Analyse der Epidemischen Covid-19 Kurve in Bayern durch Regressionsmodelle mit Bruchpunkten

Helmut Küchenhoff[1], Felix Günther[1,2], Andreas Bender[1], Michael Höhle[3]
[1] Statistisches Beratungslabor StaBLab, LMU München, Deutschland
[2]Lehrstuhl für Genetische Epidemiologie, Universität Regensburg, Deutschland
[3]Institut für Mathematik, Universität Stockholm, Schweden
E–Mail: kuechenhoff@stat.uni-muenchen.de
URL: `corona.stat.uni-muenchen.de`

Neue Fassung mit den Erkrankungsdaten bis zum 1.5.

7. Mai 2020

Zusammenfassung

Die Diskussion um den Verlauf der COVID–19 Epidemie und die Wirkung von Maßnahmen zur Eindämmung der Ausbreitung wird aktuell kontrovers geführt. Häufig wird dazu die Reproduktionszahl $R(t)$ herangezogen. Für diese existieren jedoch unterschiedliche Definitionen und Interpretationen, was zu Missverständnissen führen kann. Wir plädieren daher für eine direkte Analyse des Verlaufs der Anzahl der täglichen Neuerkrankungen. Dazu verwenden wir neben der grafischen Darstellung ein Regressionsmodell mit Bruchpunkten (engl. changepoints), das den Epidemieverlauf datengesteuert anhand der Infektionsschutzgesetz-Meldedaten in verschiedene Phasen unterteilt. Für Bayern schätzen wir unter Berücksichtigung der Inkubationszeit von ca. fünf Tagen vier Bruchpunkte: 5.3., 11.3., 18.3. und 28.3.2020. Besonders ausgeprägt ist der Bruch am 11.3. (in der Graphik ersichtlich am 16.3), wo ein Übergang von einem Wachstum zu einem leichten Rückgang zu sehen ist. Dieser ist assoziiert mit der Rede der Bundeskanzlerin, in der sie erstmals zur Meidung von Sozialkontakten aufrief, und Berichten zu den Zuständen in Bergamo. Eine weitere Veränderung ergibt sich um den 18.3. (entspricht dem Krankheitsbeginn am 23.3.), von dem an wir einen Rückgang der Infektionszahlen finden. Dieser Rückgang ist assoziiert mit verschiedenen Maßnahmen, insbesondere mit den Ausgangbeschränkungen. Auch wenn aufgrund des komplexen Geschehens keine kausalen, detaillierten Schlüsse gezogen werden können, liefert die Analyse wichtige Einblicke in das Infektionsgeschehen. Der geschätzte Verlauf der Anzahl der bestätigten Neuerkrankungen ist aus unserer Sicht der einfachen Analyse der gemeldeten Fälle und der Analyse des Verlaufs der zeit–variierenden Reproduktionszahl $R(t)$ vorzuziehen. Da bei unserer Analyse die nicht gemeldeten Fälle nicht berücksichtigt werden, sollten zukünftige Analysen um weitere Informationen, wie Zahlen zu Todesfällen und Aufnahmen ins Krankenhaus, Testverhalten und repräsentativen Studien ergänzt werden.

1 Einführung

Die Diskussion über die Interpretation des Verlaufs der COVID–19 Epidemie wird kontrovers geführt. Dies betrifft insbesondere die wichtige Frage der Wirksamkeit von verschiedenen Maßnahmen zur Eindämmung der Epidemie. Herausfordernd ist dabei die Verwendung der zeit–variierenden Reproduktionszahl $R(t)$. Diese ist eine wichtige Maßzahl zur Charakterisierung des Verlaufs der Epidemie. Es gibt in der Literatur verschiedene Ansätze zur Berechnung von $R(t)$. Zu beachten ist dabei, dass die unterschiedlichen Ansätze auch unterschiedliche Interpretationen implizieren (siehe Cori et al. (2013); Lipsitch et al. (2020)).

