Extreme value modelling of SARS-CoV-2 community transmission using discrete Generalised Pareto distributions

Abdelaati Daouia, Gilles Stupfler, Antoine Usseglio-Carleve

On the tail heaviness of secondary case numbers and cluster sizes for SARS-CoV-2

Abdelaati Daouia[a], Gilles Stupfler[b] & Antoine Usseglio-Carleve[a,b,c]

[a] University of Toulouse Capitole, TSE - Decision Mathematics and Statistics, France
[b] Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
[c] Avignon Université, Laboratoire de Mathématiques d'Avignon EA 2151, 84000 Avignon, France

Abstract. Superspreading has been suggested to be a major driver of overall transmission in the case of SARS-CoV-2. It is therefore important to investigate statistically the tail features of superspreading events (SSEs) in order to have a better understanding of virus propagation and control. Our extreme value analysis of different sources of secondary cases data, including a very recent and large database, indicates that SSEs associated with SARS-CoV-2 have a fat-tailed nature substantially less severe than predicted recently in the literature, but also less important relative to SSEs associated with SARS-CoV. This may provide policy- and decision-makers with a more reliable assessment of the real tail exposure to SARS-CoV-2 contamination. Going further, we consider the broader problem of large community transmission. We study the tail behavior of SARS-CoV-2 cluster cases documented both in official reports and the media. Our results suggest that the observed cluster sizes have been fat-tailed in the vast majority of surveyed countries. We also give estimates and confidence intervals of the extreme potential risk for those countries. A key component of our methodology is up-to-date discrete generalized Pareto approximations which allow for maximum-likelihood based inference of data having a high degree of discreteness.

Keywords. COVID-19, Superspreading, Cluster size, Secondary cases, Extreme value theory, Discrete extremes.

Superspreading events (SSEs) have been recognized as a significant source of disease transmission, including for respiratory coronaviruses such as SARS-CoV and SARS-CoV-2 [1, 2]. SSEs are generally defined as outbreaks in which a small number of cases infect a number of secondary cases well above the expected average [3]. The number of secondary cases resulting directly from an index case of SARS-CoV or SARS-CoV-2 can be viewed as a random variable, say Z. For both coronaviruses, with a basic reproductive number (R_0) estimated to be 3 to 6 [1, 4], events having triggered more than 6 secondary cases have been suggested to constitute SSEs [5]. Data on such SSEs that was reported in [5] is necessarily scarce, as it was documented in scientific studies during the period February-April 2003 for SARS-CoV and January-June for SARS-CoV-2. It consists only of 15 SSEs associated with SARS-CoV and 45 SSEs associated with SARS-CoV-2. Each of these data points Z_i corresponds to the number of secondary cases resulting from a single given index case in Europe, Asia or North America. The particularity of these datasets is that all the observations Z_i are rather high exceeding the aforementioned threshold of 6 secondary cases.

The natural framework for the analysis of SSEs, and more generally of atypical obser-

vations far away from the mean, is extreme value theory. Doing so, it was argued in [5] that SSEs are fat-tailed, although this was done by incorrectly pooling the 60 available SSEs from SARS-CoV and SARS-CoV-2. Instead, by focusing directly on the raw SARS-CoV-2 data considered in [5], we provide evidence of a lighter upper tail for SSEs with significantly less tail exposure than predicted in their study. We arrive at the same conclusion by making use of a more recent and much larger publicly available dataset containing the number of secondary cases Z_i for $n = 88{,}527$ index cases in the Indian states of Andhra Pradesh and Tamil Nadu.

We also consider the broader problem of large community transmission, being the other fundamental source of pandemic risk. Large infection clusters, along with SSEs, have been argued to play an important role in the transmission of SARS-CoV-2 [2]. We define a cluster of SARS-CoV-2 cases in our analysis as a local outbreak involving a minimum of two cases, including confirmed close contacts with epidemiological linkage over a limited period of time. We consider two databases constructed from government reports [6, 7, 8, 9] and media sources [10], comprising 15 samples of SARS-CoV-2 cluster sizes recorded in 11 countries and 4 US states. Our results show that 13 out of these 15 countries and states have fat-tailed cluster size distributions, and allow to infer their risk category in terms of large community transmission. The recent theory of discrete extremes [11, 12, 13, 14] is our basic tool to address the highly discrete nature of SARS-CoV-2 secondary transmission data and cluster sizes, and to estimate and infer more accurately the tail index and extreme percentiles of the underlying fat-tailed distributions. This allows us to better understand the drivers of superspreading and cluster formation in the ongoing COVID-19 pandemic.

Results and Discussion

Analysis of secondary individual cases. Pooling the 15 SSEs associated with SARS-CoV (Dataset S1) and 45 SSEs associated with SARS-CoV-2 (Dataset S2), that were reported in [5], into a single sample and making use of a Generalized Pareto approximation, [5] has suggested that the distribution of the number of secondary cases Z belongs to the Fréchet maximum domain of attraction MDA_ξ [15], that is, the set of Pareto-type distributions, with tail index ξ between 0.5 and 1 (the estimate provided in [5, Fig. 1 (E)] is $\hat{\xi} \approx 0.6$). The index ξ tunes the tail heaviness of the distribution, with higher positive values indicating a heavier upper tail: moments of order higher than or equal to $1/\xi$ do not exist. An estimate of ξ around 0.6 means that the second moment of Z does not exist, reflecting the outsized contribution of SSEs to overall transmission. Most importantly perhaps, these findings on the tail heaviness of Z make the conventional assumption that Z follows a negative binomial distribution no longer valid for either coronavirus, whereas this assumption was adopted in the literature on disease transmission since the influential work [1] on SARS-CoV, and it is still employed for SARS-CoV-2, see *e.g.* [16].

