- Basic reproduction number:

$$R_0 = \beta/\gamma$$

  - $R_0 > 1$: virus spreads among population
  - $R_0 = 1$: virus stabilizes
  - $R_0 < 1$: virus dies out

TU/e

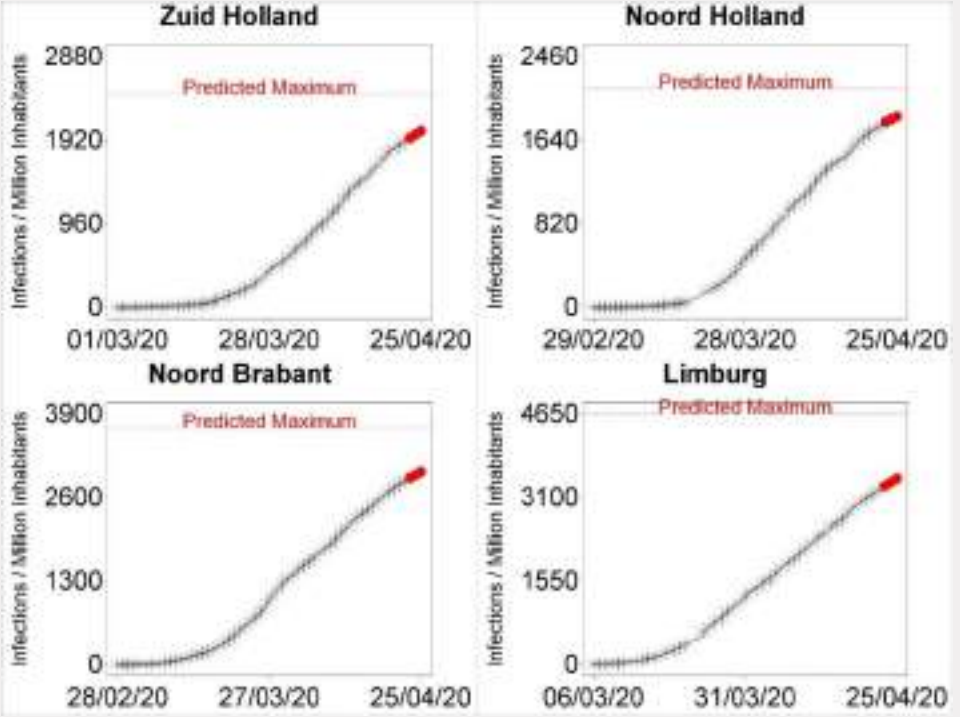
Epidemic Disease Models

Generalized logistic curves

- Flattening in European countries did not follow Verhulst growth model
- Generalized logistic curves

$$\frac{d}{dt}\mathbb{E}[Y(t)] = \beta[Y(t)]^{\gamma}\left[1-\left(\frac{Y(t)}{M}\right)^{\delta}\right]^{\eta}$$

  - Turning point not so easy
- Non-linear Poisson regression
  - We selected $\delta = 1$ (numerical issues)
  - Estimated η when necessary
  - $\Delta Y(t+1)|Y(t) \sim \text{Poisson}(\lambda_t)$
  - $\lambda_t = \beta[Y(t)]^{\gamma}\,[1 - Y(t)/M]^{\eta}$



TU/e

Epidemic Disease Models

Generalized logistic curves

Estimates of parameters

- Data up to April 30, 2020

Country	γ	$\log(\beta)$
BE	0.736 [0.726; 0.747]	0.479 [0.384; 0.574]
CA	0.715 [0.705; 0.725]	0.546 [0.459; 0.634]
DK	0.579 [0.550; 0.608]	1.017 [0.805; 1.229]
FR	0.782 [0.776; 0.788]	0.349 [0.289; 0.409]
DE	0.720 [0.715; 0.724]	1.101 [1.052; 1.150]
IR	0.676 [0.668; 0.685]	1.038 [0.953; 1.123]
IT	0.651 [0.647; 0.656]	1.632 [1.582; 1.682]
NL	0.715 [0.702; 0.727]	0.586 [0.475; 0.696]
KR	0.609 [0.597; 0.622]	1.363 [1.259; 1.467]
SE	0.684 [0.664; 0.704]	0.286 [0.128; 0.444]
US	0.664 [0.662; 0.666]	2.133 [2.107; 2.158]
UK	0.746 [0.741; 0.752]	0.635 [0.576; 0.693]

