



HAL
open science

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

Abdelaati Daouia, Gilles Stupfler, Antoine Usseglio-Carleve

► To cite this version:

Abdelaati Daouia, Gilles Stupfler, Antoine Usseglio-Carleve. Extreme value modelling of SARS-CoV-2 community transmission using discrete Generalised Pareto distributions. 2023. hal-03392044v3

HAL Id: hal-03392044

<https://hal.science/hal-03392044v3>

Preprint submitted on 23 Feb 2023 (v3), last revised 14 Mar 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

^a University of Toulouse Capitole, Toulouse School of Economics, France (ORCID:
0000-0003-2621-8860)

^b Univ Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France (ORCID:
0000-0003-2497-9412)

^c Avignon Université, Laboratoire de Mathématiques d'Avignon EA 2151, 84000 Avignon,
France (ORCID: 0000-0002-8148-3758)

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 statistically investigate the tail features of superspreading events (SSEs) to better understand virus propagation and control. Our extreme value analysis of different sources of secondary case data indicates that case numbers of SSEs associated with SARS-CoV-2 may be fat-tailed, although substantially less so than predicted recently in the literature, but also less important relative to SSEs associated with SARS-CoV. The results caution against pooling data from both coronaviruses. This could provide policy- and decision-makers with a more reliable assessment of the tail exposure to SARS-CoV-2 contamination. Going further, we consider the broader problem of large community transmission. We study the tail behaviour of SARS-CoV-2 cluster cases documented both in official reports and in 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 Generalised Pareto models which allow for maximum-likelihood based inference of data with a high degree of discreteness.

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

Introduction

Superspreading events (SSEs) have been recognised as a significant source of disease transmission for respiratory coronaviruses such as SARS-CoV and SARS-CoV-2 [1, 2]. SSEs may be defined as outbreaks in which a given individual (the index case)

24 infects a number of people (secondary cases) well above a certain measure, such
25 as the average or median number of infections. The number of secondary cases
26 resulting directly from an index case can be viewed as a random variable, say Z ,
27 defining the so-called offspring distribution. For both coronaviruses, events having
28 triggered more than 6 secondary cases have been suggested to constitute SSEs [3].
29 Data on such SSEs that was curated and reported in [3] in the early stages of the
30 COVID-19 pandemic is necessarily scarce: it consists mainly of 15 SSEs associated
31 with SARS-CoV and 45 SSEs associated with SARS-CoV-2, each represented by a
32 number of secondary cases Z_i resulting from a single given index case in Europe,
33 Asia or North America. The natural framework for the analysis of SSEs, and more
34 generally of atypical observations far away from the mean, is extreme value theory.
35 Following this framework, it was argued in [3] that SSEs are fat-tailed, although
36 this was done by pooling the 60 available SSEs from SARS-CoV and SARS-CoV-
37 2. A careful investigation of these SARS-CoV and SARS-CoV-2 datasets reveals
38 that the two largest observations in the pooled data are SARS-CoV SSEs; given the
39 small sample size, one may wonder whether the reported estimate of tail heaviness
40 is representative of the tail behaviour of SARS-CoV-2 SSEs.

41 This constitutes the motivation for this work, whose overarching goals are to
42 show how to conduct a statistically rigorous extreme value analysis of community
43 transmission parameters, and to carry out such an analysis in the example of SARS-
44 CoV-2. By focusing directly on the raw SARS-CoV-2 data considered in [3], we
45 provide evidence of a lighter upper tail for SSEs with significantly less tail exposure
46 than predicted in their study. We arrive at the same conclusion by making use of
47 a more recent and much larger publicly available surveillance and contact-tracing
48 database containing the number of secondary cases Z_i for 88,527 index cases in the
49 Indian states of Andhra Pradesh and Tamil Nadu [4]. We also analyse two other
50 South Korean contact-tracing datasets, one collected in the first half of 2020 [3],
51 the other during the summer of 2021 when the Delta variant of SARS-CoV-2 was
52 responsible for the majority of positive cases [5]. The fat-tailedness of the secondary
53 cases distribution is found to be rather clear in the 2021 sample of data, while the
54 analysis of the 2020 data is less conclusive. In all these samples of data we find point
55 estimates of the extreme value index suggesting that the secondary cases distribution
56 has a finite third moment, which stands in contrast with the earlier finding of [3] of
57 a distribution with an infinite variance.

58 In addition to that, we consider the broader problem of large community trans-
59 mission, as it represents the other fundamental source of pandemic risk. Large
60 infection clusters, along with SSEs, have been argued to play an important role in
61 the transmission of SARS-CoV-2 [2]. In a similar spirit to [2], we define a cluster
62 of SARS-CoV-2 cases in our analysis as a local outbreak involving a minimum of
63 two cases, including confirmed close contacts with epidemiological linkage over a
64 limited period of time. We consider two databases constructed from government
65 reports [6, 7, 8, 9] and media sources [10], comprising 15 samples of SARS-CoV-2
66 cluster sizes recorded in 11 countries and 4 US states. Our results show that 13
67 of these 15 countries and states have fat-tailed cluster size distributions, thus fa-
68 cilitating the process of inferring their risk category in terms of large community
69 transmission. This allows us to better understand the drivers of superspreading and

70 cluster formation in the ongoing COVID-19 pandemic. The recent theory of discrete
71 extremes [11, 12, 13, 14] is our basic tool to address the highly discrete nature of
72 SARS-CoV-2 secondary transmission data and cluster sizes. Its use constitutes our
73 main statistical contribution to the study of the transmission of the SARS-CoV-2
74 virus. As we illustrate throughout the paper, estimating and inferring the extreme
75 value index and extreme percentiles of the underlying discrete distributions with this
76 methodology is much easier and accurate than with classical extreme value meth-
77 ods such as the Hill and Generalised Pareto maximum likelihood estimators, which
78 heavily rely on the continuous data assumption.

79 The structure of the paper is as follows. We first describe the methods employed
80 throughout our study, including the discrete Generalised Pareto Distribution fitted
81 to exceedances over a high threshold by means of the maximum likelihood estimator.
82 We then analyse our datasets, first on SARS-CoV-2 secondary case numbers and
83 then on cluster sizes, using these methods. The final section gathers and contrasts
84 these findings and concludes with additional comments about the scope, limitations
85 and robustness of our results, as well as ideas for further work.

86 Methods

87 We use several methods from extreme value theory, which constitutes the correct
88 mathematical framework for the analysis of high observations from a random phe-
89 nomenon [15]. We are particularly interested in methods that can describe so-called
90 fat-tailed random variables, which infrequently but regularly generate very high
91 values and therefore appear to be relevant in the analysis of SARS-CoV-2 trans-
92 mission. A random variable X is fat-tailed if and only if its distribution function
93 $\mathbb{P}(X \leq x)$ can be, for large x , expressed as $\mathbb{P}(X \leq x) = 1 - x^{-1/\xi}\ell(x)$, where ℓ
94 satisfies $\ell(tx)/\ell(t) \rightarrow 1$ as $t \rightarrow \infty$ for any positive real number x . Informally, the
95 tail behaviour of X is controlled by the extreme value index $\xi > 0$, which must be
96 estimated to get a precise understanding of tail heaviness. A standard estimator in
97 this context is the Hill estimator [16]. For a dataset Z_1, \dots, Z_n , the Hill estimator
98 at threshold u is defined as

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

99 It is of course crucial, before using the Hill estimator, to ascertain whether the
100 distribution of the data points indeed has a heavy tail. A common diagnostic method
101 is the mean excess plot, which estimates the values of the mean excess function
102 $E(u) = \mathbb{E}[Z - u | Z > u]$ as function of u . A natural estimate of $E(u)$ is given, for
103 each threshold u , by its empirical counterpart

$$\hat{E}(u) = \frac{\sum_{i=1}^n Z_i \mathbb{1}_{\{Z_i > u\}}}{\sum_{i=1}^n \mathbb{1}_{\{Z_i > u\}}} - u.$$

104 A fat-tailed distribution will typically have mean excess plots exhibiting a linear
105 upward drift for large values of u , whose slope is a consistent estimate of $\xi/(1 - \xi)$
106 when $\xi < 1$, see for example Section 1.2.2 pp.14-19 and p.152 in [17]. In the case

107 $\xi \geq 1$, Theorem 3.4 and Remark 3.5 in [18] show that the mean excess plot converges
 108 in a suitable sense to a random curve, which in the log-log scale is a straight line
 109 with slope $1/\xi$ and random intercept term constructed upon a stable random variable
 110 with index $1/\xi$.

111 It has, however, been observed in the extreme value literature [18] that the mean
 112 excess function very often exhibits a non-linear behaviour at the right end of the
 113 mean excess plot, due to very high variability of the estimate of $E(u)$ when u is
 114 close to the highest Z_i . As a consequence, good statistical practice recommends
 115 to confirm a diagnostic of a heavy tail using other extreme value tools. One such
 116 general approach, which does not presuppose that the data is fat-tailed, consists in
 117 using the Generalised Pareto maximum likelihood estimator applied to the excesses
 118 $Z_i - u$. Recall that the Generalised Pareto distribution, with shape parameter ξ and
 119 scale parameter σ , has probability density function

$$h_{\xi,\sigma}(x) = \frac{1}{\sigma} \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi-1} \mathbb{1}_{\{x>0, 1+\xi x/\sigma>0\}}.$$

The Generalised Pareto maximum likelihood estimator is then defined as, according
 to Section 5.3.2 in [17] and [19]:

$$\begin{aligned} \left(\hat{\xi}_u^{GP}, \hat{\sigma}_u^{GP}\right) &= \arg \min_{\xi>-1, \sigma>0} \sum_{i=1}^n \log h_{\xi,\sigma}(Z_i - u) \\ &= \arg \min_{\xi>-1, \sigma>0} \sum_{i=1}^n \left[-\log \sigma - \left(\frac{1}{\xi} + 1\right) \log \left(1 + \xi \frac{Z_i - u}{\sigma}\right) \right] \mathbb{1}_{\{Z_i > u, 1 + \xi(Z_i - u)/\sigma > 0\}}. \end{aligned}$$

120 The Generalised Pareto maximum likelihood estimators are valid even when the
 121 underlying distribution is not fat-tailed, which has made them very popular in the
 122 natural sciences [20].