One may argue, however that the method of [5] is inappropriate for examining the tail behavior of their particular 60 SSEs. The sparsity of data on SSEs is addressed by combining the 15 and 45 observations associated with SARS-CoV and SARS-CoV-2 into a single sample, whereas the two datasets correspond to completely different distributions (Fig. 1 (a)) and should not be pooled accordingly. This is apparent from either a Kolmogorov-Smirnov test, with $p-$value 0.015, or the usual data analysis making the subjective assumption that Z follows a negative binomial distribution. The negative binomial fit of the probability mass function (Fig. 1 (b)) clearly suggests that the upper tail of Z for SARS-CoV appreciably dominates that for SARS-CoV-2. This is confirmed by a proper extreme value analysis of the data (Fig. 1 (c)): the ξ estimates obtained between 0.35

and 0.45 from the Hill estimator, in the special case of SARS-CoV-2, differ substantially from the various competing estimates found to vary between 0.5 and 1 in [5]. Even the 90% confidence intervals of ξ for SARS-CoV-2 (dashed red lines in Fig. 1 (c)) seem to contain only partially the estimated tail index plot for SARS-CoV (solid blue line), reflecting a net difference between the two heavy-tailed distributions of secondary cases associated with SARS-CoV and SARS-COV-2. This conclusion is corroborated by the mean excess value estimates (Fig. 1 (d)), which similarly indicate the relevance of separating the analysis for each coronavirus, pointing in particular towards a smaller effect of the superspreading phenomenon on the current COVID-19 pandemic relative to the SARS-CoV epidemic. The interpretation of this result is that, although SARS-CoV and SARS-CoV-2 belong to the same family of respiratory diseases, the two coronaviruses do not produce the same kind of large SSEs that are relevant for accurately quantifying superspreading risk. For all these reasons, pooling the data before applying extreme value tools can lead to misleading conclusions on the propagation of the SARS-CoV-2 virus.

Yet, the low sample size of this SSE dataset puts a question mark over the quality of the statistical analysis. This is why we also analyzed a very recent and much larger Indian secondary cases dataset of size $n = 88{,}527$, studied for instance in [17] and [18] (Database S3). Although the barplot of this data (Fig. 2 (a)) gives evidence of a considerable right skewness and its summary extreme value analysis (Fig. 2 (b)) suggests a heavy right tail, it should be noted that since the Z_i range from 0 to 39 with a sample size of 88,527, the data is necessarily highly discrete with a large number of tied observations. Ignoring this discrete nature of the Z_i by modeling their tail behavior with the (Generalized) Pareto distribution is inappropriate as this typically results in unreliable tail index estimates and confidence intervals [11]. We address this limitation by applying the recent theory of discrete extremes developed in [11, 12] and based on the discrete generalized Pareto distribution (D-GPD). The D-GPD, first employed by [13] to model road accidents and more recently in [14] to model hospital congestion, has been shown to outperform the continuous GPD when there are a large number of tied observations [11]. Its closed-form survival and probability mass functions allow for an exact likelihood-based inference. Using the D-GPD distribution to fit exceedances $Z_i - u$ above a varying threshold u (rather than trying to fit the whole of the distribution, as [17] did using a discrete Pareto distribution), we found an estimate of ξ around 0.239 with the 90% confidence intervals overwhelmingly suggesting an estimate greater than 0, thus confirming the heavy-tailed character of SARS-CoV-2 SSEs (Fig. 2 (c)). Interestingly, revisiting the small SSE dataset (Dataset S2) of size 60 using the D-GPD maximum likelihood estimation method (Fig. 1 (e)) results in an estimate of around 0.25, in agreement with what is found on the Indian SSE data. This suggests that the distribution of SARS-CoV-2 SSEs has a finite third moment and possibly even a fourth moment. These results are different from those obtained for the SARS-CoV SSEs. The latter rather point towards a distribution with infinite variance and thus a much heavier right tail. This is confirmed by considering the fitted D-GPD probability mass functions for secondary cases (Fig. 1 (f)) that decrease much more rapidly for SARS-CoV-2 than for SARS-CoV. As such, compared to [5], our findings could have significantly different implications when integrated to models of disease transmission that can be relied upon for making policy decisions.

