Analyse der Coronastatistiken

Hans-Gert Gräbe, Leipzig

Version vom 1. April 2020

Dieser Text bezieht sich auf die im Verzeichnis `http://leipzig-data.de/demo/Corona-20/Code` verfügbaren Materialien.

1 Datenbasis

Als Datenbasis werden die von der John Hopkins Universität (JHU) als Excel-Datei veröffentlichten Daten[1] zur Entwicklung der weltweit registrierten Covid-19-Fälle (Stand 28.03.2020) verwendet. Die Statistik listet pro Land und Tag die kumulierte Zahl der Infizierten, die Zahl der Genesenen und die Zahl der Todesfälle auf.

2 Installation

Zunächst muss das git Repo der JHU lokal geklont und der Pfad im Skript `extractData.pl` eingetragen werden. Die Daten werden für ausgewählte Länder mit diesem Perl-Skript für die weitere Verarbeitung aufbereitet und in einer Datei `BasicData.txt` gespeichert, um dann mit dem freien CAS *Maxima*[2] weiterverarbeitet zu werden.

3 Datentransformation

In der Datei `BasicData.txt` sind die Daten für jedes der ausgewählten Länder in einem Array mit drei Einträgen (infected, recovered, dead) gespeichert, die im Maxima-Skript `skript.m`[3] in einer Funktion `getland(Land)` zunächst einmal zu Paaren (t, y_t) ergänzt werden, wobei t für den Tag des Jahres ($0 = 01.01.2020$) und y_t für die Zahl der Fälle aus dem jeweiligen Record stehen.

[1] Siehe deren github-Projekt `https://github.com/CSSEGISandData/COVID-19`.

[2] Siehe dazu `http://maxima.sourceforge.net/de/`, das CAS ist in Linux-Distributionen über den Paketmanager leicht zu installieren.

[3] Dies ist eine reine Textdatei mit Code-Schnipseln, die nicht für den Batchbetrieb konzipiert ist.

4 Fitting

Alle Grafiken zu Prognosen der Daten, die ich bisher gesehen habe, gehen von einer „Glockenkurve“ aus. Das kann natürlich nur die Entwicklung der Zuwächse pro Tag abbilden, die aus den kumulierten Daten zunächst als $d_t = y_t - y_{t-1}$ extrahiert werden müssen. Dies geschieht mit der im Skript definierten Maxima-Funktion `Delta`.

Zur Abschätzung des Verlaufs längs einer Glockenkurve wird üblicherweise die Statistikfunktion $C \cdot \exp\left(-\left(\frac{t-m}{s}\right)^2\right)$ verwendet, die ich als Kurvenschar $f(t) = \exp\left(c - \left(\frac{t-m}{s}\right)^2\right)$ zur Parameterschätzung auf die Daten (t, d_t) ansetze. Die (kumulierten) Originaldaten (t, y_t) sollten dann auf die Funktion

$$\begin{aligned} h(x) &= \int_0^x \exp\left(c - \left(\frac{t-m}{s}\right)^2\right) dt \\ &= \exp(c) \cdot s \cdot \int_{-\frac{m}{s}}^{\frac{x-m}{s}} \exp\left(-u^2\right) du \\ &= \frac{1}{2}\sqrt{\pi} \cdot \exp(c) \cdot s \cdot \left(\operatorname{erf}\left(\frac{x-m}{s}\right) + \operatorname{erf}\left(\frac{m}{s}\right)\right) \end{aligned}$$

matchen, wobei $\operatorname{erf}(x)$ für die Fehlerfunktion steht und im zweiten Schritt die Variablensubstitution $u = \frac{x-m}{s}$ mit $du = s \cdot dt$ erfolgte.

Nun sind die Parameter (c, s, m) dieser Kurvenschar so zu fitten, dass die ermittelte Kurve besonders gut auf die Daten passt. In *Maxima* kann dazu das Paket *lsquares* verwendet werden.

Wir hätten natürlich auch versuchen können, die Originaldaten auf die Schar $h(t)$ zu fitten, aber Fitting auf nicht polynomialen Kurvenscharen ist eine schwierige und numerisch wenig stabile Angelegenheit, bei der Maxima schnell an seine Grenzen kommt (und die Ergebnisse anderer CAS sehr genau zu analysieren sind, da die Fitting-Ergebnisse stark von Startwerten der dabei eingesetzten Verfahren abhängen).