Predictions performance

infections

Deviation	Percentage
≤ 5%	99.65%
5% - 10%	0.35%
> 10%	0%

Deaths

Deviation	Percentage
≤ 5%	99.65%
5% - 10%	0.35%
> 10%	0%

- Better performance than Verhulst
  - In particular for the number of deaths
  - Although accumulated data is also larger
  - Improved model shows similar bias issues with parameter estimation

TU/e

Governmental interventions

Discrete Susceptible-Exposed-Infected-Removed Model

Generalized logistic curves

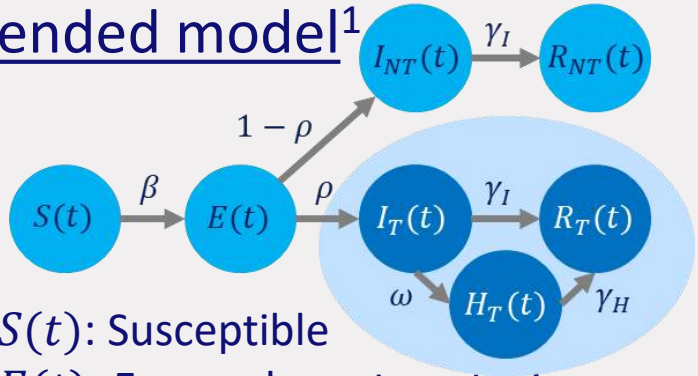
- Assume constant contact rate β, but governments try to influence this rate
- Useful for short-term predictions, but long-term predictions more difficult
- Can include additional factors in model, but does not address data issues
  - Observed data does not include asymptomatic patients
  - Delay in infection
  - Changing testing policies affect numbers

TU/e

Governmental interventions

Discrete Susceptible-Exposed-Infected-Removed Model

Extended model[1]



- $S(t)$: Susceptible
- $E(t)$: Exposed – not contagious
- $I(t) = I_{NT}(t) + I_T(t)$: Infectious
- $H_T(t)$: Hospitalized and tested
- $R(t) = R_T(t) + R_{NT}(t)$: removed
- Weibull (2.32; 6.5) incubation time
- Exponential (2.3) infectious period

- We observe the confirmed number of total infected:

$$Y(t) = I_T(t) + H_T(t) + R_T(t)$$

- Poisson regression on $\Delta Y(t)$

$$\mathbb{E}[\Delta Y(t)] = \sum_{i=0}^{t} \rho \Delta E^+(t) p_{t-i}$$
$$= \sum_{i=1}^{t} \rho \beta(t) I(t) \frac{S(t)}{M} p_{t-i}$$

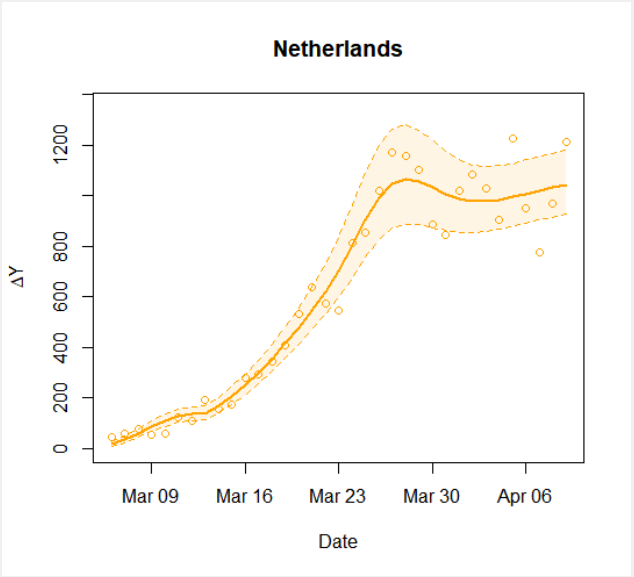
  - $E^+(t) = E(t) + Y(t) + I_{NT}(t) + R_{NT}(t)$: cumulative exposed individuals
  - $\beta(t)$: time dependent contact rate
  - M: total population
  - $S(t)$ and $I(t)$ are iteratively solved
  - Notation: $\Delta U(t) = U(t) - U(t-1)$