Analysis of cluster cases. We broaden our analysis by examining whether SARS-CoV-2 cluster sizes are fat-tailed too. We consider a database of 15 samples of cluster sizes recorded in 11 countries and 4 US states. The number of reported clusters per country or state varies from 29 (France) to 4,769 (Colorado, USA). The database is constructed from

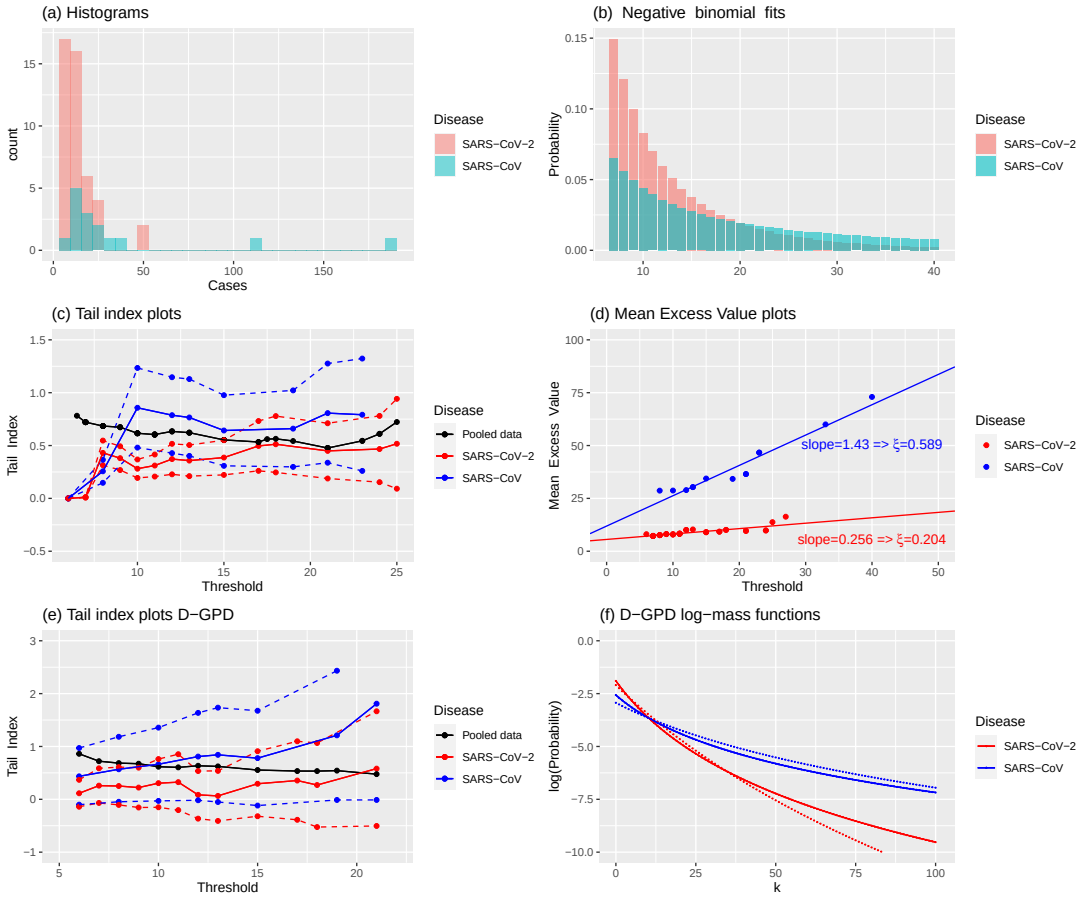

Figure 1: Secondary cases data from [5]. (a) Histogram of the number of secondary cases for SARS-CoV (blue, $n = 15$) and SARS-CoV-2 (red, $n = 45$) SSEs. (b) Fitted probability mass function, conditional on $Z > 6$, of the negative binomial distribution for SARS-CoV (blue) and SARS-CoV-2 (red) SSEs. (c) Hill estimates of ξ for SSEs associated with SARS-CoV (solid blue), SARS-CoV-2 (solid red), and the pooled data (solid black), obtained from the exceedance values $Z_i - u$ given $Z_i \geq u$, as function of the threshold u, along with the resulting 90% confidence intervals for SARS-CoV (dashed blue) and SARS-CoV-2 (dashed red) SSEs. (d) Mean excess plots of SARS-CoV (blue) and SARS-CoV-2 (red) SSEs, quantified by the average of the exceedances $Z_i - u$ given $Z_i \geq u$, as function of u. (e) Discrete GPD maximum likelihood estimates of ξ for SARS-CoV (solid blue) and SARS-CoV-2 (solid red) SSEs, calculated from the exceedances $Z_i - u$ given $Z_i \geq u$, as function of u, along with their corresponding 90% confidence intervals (dashed lines), and the Hill plot produced by combining SARS-CoV and SARS-CoV-2 SSEs. (f) Logarithm of the probability mass functions of the D-GPD fits to the exceedance values $Z_i - u$ given $Z_i \geq u$, for the thresholds $u = 6$ (dotted lines) and $u = 10$ (solid lines), for SARS-CoV (blue) and SARS-CoV-2 (red).