Maxima kommt auch beim Fittung der Schar $f(t)$ zu keinem Ergebnis. Einen einfacheren, nämlich quadratischen Zusammenhang $g(t) = c - \left(\frac{t-m}{s}\right)^2$ erhält man, wenn man zu Paaren $(t, \log(d_t))$ übergeht. Damit lassen sich dann die Fittingparameter stabil berechnen. Dafür müssen aber vorher Datenpunkte aussortiert werden, wo $d_t = 0$ ist.

Genrell kann es sinnvoll sein, für ein gutes Fitting Datenpunkte unterhalb einer Schwelle auszusortieren. Eine solche Schwelle ist als weiterer Parameter im Skript in der Funktion `FittingDelta` vorgesehen. Für die meisten Datensätze ist die Schwelle 50 eine gute Wahl[4].

Details sind im Skript `skript.m` zu finden.

Weitere Versuche, etwa mit einer Schar von Logistik-Funktionen, lassen sich mit dem CAS Maxima nicht erfolgreich zu Ende bringen.

[4]Es ist zu beachten, dass die Schwelle die Dateninkremente d_t auswertet, nicht die kumulierten Daten y_t. Ein kleines Land wie Österreich bereitet hier besondere Schwierigkeiten.

5 Ergebnisse

Die Rechnungen werden für jedes der Länder nun wie folgt ausgeführt:

1. Fasse mit `getData` die drei Datensätze für das Land als Tripel von Listen (t, y_t) zusammen.
2. Berechne für jeden der drei Datensätze mit `getFittingFunctions` das Fitting auf $(t, \log(d_t))$ gegen die Funktion $g(t)$ und verwende das gefundene Fitting, um Funktionen $h_1(t)$ (für infected), $h_2(t)$ (für recovered) und $h_3(t)$ (für dead) zu schätzen.
3. Erzeuge daraus einen Plot, welcher die Datenpunkte und die drei Kurven in verschiedenen Farben (rot für infected, grün für recovered, blau für dead) ausgibt.

Bei erfolgreichem Fitting ist eine gute Übereinstimmung der jeweiligen Kurve

$$h(t) = A\left(\operatorname{erf}\left(B(t-m)+1\right)\right)$$

mit den Datenpunkten zu verzeichnen. Die berechneten Parameter haben folgende Bedeutung:

- m – Tag, an dem die Spitze in den Inkrementdaten erreicht ist.
- $B = \frac{1}{s}$ – $[m-s \ldots m+s]$ ist das kritische Intervall.
- $2 * A$ – Zahl der am Ende insgesamt betroffenen Personen.

Auf der Basis der Daten vom 29.03.2020 lassen sich folgende Szenarien (!) für die Länder Deutschland, Italien, Spanien und Österreich erstellen, siehe auch die Abbildungen.

	m	s	$2A$
Deutschland			
infected	94	14	171274
recovered	90	5	20730
dead	95	10	2624
Italien			
infected	85	15	144893
recovered	94	18	39909
dead	90	15	22893
Spanien			
infected	89	12	156889
recovered	bad fitting		
dead	95	12	23378
Österreich			
infected	88	11	13747
recovered	bad fitting		
dead	bad fitting		

Natürlich müsste am Ende die Zahl der Infizierten mit der Summe der Zahlen der Genesenen und der Verstorbenen übereinstimmen. Die Schätzungen sind von einer solchen Invarianzforderung weit entfernt, was ein Licht auf die prognostische Qualität der Schätzungen für die weitere Zukunft (drei bis vier Wochen, also 30 Tage) wirft.

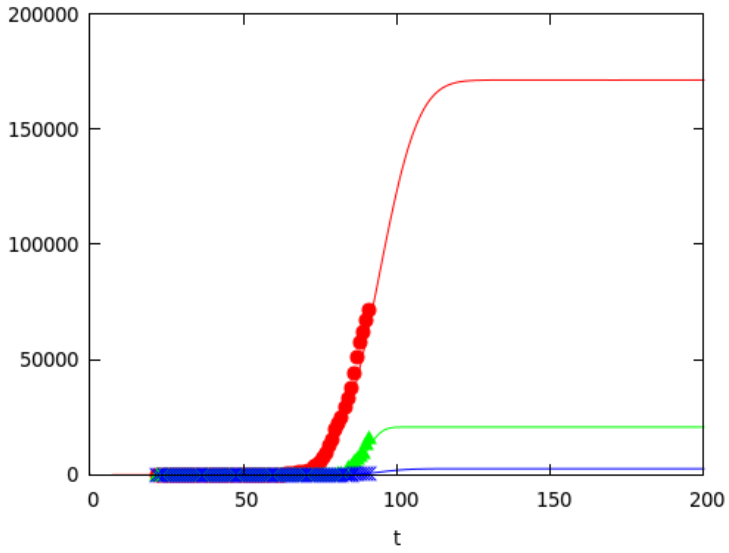

Abbildung 1: Szenario für Deutschland (82.9 Mio. Einwohner)

