RESEARCH ARTICLE

Anti-clustering in the national SARS-CoV-2 daily infection counts

Boudewijn F. Roukema[a]

[a]Institute of Astronomy, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Grudziadzka 5, 87-100 Toruń, Poland; [b]Univ Lyon, Ens de Lyon, Univ Lyon1, CNRS, Centre de Recherche Astrophysique de Lyon UMR5574, F–69007, Lyon, France

ARTICLE HISTORY
Compiled January 11, 2021

KEYWORDS
COVID-19, Data validation, Poisson point process

ABSTRACT
The noise in daily infection counts of an epidemic should be super-Poissonian due to intrinsic epidemiological and administrative clustering. Here, we use this clustering to classify the official national SARS-CoV-2 daily infection counts and check for infection counts that are unusually anti-clustered. We adopt a one-parameter model of ϕ_i' infections per cluster, dividing any daily count n_i into n_i/ϕ_i' 'clusters', for 'country' i. We assume that n_i/ϕ_i' on a given day j is drawn from a Poisson distribution whose mean is robustly estimated from the 4 neighbouring days, and calculate the inferred Poisson probability P_{ij}' of the observation. The P_{ij}' values should be uniformly distributed. We find the value ϕ_i that minimises the Kolmogorov–Smirnov distance from a uniform distribution. We investigate the (ϕ_i, N_i) distribution, for total infection count N_i. While all the daily infection count sequences are found to be consistent with the ϕ_i model, the 28-, 14- and 7-day least noisy sequences for several countries are best modelled as sub-Poissonian, suggesting a distinct epidemiological family. The 28-day sequences of Algeria, Belarus, Turkey, and the UAE have strongly sub-Poissonian preferred models, with $\phi_i^{28} < 0.5$; these are difficult to explain naturally.

1. Introduction

The daily counts of new, laboratory-confirmed infections with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) constitute one of the key statistics followed by citizens and health agencies around the world in the ongoing 2019–2020 coronavirus disease 2019 (COVID-19) pandemic[10, 22]. Can these counts be classified in a way that makes as few epidemiological assumptions as possible, as motivation for deeper analysis to either validate or invalidate the counts? While full epidemiological modelling and prediction is a vital component of COVID-19 research[3, 7, 13, 16, 25], these cannot be accurately used to study the pandemic as a whole – a global phenomenon by definition – if the data at the global level is itself inaccurate. Knowledge of the global state of the current pandemic is weakened if any of the national-level SARS-CoV-2 infection data have been artificially interfered with by the health agencies providing that data

CONTACT Boud Roukema. Email: boud astro.umk.pl

or by other actors involved in the chain of data lineage [37]. Since personal medical data are private information, only a limited number of individuals at health agencies are expected to be able to check the validity of these counts based on original records. Nevertheless, artificial interventions in the counts could potentially reveal themselves in statistical properties of the counts. Unusual statistical properties in a wide variety of quantitative data sometimes appear, for example, as anomalies related to Benford's law [26, 27], as in the 2009 first round of the Iranian presidential election [23, 31, 32]. Benford's law analysis has been used to argue that countries with higher democracy indices, high gross domestic product, and better health system indices tend to have a lower probability of having manipulated their key COVID-19 related cumulative counts (confirmed cases and deaths [5]) For other Benford's law COVID-19 count analyses, see [17, 20]. For the politics of organisational strategies regarding open government data, see [33].

Here, we check the compatibility of noise in the official national SARS-CoV-2 daily infection counts, $N_i(t)$, for country[1] i on date t, with expectations based on the Poisson distribution [28]. It is unlikely that any real count data will quite match the theoretical Poisson distribution, both due to the complexity of the logical tree of time-dependent intrinsic epidemiological infection as well as administrative effects in the SARS-CoV-2 testing procedures, and the sub-national and national level procedures for collecting and validating data to produce a national health agency's official report. In particular, clusters of infections on a scale of ϕ_i' infections per cluster, either intrinsic or in the testing and administrative pipeline, would tend to cause relative noise to increase from a fraction of $1/\sqrt{N_i}$ for pure Poisson noise up to $\sqrt{\phi_i'/N_i}$, greater by a factor of $\sqrt{\phi_i'}$. This overdispersion has been found, for example, for COVID-19 death rate counts in the United States [16].

In contrast, it is difficult to see how anti-Poissonian smoothing effects could occur, unless they were imposed administratively. For example, an administrative office might impose (or have imposed on it by political authorities) a constraint to validate a fixed or slowly and smoothly varying number of SARS-CoV-2 test result files per day, independently of the number received or queued; this would constitute an example of an artificial intervention in the counts that would weaken the epidemiological usefulness of the data.

A one-parameter model to allow for the clustering is proposed in this paper, and used to classify the counts. We allow the parameter to take on an effective anti-clustering value, in order to allow the data to freely determine its optimal value. For more in-depth models of clustering, called "burstiness" in stochastic models of discrete event counts, power-law models have also been proposed [6, 9].

The method is presented in §2. Section §2.1 describes the choice of data set and the definition, for any given country, of a consecutive time sequence that has high enough daily infection counts for Poisson distribution analysis to be reasonable. The method of analysis is given in §2.2. Results are presented in §3. Qualitative discussion of the results is given in §4 and conclusions are summarised in §5. This work is intended to be fully reproducible by independent researchers using the MANEAGE framework; it was produced using commit f7e7b46 of the GIT repository `https://codeberg.org/boud/subpoisson` and the archive zenodo.4432080, on a computer with Little Endian x86_64 architecture.

[1]No position is taken in this paper regarding jurisdiction over territories; the term 'country' is intended here as a neutral term without supporting or opposing the formal notion of state. Apart from minor changes for technical reasons, the 'countries' are defined by the data sources.

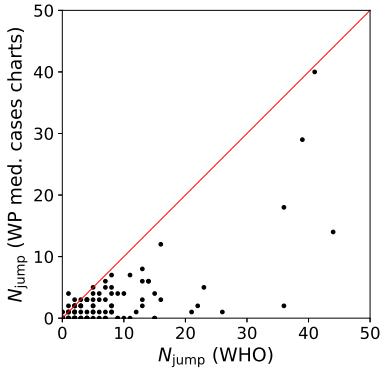

Figure 1. Number N_{jump} of sudden jumps or drops in counts on adjacent days in WHO and Wikipedia WikiProject COVID-19 Case Count Task Force medical cases chart national daily SARS-CoV-2 infection counts for countries present in both data sets. A line illustrates equal quality of the two data sets. The C19CCTF version of the data is clearly less affected by sudden jumps than the WHO data. Plain text table: zenodo.4432080/WHO_vs_WP_jumps.dat.

2. Method

2.1. SARS-CoV-2 infection data

Two obvious choices of a dataset for national daily SARS-CoV-2 counts would be those provided by the World Health Organization (WHO)[2] or those curated by the Wikipedia WikiProject COVID-19 Case Count Task Force[3] in medical cases chart templates (hereafter, C19CCTF). While WHO has published a wide variety of documents related to the COVID-19 pandemic, it does not appear to have published details of how national reports are communicated to it and collated. Given that most government agencies and systems of government procedures tend to lack transparency, despite significant moves towards forms of open government[39] in many countries, data lineage tracing from national governments to WHO is likely to be difficult in many cases. In contrast, the curation of official government SARS-CoV-2 daily counts by the Wikipedia WikiProject COVID-19 Case Count Task Force follows a well-established technology of tracking data lineage. The Wikipedia community high-tempo collaborative editing that has taken place in response to the COVID-19 pandemic is well quantified[15]. The John Hopkins University Center for Systems Science and Engineering curated set of official COVID-19 data is discussed below.

Unfortunately, it is clear that in the WHO data, there are several cases where two days' worth of detected infections appear to be listed by WHO as a sequence of two days j and $j+1$ on which all the infections are allocated to the second of the two days, with zero infections on the first of the pair. There are also some sequences in the WHO data where the day listed with zero infections is separated by several days from a nearby day with double the usual amount of infections. This is very likely an effect of difficulties in correctly managing world time zones, or time zone and sleep schedule effects, in any of several levels of the chains of communication between health agencies and WHO. In other words, there are several cases where a temporary sharp jump or drop in the counts appears in the data but is most likely a timing artefact. Whatever

[2] `https://covid19.who.int/WHO-COVID-19-global-data.csv`; (archive)

[3] `https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_COVID-19/Case_Count_Task_Force&oldid=967874960`

the reason for the effect, this effect will tend to confuse the epidemiological question of interest here: the aim is to globally characterise the noise and to highlight countries where unusual smoothing may have taken place.