123 However, both the Hill and Generalised Pareto estimators of ξ suffer from jagged
 124 sample paths when the data points Z_i feature a substantial number of ties, that is,
 125 they come from a distribution with a high degree of discreteness. This behaviour
 126 makes it extremely difficult to choose an accurate estimate of ξ , which renders the
 127 two methods highly unsatisfactory. The essential reason behind this phenomenon
 128 is that both estimators are built under the – generally incorrect – assumption that
 129 the data points come from a pure (Generalised) Pareto distribution, which is con-
 130 tinuous, and as such, they cannot be expected to handle a substantial degree of
 131 discreteness. We exemplify this phenomenon in Fig. 1: notice, in the top panels, the
 132 stark difference in stability and smoothness of sample paths between a plot of the
 133 Hill estimator as a function of the threshold value (henceforth referred to as a Hill
 134 plot) for continuous data Z_i and its counterpart for data rounded to the nearest in-
 135 teger up. Crucially in applied setups, the asymptotic Gaussian confidence intervals
 136 constructed by approximating the distribution of $\sqrt{n\mathbb{P}(X > u_n)}(\hat{\xi}_{u_n}^H - \xi)$ by a Gaus-
 137 sian distribution with expectation 0 and variance ξ^2 , which is valid when $u_n \rightarrow \infty$
 138 satisfies reasonable conditions [21], are highly unstable when the data features a
 139 large number of ties, thus making inference using the Hill estimator inadvisable.
 140 The bottom panels further show the impact of these data ties: the Hill estimator for

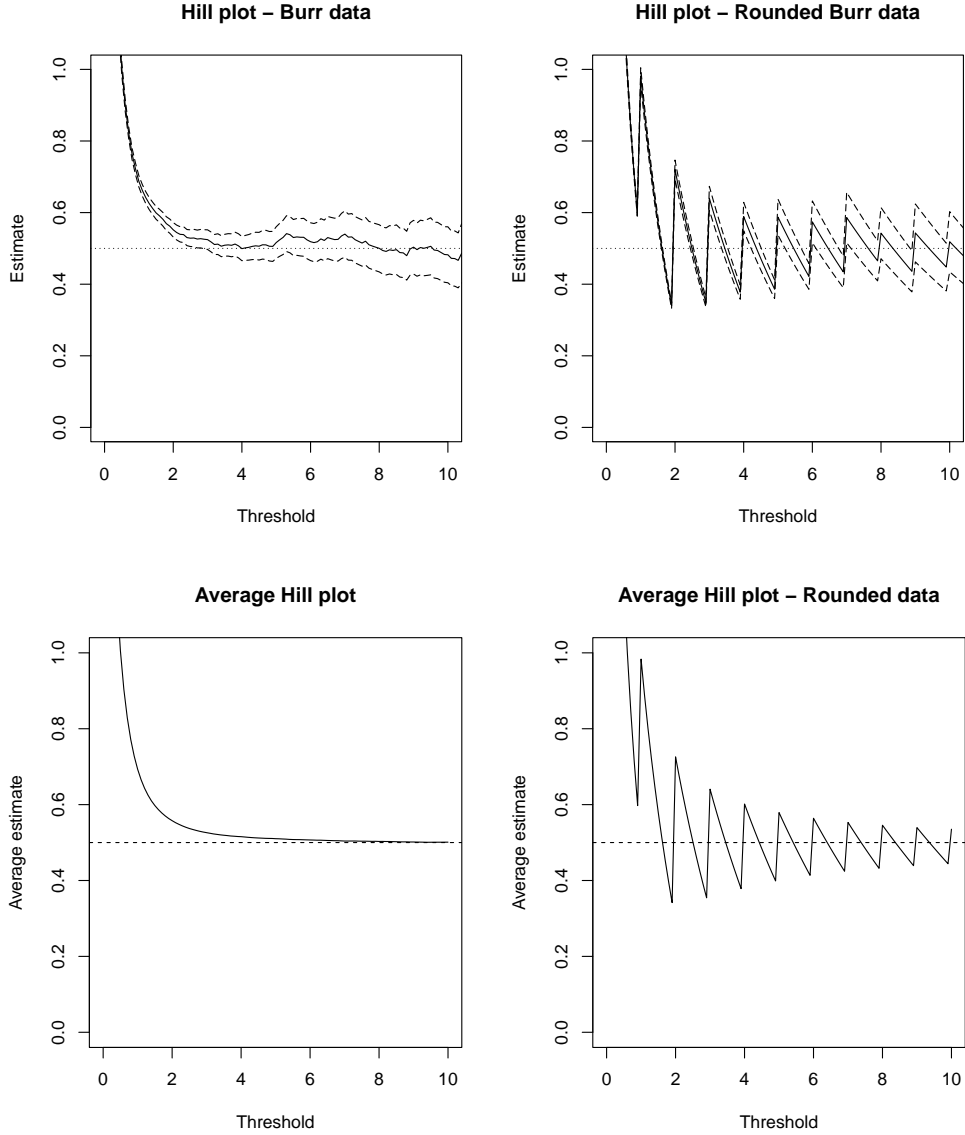


Figure 1: Top panels: Hill plots (solid lines) and corresponding 90% Gaussian asymptotic confidence intervals (dashed lines) as functions of the threshold value u , for $n = 10,000$ simulated data points Z_i from the Burr distribution with probability density function $f(x) = \xi^{-1}x^{-\rho/\xi-1}(1 + x^{-\rho/\xi})^{1/\rho-1}$ (for $x > 0$) with $\xi = 1/2$ and $\rho = -1$ in the left panel, and for the data $\lceil Z_i \rceil$ (*i.e.* the smallest integer larger than or equal to Z_i) in the right panel. Bottom panels: Averaged Hill plots when this experiment is repeated $N = 1,000$ times.

141 discrete data tends to be strongly biased and much more so than the Hill estimator
 142 for continuous data.

143 An alternative option properly taking the discreteness of the data into account
 144 is to employ discrete models to construct an estimator of the extreme value index.
 145 This was pursued by [11, 13], which used so-called D-GPD (for Discrete-Generalised
 146 Pareto Distribution) models, first employed by [12] to model road accidents and
 147 more recently in [14] to model hospital congestion. The D-GPD, whose probability
 148 mass function is

$$p_{\xi,\sigma}(x) = \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi} - \left(1 + \xi \frac{x+1}{\sigma}\right)^{-1/\xi} \quad \text{for } x = 0, 1, 2, \dots \text{ with } p_{\xi,\sigma}(x) > 0$$

149 for $\xi \geq 0$ or $\xi < 0$ and σ/ξ a negative integer, has been shown to outperform
 150 the continuous GPD when there are a large number of tied observations: see the
 151 simulated Poisson and discrete Inverse-Gamma examples in Section 3.1 of [13], which
 152 respectively show that the GPD provides poor model fits and poor tail estimates
 153 when the data is highly discrete, while the D-GPD distribution performs well. Its
 154 closed-form survival and probability mass functions allow for an exact likelihood-
 155 based inference constructed upon the maximum likelihood estimators

$$\left(\hat{\xi}_u, \hat{\sigma}_u\right) = \arg \min_{\xi > -1, \sigma > 0} \sum_{i=1}^n \log p_{\xi,\sigma}(Z_i - u).$$

156 When $\xi = 0$, the convention we adopt is that $(1 + \xi z)^{-1/\xi} = \exp(-z)$, for any
 157 $z \in \mathbb{R}$. These maximum likelihood estimators of the extreme value index ξ and scale
 158 parameter σ of the D-GPD model are readily obtained through the R maximisa-
 159 tion routine `optim`. Using the classical theory of maximum likelihood estimators,
 160 confidence intervals for ξ may be derived from $\hat{\xi}_u$ by estimating the total Fisher
 161 information matrix $I(\xi, \sigma)$ using a finite difference method and then deducing the
 162 following $100\alpha\%$ -confidence interval for ξ :

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

163 where Φ denotes the standard normal distribution function and Φ^{-1} its inverse
 164 (quantile function). Modelling $Z - u$ conditional on $Z \geq u$ by a D-GPD distribution
 165 with parameter estimates $(\hat{\xi}_u, \hat{\sigma}_u)$ suggests the following estimate of the 100α th
 166 percentile of Z adapted from [12, Formula (5) p.41]:

$$\hat{q}_\alpha = \left\lceil \frac{\hat{\sigma}_u}{\hat{\xi}_u} \left(\left(\frac{n(1-\alpha)}{\sum_{i=1}^n \mathbb{1}_{\{Z_i \geq u\}}} \right)^{-\hat{\xi}_u} - 1 \right) + u - 1 \right\rceil,$$

167 for $\alpha \in (0, 1)$ large enough. Here, $\lceil \cdot \rceil$ denotes the ceiling function, that is, $\lceil x \rceil$ denotes
 168 the smallest integer larger than or equal to x . Estimating this quantile by plugging in
 169 the aforementioned estimates of ξ and σ makes it possible to infer extreme quantile
 170 levels and therefore get precise information on the tail behaviour of a distribution
 171 with a large degree of discreteness. For each of the extreme value estimators we have

172 introduced (Hill estimator and GPD and D-GPD maximum likelihood estimators), a
 173 common practice for selecting a suitable pointwise estimate of ξ , colloquially referred
 174 to as “eyeballing”, is to pick out a sufficiently high threshold u corresponding to a
 175 stable region of the plot [15]. We shall indeed also adopt this practice and will
 176 clearly indicate selected threshold or threshold regions in our analyses.

177 For comparison purposes, we will contrast the resulting extreme quantile estimates
 178 with those provided by the (conditioned) negative binomial distribution. Recall that
 179 the probability mass function of the negative binomial distribution (with parameters
 180 $r > 0$ and $p \in (0, 1)$) conditional on $Z > u$, is given by

$$\mathbb{P}_{p,r,u}(Z = k) = \frac{\frac{\Gamma(k+r)}{k! \Gamma(r)} p^r (1-p)^k}{1 - \sum_{i=0}^u \Gamma(i+r)/(i! \Gamma(r)) p^r (1-p)^i}, \text{ for all } k > u.$$

181 Here Γ denotes Euler’s Gamma function. With a dataset z_1, \dots, z_n , the parameter
 182 estimators are therefore obtained as the maximum log-likelihood solution

$$\arg \max_{(p,r) \in (0,1) \times (0,\infty)} \sum_{i=1}^n \log \mathbb{P}_{p,r,u}(Z = z_i).$$

183 Ever since the seminal work of [1], the negative binomial distribution has been widely
 184 used to describe the number of secondary cases resulting from an index case of SARS-
 185 CoV. As suggested in [3, 23], this model has exponentially decreasing probability
 186 mass functions and thus cannot be expected to accurately represent tail heaviness in
 187 SARS-CoV-2 transmission data. We provide below further evidence for this claim,
 188 and for the suitability of D-GPD maximum likelihood estimates in the context of
 189 discrete data, through several datasets gathering numbers of SARS-CoV-2 secondary
 190 cases and cluster sizes in different settings.