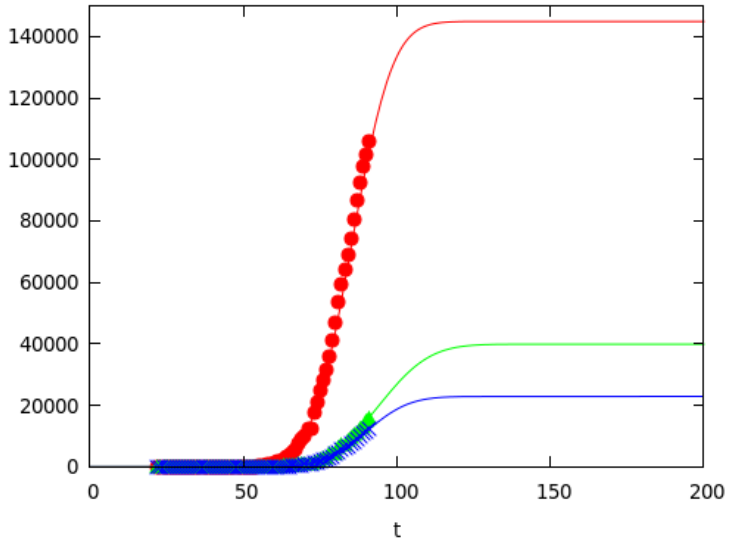

Abbildung 2: Szenario für Italien (60.4 Mio. Einwohner)

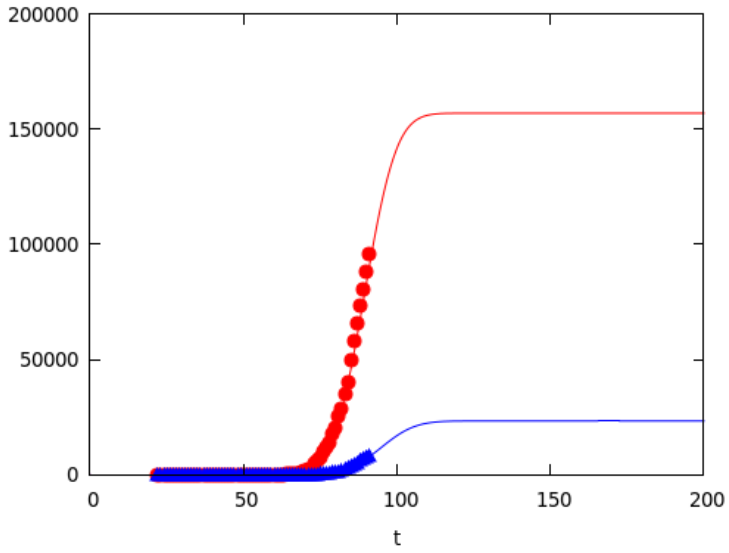

Abbildung 3: Szenario für Spanien (46.7 Mio. Einwohner)

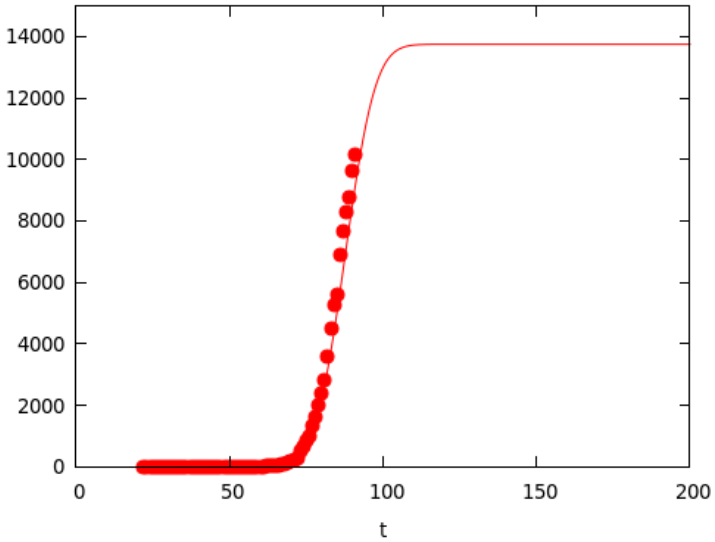

Abbildung 4: Szenario für Österreich (8.85 Mio. Einwohner)