We quantify this jump/drop problem as follows. We consider a pair of days j, $j+1$ for a given country to be a jump if the absolute difference in counts, $|n_i(j+1)-n_i(j)|$, is greater than the mean, $(n_i(j+1)+n_i(j))/2$. In the case of a pair in which one value is zero, the ratio is two, and the condition is satisfied. We evaluate the number of jumps N_{jump} for both the WHO data and the C19CCTF medical cases chart data, starting, for any given country, from the first day with at least 50 infections. Figure 1 shows N_{jump} for the 130 countries in common to the two data sets; there are 216 countries in the WHO data set and 132 in the C19CCTF data. It is clear that most countries have fewer jumps or drops in the Wikipedia data set than in the WHO data set.

Thus, at least for the purposes of understanding intrinsic and administrative clustering, the C19CCTF medical cases chart data appear to be the better curated version of the national daily SARS-CoV-2 infection counts as reported by official agencies. The detailed download and extraction script of national daily SARS-CoV-2 infection data from these templates and the resulting data file zenodo.4432080/WP_C19CCTF_SARSCoV2.dat are available in the reproducibility package associated with this paper (§Code availability). Dates without data are omitted; this should have an insignificant effect on the analysis if these are due to low infection counts.

Another global collection of daily SARS-CoV-2 counts that could be considered is the John Hopkins University Center for Systems Science and Engineering (JHU CSSE) git repository. Unfortunately, for several countries, the JHU CSSE data are provided for sub-national divisions rather than as official national statistics, making the dataset inhomogeneous for the purposes of this study. Artificial interference in the data at the national level will not be shown in data that is the sum of data obtained directly from sub-national geographical/political divisions. Moreover, detailed data provenance analysis (which exact government URL did a particular count come from? where is the archived version of the data of the original URL?) appears to be more difficult for the JHU CSSE data than for the C19CCTF data. Nevertheless, for completeness, the JHU CSSE data is analysed using the same method as the main analysis, with results presented as tables in Appendix A.

The full set of C19CCTF data includes many days, especially for countries or territories (as defined by the data source) of low populations, with low values, including zero and one. The standard deviation of a Poisson distribution[28] of expectation value N is $\sqrt{N}$, giving a fractional error of $1/\sqrt{N}$. Even taking into account clustering or anticlustering of data, inclusion of these periods of close to zero infection counts would contribute noise that would overwhelm the signal from the periods of higher infection rates for the same or other countries. In the time sequences of SARS-CoV-2 infection counts, chaos in the administrative reactions to the initial stages of the pandemic will tend to create extra noise, so it is reasonable to choose a moderately high threshold at which the start and end of a consecutive sequence of days should be defined for analysis. Here, we set the threshold for a sequence to start at a minimum of 50 infections in a single day. The sequence is continued for at least 7 days (if available in the data), and stops when the counts drop below the same threshold for 2 consecutive days. The cutoff criterion of 2 consecutive days avoids letting the analysable sequence be too sensitive to individual days of low fluctuations. If the resulting sequence includes less than 7 days, the sequence is rejected as having insufficient signal to be analysed.

2.2. Analysis

2.2.1. Poissonian and ϕ'_i models: full sequences

We first consider the full count sequence $\{n_i(j), 1 \leq j \leq T_i\}$ for each country i, with T_i valid days of analysis as defined in §2.1. Our one-parameter model assumes that the counts are predominantly grouped in clusters, each with ϕ'_i infections per cluster. Thus, the daily count $n_i(j)$ is assumed to consist of $n_i(j)/\phi'_i$ infection events. We assume that $n_i(j)/\phi'_i$ on a given day is drawn from a Poisson distribution of mean $\widehat{\mu}_i(j)/\phi'_i$. We set $\widehat{\mu}_i(j)$ to the median of the 4 neighbouring days, excluding day j and centred on it. For the initial sequence of 2 days, $\widehat{\mu}_i(j)$ is set to $\widehat{\mu}_i(3)$, and $\widehat{\mu}_i(j)$ for the final 2 days is set to $\widehat{\mu}_i(T_i - 2)$. By modelling $\widehat{\mu}_i$ as a median of a small number of neighbouring days, our model is almost identical to the data itself and statistically robust, with only mild dependence on the choices of parameters. This definition of a model is more likely to bias the resulting analysis towards underestimating the noise on scales of several days rather than overestimating it; this method will not detect oscillations on the time scale of a few days to a fortnight that are related to the SARS-CoV-2 incubation time [11]. For any given value ϕ'_i, we calculate the cumulative probability P'_{ij} that $n_i(j)/\phi'_i$ is drawn from a Poisson distribution of mean $\widehat{\mu}_i(j)/\phi'_i$. For country i, the values P'_{ij} should be drawn from a uniform distribution if the model is a fair approximation. In particular, for ϕ'_i set to unity, P'_{ij} should be drawn from a uniform distribution if the intrisic data distribution is Poissonian. Individual values of P'_{ij} (close to zero or one) could, in principle, be used to identify individual days that are unusual, but here we do not consider these further.

We allow a wide logarithmic range in values of ϕ'_i, allowing the unrealistic domain of $\phi'_i < 1$, and find the value ϕ_i that minimises the Kolmogorov–Smirnov (KS) distance [18, 35] from a uniform distribution, i.e. that maximises the KS probability that the data are consistent with a uniform distribution, when varying ϕ'_i. The one-sample KS test is a non-parametric test that compares a data sample with a chosen theoretical probability distribution, yielding the probability that the sample is drawn randomly from the theoretical distribution. We label the corresponding KS probability as P_i^{KS}. We write $P_i^{\mathrm{Poiss}} := P_i^{\mathrm{KS}}(\phi'_i = 1)$ to check if any country's daily infection rate sequence is consistent with Poissonian, although this is likely to be rare, as stated above: super-Poissonian behaviour seems reasonable. Of particular interest are countries with low values of ϕ_i. Allowing for a possibly fractal or other power-law nature of the clustering of SARS-CoV-2 infection counts, we consider the possibility that the optimal values ϕ_i may be dependent on the total infection count N_i. We investigate the (ϕ_i, N_i) distribution and see whether a scaling type relation exists, allowing for a corrected statistic ψ_i to be defined in order to highlight the noise structure of the counts independent of the overall scale N_i of the counts.

Standard errors in ϕ_i for a given country i are estimated once ϕ_i has been obtained by assuming that $\widehat{\mu}_i(j)$ and ϕ_i are correct and generating 30 Poisson random simulations of the full sequence for that country. Since the scales of interest vary logarithmically, the standard deviation of the best estimates of $\log_{10} \phi_i$ for these numerical simulations is used as an estimate of $\sigma(\log_{10} \phi_i)$, the logarithmic standard error in ϕ_i.

2.2.2. Subsequences

Since artificial interference in daily SARS-CoV-2 infection counts for a given country might be restricted to shorter periods than the full data sequence, we also analyse 28-,

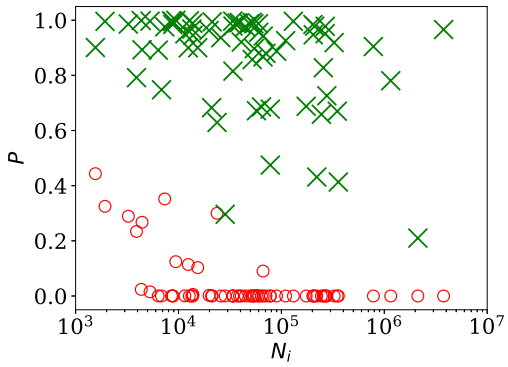

Figure 2. Probability of the noise in the country-level daily SARS-CoV-2 counts being consistent with a Poisson point process, P_i^{Poiss}, shown as red circles; and probability $P_i^{\mathrm{KS}}(\phi_i)$ for the ϕ_i clustering model proposed here (§2.2.1), shown as green `X` symbols, versus N_i, the total number of officially recorded infections for that country. The horizontal axis is logarithmic. As discussed in the text (§3.2.1), the Poisson point process is unrealistic for most of these data, while the ϕ_i clustering model is consistent with the data for all countries. Plain text table: zenodo.4432080/phi_N_full.dat.

14- and 7-day subsequences. These analyses are performed using the same methods as above (§2.2.1), except that the 28-, 14- or 7-day subsequence that minimises ϕ_i is found. The search over all possible subsequences would require calculation of a Šidàk-Bonferonni correction factor [1] to judge how anomalous they are. The KS probabilities that we calculate need to be interpreted keeping this in mind. Since the subsequences for a given country overlap, they are clearly not independent from one another. Instead, the a posteriori interpretation of the results of the subsequence searches found here should at best be considered indicative of periods that should be considered interesting for further verification.

3. Results

3.1. Data

The 132 countries and territories in the C19CCTF counts data have 19 negative values out of the total of 16367 values. These can reasonably be interpreted as corrections for earlier overcounts, and we reset these values to zero with a negligible reduction in the amount of data. Consecutive day sequences satisfying the criteria listed in §2.1 were found for 68 countries.