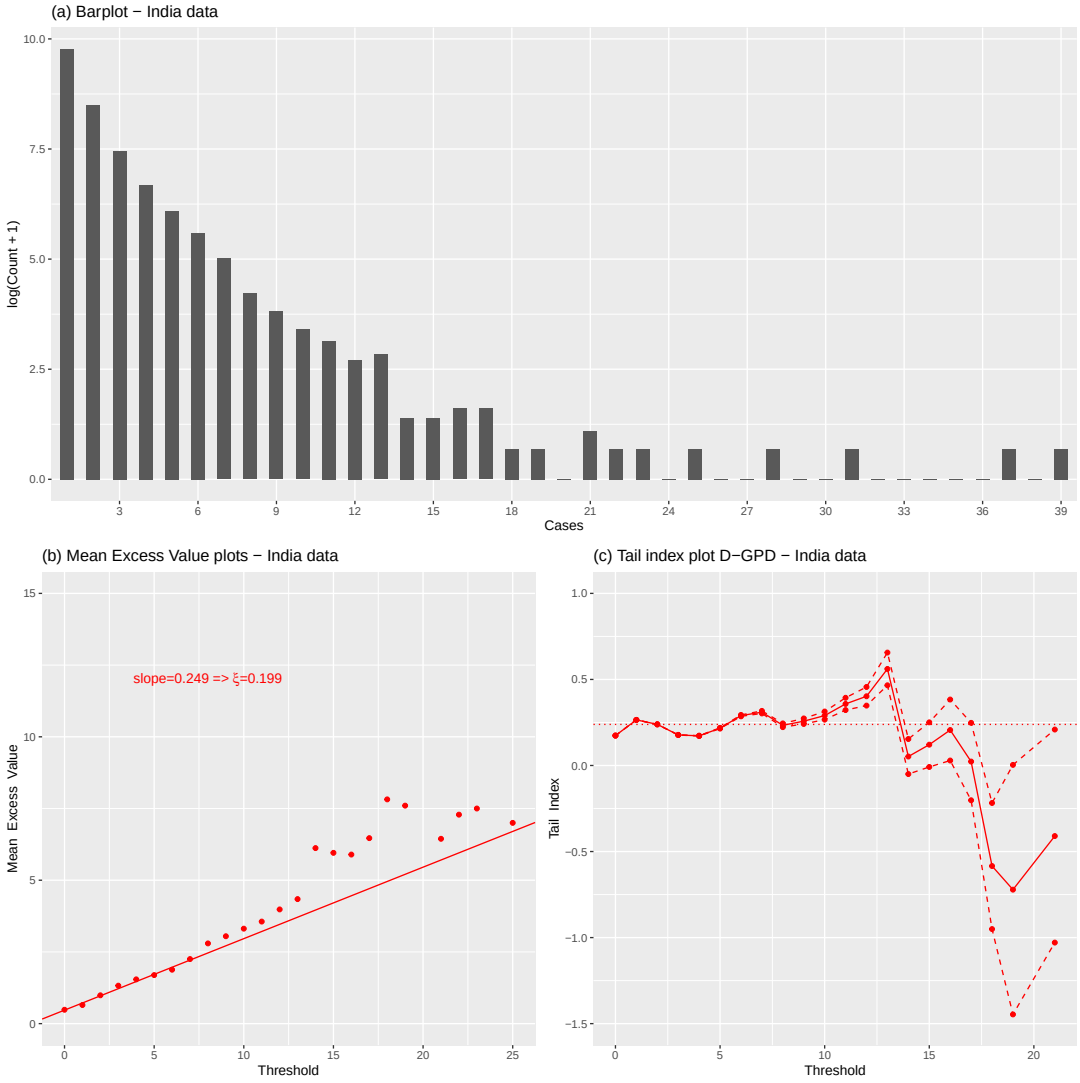

Figure 2: Secondary cases data for SARS-CoV-2 from Andhra Pradesh and Tamil Nadu (India). (a) Bar plot of the $\log(Z_i + 1)$ ($n = 88{,}527$). (b) Mean excess plots of secondary cases. (c) Discrete GPD maximum likelihood estimates of ξ (solid red) and their associated 90% confidence intervals (dashed red), with the averaged estimate $\widehat{\xi} = 0.239$ over the stable region $u \in [0, 10]$ being indicated in dotted horizontal line.

government reports [6, 7, 8, 9] (Database S4) and media sources [10] (Database S5). The median cluster sizes were 5 (Database S4) and 33 (Database S5), and the largest clusters had size 1,761 (Database S4, in a Colorado prison) and 7,000 (Database S5, in an Italian football stadium). Here we denote by Y_i the number of SARS-CoV-2 cases in cluster i. The ξ estimates from each sample of cluster sizes allow to infer the risk category of the corresponding country/state in terms of local community transmission.

Figs. 3 and 4 display the D-GPD maximum likelihood estimates of ξ as functions of the cluster size u. A common practice for selecting a suitable pointwise estimate of ξ is to pick out a sufficiently high threshold u corresponding to a stable region of the plot [19], as indicated by the vertical dashed lines in Figs. 3 and 4. The final selected estimates are reported in Table 1, where 13 out of the 15 countries or states appear to have fat-tailed cluster size distributions (confirmed at the 90% confidence level except for China). The analysis for California and UK & Ireland was inconclusive. For the California dataset, this is possibly due to a strong degree of heterogeneity (see the histogram in the bottom left panel of Fig. 4). A stratified study of the Californian data might be more conclusive. For the UK & Ireland dataset, the fact that the sample is so small (30 clusters) in two countries with highly developed healthcare and contact tracing systems is suspicious and may suggest reporting issues.

Using the D-GPD model, one can gain further insight into large cluster sizes by providing extrapolated estimates of extreme percentiles q_α potentially beyond the sample maximum [13, Formula (5) p.41]. Estimated 95th and 99th percentiles are given in Table 1. One may also match the estimated percentiles with actual observations to get a sense of what would be a conducive environment for the formation of large SARS-CoV-2 clusters. For example, the estimated 95th percentile of 120 cases in Kerala is close to two clusters of 113 cases (nursing home) and 132 cases (local transmission) already observed in Kerala. Likewise, the estimate $\hat{q}_{0.95} = 272$ cases in Canada is fairly close to a cluster of 324 cases in Canadian nursing homes. In Oregon, the estimated 99th percentile $\hat{q}_{0.99} = 124$ cases is in the vicinity of a cluster of 134 cases in a care home setting. In Colorado, the estimate $\hat{q}_{0.99} = 140$ cases is close to a cluster of 134 cases in a nursing home. All of these clusters bar one (the local transmission cluster in Kerala) correspond to indoor environments where social distancing is difficult to practice.