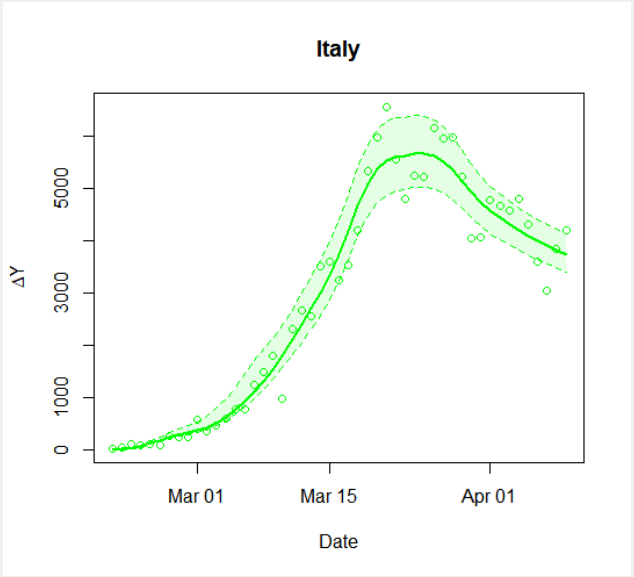
[1] Post RA, Regis M, Zhan Z, van den Heuvel ER. How did governmental interventions affect the spread of COVID-19 in European countries?, *BMC public health*, 2021, **21**(1):1-1.