3.2. Clustering of SARS-CoV-2 counts

3.2.1. Full infection count sequences

Figure 2 shows, unsurprisingly, that only a small handful of the countries' daily SARS-CoV-2 counts sequences have noise whose statistical distribution is consistent with the Poisson distribution, in the sense modelled here: P_i^{Poiss} (red circles) is close to zero in most cases. On the contrary, the introduction of the ϕ_i' parameter, optimised to ϕ_i for country i, provides a sufficient fit in all cases; none of the probabilities ($P_i^{\mathrm{KS}}(\phi_i)$, green `X` symbols) in Fig. 2 is low enough to be considered a significant rejection.

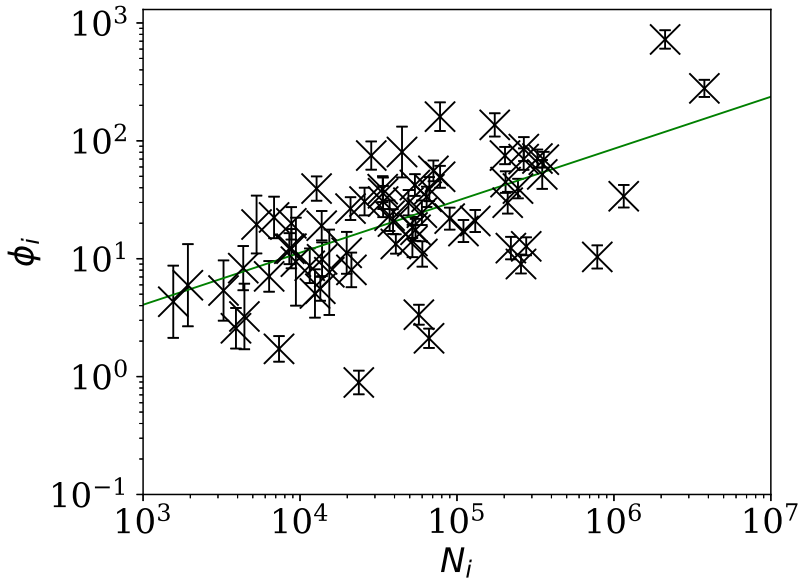

Figure 3. Noisiness in daily SARS-CoV-2 counts, showing the clustering parameter ϕ_i (§2.2.1) that best models the noise, versus the total number of counts for that country N_i. The error bars show standard errors derived from numerical (bootstrap) simulations based on the model. The axes are logarithmic, as indicated. Values of the clustering parameter ϕ_i below unity indicate sub-Poissonian behaviour – the counts in these cases are less noisy than expected for Poisson statistics. A robust (Theil–Sen [34, 36]) linear fit of $\log_{10}\phi_i$ against $\log_{10} N_i$ is shown as a thick green line (§3.2.1). Plain text table: zenodo.4432080/phi_N_full.dat.

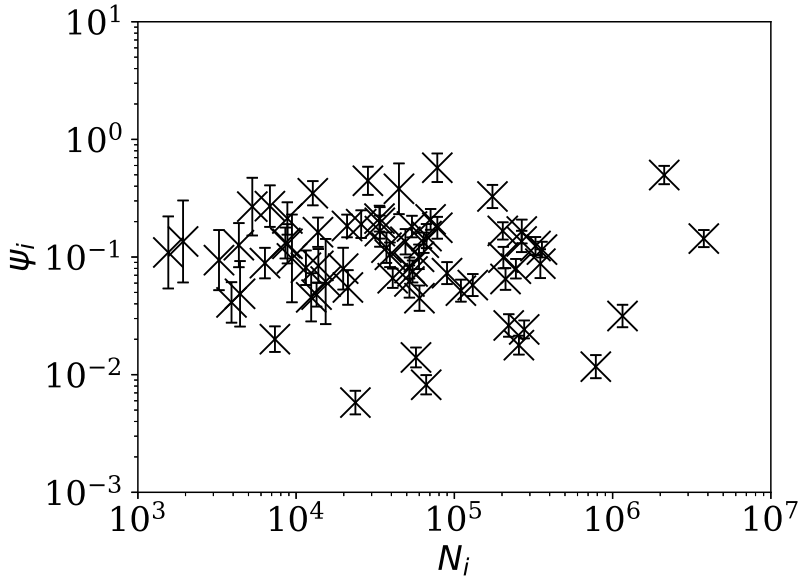

Figure 4. Normalised noisiness ψ_i (Eq. (1)) for daily SARS-CoV-2 counts versus total counts N_i. The error bars are as in Fig. 3, assuming no additional error source contributed by N_i. The axes are logarithmic. A few low ψ_i values appear to be outliers of the ψ_i distribution.

Table 1. Clustering parameters for the countries with the 10 lowest ϕ_i and 10 lowest ψ_i values (least noise); extended version of table: zenodo.4432080/phi_N_full.dat.

Country	N_i	P_i^{Poiss}	P_i^{KS}	ϕ_i	ψ_i
DZ	23691	0.30	0.63	0.89	0.005
FI	7347	0.35	0.98	1.72	0.020
BY	66348	0.09	0.87	2.11	0.008
AL	3906	0.23	0.79	2.57	0.041
HR	4422	0.27	0.89	3.24	0.048
AE	57193	0.00	0.67	3.35	0.014
NZ	1557	0.44	0.90	4.32	0.109
AU	12450	0.11	0.90	5.07	0.045
TH	3255	0.29	0.99	5.37	0.094
DK	13466	0.00	0.97	5.56	0.047
DZ	23691	0.30	0.63	0.89	0.005
BY	66348	0.09	0.87	2.11	0.008
RU	783328	0.00	0.91	10.35	0.011
AE	57193	0.00	0.67	3.35	0.014
SA	255825	0.00	0.83	9.02	0.017
FI	7347	0.35	0.98	1.72	0.020
IR	276202	0.00	0.73	12.73	0.024
TR	220572	0.00	0.43	12.30	0.026
IN	1155191	0.00	0.78	33.88	0.031
AL	3906	0.23	0.79	2.57	0.041

The consistency of the ϕ_i model with the data justifies continuing to Figure 3, which clearly shows a scaling relation: countries with greater overall numbers N_i of infections also tend to have greater noise in the daily counts $n_i(j)$. A Theil–Sen linear fit [34, 36] to the relation between $\log_{10}\phi_i$ and $\log_{10} N_i$ has a zeropoint of -0.71 ± 0.27 and a slope of 0.44 ± 0.07, where the standard errors (68% confidence intervals if the distribution is Gaussian) are conservatively generated for both slope and zeropoint by 100 bootstraps. By using a robust estimator, the low ϕ_i cases, which appear to be outliers, have little influence on the fit. The fit is shown as a thick green line in Fig. 3.

This ϕ_i–N_i relation is consistent with $\phi_i \propto \sqrt{N_i}$. To adjust the ϕ_i clustering value to take into account the dependence on N_i, and given that the slope is consistent with this simple relation, we propose the empirical definition of a normalised clustering parameter

$$\psi_i := \phi_i / \sqrt{N_i}\,, \tag{1}$$

so that ψ_i should, by construction, be approximately constant. While the estimated slope of the relation could be used rather than this half-integer power relation, the fixed relation in Eq. (1) offers the benefit of simplicity.

This relation should not be confused with the usual Poisson error. By the divisibility of the Poisson distribution, the relation $\phi_i \propto \sqrt{N_i}$ found here can be used to show

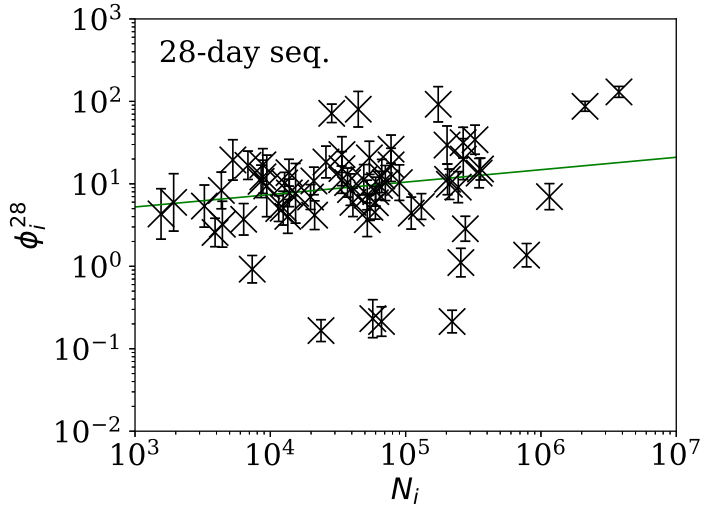

Figure 5. Clustering parameter ϕ_i for 28-day sequence with lowest ϕ_i, as in Fig. 3. The vertical axis range is expanded from that in Fig. 3, to accommodate lower values A robust (Theil–Sen [34, 36]) linear fit of $\log_{10}\phi_i^{28}$ against $\log_{10} N_i$ is shown as a thick green line (§3.2.1). Plain text table: zenodo.4432080/phi_N_28days.dat.