In summary, we have provided evidence that SSEs and cluster sizes for SARS-CoV-2 were fat-tailed, albeit (for SSEs) less so than for SARS-CoV, and less so than argued in [5]. We have not discussed quality checks for our likelihood-based confidence intervals in Table 1 as the D-GPD confidence intervals have already been established to be quite accurate [11]. A limitation of our study lies in the quality of the data, as it is not obvious whether all SSEs or clusters over a given time period were available. To check robustness against such reporting issues, we have reproduced part of our analysis of cluster data by removing at random 10% of observations in each sample containing at least 100 data points, and replicating this experiment 10,000 times. There is indeed some variation in the resulting estimates of ξ (Fig. 5) but this does not affect our conclusion on the fat-tailed behavior of the data, except in rare situations when almost all the large values in the data go missing. This highlights the importance of accurate data reporting as a prerequisite to such analyses.

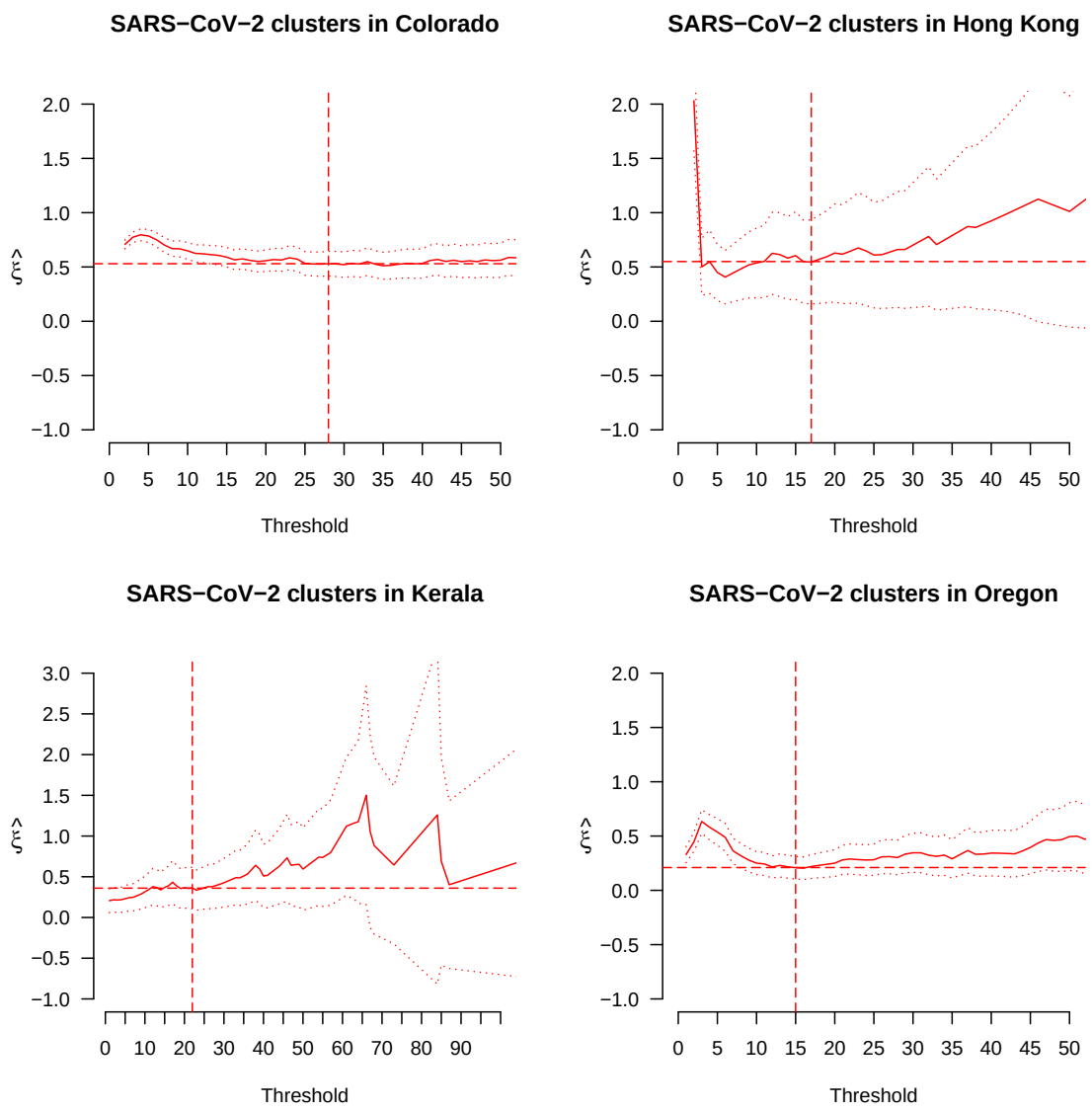

Figure 3: Analysis of cluster cases. Plots of discrete GPD maximum likelihood estimates of ξ (solid lines), along with their 90% confidence intervals (dotted lines) and the final selected estimates (horizontal dashed lines) and thresholds (vertical dashed lines), for the four countries or US states where the source is official data.