Hier möchten wir einen einfachen Ansatz vorstellen, um den Zeitverlauf der Anzahl der täglichen Neuerkrankungen zu analysieren. Diese Kurve kann per sog. *nowcast* (Höhle and an der Heiden, 2014) aus den Meldedaten geschätzt werden und wird vom Robert Koch-Institut für Deutschland (`https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt.html`) und von uns für Bayern und München (`https://corona.stat.uni-muenchen.de/nowcast/`) aktualisiert zur Verfügung gestellt. Um genauere Einblicke in den Verlauf in Bayern zu bekommen, analysieren wir die Kurve mit einem Regressionsmodell mit Bruchpunkten (changepoints). Die Bruchpunkte teilen den Verlauf der Epidemie, wie sie anhand der Meldedaten abgebildet wird, in mehrere Phasen ein. Die Bruchpunkte werden anhand der Daten bestimmt und können, *unter Berücksichtigung der Inkubationszeit*, als Zeitpunkte der Veränderung des Infektionsgeschehens interpretiert werden und Hinweise auf die Bewertung von Maßnahmen zur Eindämmung der Epidemie liefern.

2 Methoden

Basis sind die Covid-19 Meldedaten des Bayerischen Landesamtes für Gesundheit und Lebensmittelsicherheit (LGL), welche im Rahmen des Infektionsschutz-Gesetzes (IfSG) erhoben werden. Diese enthalten auf Fallebene neben dem Meldedatum auch den Zeitpunkt des Krankheitsbeginns (i.S.v. erstes Auftreten von Symptomen). Der Zeitpunkt ist jedoch nicht immer bekannt: teils weil er nicht ermittelt werden konnte, Teils weil der Fall zum Zeitpunkt der Erfassung (noch) keine Symptome hatte. Es besteht ferner eine zeitliche Verzögerung, bis die Informationen über den Erkrankungsbeginn beim LGL eingehen. Zunächst muss der Patient einen Arzt aufsuchen, eine Labortestung muss vorgenommen werden, anschließend wird das Testergebnis vom Labor über das Gesundheitsamt an das LGL geschickt. Aufgrund dieses Meldeverzugs sind die aktuellen Daten bzgl. Erkrankungsbeginn nur unvollständig erhoben (Rechts-Trunkierung). Daher verwenden wir ein sog. *nowcasting*-Verfahren (Höhle and an der Heiden, 2014), welches für einen aktuellen Datenstand eine Schätzung der aktuellen Anzahl von Neuerkrankungen liefert (Günther et al., 2020). Das Verfahren beinhaltet auch die Imputation von fehlenden Werten bezüglich des Krankheitsbeginns. Hierzu wird ein Modell für die Verzögerungszeit angepasst. Dabei wird angenommen, dass die Verteilung der Werte des Krankheitsbeginns sich bei den Personen mit fehlenden Werten nicht wesentlich von denen mit dokumentiertem Krankheitsbeginn unterscheidet (missing at random Annahme). Details hierzu sind in (Günther et al., 2020) zu finden. Die Ergebnisse des *nowcasting* liefern die Basis für die hier vorgeschlagene Modellierung.

Zur Analyse des zeitlichen Verlaufes des Infektionsgeschehens verwenden wir folgendes Poisson-Regressionsmodell mit Überdispersion und Bruchpunkten (siehe Muggeo (2003), Muggeo et al. (2020)):

$$E(Y_t) = \exp\left(\beta_0 + \beta_1 t + \sum_{k=1}^{K} \gamma_k (t - CP_k)_+\right) \tag{1}$$

Dabei bezeichnet $E(Y_t)$ die erwartete Anzahl der Neuerkrankungen zum Zeitpunkt t, und K die Anzahl der Bruchpunkte. Mit $x_+ = \max(x, 0)$ wird der positive Teil von x bezeichnet. Mit Hilfe der Bruchpunkte wird der Verlauf von Y_t in $K + 1$ Phasen aufgeteilt. Diese sind durch unterschiedliche Wachstumsparameter charakterisiert. In der Phase vor dem ersten Bruchpunkt CP_1 ist das Wachstum durch den Parameter β_1 gekennzeichnet, in der 2. Phase zwischen CP_1 und CP_2 durch $\beta_2 = \beta_1 + \gamma_1$. Die nächste Änderung gibt es dann zum Zeitpunkt CP_2. In der 3. Phase zwischen CP_2 und CP_3 ist der Wachstumsparameter durch $\beta_3 = \beta_1 + \gamma_1 + \gamma_2$ gegeben. Dies gilt entsprechend bis zur letzten Phase nach CP_4. Die Größen $\exp(\beta_j)$, $j = 1, \ldots, K + 1$ können als tägliche Wachstumsfaktoren interpretiert werden. Die Schätzung des Modells erfolgt mit dem R-Paket `segmented`, siehe Muggeo (2008). Die Anzahl der Bruchpunkte K wird schrittweise bis zu einer Maximalzahl von $K = 4$ variiert und es wird geprüft, ob die Erhöhung der Zahl der Bruchpunkte zu einer relevanten Verbesserung der Modellanpassung (Überdispersionsparameter) führt. Modelle mit einer höheren Anzahl von Bruchpunkten als vier sind für den aktuell betrachteten Zeitraum sehr instabil und werden daher nicht betrachtet.