191 Data and results

192 **Analysis of secondary case data.** Our first two datasets were reported in [3].
 193 They consist of 15 SSEs associated with SARS-CoV (Dataset S1) and 45 SSEs
 194 associated with SARS-CoV-2 (Dataset S2), each resulting in more than 6 secondary
 195 cases, along with month of occurrence and location of the superspreading event,
 196 and its setting. We refer to [3] for further details about the construction of these
 197 datasets. Pooling the 15 SSEs associated with SARS-CoV and 45 SSEs associated
 198 with SARS-CoV-2 into a single sample and making use of a Generalised Pareto
 199 approximation, [3] has suggested that the distribution of the number of secondary
 200 cases Z belongs to the Fréchet maximum domain of attraction [22], that is, the set
 201 of Pareto-type distributions, with extreme value index ξ between 0.5 and 1 (the
 202 estimate provided in [3, Fig. 1 E] is $\hat{\xi} \approx 0.6$). The index ξ tunes the tail heaviness of
 203 the distribution, with higher positive values indicating a heavier upper tail: moments
 204 of order higher than or equal to $1/\xi$ do not exist. An estimate of ξ around 0.6 means
 205 that the second moment of Z does not exist, reflecting the outsized contribution of
 206 SSEs to overall transmission. Most importantly perhaps, these findings on the tail
 207 heaviness of Z invalidate the conventional assumption that Z follows a negative

208 binomial distribution for either coronavirus, whereas this assumption was widely
209 adopted in the literature on disease transmission ever since the influential work [1]
210 on SARS-CoV, and it is still widely employed for SARS-CoV-2, see *e.g.* [5, 24, 25].

211 Based on our statistical analysis of these datasets, summarised in Fig. 2, one
212 may however argue that the method of [3] is inappropriate for examining the tail
213 behaviour of their particular 60 SSEs. The sparsity of data on SSEs is addressed by
214 combining the 15 and 45 observations associated with SARS-CoV and SARS-CoV-2
215 into a single sample, whereas the two datasets correspond to completely different
216 distributions (Fig. 2 (a)) and should not be pooled accordingly. This is apparent
217 from either a Kolmogorov-Smirnov test, with p -value 0.015, or the more common
218 approach making the questionable assumption that Z follows a negative binomial
219 distribution. The conditional (given $Z > 6$) negative binomial fit of the probability
220 mass function to the Z_i (by construction larger than 6), calculated as described
221 in the last paragraph of the Methods section (Fig. 2 (b)), already suggests that
222 the upper tail of Z for SARS-CoV appreciably dominates that for SARS-CoV-2.
223 In other words, even a naive analysis of the SSE distributions, using the classical
224 negative binomial distribution and not accounting for the heavy tail in the data,
225 indicates that the SSEs for SARS-CoV and those for SARS-CoV-2 exhibit different
226 statistical behaviour. This is confirmed by an analysis of the data properly taking
227 into account its extremes (Fig. 2 (c)): the ξ estimates obtained from the Hill esti-
228 mator in the special case of SARS-CoV-2 vary between 0.35 and 0.45, and as such
229 differ substantially from the various competing estimates found to vary between 0.5
230 and 1 in [3]. Even the 90% confidence intervals of ξ for SARS-CoV-2 (dashed red
231 lines in Fig. 2 (c)) only partially contain the estimated extreme value index plot for
232 SARS-CoV (solid blue line), reflecting a net difference between the two fat-tailed
233 distributions of secondary cases associated with SARS-CoV and SARS-CoV-2. This
234 conclusion is corroborated by the mean excess function estimates (Fig. 2 (d)), which
235 similarly indicate the relevance of separating the analysis for each coronavirus. This
236 suggests that although SARS-CoV and SARS-CoV-2 belong to the same family
237 of respiratory diseases, superspreading events are larger in scale for SARS-CoV in
238 comparison to SARS-CoV-2. For all these reasons, pooling the data before applying
239 extreme value tools can lead to misleading conclusions on the propagation of the
240 SARS-CoV-2 virus.

241 Yet, the low sample size of this SSE dataset puts a question mark over the quality
242 of the statistical analysis. Trustworthy extreme value inference may require a larger
243 sample size, of the order of at least several thousands. This is why we also analysed
244 a much larger Indian secondary case dataset of size $n = 88,527$ (Database S3). This
245 comprehensive surveillance and contact-tracing database was collected in 2020 by
246 the public health authorities of the two Indian states of Andhra Pradesh and Tamil
247 Nadu, whose residents total about 10% of India's population. It was studied for
248 instance in [4] and [23], and we refer to the latter for more information about the
249 database's construction and contents. Results are reported in Fig. 3. Although the
250 barplot of this data (Fig. 3 (a)) gives evidence of a considerable right skewness and
251 its summary extreme value analysis (Fig. 3 (b)) suggests a heavy right tail, it should
252 be noted that since the Z_i range from 0 to 39 with a sample size of 88,527, the data
253 is necessarily highly discrete with a large number of tied observations (see Table 1).

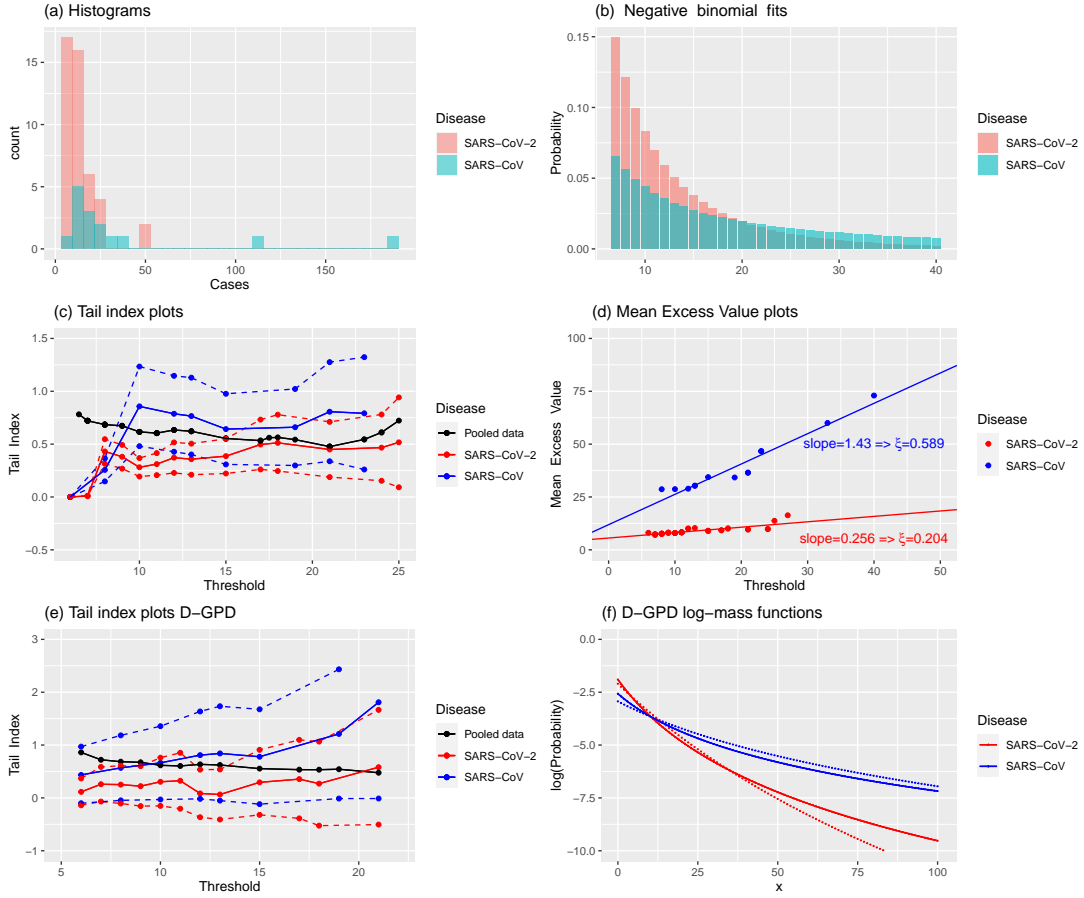


Figure 2: Secondary case data from [3] (Datasets S1 and S2). (a) Histogram of the number of secondary cases for SARS-CoV (blue, $n = 15$) and SARS-CoV-2 (red, $n = 45$) SSEs. (b) Fitted probability mass function, conditional on $Z > 6$, of the negative binomial distribution for SARS-CoV (blue) and SARS-CoV-2 (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 (black line). (f) Logarithm of the probability mass functions $\mathbb{P}_{\sigma, \xi}(X = x)$ 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).

Z	0	1	2	3	4	5	6	7	8	9	10	11				
Count	62,540	17,493	4,885	1,730	802	444	267	149	67	44	29	22				
Z	12	13	14	15	16	17	18	19	21	22	23	25	28	31	37	39
Count	14	16	3	3	4	4	1	1	2	1	1	1	1	1	1	1

Table 1: Secondary case data (Database S3) for SARS-CoV-2 from Andhra Pradesh and Tamil Nadu (India).

254 Ignoring the discrete nature of the Z_i by modelling their tail behaviour with the
255 (Generalised) Pareto distribution is inappropriate as this typically results in unre-
256 liable extreme value index estimates and confidence intervals [13]. This becomes
257 obvious here by superimposing both the classical Hill and continuous Generalised
258 Pareto maximum likelihood estimators of the extreme value index, as functions of a
259 varying threshold u in Fig. 3 (c). Clearly, both plots are so volatile and jagged that
260 it is hard to identify any stable region and therefore a reasonable point estimate of ξ
261 cannot easily be determined. Using the D-GPD distribution to fit exceedances $Z_i - u$
262 above the threshold u (rather than trying to fit the whole of the distribution, as [23]
263 did using a discrete Pareto distribution) results in a much smoother and stable fit
264 (Fig. 3 (c)), and leads to an estimate of ξ around 0.24 with the 90% confidence
265 intervals overwhelmingly suggesting an estimate greater than 0, thus confirming the
266 fat-tailed nature of SARS-CoV-2 SSEs (Fig. 3 (d)) in this sample. Interestingly,
267 revisiting the small SARS-CoV-2 SSE dataset (Dataset S2) of size 45 using the D-
268 GPD maximum likelihood estimation method (Fig. 2 (e)) results in an estimate of
269 around 0.25, in agreement with the results from the Indian secondary case data.
270 This suggests that the distribution of SARS-CoV-2 SSEs has a finite third moment
271 and possibly even a fourth moment. These results are different from those obtained
272 for the SARS-CoV SSEs. The latter rather point towards a distribution with infi-
273 nite variance and thus a much heavier right tail. This is confirmed by considering
274 the fitted D-GPD probability mass functions for secondary cases (Fig. 2 (f)) that
275 decrease much more rapidly for SARS-CoV-2 than for SARS-CoV.