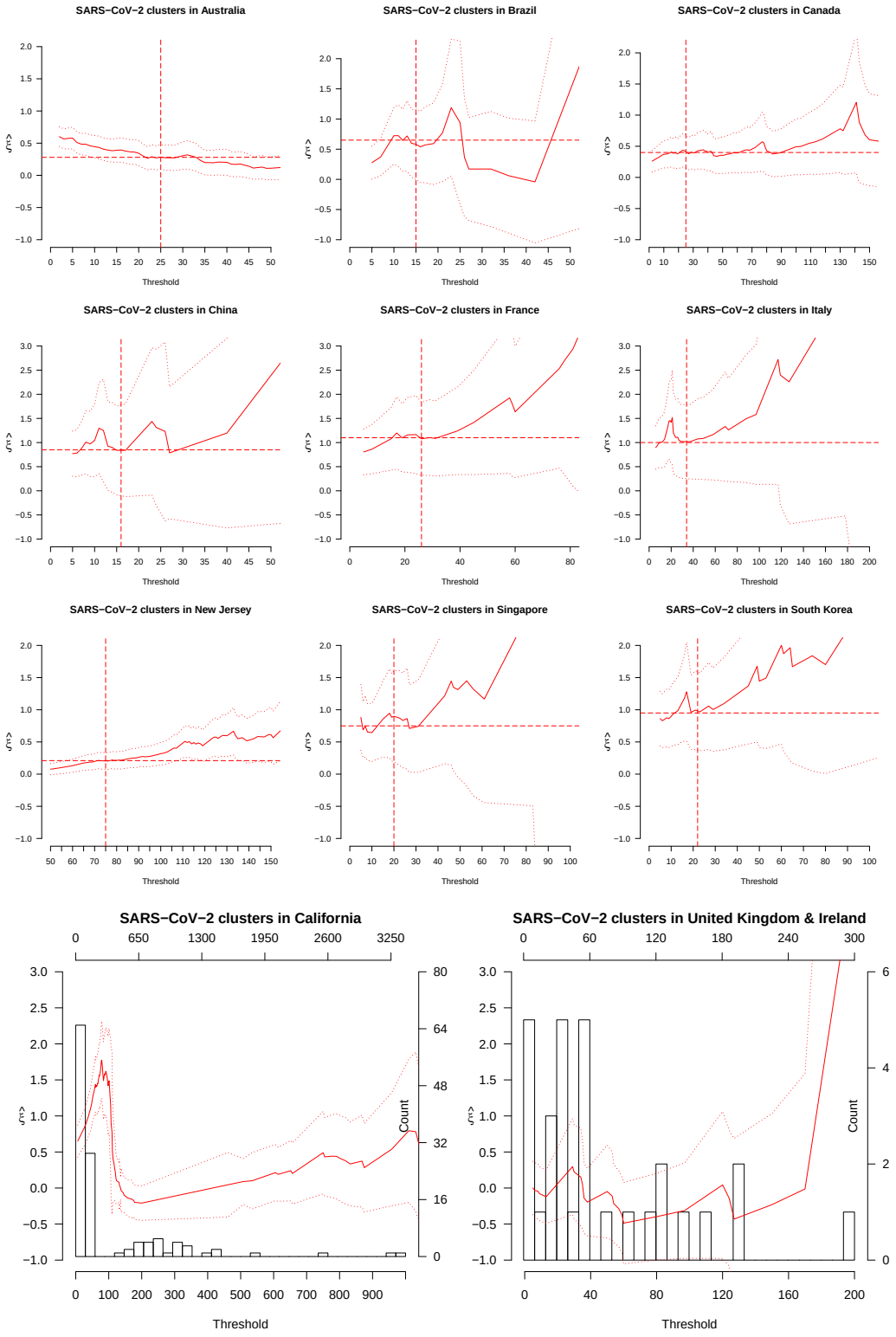

Figure 4: Analysis of cluster cases. As in Figure 3 with the results obtained from the data whose sources were the media. The top 9 plots refer to those countries and states for which the extreme value analysis was conclusive. The bottom 2 plots refer to those for which the extreme value analysis was inconclusive.

Database S4

Location	n	$\hat{\xi}$ [95% CI]	u (n_u)	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	Max. Y_i (Setting)
Colorado, USA	4,769	0.53 [0.41, 0.64]	27 (474)	48	140	1,761 (Prison)
Hong Kong	54	0.55 [0.16, 0.93]	17 (34)	119	310	732 (Dancing)
Kerala, India	113	0.36 [0.11, 0.62]	22 (60)	120	255	580 (Unknown)
Oregon, USA	795	0.21 [0.10, 0.31]	15 (254)	64	124	639 (Prison)