that

$$\begin{aligned}&\sigma[\widehat{\mu}_i(j)/\phi_i] \sim \sqrt{\widehat{\mu}_i(j)/\phi_i}\\ &\Rightarrow \sigma[\widehat{\mu}_i(j)] \sim \phi_i\sqrt{\widehat{\mu}_i(j)/\phi_i} \propto N_i^{1/4}\widehat{\mu}_i(j)^{1/2}\,,\end{aligned} \tag{2}$$

where $\sigma[x]$ is the standard deviation of random variable x. If we accept $\widehat{\mu}_i(j)$ as a fair model for $n_i(j)$ and that $n_i(j)$ is proportional to N_i, then we obtain

$$\sigma[n_i(j)] \propto n_i^{3/4}\,. \tag{3}$$

Figure 4 shows visually that ψ_i appears to be scale-independent, in the sense that the dependence on N_i has been cancelled, by construction. The countries with the 10 lowest values of ψ_i are those with ISO 3166-1 alpha-2 codes DZ, BY, RU, AE, SA, FI, IR, TR, IN, AL. Detailed SARS-CoV-2 daily count noise characteristics for the countries with lowest ϕ_i and ψ_i are listed in Table 1, including Kolmogorov–Smirnov probability that the data are drawn from a Poisson distribution, $P_i^{\rm Poiss}$, the probability of the optimal ϕ_i model, $P_i^{\rm KS}$, and ϕ_i and ψ_i.

The approximate proportionality of ϕ_i to $\sqrt{N_i}$ for the full sequences is strong and helps separate low-noise SARS-CoV-2 count countries from those following the main trend. However, the results for subsequences shown below in §3.2.2 suggest that this N_i dependence may be an effect of the typically longer durations of the pandemic in countries where the overall count is higher.

3.2.2. Subsequences of infection counts

Figures 5–7 show the equivalent of Fig. 3 for sequences of lengths 28, 14 and 7 days, respectively. The Theil–Sen robust fits to the logarithmic (ϕ_i^{28}, N_i); (ϕ_i^{14}, N_i); and (ϕ_i^{7}, N_i) relations are zeropoints and slopes of 0.27 ± 0.33 and 0.15 ± 0.07; 0.20 ± 0.64

Table 2. Least noisy 28-day sequences – clustering parameters for the countries with the 10 lowest ϕ_i^{28} values; extended table: zenodo.4432080/phi_N_28days.dat.

country	N_i	$\langle n_i^{28} \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^{28}	starting date
DZ	23691	154.1	0.10	0.75	0.17	2020-05-13
BY	66348	921.9	0.14	0.89	0.21	2020-05-08
TR	220572	1131.2	0.08	0.82	0.21	2020-06-23
AE	57193	512.8	0.08	0.23	0.23	2020-04-14
FI	7347	83.4	0.97	0.99	0.92	2020-04-15
SA	255825	1182.2	0.47	0.54	1.11	2020-04-12
RU	783328	6946.0	0.78	0.92	1.36	2020-06-17
AL	3906	74.6	0.23	0.79	2.57	2020-06-21
IR	276202	1863.3	0.20	0.97	2.85	2020-03-30
HR	4422	60.2	0.27	0.89	3.24	2020-03-28

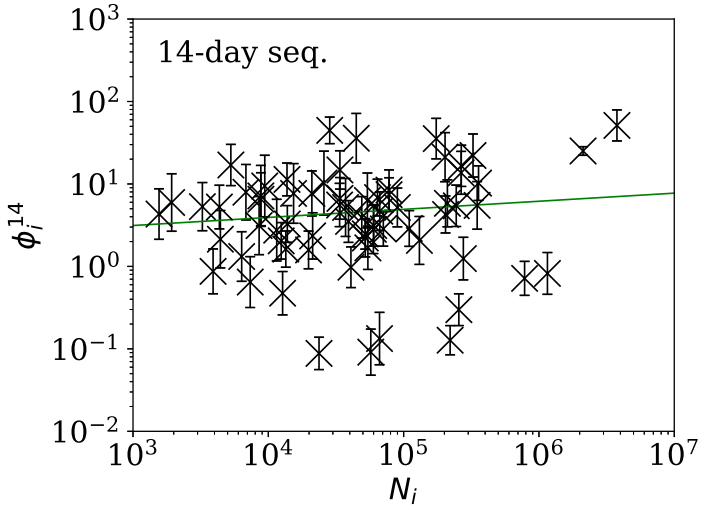

Figure 6. Clustering parameter ϕ_i for 14-day sequence with lowest ϕ_i, as in Fig. 5. Plain text table: zenodo.4432080/phi_N_14days.dat.

Table 3. Least noisy 14-day sequences – clustering parameters for the countries with the 10 lowest ϕ_i^{14} values; extended version of table: zenodo.4432080/phi_N_14days.dat.

country	N_i	$\langle n_i^{14} \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^{14}	starting date
AE	57193	521.2	0.11	0.56	0.09	2020-04-19
DZ	23691	144.1	0.11	0.48	0.09	2020-05-23
BY	66348	945.6	0.22	1.00	0.13	2020-05-12
TR	220572	991.6	0.12	0.95	0.13	2020-07-06
SA	255825	1227.5	0.38	0.96	0.30	2020-04-19
KE	12750	126.2	0.22	0.64	0.47	2020-06-03
FI	7347	95.1	0.62	0.96	0.65	2020-04-16
RU	783328	6522.9	0.37	0.42	0.72	2020-07-04
IN	1155191	9409.7	0.61	0.65	0.82	2020-05-30
AL	3906	70.8	0.57	0.88	0.87	2020-06-24

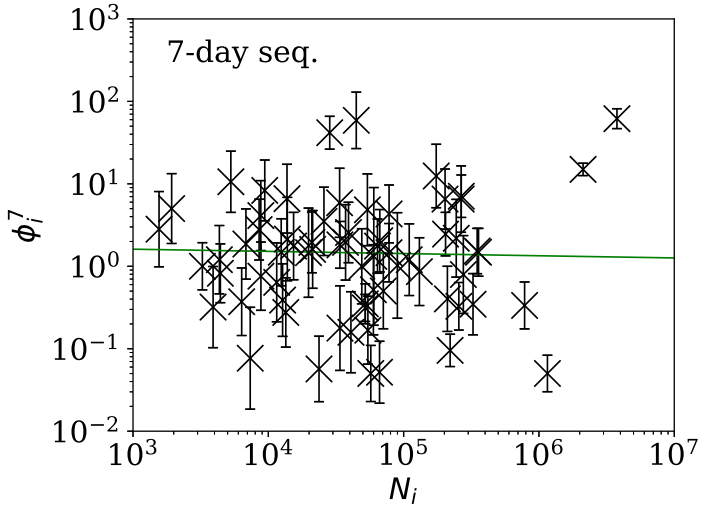

Figure 7. Clustering parameter ϕ_i for 7-day sequence with lowest ϕ_i^7, as in Fig. 5. There is clearly a wider overall scatter and bigger error bars compared to Figs 5 and 6; a low ϕ_i^7 is a weaker indicator than ϕ_i^{28} and ϕ_i^{14}. Plain text table: zenodo.4432080/phi_N_07days.dat.

Table 4. Least noisy 7-day sequences – clustering parameters for the countries with the 10 lowest ϕ values; extended table: zenodo.4432080/phi_N_07days.dat.

country	N_i	$\langle n_i^7 \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^7	starting date
AE	57193	544.9	0.24	0.99	0.05	2020-04-27
BY	66348	947.9	0.60	0.94	0.05	2020-05-13
IN	1155191	10109.3	0.34	0.60	0.05	2020-06-06
DZ	23691	188.6	0.20	0.99	0.06	2020-05-20
FI	7347	94.9	0.42	0.55	0.08	2020-04-20
TR	220572	1022.4	0.43	0.94	0.10	2020-07-07
PL	40782	297.7	0.31	0.96	0.16	2020-06-20
PA	54426	171.1	0.82	0.96	0.17	2020-05-09
HN	33835	160.7	0.89	0.99	0.18	2020-06-01
DK	13466	71.1	0.48	0.94	0.28	2020-05-11

and 0.10 ± 0.13; and 0.29 ± 0.68 and -0.03 ± 0.15, respectively. There is clearly no significant dependence of ϕ_i^d on N_i for any of these fixed length subsequences, in contrast to the case of the ϕ_i dependence on N_i for the full count sequences. Thus, the empirical motivation for using ψ (Eq. (1)) to discriminate between the countries' full sequences of SARS-CoV-2 data is not justified for the subsequences. Tables 2–4 show the countries with the least noisy sequences as determined by ϕ_i^{28}, ϕ_i^{14} and ϕ_i^7, respectively.