3 Ergebnisse

In Abbildung 1 ist die Schätzung des Verlaufs der Neuerkrankungen (epidemische Kurve) nach Günther et al. (2020) unter Verwendung der Daten bis zum 22.4.2020 dargestellt. Es ist klar ersichtlich, dass für einen großen Anteil der Fälle (52%) der Tag des Krankheitsbeginns nicht beobachtet, sondern imputiert wurde (Günther et al., 2020). Diese geschätzte epidemische Kurve wird als Datengrundlage zur Berechnung der Bruchpunktanalyse verwendet.

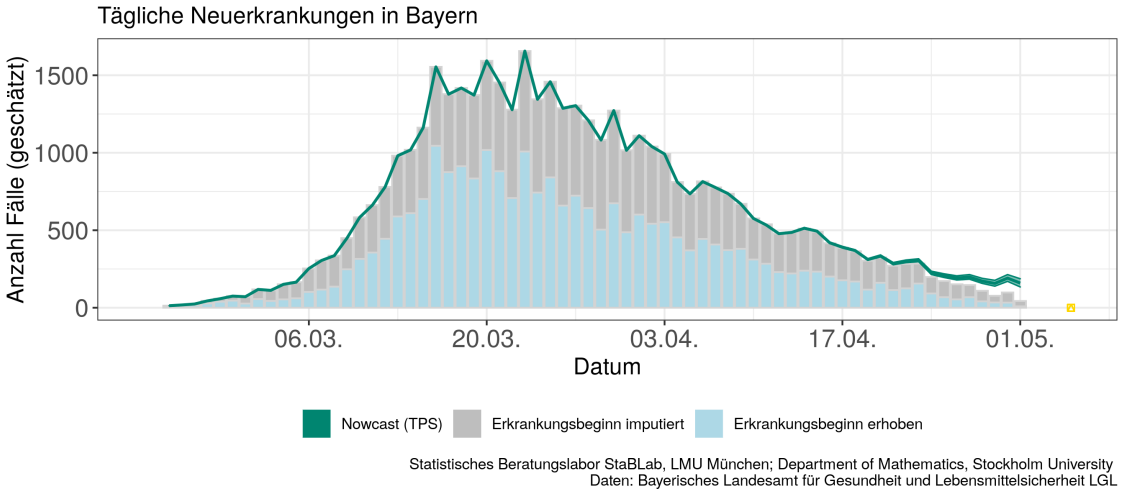

Abbildung 1: Geschätzte Kurve der Neuerkrankungen in Bayern durch den Nowcast unter Verwendung von Daten des LGL bis zum 5.5.

Das Modell mit 4 Bruchpunkten liefert das beste Ergebnis mit einer Schätzung des Überdispersionsparameters von 4.1, d.h. die Varianz von Y_t ist 4.1 Mal höher als der durch die Poisson-Verteilung festgelegte Wert von $\text{Var}(Y_t) = E(Y_t)$. Die Überdispersion bei einem Modell mit 3 Bruchpunkten ist deutlich höher (4.7), was für die Verwendung des Modells mit 4 Bruchpunkten spricht. Die Ergebnisse der Modellierung sind in Abbildung 2 dargestellt und in Tabelle 1 (Koeffizienten und Bruchpunkte) zusammengefasst.