276 To examine the extreme value behaviour of the SARS-CoV-2 offspring distribu-
277 tion in different conditions, we turn to the analysis of two contact-tracing datasets in
278 South Korea, a country which has a similar population density to the Indian state of
279 Tamil Nadu, but did not resort to any full lockdown and has one of the largest and
280 best-organised epidemic control programmes in the world. The first dataset was col-
281 lected in the first half of 2020 (Database S4), while the second was collected during
282 the fourth community epidemic in the summer of 2021 (Database S5) in the context
283 of the assessment of transmission dynamics for the Delta variant of SARS-CoV-2.
284 The first dataset, which consists of $n = 5,165$ numbers of SARS-CoV-2 secondary
285 cases Z_i , was analysed in [3]. See Table 2.

286 We revisit the estimation of, and inference about, the underlying extreme value
287 index by comparing the D-GPD estimates with the classical GPD and Hill estimates.
288 Results are displayed in Fig. 4. A least squares fit to the first part of the mean excess
289 plot (Fig. 4 (b)) suggests a linearly increasing fit to the mean excess function with
290 a slope of around 0.85, but this ignores the flat or even slightly linearly decreasing
291 right-hand part of the data cloud. This throws the assumption that the offspring

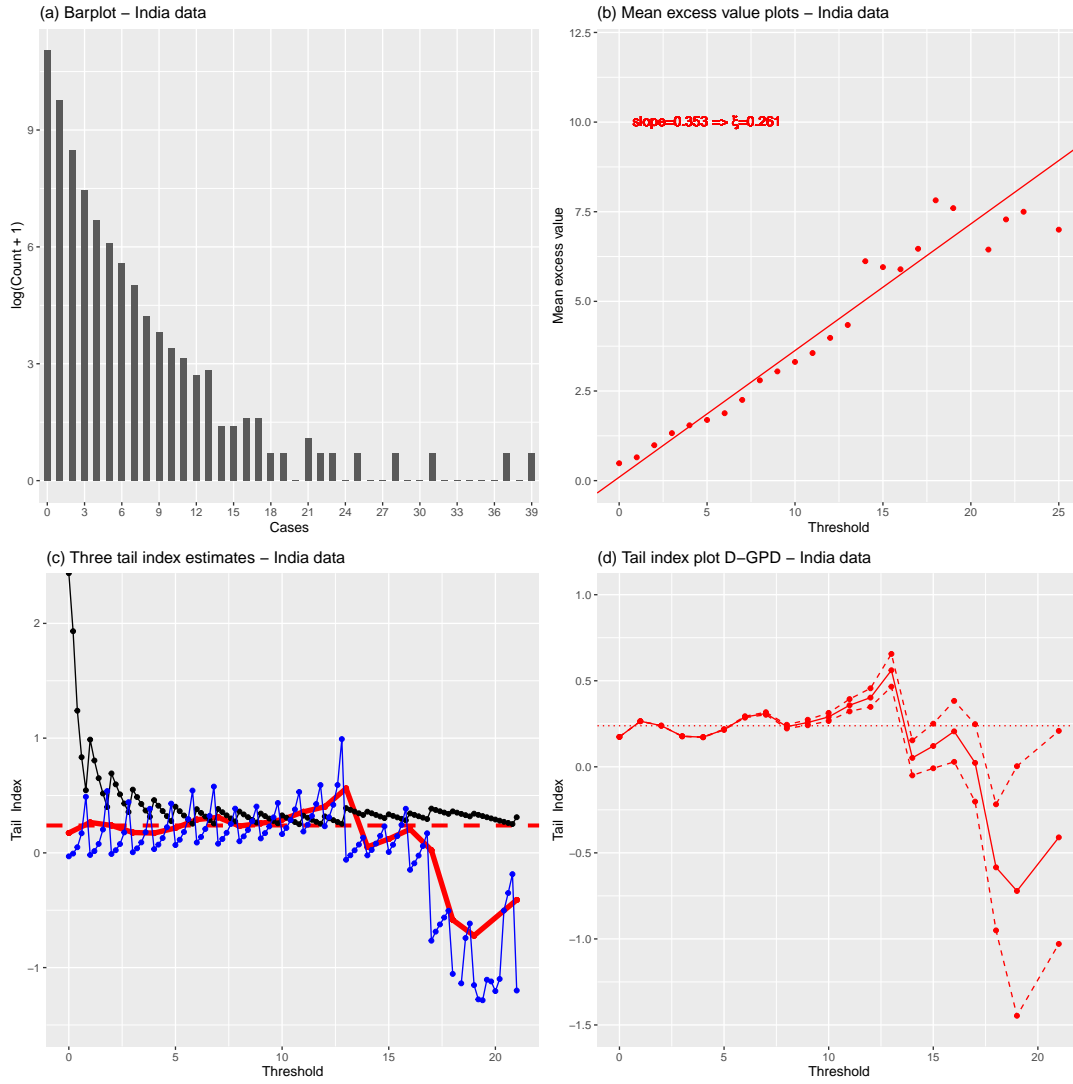


Figure 3: Secondary case data (Database S3) for SARS-CoV-2 from Andhra Pradesh and Tamil Nadu (India). (a) Bar plot of the $\log(Z_i + 1)$ ($n = 88,527$). (b) Mean excess plots of secondary cases. (c) Hill (solid black), continuous GPD maximum likelihood (solid blue) and discrete GPD maximum likelihood (solid bold red) estimates of ξ . (d) Discrete GPD maximum likelihood estimates of ξ (solid red) and their associated 90% confidence intervals (dashed red). In panels (c) and (d), the averaged discrete GPD estimate $\hat{\xi} = 0.239$ over the stable region $u \in [0, 10]$ is indicated with the horizontal red line.

Z	0	1	2	3	4	5	6	7	8	9
Count	4,558	364	114	62	27	7	7	4	4	1
Z	10	11	12	15	17	18	21	24	27	51
Count	2	3	1	2	2	1	2	2	1	1

Table 2: Secondary case data (Database S4) for SARS-CoV-2 collected in South Korea in the first half of 2020.

292 distribution is fat-tailed in doubt, although the barplot of the data (Fig. 4 (a)) would
 293 tentatively back the heavy tail assumption. The Hill estimator, which presupposes
 294 that the data is fat-tailed and graphed as a black line in Fig. 4 (c), does not ex-
 295 hibit any stable region which would allow to produce a reasonable point estimate.
 296 In such scenarios, best practice in extreme value theory requires calculating alter-
 297 native extreme value estimators whose consistency does not rest upon the heavy
 298 tail assumption (unlike the Hill estimator), such as the general GPD and D-GPD
 299 estimators. These are also represented in Fig. 4 (c). Clearly, the paths of these two
 300 estimates follow a similar trajectory which is very different from that of the Hill
 301 plot. They point towards substantially lower estimates of ξ , and even though the
 302 estimates are overall larger than 0, the validity of the heavy tail assumption $\xi > 0$
 303 is not obvious for this dataset. Fig. 4 (d) further supports this observation: in the
 304 (somewhat) stable region around the threshold $u = 10$, the 90% confidence interval
 305 produced through maximum likelihood theory contains the value 0. Our conclusion
 306 from the analysis of this dataset is that the distribution of the number of secondary
 307 cases is either fat-tailed but with a low extreme value index, or perhaps even has an
 308 exponential-type tail. As a consequence, our finding is qualitatively different from
 309 that of [3], since we do not obtain ξ estimates similar to those found by merging
 310 Datasets S1 and S2.

311 The second South Korean contact-tracing dataset comprises $n = 33,903$ SARS-
 312 CoV-2 numbers of secondary cases Z_i (Database S5) detected between 25th July
 313 2021 and 15th August 2021. It was initially explored in [5], where it was highlighted
 314 that the Delta variant accounted for the majority of those cases. We therefore inves-
 315 tigate this dataset to ascertain whether the tail behaviour of SSEs is substantially
 316 different for the Delta variant. The data is presented in Table 3 below. The re-
 317 sults we obtain for this dataset are displayed in Fig. 5. The barplot of the data in
 318 Fig. 5 (a) again backs the assumption of a heavy tail, but here, the mean excess plot
 319 in Fig. 5 (b) suggests a more convincing linearly increasing fit to the mean excess
 320 function with a slope of around 0.3. The Hill estimator and both continuous and
 321 discrete GPD maximum likelihood estimators, represented in Fig. 5 (c), appear to
 322 support the fat tail assumption of the offspring distribution which is mainly dom-
 323 inated here by the Delta variant. Once again, the D-GPD estimate has a much
 324 smoother and more stable sample path, with a stable zone over $u \in [1, 10]$ indicat-
 325 ing a point estimate of around 0.21. The 90% confidence interval of the D-GPD
 326 estimate over that region, provided in Fig. 5 (d), does not contain 0 and offers fur-
 327 ther justification of the assumption that the offspring distribution is fat-tailed in
 328 this dataset, in contrast to the 2020 South Korea data where the validity of this
 329 conclusion is much less clear.

Z	0	1	2	3	4	5	6	7	8	9	10		
Count	29,193	2,154	1,121	594	332	207	113	53	53	21	21		
Z	11	12	13	14	15	16	17	18	19	21	22	24	32
Count	6	8	5	3	3	2	2	3	3	1	2	2	1

Table 3: Secondary case data (Database S5) for SARS-CoV-2 collected in South Korea from 25th July 2021 to 15th August 2021.