Tables 2 and 3 show that the lists of countries with the strongest anti-clustering are similar. Thus, Fig. 8 shows the SARS-CoV-2 counts curves for countries with the lowest ϕ_i^{28}, and Fig. 9 the curves for those with the lowest ϕ_i^7. Both figures exclude countries with total counts $N_i \leq 10000$, in which low total counts tend to give low clustering. It is clear in these figures that several countries have subsequences that are strongly sub-Poissonian – with some form of anti-clustering, whether natural or artificial.

Countries in the median of the ϕ_i^{28} and ϕ_i^7 distributions have their curves shown in Fig. 10 for comparison. It is visually clear in the figure that the counts are dispersed widely beyond the Poissonian band, and that the ϕ_i^{28} and ϕ_i^7 models are reasonable as a model for representing about 68% of the counts within one standard deviation of the model values.

4. Discussion

Figures 3 and 4 clearly show that some groups of countries are unusual in terms of the characteristics of their location in the (N_i, ψ_i) plane.

4.1. High total infection count

Brazil (BR) and the United States (US) are separated from the majority of other countries by their high total infection count. They have correspondingly higher clustering values ϕ_i, although their normalised clustering values ψ_i are in the range of about $0.4 < \psi_i < 10$ covered by the majority of countries in Fig. 4.

It does not seem realistic that these two countries' ϕ_i values greater than 300 are

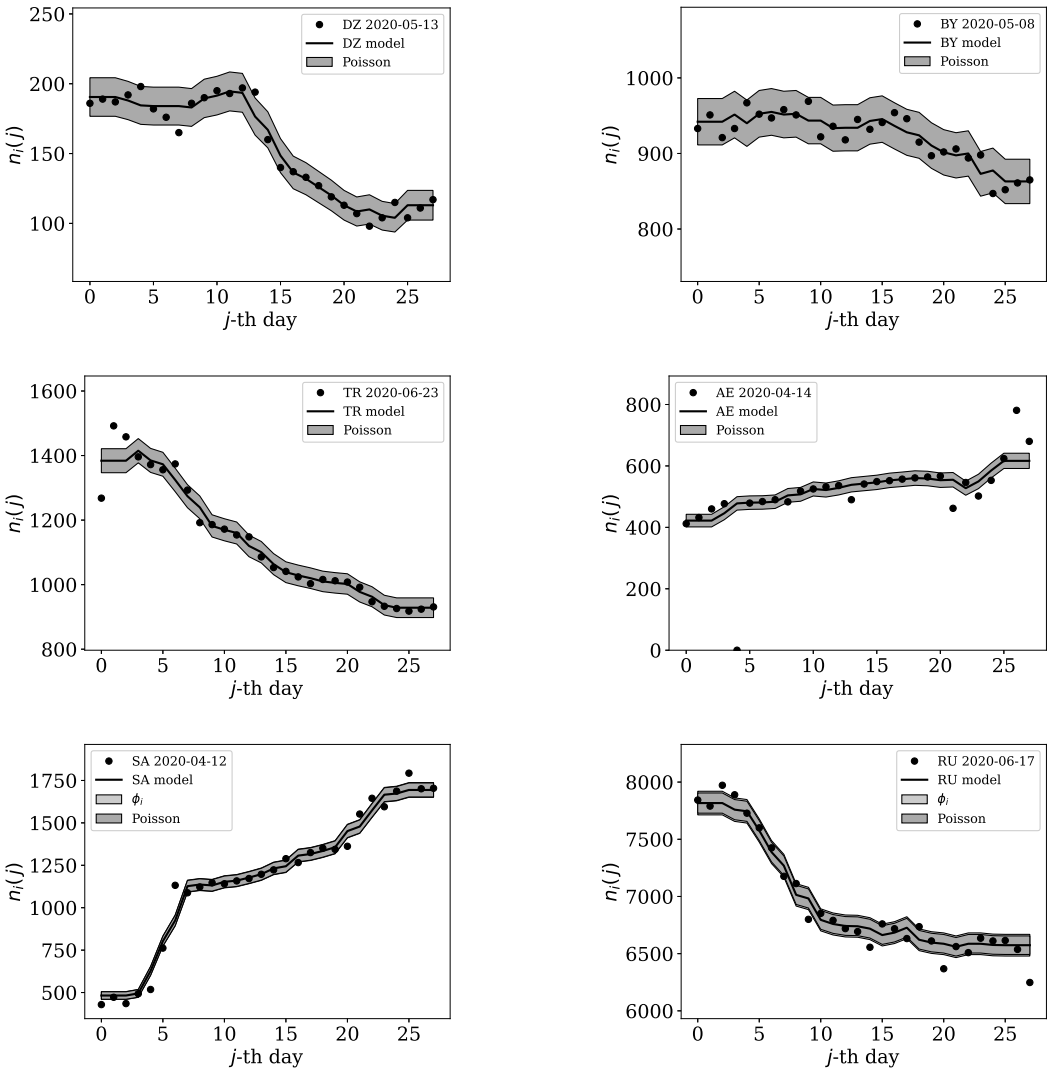

Figure 8. Least noisy 28-day official SARS-CoV-2 national daily counts for countries with total counts $N_i > 10000$ (see Fig. 5 and Table 2), shown as dots in comparison to the $\widehat{\mu}_i(j)$ model (median of the 4 neighbouring days) and 68% error band for the Poisson point process. The ranges in daily counts (vertical axis) are chosen automatically and in most cases do not start at zero. About nine (32%) of the points should be outside of the shaded band unless the counts have an anti-clustering effect that weakens Poisson noise. A faint shaded band shows the ϕ_i^{28} model for the one country here with ϕ_i (slightly) greater than one (RU), but is almost indistinguishable from the Poissonian band. The dates indicate the start date of each sequence.

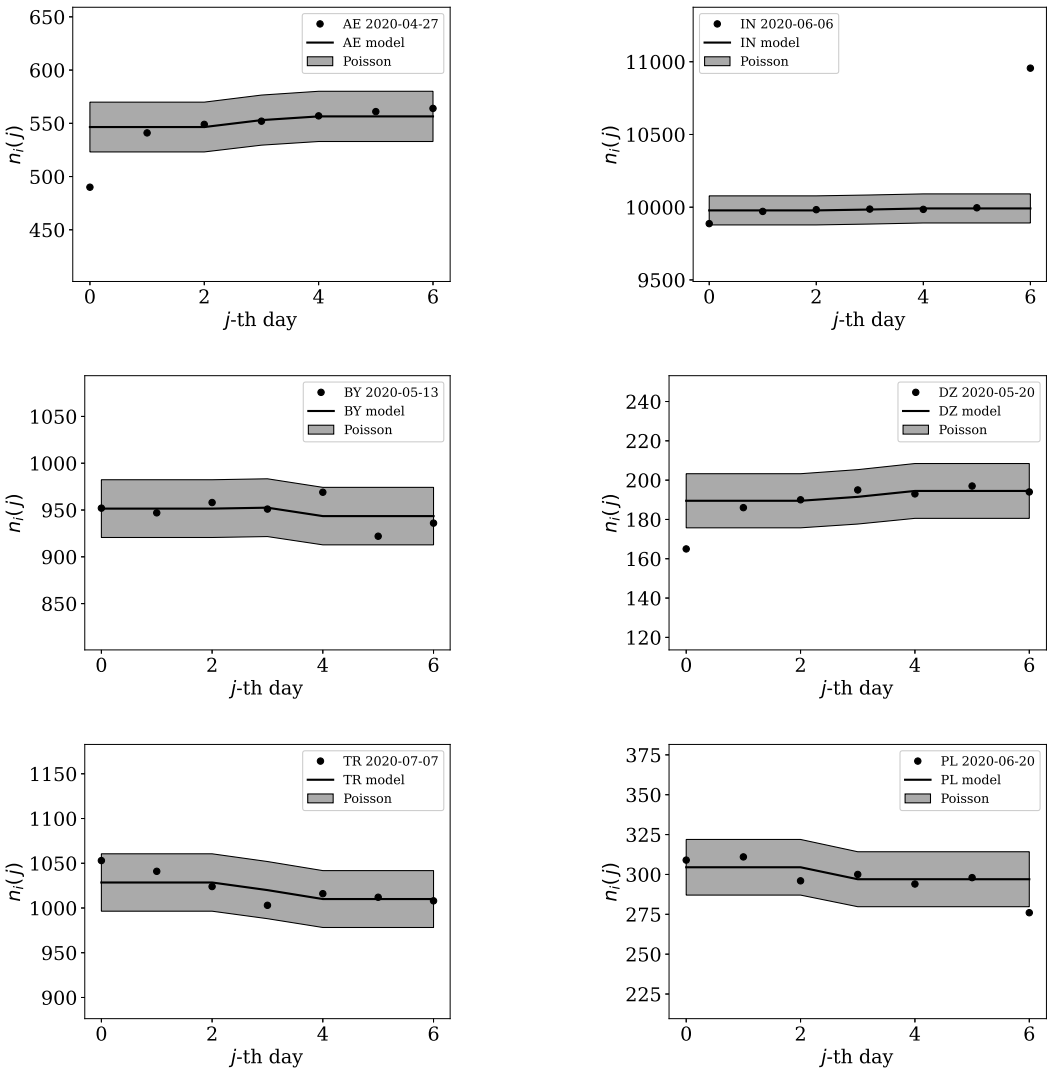

Figure 9. Least noisy 7-day daily counts for countries with total counts $N_i > 10000$, as in Fig. 8. Concentration of points close to the model indicates an anti-clustering effect; about 68% (two) of the points should scatter up and down throughout the shaded band if the counts are Poissonian. In several cases, the data points appear to be mostly stuck to the model, with almost no scatter.