Database S5

Location	n	$\hat{\xi}$ [95% CI]	u (n_u)	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	Max. Y_i (Setting)
Australia	355	0.28 [0.09, 0.48]	25 (145)	157	326	662 (Cruise ship)
Brazil	42	0.58 [0.00, 1.16]	15 (22)	82	220	191 (Hospital)
Canada	100	0.42 [0.15, 0.69]	25 (74)	272	624	1,500 (Meat processing plant)
China	34	0.84 [-0.12, 1.80]	16 (10)	99	401	368 (Market)
France	29	1.08 [0.32, 1.83]	26 (17)	443	2,530	2,500 (Religious gathering)
Italy	41	1.02 [0.25, 1.79]	34 (15)	378	2,013	7,000 (Stadium)
New Jersey, USA	183	0.20 [0.08, 0.33]	75 (157)	299	496	1,042 (Prison)
Singapore	45	0.90 [0.19, 1.61]	20 (21)	156	661	797 (Worker housing)
South Korea	45	0.98 [0.37, 1.59]	22 (24)	324	1,616	5,016 (Religious gathering)

Table 1: Final results for SARS-CoV-2 cluster sizes by country (first column), the corresponding sample size n (second column), D-GPD maximum likelihood ξ estimate and 90% confidence interval (third column), selected cluster size threshold u and associated number n_u of exceedance values $Y_i - u$ given $Y_i \geq u$ upon which the ξ estimate is calculated (fourth column), D-GPD maximum likelihood 95% and 99% percentile estimates of cluster size (fifth and sixth columns), and the sample maximum (last column). The top table corresponds to data from official sources, and the bottom table to data from media sources. The results reported in the latter table only concern the 9 countries and states for which the extreme value analysis was conclusive.

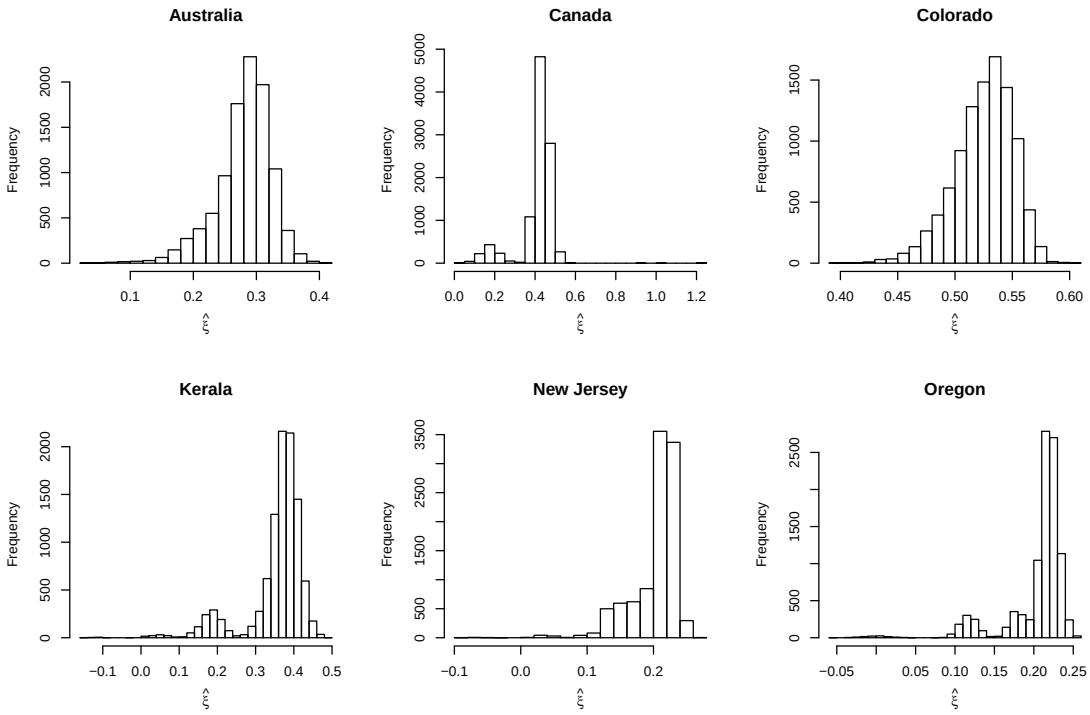

Figure 5: Robustness check for the analysis of cluster cases. Histograms of the 10,000 estimates of ξ obtained by omitting at random 10% of the data. This was done only for the six samples containing at least 100 data points.

Methods

Hill estimator For a dataset $Z_1, \ldots, Z_n$, the Hill estimator at threshold u is defined as

$$\widehat{\xi}_u^H = \frac{1}{\sum_{i=1}^n \mathbb{1}_{\{Z_i \geq u\}}} \sum_{i=1}^n \log\left(\frac{Z_i}{u}\right) \mathbb{1}_{\{Z_i \geq u\}}.$$

D-GPD maximum likelihood estimators The D-GPD maximum likelihood estimators of the shape and scale parameters are obtained by maximizing (with the R function `optim`) the D-GPD log-likelihood function introduced in [11]:

$$\left(\widehat{\xi}_u, \widehat{\sigma}_u\right) = \underset{(\xi,\sigma)\in\mathbb{R}\times(0,\infty)}{\arg\min} \sum_{i=1}^n \log\left(\left(1+\xi\frac{Z_i-u}{\sigma}\right)^{-1/\xi} - \left(1+\xi\frac{Z_i-u+1}{\sigma}\right)^{-1/\xi}\right) \mathbb{1}_{\{Z_i\geq u\}}.$$

Using the classical theory of maximum likelihood estimators, confidence intervals for ξ may be derived from $\widehat{\xi}_u$. Indeed, we first estimate the total Fisher information matrix $I(\xi,\sigma)$ using a finite difference method (with a step $h = 0.001$), and then deduce the following α%-confidence intervals for ξ:

$$\left[\widehat{\xi}_u + \sqrt{\left(\widehat{I}(\xi,\sigma)^{-1}\right)_{1,1}}\Phi^{-1}\left(\frac{1-\alpha}{2}\right),\ \widehat{\xi}_u + \sqrt{\left(\widehat{I}(\xi,\sigma)^{-1}\right)_{1,1}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right],$$

where Φ denotes the standard normal distribution function.