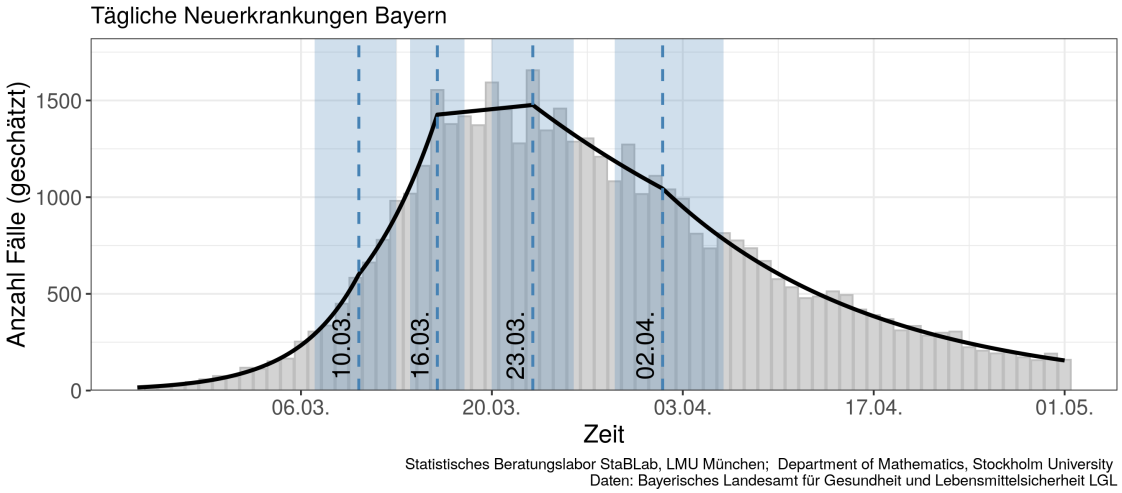

Abbildung 2: Poisson-Modell für Bayern mit 4 geschätzten Bruchpunkten. Die durchgezogene Linie ist die angepasste Kurve gemäß Modell. Die Balken sind die Werte der geschätzten täglichen Neuerkrankungen laut Nowcast. Die Schattierungen geben approximative 95%-Konfidenzintervalle für die Bruchpunkte an.

Tabelle 1: Ergebnisse des Poissonmodells mit 4 Bruchpunkten. Angegeben sind die Zeitpunkte (in Tagen, Zeitpunkt 1 entspricht dem 24.2.2020) der geschätzten Bruchpunkte und die zugehörigen 95%-Konfidenzintervalle. Für das Datum der unteren/oberen Grenze der Konfidenzintervalle wurde jeweils auf den extremeren Wert ab-, bzw. aufgerundet. Im zweiten Teil der Tabelle sind die geschätzten Multiplikationsfaktoren der Fallzahlen pro Tag mit den Konfidenzintervallen für die 5 Phasen angegeben.

Bruchpunkte

Bruchpunkt	Zeitpunkt	95%-KI Untergrenze	95%-KI Obergrenze
1	16.2 (2020-03-10)	13.8 (2020-03-07)	18.7 (2020-03-13)
2	22.0 (2020-03-16)	21.0 (2020-03-14)	23.0 (2020-03-18)
3	29.0 (2020-03-23)	26.5 (2020-03-20)	31.5 (2020-03-26)
4	38.5 (2020-04-02)	35.0 (2020-03-29)	42.0 (2020-04-06)

Geschätzter Multiplikationsfaktor

Phase (j)	Faktor $\exp(\beta_j)$	95%-KI Untergrenze	95%-KI Obergrenze
1	1.250	1.223	1.277
2	1.162	1.113	1.212
3	1.005	0.989	1.021
4	0.964	0.950	0.979
5	0.937	0.933	0.942

Insgesamt ergeben sich aus dem Modell 5 Phasen der epidemischen Kurve:

1.Phase Es gibt es einen deutlichen Anstieg der Neuerkrankungen bis zum 10.3. Hier wird ein Multiplikationsfaktor von 1.25 pro Tag geschätzt.

2.Phase Der exponentielle Anstieg verlangsamt sich zwischen dem 10.3. und dem 16.3.. Der Multiplikationsfaktor sinkt auf einen Wert von 1.16, ist aber immer noch deutlich größer als 1.

3.Phase Am 16.3. kommt es zu einem deutlichen Knick in der Kurve. Die Zahlen bleiben etwa konstant. Der Multiplikationsfaktor liegt zwischen 0.99 und 1.02.

4.Phase Ab dem 23.3. kommt es zu einem Rückgang der Neuerkrankungen. Der Multiplikationsfaktor liegt jetzt bei 0.96 (Konfidenzintervall 0.95-0.98).

5.Phase Ab dem 2.4. verstärkt sich der Rückgang der Neuerkrankungen leicht (Multiplikationsfaktor 0.94). Dieser setzt sich bis zum 1.5. fort.