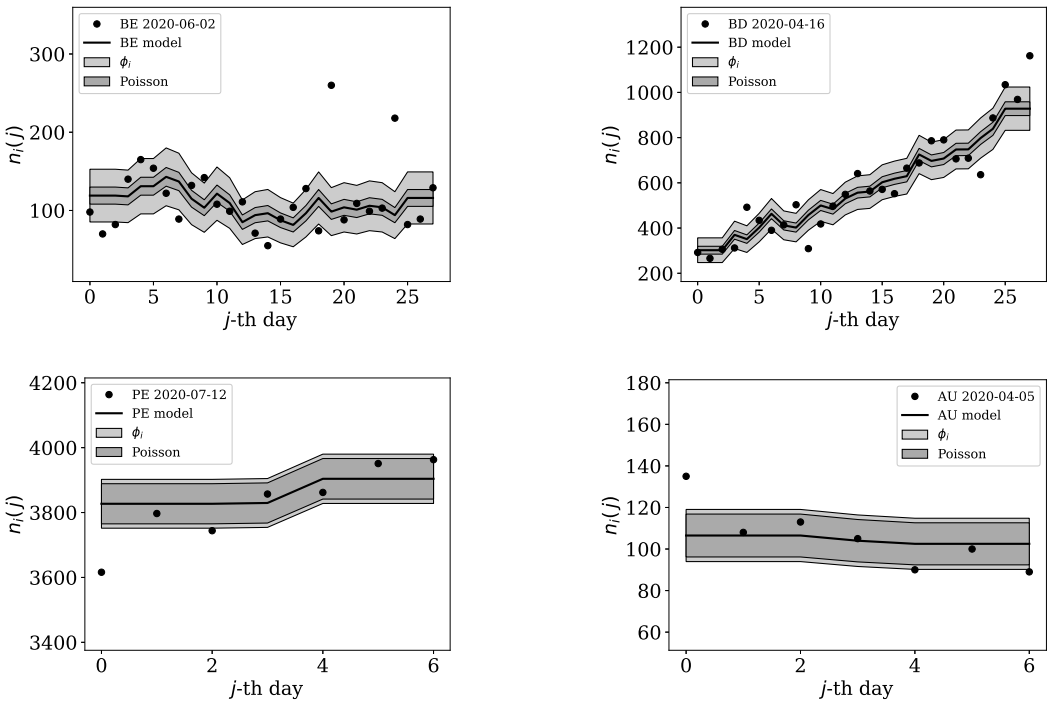

Figure 10. Typical (median) 28-day (above) and 7-day (below) daily counts, as in Figs 8 and 9. The dark shaded band again shows a Poissonian noise model, which underestimates the noise. A faint shaded band shows the ϕ_i models for these countries' SARS-CoV-2 daily counts, and should contain about 68% of the infection count points.

purely an effect of intrinsic infection events – 'superspreader' events in crowded places or nursing homes. While individual big clusters may occur given the high overall scale of infections, it seems more likely that this is administrative clustering. Both countries are federations, and have numerous geographic administrative subdivisions with a diversity of political and administrative methods. A plausible explanation for the dominant effect yielding $\phi_i > 300$ in these two countries is that on any individual day, the arrival and full processing of reports depends on a number of sub-national administrative regions, each reporting a few hundred new infections.

For example, if there are 10 reporting regions, each typically reporting 300 infections, then typically (on about 68% of days) there will be about 7 to 13 reports per day. This would give a range varying from about 2100 to 3900 cases per day, rather than 2945 to 3055, which would be the case for unclustered, Poissonian counts (since $\sqrt{3000} \approx 55$). Lacking a system that obliges sub-national divisions – and laboratories – to report their test results in time-continuous fashion and that validates and collates those reports on a time scale much shorter than 24 hours, this type of clustering seems natural in the sociological sense.

4.2. Low normalised clustering ψ_i

In Fig. 4, there appears to be a group of eight countries that are also separated from the main group of countries, but by having low normalised noise ψ_i rather than just having a high total count N_i.

4.2.1. Low ψ_i, low N_i, high P_i^{Poiss}

Classifying the countries by ψ_i alone (Table 1) would add Finland (FI) to this group, but in Fig. 4, Finland appears better grouped with the main body of countries in the (ψ_i, N_i) plane. This could be interpreted as Eq. (1) providing insufficient correction for the ϕ_i–N_i relation. Alternatively, looking at Finland's entry in Table 2 for 28-day sequences, we see that Finland is among the three with the lowest total (or mean) daily infection counts in the table, and has the highest consistency with a Poisson distribution (P_i^{Poiss}). Having a low total infection count, it seems credible that Finland lacks the intrinsic, testing and administrative clustering of countries with higher infection counts.

4.2.2. Low ψ_i, high N_i

India (IN) and Russia (RU) have total infection counts nearly as high (logarithmically) as Brazil and the US, but have managed to keep their daily infection rates much less noisy – by about a factor of 10 to 100 – than would be expected from the general pattern displayed in the diagram. Despite having of the order of a million total official SARS-CoV-2 infections each, these two countries have, as of the download date of the data, 21 July 2020, avoided having the clustering effects present in Brazil and the US.

The most divergent case in the high-N_i part of this group (see Fig. 4 and Table 1) is Russia, which has only a very modest value of $\phi_i = 10.4 \times 10^{\pm 0.098}$ for its total infection count of over a million. This would require that both intrinsic clustering of infection events and administrative procedures work much more smoothly in Russia than in the United States, Brazil and, to a lesser degree, India. Tables 2 and 3 and Fig. 8 show that the Russian official SARS-CoV-2 counts indeed show very little noise compared to more typical cases (Fig. 10). At the intrinsic epidemiological level,

this means that if the Russian counts are to be considered accurate, then very few clusters – in nursing homes, religious gatherings, bars, restaurants, schools, shops – can have occurred. Moreover, laboratory testing and transmission of data through the administrative chain from local levels to the national (federal) health agency must have occurred without the clustering effects present in the United States and Brazil and in countries with more typical clustering values ϕ_i, characterising their daily infection counts. International media interest in Russian COVID-19 data has mostly focussed on controversy related to COVID-19 death counts [8], with apparently no attention given so far to the modestly super-Poissonian nature of the daily counts, in contrast to the strongly super-Poissonian counts of other countries with high total infection counts.

India's overall position in the (ψ_i, N_i) plane (Fig. 4 and Table 1) is less extreme than that of Russia, with an unnormalised clustering parameter $\phi_i = 34 \times 10^{\pm 0.096}$. However, Table 3 shows that despite its large overall infection count, India achieved a 14-day sequence with a preferred ϕ_i value close to unity. Moreover, it has a very low-ranked ϕ_i^7 value, as given in Table 4 and illustrated in Fig. 9. Five values appear almost exactly on the model curve rather than scattering above and below. Moreover, the value is just below 10,000. Epidemiologically, it is not credible to believe that 10,000 officially reported cases per day should be an attractor resulting from the pattern of infections and system of reporting. Given that the value of 10,000 is a round number in the decimal-based system, a reasonable speculation would be that the daily counts for India were artificially held at just below 10,000 for several days. The crossing of the 10,000 psychological threshold of daily infections was noted in the media [29], but the lack of noise in the counts during the week preceding the crossing of the threshold appears to have gone unnoticed. After crossing the 10,000 threshold, the daily infections in India continued increasing, as can be seen in the full counts (zenodo.4432080/WP_C19CCTF_SARSCoV2.dat).

4.2.3. Low ψ_i, low ϕ_i, medium N_i

Among the group of eight low ψ_i countries, Table 1 shows that only one country has its full data set (as defined here) best modelled by the ordinary Poisson point process. Algeria (DZ) appears to have completely avoided clustering effects, with ϕ_i close to unity. Figure 8 shows the least noisy 28-day sequence for Algeria. Only one day of SARS-CoV-2 recorded infections appears to have diverged beyond the Poissonian 68% band, rather than about nine, the expected number for a Poissonian distribution. Most of the points appear to stick very closely to the model. It is difficult to imagine a natural process for obtaining this sub-Poissonian noise (as preferred by the ϕ_i model), especially in the context where most countries have super-Poissonian daily counts. In a frequentist interpretation, the least noisy Algerian 28-day count sequence would be considered only mildly, not significantly, unusual, since it is consistent with a Poisson distribution, with only a weak rejection (Tables 2–4). However, as a member of the general class of countries' SARS-CoV-2 daily infection count curves, use of the ϕ_i model would appear to be justified. It is in this sense that the sequence can be considered sub-Poissonian. Moreover, a full Bayesian analysis would need to consider independent credibility criteria[5]. Compartmental epidemic modelling of the Algerian data, which has been published for the period ending 24 May 2020[30], could also be included in an extended analysis.