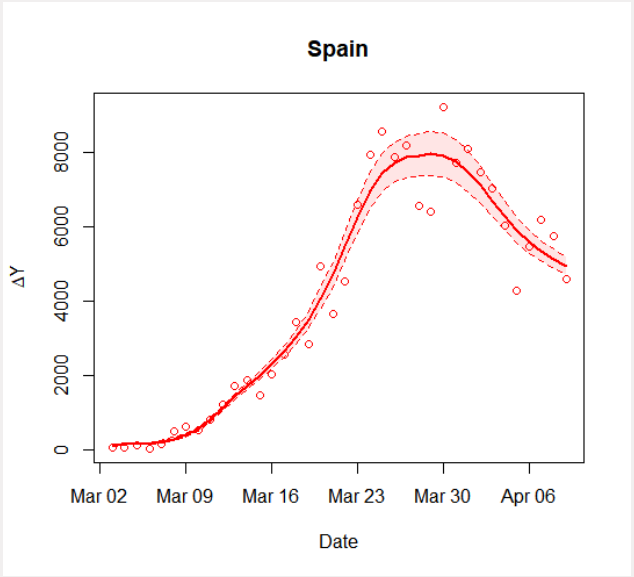
TU/e

Governmental interventions

Goodness-of-Fit



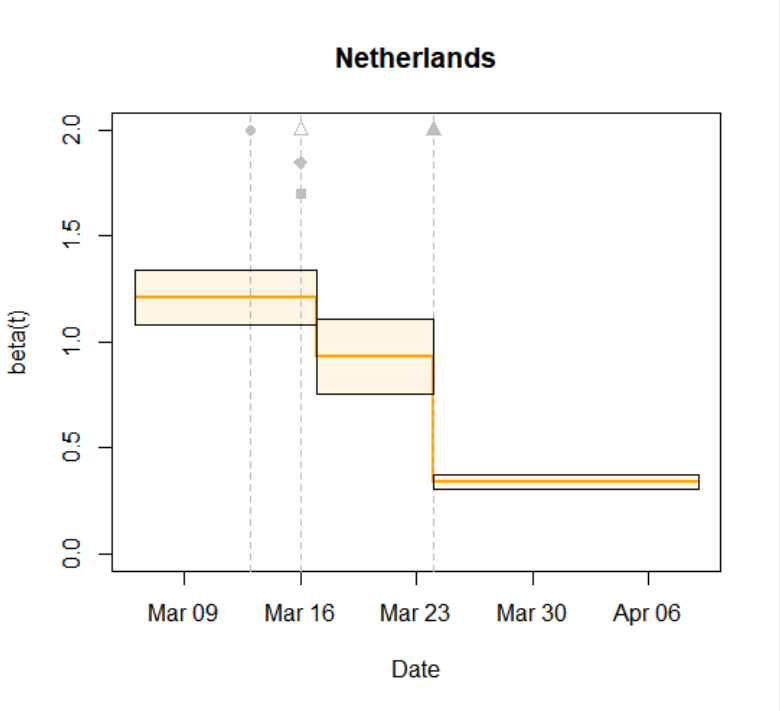




Governmental interventions

Daily effective contact-rate profile

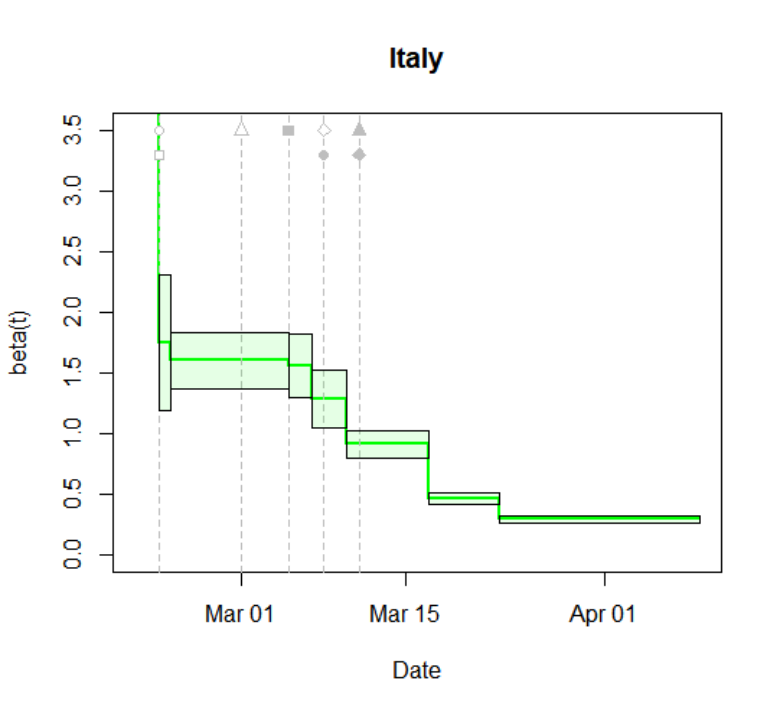
- Determine data-driven change points in daily contact-rates $\beta(t)$
- Restriction $\beta(t) = \sum_{i=t}^{n} \exp(\beta_i)$
  - Data is very noisy – outlier estimates
  - Outlier estimates affect all future estimates quite strongly
- Compare change-points in $\beta(t)$ with governmental measures
  - Lockdown (▲)
  - Closing schools (■)
  - Closing restaurants (♦)
  - Banning events (●)