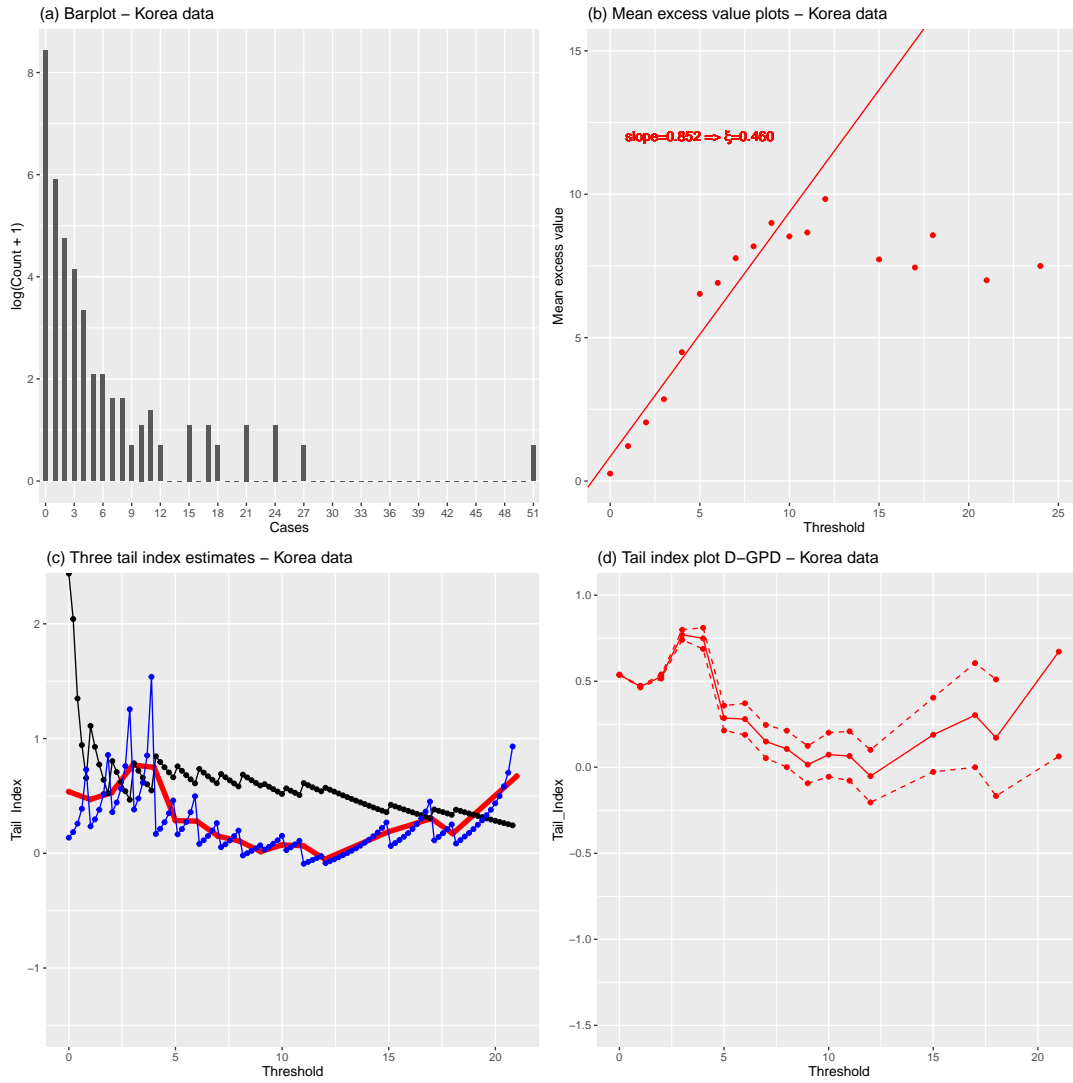


Figure 4: Secondary case data (Database S4) for SARS-CoV-2 from South Korea (first half of 2020). (a) Bar plot of the $\log(Z_i + 1)$ ($n = 5,165$). (b) Mean excess plots of secondary cases. (c) Hill (solid black), continuous GPD maximum likelihood (solid blue) and discrete GPD maximum likelihood (solid bold red) estimates of ξ . (d) Discrete GPD maximum likelihood estimates of ξ (solid red) and their associated 90% confidence intervals (dashed red).

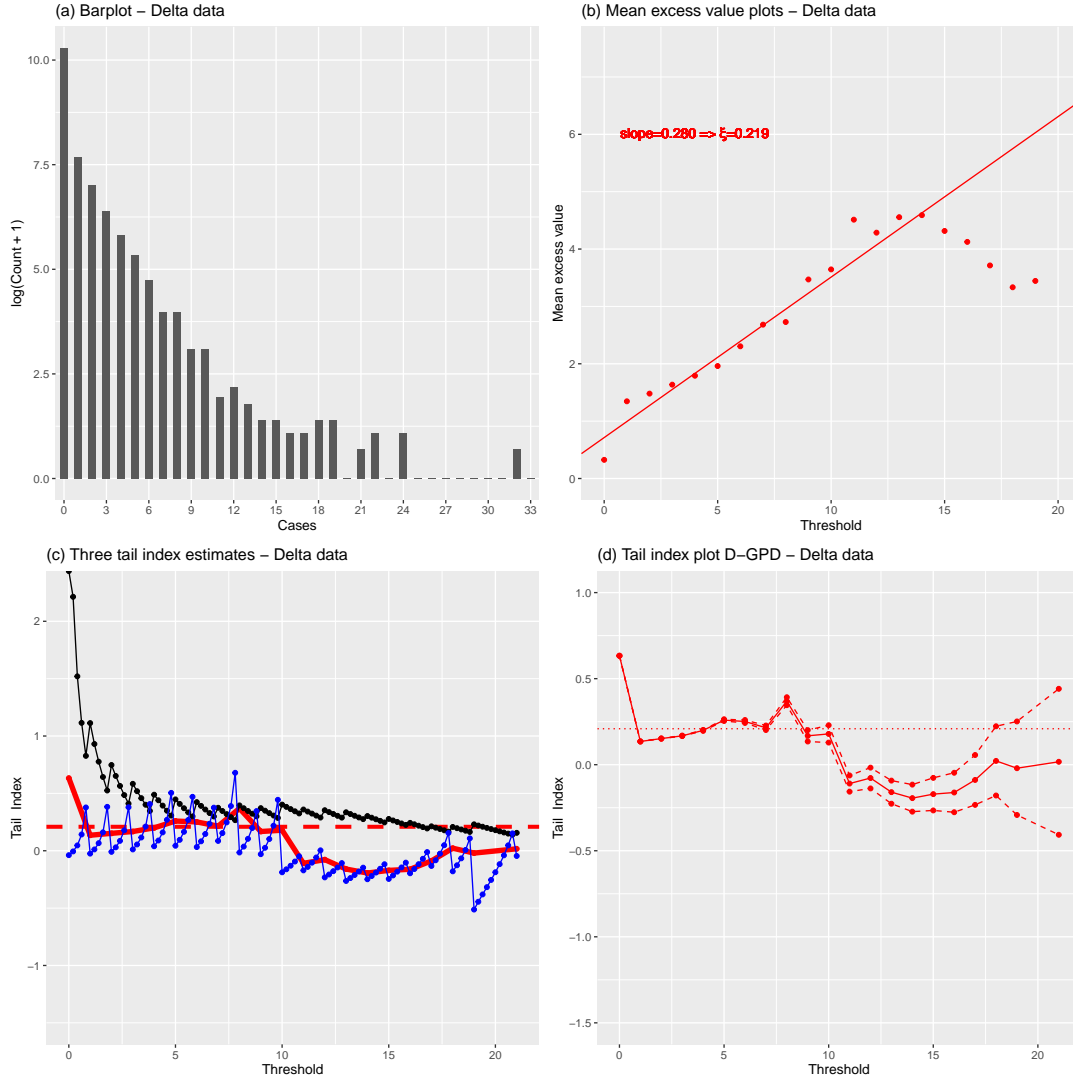


Figure 5: Secondary case data (Database S5) for SARS-CoV-2 from South Korea (July-August 2021). (a) Bar plot of the $\log(Z_i + 1)$ ($n = 33,903$). (b) Mean excess plots of secondary cases. (c) Hill (solid black), continuous GPD maximum likelihood (solid blue) and discrete GPD maximum likelihood (solid bold red) estimates of ξ . (d) Discrete GPD maximum likelihood estimates of ξ (solid red) and their associated 90% confidence intervals (dashed red). In panels (c) and (d), the averaged discrete GPD estimate $\hat{\xi} = 0.209$ over the stable region $u \in [1, 10]$ is indicated with the horizontal red line.

330 **Analysis of cluster size data.** We broaden our analysis by examining whether
331 SARS-CoV-2 cluster sizes are fat-tailed. We consider a database of 15 samples of
332 cluster sizes recorded in 11 countries and 4 US states. We define a cluster as a
333 local outbreak involving a minimum of two cases, including confirmed close contacts
334 with epidemiological linkage observed up to extinction of the outbreak. This differs
335 from the number of secondary cases linked to a single, given index case in an SSE,
336 since the cluster size is now the total number of infected people over the duration of
337 the outbreak. The number of reported clusters per country or state varies from 29
338 (France) to 4,769 (Colorado, USA). The database is constructed from government
339 reports [6, 7, 8, 9] (Database S6) and media sources [10] (Database S7). The median
340 cluster sizes were 5 (Database S6) and 33 (Database S7), and the largest clusters had
341 sizes 1,761 (Database S6, in a Colorado prison) and 7,000 (Database S7, in an Italian
342 football stadium). We denote by Y_i the number of SARS-CoV-2 cases in cluster i .
343 The ξ estimates from each sample of cluster sizes allow to infer the risk category of
344 the corresponding country/state in terms of local community transmission.

345 Figs. 6 and 7 display the D-GPD maximum likelihood estimates of ξ as functions
346 of the cluster size u . Eyeballed thresholds are indicated by the vertical dashed lines
347 in Figs. 6 and 7. The final selected estimates are reported in Table 4, where 13
348 out of the 15 countries or states appear to have fat-tailed cluster size distributions
349 (confirmed at the 90% confidence level except for China). We note that there is
350 strong variation in point estimates of ξ across countries and states. The low sample
351 sizes of the data available in each case (except for the two US states of Colorado
352 and Oregon) certainly play an important role in that variation. Heterogeneity in
353 population density and healthcare policies may also be substantial factors, although
354 this would have to be cross-checked using complete demographic and public health
355 data. The analysis for California and the United Kingdom was inconclusive. For the
356 California dataset, this is possibly due to a strong degree of heterogeneity (see the
357 histogram in the bottom left panel of Fig. 7). A stratified study of the Californian
358 data might be more conclusive. For the UK dataset, the fact that the sample is
359 so small (26 clusters) in a country with a highly developed healthcare and contact
360 tracing system is suspicious and may suggest reporting issues.