In line with the counts for India that appeared to be smooth just below a round-number boundary of 10,000 infections per day, the least noisy 7-day sequence for

Algeria, shown in Fig. 9, might appear to have been affected by a similar psychological boundary of 200 infections per day. Medical specialists interviewed by the media interpreted the 200 daily infections period as representing stability and resulting from partial lockdown measures, without providing an explanation for why Poisson noise was nearly absent [14]. While lockdown measures should reduce intrinsic epidemiological clustering down towards the Poissonian level, it is difficult to see how they could reduce testing and administrative pipeline clustering. A coincidence that occurred during this least-noisy 7-day period, on 24 May 2020, was that a full COVID-19 lockdown was implemented in Algeria [24].

The Belarus (BY) case is present in all four tables (Tables 1–4). The least noisy Belarusian counts curve appears in Fig. 8. As with the other panels in the daily counts figures, the vertical axis is set by the data instead of starting at zero, in order to best display the information on the noise in the counts. With the vertical axis starting at zero, the Belarus daily counts would look nearly flat in this figure. They appear to be bounded above by the round number of 1000 SARS-CoV-2 infections per day, which, again, appears to be a psychologically preferred barrier. Media have expressed scepticism of Belarusian COVID-19 related data [2, 19].

One remaining case of a coincidence is that the lowest noise 7-day sequence listed for Poland (PL, Table 4) is for the 7-day period starting 20 June 2020, with $\phi_i^7 = 0.16 \times 10^{\pm 0.49}$. This is a factor of about 100 (or at least 10 at about 95% confidence) below Poland's clustering value for the full sequence of its SARS-CoV-2 daily infection counts, $\phi_i = 13 \times 10^{\pm 0.082}$, which Fig. 3 shows is typical for a country with an intermediate total infection count. On 28 June 2020, there was a de facto (of disputed constitutional validity [21, 38]) first-round presidential election in Poland. Figure 9 shows that the counts for Poland during the final pre-first-round-election week did not scatter widely throughout the Poissonian band. A decimal-system round number also appears in this figure: the daily infection rate is slightly above about 300 infections per day and drops to slightly below that. For an unknown reason that does not previously appear to have been studied, the intrinsic clustering of SARS-CoV-2 infections in Poland together with testing and administrative clustering of the confirmed cases appears to have temporarily disappeared just prior to the election date, yielding what is best modelled as sub-Poissonian counts.

4.2.4. JHU CSSE data

The JHU CSSE data give mostly similar results to the C19CCTF data. These are presented and briefly discussed in Appendix A.

5. Conclusion

Given the overdispersed, one-parameter Poissonian ϕ_i model proposed, the noise characteristics of the daily SARS-CoV-2 infection data suggest that most of the countries' data form a single family in the (ϕ_i, N_i) plane. The clustering – whether epidemiological in origin, or caused by testing or administrative pipelines – tends to be greater for greater numbers of total infections. Several countries appear, however, to show unusually anti-clustered (low-noise) daily infection counts.

Since these daily infection counts data constitute data of high epidemiological interest, the statistical characteristics presented here and the general method could be used as the basis for further investigation into the data of countries showing exceptional

characteristics. The relations between the most anti-clustered counts and the psychologically significant decimal system round numbers (IN: 10,000 daily, DZ: 200 daily, BY: 1000 daily, PL: 300 daily), and in relation to a de facto national presidential election, raise the question of whether or not these are just coincidences. The suspicious periods of data found here are mostly complementary to those studied by Balashov et al., since those authors' Benford's law analysis mainly focuses on the first-digit Benford's law[5].

It should be straightforward for any reader to extend the analysis in this paper by first checking its reproducibility with the free-licensed source package provided using the MANEAGE framework [4], and then extending, updating or modifying it in other appropriate ways; see §Code availability below. Reuse of the data should be straightforward using the files archived at zenodo.4432080.

FUNDING No funding has been received for this project.

DATA AVAILABILITY As described above in §2.1, the source of curated SARS-CoV-2 infection count data used for the main analysis in this paper is the C19CCTF data, downloaded using the script `download-wikipedia-SARS-CoV-2-charts.sh` and stored in the file `Wikipedia_SARSCoV2_charts.dat` in the reproducibility package available at zenodo.4432080. The data file is archived at zenodo.4432080/WP_C19CCTF_SARSCoV2.dat. The WHO data that was compared with the C19CCTF data via a jump analysis (Fig. 1) was downloaded from `https://covid19.who.int/WHO-COVID-19-global-data.csv` and was archived on 15 July 2020.

CODE AVAILABILITY In addition to the SARS-CoV-2 infection count data for this paper, the full calculations, production of figures, tables and values quoted in the text of the pdf version of the paper are intended to be fully reproducible on any POSIX-compatible system using free-licensed software, which, by definition, the user may modify, redistribute and redistribute in modified form. The reproducibility framework is technically a GIT branch of the MANEAGE package [4][4], earlier used to produce reproducible papers[12]. The GIT repository commit ID of this version of this paper is subpoisson-f7e7b46. The primary GIT repository is `https://codeberg.org/boud/subpoisson`. Bug reports and discussion are welcome at `https://codeberg.org/boud/subpoisson/issues`.

CONFLICT OF INTEREST The author of this paper is aware of no financial or similar conflicts of interests.

ORCID *Boudewijn F. Roukema* ORCID: https://orcid.org/0000-0002-3772-0250

REFERENCES

[1] S. Abdi, *Bonferroni and Sidak corrections for multiple comparisons*, in *Encyclopedia of Measurement and Statistics*, N. Salkind, ed., Thousand Oaks, Sage, USA (2007). Available at `http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf`.

[4] `https://maneage.org`

[2] AFN, *Nexta channel accuses the Ministry of Health of the Republic of Belarus of publishing censored data on coronavirus (in Russian)*, AFN (2020). Available at `https://afn.by/news/i/275882`, Archived at Wayback.

[3] N. Afshordi, B. Holder, M. Bahrami, and D. Lichtblau, *Diverse local epidemics reveal the distinct effects of population density, demographics, climate, depletion of susceptibles, and intervention in the first wave of COVID-19 in the United States*, arXiv e-prints (2020), (arXiv:2007.00159).

[4] M. Akhlaghi, R. Infante-Sainz, B.F. Roukema, D. Valls-Gabaud, and R. Baena-Gallé, *Towards Long-term and Archivable Reproducibility*, CiSE, submitted (2020), (arXiv:2006.03018).

[5] V.S. Balashov, Y. Yan, and X. Zhu, *Are Less Developed Countries More Likely to Manipulate Data During Pandemics? Evidence from Newcomb-Benford Law*, arXiv e-prints (2020), (arXiv:2007.14841).

[6] A.L. Barabási, *The origin of bursts and heavy tails in human dynamics*, Nature 435 (2005), pp. 207–211, DOI:10.1038/nature03459, (arXiv:cond-mat/0505371).

[7] R. Chowdhury, K. Heng, M.S.R. Shawon, G. Goh, D. Okonofua, C. Ochoa–Rosales, V. Gonzalez–Jaramillo, A. Bhuiya, and et al., *Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries*, Eur. J. Epidemiol. 35 (2020), pp. 389–399, DOI:10.1007/s10654-020-00649-w.

[8] B. Cole, *Russia accuses media of false coronavirus death numbers as Moscow officials say 60 percent of fatalities not included*, Newsweek (2020). Available at `https://www.newsweek.com/russia-accuses-media-false-coronavirus-death-numbers-1503932`, Archived at Archive Today.

[9] K.I. Goh and A.L. Barabasi, *Burstiness and Memory in Complex Systems*, Europhys. Lett. Assoc.81 (2006), p. 4, (arXiv:physics/0610233).

[10] C. Huang, Y. Wang, X. Li, R. L., J. Zhao, Y. Hu, and et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China*, Lancet 395 (2020), pp. 97–506, DOI:10.1016/s0140-6736(20)30183-5.

[11] L. Huang, X. Zhang, X. Zhang, W. Zhijian, L. Zhang, J. Xu, and et al., *Rapid asymptomatic transmission of COVID-19 during the incubation period demonstrating strong infectivity in a cluster of youngsters aged 16–23 years outside Wuhan and characteristics of young patients with COVID-19: A prospective contact-tracing study*, J. Infection 80 (2020), pp. e1–e13, DOI:10.1016/j.jinf.2020.03.006.