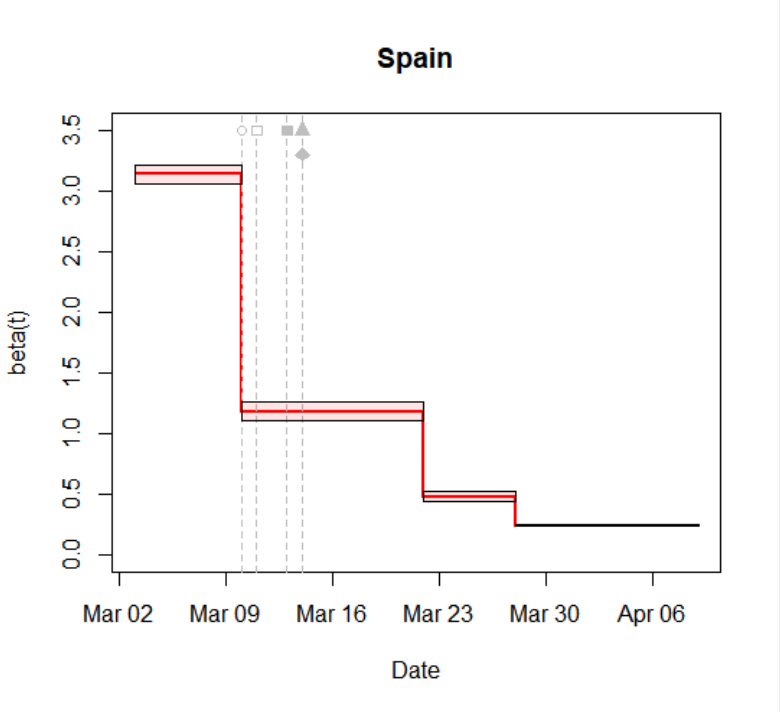


TU/e

Governmental interventions

Daily effective contact-rate profile





Country	Date	Measures	Growth-rate change	Relative change	Absolute change
IT	23-2	closure of schools + banning events in red zone	$9.14 \rightarrow 1.60$	0.83	7.54
	7-3	lockdown in red zone	$1.56 \rightarrow 1.28$	0.18	0.28
	10-3	full lockdown	$1.28 \rightarrow 0.91$	0.29	0.37
	17-3	none (enforced police force)	$0.91 \rightarrow 0.46$	0.49	0.45
ES	10-3	closure of schools + banning events in red zone	$3.14 \rightarrow 1.18$	0.62	1.96
	22-3	none (enforced police force)	$1.18 \rightarrow 0.48$	0.59	0.70
DE	15-3	closure of schools + event banning (+ railway reduction)	$2.02 \rightarrow 0.68$	0.66	1.34
	30-3	lockdown	$0.68 \rightarrow 0.31$	0.54	0.37
UK	15-3	none (increased to high risk level)	$1.71 \rightarrow 1.08$	0.37	0.63
	24-3	Lockdown	$1.08 \rightarrow 0.74$	0.32	0.34
	31-3	none (information to public)	$0.74 \rightarrow 0.37$	0.50	0.37
NL	17-3	closure of schools + restaurants (+ request to stay inside)	$1.21 \rightarrow 0.93$	0.23	0.28
	24-3	Lockdown (fines + enforced police force)	$0.93 \rightarrow 0.34$	0.63	0.59
BE	13-3	closure of schools + banning events	$1.44 \rightarrow 1.30$	0.10	0.14
	25-3	none (obligated quarantine for flight passengers + lockdown extended)	$1.30 \rightarrow 0.34$	0.74	0.96
SE	12-3	banning events + warnings to public	$0.88 \rightarrow 0.57$	0.35	0.21
	29-3	stricter measures banning events	$0.57 \rightarrow 0.45$	0.21	0.12

Governmental interventions

Daily effective contact-rate profile

Conclusions:

- Closing of schools/banning events seem to have direct effect
  - We do observe a combined effect
  - Effect sizes are heterogeneous across countries
- Lockdown not always direct effect
  - In some countries it needed police enforcement before changing profile
- Closing of restaurants did not show a clear effect

Parameter estimates:

	ρ	β_{start}	β_{end}
IT	0.454 (0.104)	9.031 (0.333)	0.290 (0.010)
ES	0.365 (0.047)	3.266 (0.038)	0.240 (0.004)
DE	0.780 (0.432)	2.017 (0.068)	0.271 (0.016)
UK	0.043 (0.030)	1.688 (0.037)	0.370 (0.005)
NL	0.207 (0.123)	1.212 (0.067)	0.340 (0.019)
BE	0.369 (0.194)	1.440 (0.057)	0.340 (0.012)
SE	0.028 (0.010)	0.880 (0.032)	0.452 (0.011)

- Variability at start is large
- Countries converge to same rate
  - Group DE, ES, IT: ≈ 0.26
  - Group BE, NL, UK: ≈ 0.35
  - Group SE: ≈ 0.45

TU/e

Data Science within a pandemic

Why we need to change statistical inference

Traditional Approach Fails:

- Statistical model is central
  - Synergy between domain knowledge and statistical model
  - Synergy between data and model
  - Parameter estimates are directly interpretable to population
  - Limited sensitivity analysis and discussion on model weaknesses


Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Data Oriented Approach:

- Statistical models are used to understand data – feature selection
  - Suitability and characteristics of models are being evaluated
  - Multiple data sets for verification – heterogeneity versus robustness
  - Simulations are used to understand approach under well-known conditions
  - Data is used to investigate sensitivity
- Thinking in line with Leo Breimann

TU/e