361 Using the D-GPD model, one can gain further insight into large cluster sizes
362 by providing extrapolated estimates of extreme percentiles q_α potentially beyond
363 the sample maximum, through the estimate \hat{q}_α described in the Methods section.
364 Estimated 95th and 99th percentiles are given in Table 4. One may also match
365 the estimated percentiles with actual observations to get a sense of what would
366 constitute a conducive environment for the formation of large SARS-CoV-2 clusters.
367 For example, the estimated 95th percentile of 120 cases in Kerala is close to two
368 clusters of 113 cases (nursing home) and 132 cases (local transmission) already
369 observed in Kerala. Likewise, the estimate $\hat{q}_{0.95} = 272$ cases in Canada is fairly
370 close to a cluster of 324 cases in Canadian nursing homes. In Oregon, the estimated
371 99th percentile $\hat{q}_{0.99} = 124$ cases is in the vicinity of a cluster of 134 cases in a care
372 home setting. In Colorado, the estimate $\hat{q}_{0.99} = 140$ cases is close to a cluster of 134
373 cases in a nursing home. All of these clusters bar one (the local transmission cluster
374 in Kerala) correspond to indoor environments where social distancing is difficult to
375 practice.

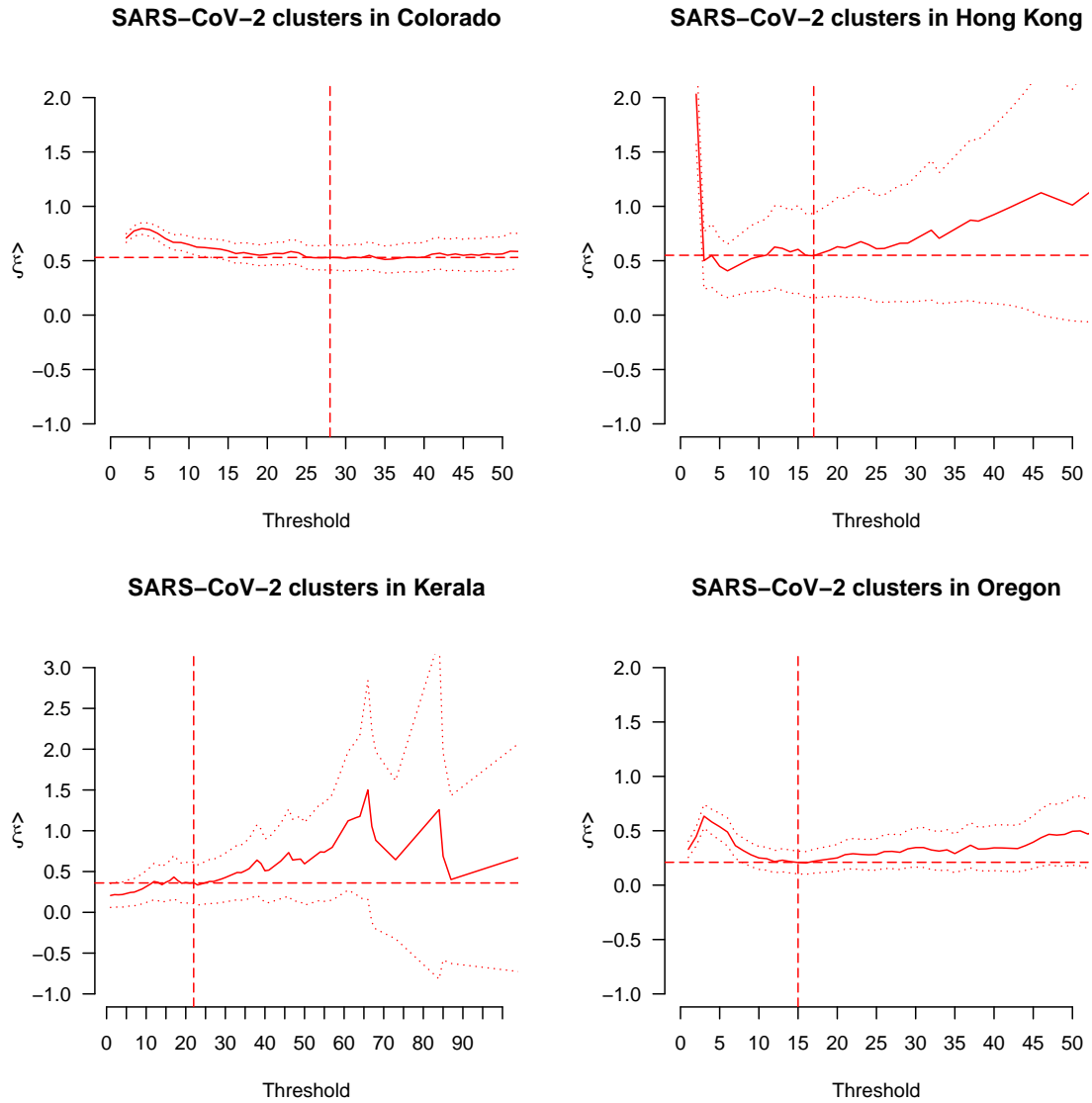


Figure 6: Analysis of cluster cases, for the four countries/states where the source is official data (Database S6). Plots of discrete GPD maximum likelihood estimates of ξ (solid lines), along with their 90% confidence intervals (dotted lines) and the final selected estimates (horizontal dashed lines) and thresholds (vertical dashed lines).

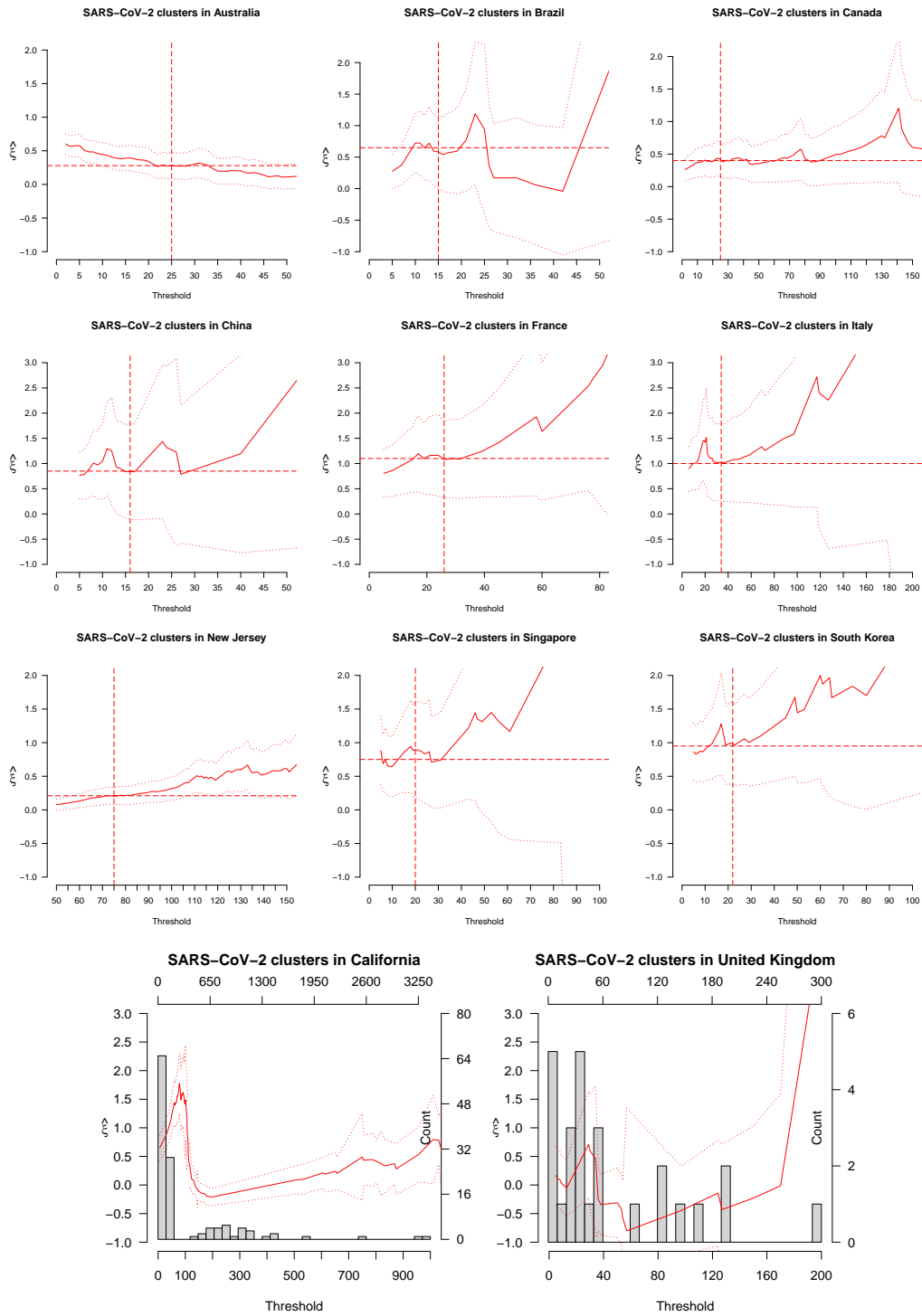


Figure 7: Analysis of cluster cases as in Fig. 6, with the results obtained from the data whose sources were the media (Database S7). The top 9 plots refer to those countries and states for which the extreme value analysis was conclusive. The bottom 2 plots refer to those for which the extreme value analysis was inconclusive.

Database S6

Location	n	$\hat{\xi}$ [90% 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 S7

Location	n	$\hat{\xi}$ [90% 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 4: 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 $Y_i \geq u$ 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 (Database S6), and the bottom table to data from media sources (Database S7). The results reported in the latter table only concern the 9 countries and states for which the extreme value analysis was conclusive.