Finally, the 100αth percentile of the D-GPD distribution having location parameter u, scale parameter σ and shape parameter ξ, adapted from [13], is given by

$$q_\alpha = \left\lceil \frac{\sigma}{\xi}\left(\left(\frac{1-\alpha}{\mathbb{P}(Z\geq u)}\right)^{-\xi} - 1\right) + u - 1 \right\rceil.$$

Mean excess plot The mean excess plots represent the values $E(u) = \mathbb{E}\left[Z-u|Z\geq u\right]$ as function of u. Using a dataset $Z_1, \ldots, Z_n$, the estimate of $E(u)$ is given, for each threshold u, by its empirical counterpart

$$\widehat{E}(u) = \frac{\sum_{i=1}^n Z_i \mathbb{1}_{\{Z_i\geq u\}}}{\sum_{i=1}^n \mathbb{1}_{\{Z_i\geq u\}}} - u$$

Conditioned negative binomial distribution The probability mass function of the negative binomial distribution (with parameters $r > 0$ and $p \in [0,1]$) conditional on $Z > u$, is given by

$$\mathbb{P}(Z=k) = \frac{\frac{\Gamma(k+r)}{k!\Gamma(r)}p^r(1-p)^k}{1-\sum_{i=1}^u \frac{\Gamma(i+r)}{i!\Gamma(r)}p^r(1-p)^i}, \text{ for all } k > u.$$

With a dataset $z_1, \ldots, z_n$, the parameter estimators are therefore obtained as the maximum log-likelihood solution

$$\underset{(p,r)\in[0,1]\times(0,\infty)}{\arg\max} \sum_{i=1}^n \log \mathbb{P}(Z=z_i).$$

Code and data availability

The R code for the numerical analysis and datasets are available at GitHub, `https://github.com/AntoineUC`.

Acknowledgements

This research was supported by the French National Research Agency under the grants ANR-19-CE40-0013 and ANR-17-EURE-0010. G. Stupfler also acknowledges support from an AXA Research Fund Award on "Mitigating risk in the wake of the COVID-19 pandemic".

References

[1] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355-359 (2005).

[2] D.C. Adam *et al.*, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.*, 10.1038/s41591-020-1092-0 (2020).

[3] A. P. Galvani, R.M. May, Dimensions of superspreading. *Nature* **438** (7066), 293-5 (2005).

[4] Y. M. Bar-On, A. Flamholz, R. Phillips, R. Milo, SARS-CoV-2 (COVID-19) by the numbers. *eLife* **9**, e57309 (2020).

[5] F. Wong, J. J. Collins, Evidence that coronavirus superspreading is fat-tailed. *PNAS*, `www.pnas.org/cgi/doi/10.1073/pnas.2018490117` (2020).

[6] State of Colorado: `https://covid19.colorado.gov/covid19-outbreak-data`, last updated on 2 June 2021 (resolved outbreaks only). Accessed 27th September 2021.

[7] Government of Hong Kong, `https://www.chp.gov.hk/files/pdf/local_situation_covid19_en.pdf`, last updated on 6 September 2021. Accessed 6th September 2021.

[8] Government of Kerala, `https://covid19jagratha.kerala.nic.in/home/clusterList`. Accessed 21st July 2021.

[9] State of Oregon, `https://www.oregon.gov/oha/covid19/Documents/DataReports/Weekly-Outbreak-COVID-19-Report-2021-08-25-FINAL.pdf?utm_medium=email&utm_source=govdelivery`. Accessed 25 August 2021.

[10] K. Swinkels, SARS-CoV-2 Superspreading Events Database, `https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b`. Accessed 21st July 2021.

[11] A. Hitz, R. Davis, G. Samorodnitsky, Discrete Extremes. arXiv [Preprint] (2017). `https://arxiv.org/abs/1707.05033`

[12] T. Shimura, Discretization of distributions in the maximum domain of attraction. *Extremes*, **15(3)**, 299-317 (2012).

[13] F. Prieto, E. Gómez-Déniz, J. M. Sarabia, Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention* **49**, 71:38 (2014).

[14] S. Ranjbar *et al.*, Modelling the extremes of seasonal viruses and hospital congestion: the example of flu in a Swiss hospital. arXiv [Preprint] (2020). `https://arxiv.org/abs/2005.05808`

[15] P. Cirillo, N. N. Taleb, Tail risk of contagious diseases. *Nat. Phys.* **16**, 606-613 (2020).

[16] N. Islam *et al.*, COVID-19 and climatic factors: A global analysis. *Environmental Research* **193**, 110355 (2021).

[17] C. Kremer *et al.*, Quantifying superspreading for COVID-19 using Poisson mixture distributions. *Sci. Rep.* **11**, 14107 (2021).

[18] R. Laxminarayan *et al.*, Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science* **370**, 691–697 (2020).

[19] L. de Haan, A. Ferreira, *Extreme Value Theory: An Introduction*, Springer-Verlag, New York (2006).