Zu beachten ist hier noch, dass bei der Beurteilung von Maßnahmen eher die Zahl der Neuinfektionen und nicht die der Neuerkrankungen relevant ist. Diese können mit Hilfe der Inkubationszeit geschätzt werden. Bei der Annahme von einer durchschnittlichen Inkubationszeit von 5 Tagen (Lauer et al., 2020), ist die Kurve in Abb. 2 um 5 Tage nach links zu verschieben. **Damit ergeben sich Bruchpunkte im Infektionsgeschehen am 5.3, 11.3., 18.3. und am 28.3..**

4 Diskussion

Bei der vorliegenden Analyse handelt es sich um eine explorative Analyse der bayerischen Meldedaten.

4.1 Limitationen

In der Analyse sind Fälle, die nicht erfasst wurden, nicht berücksichtigt. Wenn sich der Anteil der nicht entdeckten Fälle über die Zeit ändert, so kann dies die Kurve und damit die Bestimmung der Bruchpunkte verzerren. Dies gilt insbesondere für die schwierige Situation bezüglich der Möglichkeit zur Durchführung von Tests in der Anfangsphase der Epidemie. Daher sollten zusätzlich Daten zu täglichen Todesfällen und

Krankenhausaufnahmen sowie der Anzahlen an durchgeführten Tests betrachtet werden. Weiter kann man mit Hilfe von repräsentativen Studien, wie sie in München gerade durchgeführt werden, den Anteil der nicht erfassten Fälle abschätzen.

Unsere Analyse basiert zu einem erheblichen Teil auf imputierten Daten, siehe dazu Günther et al. (2020), die die unbekannten Daten zum Krankheitsbeginn reflektieren.

Die Annahme einer Inkubationszeit von 5 Tagen ist eine Näherung, die durch weitere Modellierung verfeinert werden könnte. Ein Ansatz dafür wäre eine Rückprojektion der Erkrankungsbeginne durch die Verteilung der Inkubationszeit, siehe z.B. Werber et al. (2013). Die resultierende Kurve der Expositionszeitpunkte könnte dann entsprechend analysiert werden. Da Änderungen des Verhaltens nicht abrupt geschehen, ist auch die Annahme von Bruchpunkten an sich problematisch. Daher sollte die Interpretation der Bruchpunkte immer in Verbindung mit einer direkten Betrachtung der epidemischen Kurve geschehen.

4.2 Interpretation der Ergebnisse

Da bei unseren Analysen der Erkrankungsbeginn (genauer: Beginn der Symptome) betrachtet wird, reflektiert diese Analyse trotz der Limitationen das Infektionsgeschehen besser als die verbreitete Analyse der täglichen oder kumulativen gemeldeten Fallzahlen. Die Bruchpunkte im Infektionsgeschehen am 5.3., 11.3. 18.3. und 28.3. müssen mit Vorsicht interpretiert werden. Sie geben aber wichtige Einblicke in das Infektionsgeschehen in Bayern: Der erste Bruchpunkt am 5.3. (Krankheitsbeginn am 10.3.) hat eine Unsicherheit von -2 bis $+3$ Tagen und deutet auf eine erste Verlangsamung des exponentiellen Wachstums der Epidemie hin. Hier könnten erste Maßnahmen und die öffentliche Diskussion ein Rolle gespielt haben. Am deutlichsten ist der Bruchpunkt des Infektionsgeschehens am 11.3. (Erkrankungsdatum 16.3.) zu sehen. Hier gibt es nur eine Unsicherheit um höchstens $\pm$ einen Tag. Der Zeitpunkt des Bruchpunkts entspricht dem der Rede der Bundeskanzlerin, in der sie erstmals zur Meidung von Sozialkontakten aufrief, der medialen Berichterstattung aus Bergamo, sowie der freiwilligen Umstellung auf Heimarbeit und Telearbeit. Der dritte Bruchpunkt am 18.3. (Krankheitsbeginn am 23.3) hat eine höhere Schwankungsbreite ($-3/+2$ Tage) und ist nicht so deutlich ausgeprägt. Dieser Bruchpunkt liegt nach Datum der Schulschließungen, der EU-Grenzschließung vom 16.03. Diese könnten etwas verzögert zu einer weiteren Verringerung des Infektionsgeschehens geführt haben. Der Bruchpunkt am 28.3 (Erkrankungsbeginn am 2.4.) hat die höchste Streubreite. Es lässt sich hier eine weitere Verstärkung des Rückgangs beobachten.