376 Discussion

377 In summary, we have investigated four datasets of secondary case numbers Z_i for
378 SARS-CoV-2 as a way to estimate and infer the extreme value index of the related
379 underlying offspring distribution. Motivated by the highly discrete nature of such
380 data, we used the Discrete GPD (D-GPD) maximum likelihood estimation method
381 which produces smoother and more stable plots of the associated D-GPD estimator
382 than the classical continuous GPD and Hill estimators. We first provided evidence
383 that the small SSE dataset (Dataset S2) compiled by [3] during the early phase of
384 the COVID-19 pandemic was fat-tailed, thus confirming their findings, although we
385 show in various ways that this dataset should not be pooled with their 15 SSEs
386 associated with SARS-CoV (Dataset S1), since they correspond to substantially dif-
387 ferent distributions. On the other hand, as accurate extreme value inference requires
388 a large sample size in general, we also analysed an Indian secondary case dataset
389 of size 88,527 collected in 2020 (Database S3), which contains a very large number
390 of tied observations. The D-GPD estimate of the extreme value index is around
391 0.24, which is in full agreement with the estimate of around 0.25 found by revisiting
392 the small SSE dataset of size 45 from [3]. The distribution of SARS-CoV-2 SSEs
393 therefore appears to have at least a finite third moment, whereas that of SARS-CoV
394 SSEs is found to have a much heavier upper tail with infinite variance and therefore
395 stronger superspreading effect. In an effort to account for the quality of implemented
396 control programmes as well as the nature of the variant under study, we used two
397 extra South Korean contact-tracing datasets. For the first dataset (Database S4),
398 collected in the first half of 2020 and used in [3], we cannot disprove that the distri-
399 bution of the number of secondary cases has an exponential-type tail. By contrast,
400 for the second South Korean dataset (Database S5) collected during the summer of
401 2021, in which the majority of cases correspond to the Delta variant of SARS-CoV-
402 2 [5], we obtained a D-GPD estimate, $\hat{\xi} \approx 0.21$ clearly suggesting a heavier upper
403 tail for the Delta variant and therefore more pronounced superspreading potential
404 in South Korea relative to the first half of 2020.

405 We broaden our analysis by providing evidence that SARS-CoV-2 cluster sizes
406 are typically fat-tailed, based on 15 samples from 11 countries and 4 US states.
407 We infer the risk exposure and risk category of each country and state by making
408 use of D-GPD maximum likelihood estimates of both the extreme value index and
409 extreme percentiles, along with their associated confidence intervals. For the sake
410 of simplicity, we used a straightforward threshold selection rule, which is to spot a
411 stability region in the estimates (as a function of the threshold value) and choose an
412 estimate whose value is representative of those reached in this region. This practice,
413 colloquially known as “eyeballing”, is standard in applied extreme value analysis:
414 see for example the discussion in p.77 of Chapter 4 in [26]. It applies reasonably
415 well to the D-GPD sample paths, because they are overall much smoother and more
416 stable than the standard Hill and GPD maximum likelihood sample paths, which
417 are not designed to handle the discreteness of the data. The development of more
418 elaborate statistical techniques for the choice of threshold in discrete GPD maximum
419 likelihood estimation, such as methods based on asymptotic MSE minimisation or
420 the bootstrap in the spirit of the approaches outlined in Section 5.4 of [27] for Hill

421 estimation, is an open question which is beyond the scope of this paper.

422 A limitation of our study lies in the quality of the data, as it is not obvious
423 whether all SSEs or clusters over a given time period were available, or whether
424 cluster sizes were correctly recorded. To check robustness against missing data, we
425 have reproduced part of our analysis of cluster data by removing 10% of observa-
426 tions at random in each sample containing at least 100 data points, and replicating
427 this experiment 10,000 times. Robustness against poor recording was checked by
428 multiplying each observation Y_i by an independent normal variate W_i having mean
429 $\mu = 1$ and standard deviation $\sigma = 0.05$, and then reproducing our analysis of cluster
430 data on the $Y'_i = W_i Y_i$, this experiment being again replicated 10,000 times. There
431 is indeed some variation in the resulting estimates of ξ (Figs. 8 and 9), but this
432 does not affect our conclusion on the fat-tailed behaviour of the data, except in rare
433 situations when almost all the large values in the data go missing. This highlights
434 the importance of accurate data reporting as a prerequisite to such analyses. A fur-
435 ther limitation lies in the assumption of independent data that is implicitly made in
436 order to derive confidence intervals for extreme value parameters, even though the
437 data are implicitly time series. Handling serial dependence in the current setting of
438 discrete epidemiological data is obviously an interesting but very difficult question,
439 involving the hitherto open problem of extreme value dependence in discrete time
440 series, which deserves a study of its own.

441 It should be noted that, in classical epidemiological models, accurate estimation
442 of the basic reproduction number R_0 is of crucial importance as it informs the
443 extent of restrictions on social interactions and other control measures that should
444 be imposed to terminate the spread of an epidemic. The range of R_0 for SARS-CoV-
445 2 has been revised in [28] to 4.7-11.4, which is considerably higher than most early
446 estimates. This might explain why moderate restrictions that were implemented in
447 some nations, e.g. France, Italy, Spain, the UK, Australia and New Zealand, turned
448 out to be insufficient and replaced by nationwide or statewide lockdowns and/or
449 border closures. It should be clear that our results are, by construction, robust
450 to misspecified estimates of the expected number of secondary cases R_0 since they
451 solely rely on extreme values of numbers of secondary cases.

452 Our approach can be viewed as a proof of concept that transmission data from
453 a respiratory disease should not be pooled with data from a similar disease, since
454 similar R_0 numbers or parameters of average transmission do not, in general, induce
455 similar parameters of large community transmission. As such, preparing proactive
456 control measures actually requires a fine assessment of how unequal the distributions
457 of SSEs associated with different SARS-CoV-2 variants are. [29] conclude that the
458 reproductive number of the Delta variant is far higher than that of the historical
459 SARS-CoV-2 virus. Similarly, [30] estimate that the effective reproduction number
460 of the Omicron variant is more than 3 times that of the Delta variant in Denmark.
461 Our analysis of secondary case data did not, strictly speaking, allow one to conclude
462 statistically that SSEs linked to the Delta variant had a different extreme value index
463 from those linked to the original strains of SARS-CoV-2. However, in the contact-
464 tracing data recorded in South Korea, we did find a heavy tail in the offspring
465 distribution when the Delta variant made the majority of cases, as opposed to when
466 it did not. This tentative finding of a heavier tail in the data linked to the Delta

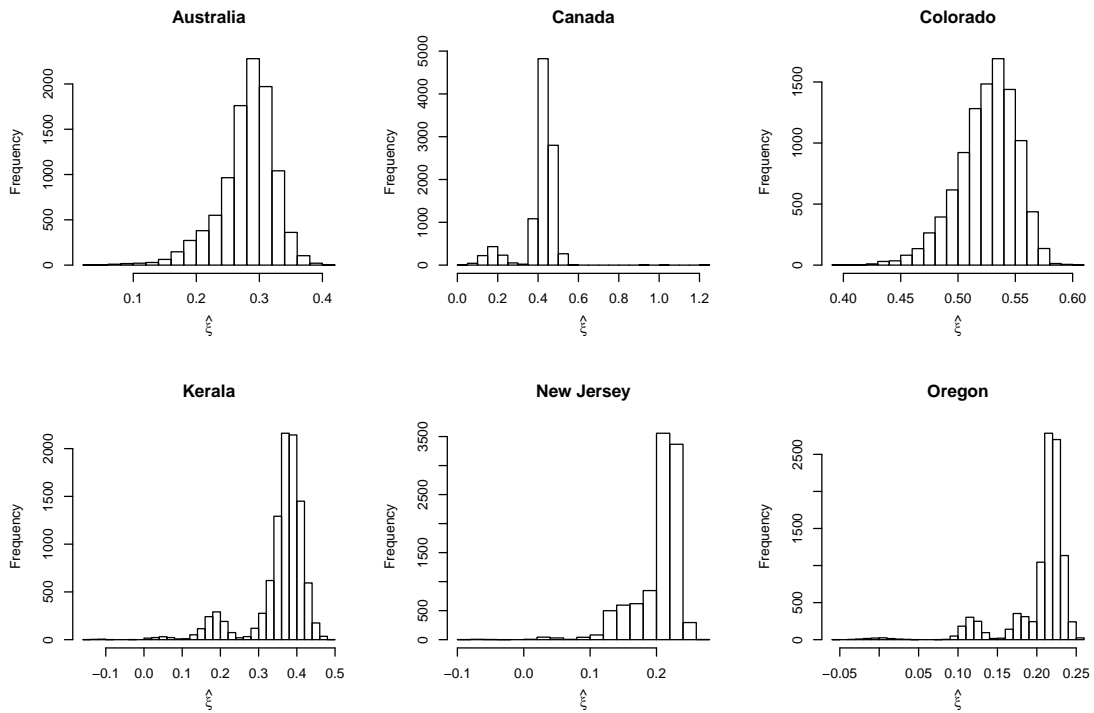


Figure 8: Robustness check (with respect to data omission) for the analysis of cluster cases (Databases S6 and S7). Histograms of the 10,000 estimates of ξ obtained by omitting at random 10% of the data. This was done only for the six samples containing at least 100 data points.

467 variant is coherent with the higher reproductive number of the Delta variant found
 468 in [29]. The question of estimating parameters of large community transmission for
 469 the Omicron variant remains open, as we could not find a dataset whose sample size
 470 would enable us to draw statistically principled conclusions about the tail behaviour
 471 of Omicron-related SSEs.

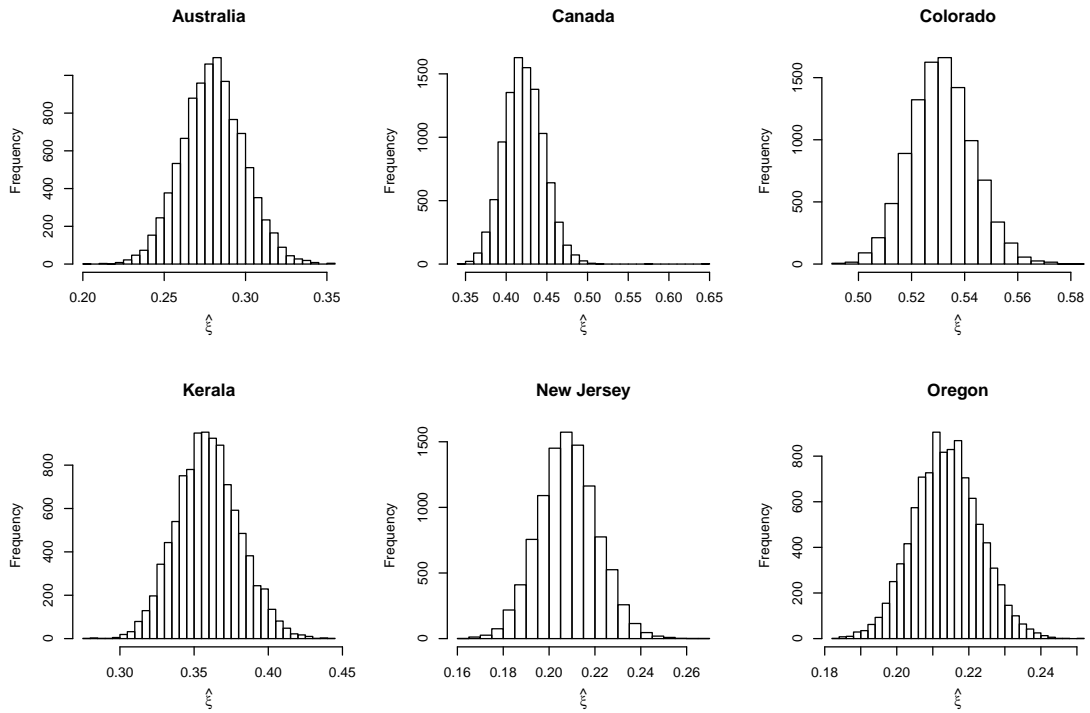


Figure 9: Robustness check (with respect to poor recording of the data) for the analysis of cluster cases (Databases S6 and S7). Histograms of the 10,000 estimates of ξ obtained by multiplying each data point by a random draw from the normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 0.05$. This was done only for the six samples containing at least 100 data points.