Der zeitliche Zusammenhang zwischen den Bruchpunkten in unserer Analyse und den Maßnahmen ist als Assoziation zu sehen und kann nicht direkt als kausaler Zusammenhang interpretiert werden. Alternative Erklärungsmöglichkeiten sind z.B. ein saisonaler Effekt auf die Coronaviren-Aktivität oder Veränderungen in der Testkapazität. Entscheidend sind nicht nur die Maßnahmen, sondern auch deren Umsetzung in der Bevölkerung. Auch zeigt sich, dass spätestens 3 Wochen nach Erkrankungsbeginn so gut wie alle Meldungen am LGL sind. Somit hängen die Ergebnisse für März nicht mehr wirklich vom Nowcasting ab, jedoch weiterhin zu einem gewissen Grad vom Imputationsverfahren.

Es erscheint trotzdem wichtig, die epidemische Kurve und die absoluten Fallzahlen direkt zu interpretieren. Die Betrachtung der zeit–variierenden Reproduktionszahl $R(t)$ kann zwar auch ein relatives Bild des Infektionsgeschehens geben, jedoch ist die genaue Bestimmung und die Interpretation dieser Zahl in Zusammenhang mit Interventionen mit einigen Schwierigkeiten verbunden, siehe z.B. Lipsitch et al. (2020). Des Weiteren sagt die Reproduktionszahl nichts darüber aus, wie viele Personen aktuell betroffen sind bzw. ob die Infizierten Risikogruppen angehören. Der von uns berechnete Verlauf der zeit–variierenden Reproduktionszahl (siehe `https://corona.stat.uni-muenchen.de/nowcast/`) für Bayern passt gut zu der Bruchpunktanalyse. Ein Wert von $R(t) > 1$ entspricht einer Steigerungsrate >1. Weiter sind die zeitlichen Verzögerungen bei der Interpretation von $R(t)$ im Auge zu behalten.

Bei der vorliegenden Analyse ist es wichtig festzuhalten, dass Maßnahmen unter einem ganz anderen Informationsstand entschieden werden müssen, als die retrospektiv erstellte epidemische Kurve suggeriert. Die einfache Betrachtung des Verlaufs der gemeldeten Fallzahlen nach Meldedatum ist auch deswegen problematisch, weil dieser Verlauf durch das Meldeverhalten und die Arbeitsweise der Testlabore stark beeinflusst wird. Typischerweise werden an Wochenenden deutlich weniger Fälle gemeldet als unter der Woche. Daher ist die Nowcasting-Schätzung ein wichtiger Schritt, um die besser interpretierbare Kurve der Neuerkrankungen schätzen zu können, hat jedoch wiederum Annahmen und Limitationen die beachtet werden müssen.

Diese Fassung der Analyse zeigt nur geringe Unterschiede zu einer früheren Analyse, die auf Daten bis zum 20.4. beruhte. Mit den neuen Daten konnte keine Verlangsamung des Rückgangs des Infektionsgeschehens am Ende des Beobachtungszeitraumes festgestellt werden.

Verfügbarkeit von Daten und Code

Daten und Code sind unter `https://github.com/FelixGuenther/nc_covid19_bavaria` verfügbar.

Acknowledgements

Wir danken Frau Katharina Katz und Herrn Manfred Wildner vom Bayerischen Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) für die Bereitstellung der Daten und nützliche Diskussionen.

Literatur

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology*, 178(9):1505–1512. Publisher: Oxford Academic.

Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2020). Nowcasting the COVID-19 Pandemic in Bavaria. `https://www.stablab.stat.uni-muenchen.de/_assets/docs/nowcasting_covid19_bavaria.pdf`. Accessed 22.4.2020.

Höhle, M. and an der Heiden, M. (2014). Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*.

Lipsitch, M., Joshi, K., and Cobey, S. (2020). Comment on Pan A, Liu L, Wang C, et al. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. JAMA. `https://github.com/keyajoshi/Pan_response`. Accessed 22.4.2020.

Muggeo, V. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071.

Muggeo, V., Sottile, G., and Porcu, M. (2020). Modelling covid-19 outbreak: segmented regression to assess lockdown effectiveness.

Muggeo, V. M. (2008). segmented: an R package to fit regression models with broken-line relationships. *R News*, 8(1):20–25.

Werber, D., King, L., Müller, L., Follin, P., Bernard, H., Rosner, B., Déléré, Y., de Valk, H., Ethelberg, S., Buchholz, U., and Höhle, M. (2013). Associations of age and sex on clinical outcome and incubation period of shiga toxin-producing Escherichia coli O104:H4 infections, 2011. *American Journal of Epidemiology*, 178(6):984–992.