472 **Ethics.** This article does not present research with ethical considerations.

473 **Data accessibility.** Data and relevant code for this research work are stored
474 in GitHub: <https://github.com/AntoineUC/SARS-CoV-2-codes> and have been
475 archived within the Zenodo repository: <https://doi.org/10.5281/zenodo.7509725>.

476 **Authors' contributions.** A.U.C. undertook data curation and wrote the code for
477 the statistical analysis and visualisation of the results. All three authors participated
478 in the statistical analysis of the data and in drafting and revising the manuscript.

479 **Competing interests.** The authors declare no competing interests.

480 **Funding.** This research was supported by the French National Research Agency
481 (grant numbers ANR-19-CE40-0013, ANR-17-EURE-0010), the TSE-HEC ACPR
482 Chair and an AXA Research Fund Award on 'Mitigating risk in the wake of the
483 COVID-19 pandemic'.

484 **Acknowledgements.** The authors acknowledge an anonymous Associate Editor
485 and three anonymous reviewers for their helpful comments.

486 References

- 487 [1] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and
488 the effect of individual variation on disease emergence. *Nature* **438**, 355-359 (2005).
489 (doi:10.1038/nature04153)
- 490 [2] D. C. Adam, P. Wu, J. Y. Wong, E. H. Y. Lau, T. K. Tsang, S. Cauchemez, G. M. Leung,
491 B. J. Cowling, Clustering and superspreading potential of SARS-CoV-2 infections in Hong
492 Kong. *Nat. Med.* **26**, 1714-1719 (2020). (doi:10.1038/s41591-020-1092-0)
- 493 [3] F. Wong, J. J. Collins, Evidence that coronavirus superspreading is fat-tailed. *PNAS* **117**,
494 29416-29418 (2020). (doi:10.1073/pnas.2018490117)
- 495 [4] R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. Mohan B, S. Neelima, K. S. Jawa-
496 har Reddy, J. Radhakrishnan, J. A. Lewnard, Epidemiology and transmission dynamics of
497 COVID-19 in two Indian states. *Science* **370**, 691-697 (2020). (doi:10.1126/science.abd7672)
- 498 [5] S. Ryu, D. Kim, J.-S. Lim, S. T. Ali, B. J. Cowling, Serial interval and transmission dynamics
499 during SARS-CoV-2 Delta variant predominance, South Korea. *Emerg. Infect. Dis.* **28**, 407-
500 410 (2022). (doi:10.3201/eid2802.211774)
- 501 [6] State of Colorado, <https://covid19.colorado.gov/covid19-outbreak-data>, last updated
502 on 2 June 2021 (resolved outbreaks only). Accessed 27th September 2021.
- 503 [7] Government of Hong Kong, [https://www.chp.gov.hk/files/pdf/local_situation_](https://www.chp.gov.hk/files/pdf/local_situation_covid19_en.pdf)
504 [covid19_en.pdf](https://www.chp.gov.hk/files/pdf/local_situation_covid19_en.pdf), last updated on 6 September 2021. Accessed 6th September 2021.
- 505 [8] Government of Kerala, <https://covid19jagratha.kerala.nic.in/home/clusterList>. Ac-
506 cessed 21st July 2021.
- 507 [9] State of Oregon, [https://www.oregon.gov/oha/covid19/Documents/DataReports/](https://www.oregon.gov/oha/covid19/Documents/DataReports/Weekly-Outbreak-COVID-19-Report-2021-08-25-FINAL.pdf?utm_medium=email&utm_source=govdelivery)
508 [Weekly-Outbreak-COVID-19-Report-2021-08-25-FINAL.pdf?utm_medium=email&utm_](https://www.oregon.gov/oha/covid19/Documents/DataReports/Weekly-Outbreak-COVID-19-Report-2021-08-25-FINAL.pdf?utm_medium=email&utm_source=govdelivery)
509 [source=govdelivery](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.

- 510 [10] K. Swinkels, SARS-CoV-2 Superspreading Events Database, [https://kmswinkels.](https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b)
511 [medium.com/covid-19-superspreading-events-database-4c0a7aa2342b](https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b). Accessed 21st
512 July 2021.
- 513 [11] T. Shimura, Discretization of distributions in the maximum domain of attraction. *Extremes*,
514 **15**, 299-317 (2012). (doi:10.1007/s10687-011-0137-7)
- 515 [12] F. Prieto, E. Gómez-Déniz, J. M. Sarabia, Modelling road accident blackspots data with the
516 discrete generalized Pareto distribution. *Accident Analysis & Prevention* **49**, 71:38 (2014).
517 (doi:10.1016/j.aap.2014.05.005)
- 518 [13] A. Hitz, R. Davis, G. Samorodnitsky, Discrete Extremes. arXiv [Preprint] (2017). <https://arxiv.org/abs/1707.05033>
519 (doi:10.48550/arXiv.1707.05033)
- 520 [14] S. Ranjbar, E. Cantoni, V. Chavez-Demoulin, G. Marra, R. Radice, K. Jatton, Modelling the
521 extremes of seasonal viruses and hospital congestion: the example of flu in a Swiss hospital.
522 *J. Roy. Stat. Ser. C* **71**, 884-905 (2022). (doi:10.1111/rssc.12559)
- 523 [15] L. de Haan, A. Ferreira, *Extreme Value Theory: An Introduction*, Springer-Verlag, New York
524 (2006). (doi:10.1007/0-387-34471-3)
- 525 [16] B. M. Hill, A simple general approach to inference about the tail of a distribution. *Ann.*
526 *Statist.* **3**, 1163-1174 (1975). (doi:10.1214/aos/1176343247)
- 527 [17] J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, *Statistics of Extremes: Theory and Applica-*
528 *tions*, John Wiley & Sons, Chichester (2004).
- 529 [18] S. Ghosh, S. Resnick, A discussion on mean excess plots. *Stoch. Proc. Appl.* **120**, 1492-1517
530 (2010). (doi:10.1016/j.spa.2010.04.002)
- 531 [19] C. Zhou, The extent of the maximum likelihood estimator for the extreme value index. *J.*
532 *Multivariate Anal.* **101**, 971-983 (2010). (doi:10.1016/j.jmva.2009.09.013)
- 533 [20] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, London
534 (2004). (doi:10.1007/978-1-4471-3675-0)
- 535 [21] C. M. Goldie, R. L. Smith, Slow variation with remainder: Theory and applications. *Quart.*
536 *J. Math. Oxford* **38**, 45-71 (1987). (doi: 10.1093/qmath/38.1.45)
- 537 [22] P. Cirillo, N. N. Taleb, Tail risk of contagious diseases. *Nat. Phys.* **16**, 606-613 (2020).
538 (doi:10.1038/s41567-020-0921-x)
- 539 [23] C. Kremer, A. Torneri, S. Boesmans, H. Meuwissen, S. Verdonshot, K. Vanden Driessche,
540 C. L. Althaus, C. Faes, N. Hens, Quantifying superspreading for COVID-19 using Poisson
541 mixture distributions. *Sci. Rep.* **11**, 14107 (2021). (doi:10.1038/s41598-021-93578-x)
- 542 [24] N. Islam, Q. Bukhari, Y. Jameel, S. Shabnam, A. M. Erzurumluoglu, M. A. Siddique, J. M.
543 Massaro, R. B. D'Agostino, COVID-19 and climatic factors: A global analysis. *Environmental*
544 *Research* **193**, 110355 (2021). (doi:10.1016/j.envres.2020.110355)
- 545 [25] H. Hwang, J.-S. Lim, S.-A. Song, C. Achangwa, W. Sim, G. Kim, S. Ryu, Transmission
546 dynamics of the Delta variant of SARS-CoV-2 infections in South Korea. *J. Infect. Dis.* **225**,
547 793-799 (2022). (doi:10.1093/infdis/jiab586)
- 548 [26] M. Jacob, C. Neves, D. Vukadinović Greetham, *Extreme Value Statistics*, in: Forecasting and
549 Assessing Risk of Individual Electricity Peaks. Mathematics of Planet Earth, Springer, Cham
550 (2020). (doi:10.1007/978-3-030-28669-9)
- 551 [27] M. I. Gomes, A. Guillou, Extreme value theory and statistics of univariate extremes: A review.
552 *Int. Stat. Rev.* **83**, 263-292 (2015). (doi:10.1111/insr.12058)
- 553 [28] M. Kočańczyk, F. Grabowski, T. Lipniacki, Super-spreading events initiated the exponential
554 growth phase of COVID-19 with \mathcal{R}_0 higher than initially estimated. *Royal Society Open*
555 *Science* **7**, 200786 (2020). (doi:10.1098/rsos.200786)

- 556 [29] Y. Liu, J. Rocklöv, The reproductive number of the Delta variant of SARS-CoV-2 is far
557 higher compared to the ancestral SARS-CoV-2 virus. *J. Travel Med.* **28**, taab124 (2021).
558 (doi:10.1093/jtm/taab124)
- 559 [30] K. Ito, C. Piantam, H. Nishiura, Relative instantaneous reproduction number of Omicron
560 SARS-CoV-2 variant with respect to the Delta variant in Denmark. *J. Med. Virol.* **94**, 2265-
561 2268 (2022). (doi:10.1002/jmv.27560)