[12] R. Infante-Sainz, I. Trujillo, and J. Román, *The Sloan Digital Sky Survey extended point spread functions*, MNRAS 491 (2020), pp. 5317–5329, DOI:10.1093/mnras/stz3111, (arXiv:1911.01430).

[13] F. Jiang, Z. Zhao, and X. Shao, *Time Series Analysis of COVID-19 Infection Curve: A Change-Point Perspective*, arXiv e-prints (2020), p. arXiv:2007.04553, (arXiv:2007.04553).

[14] T. K., *Evolution de la pandémie en Algérie : Les spécialistes relèvent une situation maîtrisée*, El Moudjahid (2020). Available at `https://www.elmoudjahid.com/fr/actualites/154163`, Archived at Archive Today.

[15] B.C. Keegan and C. Tan, *A Quantitative Portrait of Wikipedia's High-Tempo Collaborations during the 2020 Coronavirus Pandemic*, arXiv e-prints (2020), (arXiv:2006.08899).

[16] T. Kim, B. Lieberman, G. Luta, and E. Pena, *Prediction Regions for Poisson and Over-Dispersed Poisson Regression Models with Applications to Forecasting Number of Deaths during the COVID-19 Pandemic*, arXiv e-prints (2020), (arXiv:2007.02105).

[17] C. Koch and K. Okamura, *Benford's Law and COVID-19 reporting*, Econ.Lett. 196 (2020), p. 109573, DOI:10.1016/j.econlet.2020.109573, DOI:10.2139/ssrn.3586413.

[18] A.N. Kolmogorov, *Sulla Determinazione Empirica di Una Legge di Distribuzione*, Giornale dell'Istituto Italiano degli Attuari 4 (1933), pp. 83–91.

[19] A.E. Kramer, *"There Are No Viruses Here": Leader of Belarus Scoffs at Lockdowns*, The New York Times (2020). Available at `https://www.nytimes.com/2020/04/25/world/europe/belarus-lukashenko-coronavirus.html`, Archived at Archive Today.

[20] K.B. Lee, S. Han, and Y. Jeong, *COVID-19, flattening the curve, and Benford's law*, Physica A 559 (2020), p. 125090, DOI:10.1016/j.physa.2020.125090.

[21] E. Letowska and P. Pacewicz, *Prof. Łętowska: To nie były wybory, ale plebiscyt. Uchybienia wyborcze rzucają długi gęsty cień*, OKO.press (2020). Available at `https://oko.press/prof-letowska-to-nie-byly-wybory-ale-plebiscyt`, Archived at Wayback.

[22] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, *Substantial undocu-*

mented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), Science 368 (2020), pp. 489–493, DOI:10.1126/science.abb3221. Available at `https://www.medrxiv.org/content/10.1101/2020.02.14.20023127v1`.

[23] W.R.J. Mebane, *Fraud in the 2009 presidential election in iran?*, Chance 23 (2010), pp. 6–15, DOI:10.1080/09332480.2010.10739785.

[24] M. Mehenni, *Gestion du Covid durant l'Aïd-el-Adha : ce mouton qui complique tout*, TSA Algérie (2020). Available at `https://www.tsa-algerie.com/gestion-du-covid-durant-laid-el-adha-ce-mouton-qui-complique-tout`, Archived at Archive Today.

[25] E.A. Molina-Cuevas, *Choosing a growth curve to model the Covid-19 outbreak*, arXiv e-prints (2020), (arXiv:2007.03779).

[26] S. Newcomb, *Note on the Frequency of Use of the Different Digits in Natural Numbers*, American Journal of Mathematics 4 (1881), pp. 39–40.

[27] M. Nigrini and S.J. Miller, *Data diagnostics using second order tests of Benford's Law*, Auditing: J. Pract. & Theory 28 (2009), pp. 305–324, DOI:10.2308/aud.2009.28.2.305.

[28] S.D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile ; précédées des Règles générales du calcul des probabilités*, Bachelier, Imprimeur-Libraire, Paris, 1837, Available at `https://gallica.bnf.fr/ark:/12148/bpt6k110193z/f218.image`.

[29] M. Porecha, *India records over 10,000 new Covid-19 cases for first time*, The Hindu (2020). Available at `https://www.thehindubusinessline.com/news/national/india-records-over-10000-new-covid-19-cases-for-first-time/article31810421.ece`, Archived at Archive Today.

[30] M.T. Rouabah, A. Tounsi, and N.E. Belaloui, *A mathematical epidemic model using genetic fitting algorithm with cross-validation and application to early dynamics of COVID-19 in Algeria*, J.Fundam.Appl.Sci 12 (2020), pp. 1253–1276, (arXiv:2005.13516).

[31] B.F. Roukema, *A first-digit anomaly in the 2009 iranian presidential election*, Journal of Applied Statistics 41 (2014), pp. 164–199, DOI:10.1080/02664763.2013.838664, (arXiv:0906.2789v6).

[32] B.F. Roukema, *Complementing Benford's Law for small N : a local bootstrap*, in *The Theory and Applications of Benford's Law*, S.J. Miller, ed., Princeton University Press, Princeton, 2015, pp. 223–232.

[33] E. Ruijer, F. Détienne, M. Baker, J. Groff, and A.J. Meijer, *The Politics of Open Government Data: Understanding Organizational Responses to Pressure for More Transparency*, Am.Rev.Publ.Admin. (2019), DOI:10.1177/0275074019888065.

[34] P.K. Sen, *Estimates of the regression coefficient based on Kendall's tau*, J. Amer. Stat. Assoc. 63 (1968), pp. 1379–1389, DOI:10.2307/2285891.

[35] N. Smirnov, *Table for Estimating the Goodness of Fit of Empirical Distributions*, Ann. Math. Stat.19 (1948), pp. 279–281. Available at `https://projecteuclid.org/euclid.aoms/1177730256`.

[36] H. Theil, *A rank-invariant method of linear and polynomial regression analysis*, Nederl. Akad. Wetensch., Proc. 53 (1950), pp. 386–392.

[37] P. Thomas, M.K. Lau, A. Trisovic, E.R. Boose, B. Couturier, M. Crosas, A.M. Ellison, V. Gibson, C.R. Jones, and M. Seltzer, *If these data could talk*, Scientific Data 4 (2017), p. 170114, DOI:10.1038/sdata.2017.114.

[38] M. Wyrzykowski, *Former CT judge Prof. Wyrzykowski: The presidential elections in Poland will be held under the pretence of legality* (2020). Available at `https://ruleoflaw.pl/former-ct-judge-prof-wyrzykowski-the-presidential-elections-in-poland-will-be-held-under-the-pretence-of-legality`, Archived at Wayback.

[39] H. Yu and D.G. Robinson, *The New Ambiguity of "Open Government"*, UCLA L. Rev. Disc. 59 (2012), p. 178, DOI:10.2139/ssrn.2012489.

Appendix A. JHU CSSE data

The John Hopkins University Center for Systems Science and Engineering global time series data was downloaded on 2020-08-

Table A1. As in Table 1, for the JHU CSSE data: clustering parameters for the countries with the 10 lowest ϕ_i and 10 lowest ψ_i values (least noise); extended version of table: zenodo.4432080/phi_N_full_jhu.dat.

Country	N_i	P_i^{Poiss}	P_i^{KS}	ϕ_i	ψ_i
DZ	37664	0.40	0.67	0.88	0.004
BY	69308	0.02	0.72	2.29	0.008
AL	7117	0.13	0.56	2.57	0.030
HR	6258	0.27	0.89	3.24	0.040
NZ	1611	0.10	0.88	3.85	0.095
AE	63819	0.00	0.61	4.42	0.017
TH	3376	0.29	0.99	5.37	0.092
DK	15758	0.00	0.97	5.56	0.044
IS	1983	0.33	1.00	5.96	0.133
GR	6632	0.03	0.98	6.53	0.080
DZ	37664	0.40	0.67	0.88	0.004
RU	910778	0.00	0.58	7.50	0.007
BY	69308	0.02	0.72	2.29	0.008
SA	295902	0.00	0.91	8.32	0.015
TR	246861	0.00	0.16	7.67	0.015
AE	63819	0.00	0.61	4.42	0.017
IR	338825	0.00	0.71	10.35	0.017
AL	7117	0.13	0.56	2.57	0.030
HR	6258	0.27	0.89	3.24	0.040
AZ	34018	0.10	0.90	7.67	0.041

15 from `https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv`, from git commit 78AE929, and analysed using the same software and parameters as for the C19CCTF data. Tables A1–A4 show the equivalent of Tables 1–4. The rankings and ϕ_i estimates appear mostly similar between the two datasets. One difference is that the low ϕ_i^7 value for India shown in Table 4 is absent in Table A4. In other words, while the media stated that the daily confirmed count in India first went above the 10,000-per-day psychological threshold on 12 June 2020 [29], the JHU CSSE data crossed this threshold earlier, and contains noise that was unknown at that time to the media and is absent from the C19CCTF data.

Table A2. As in Table 2, for the JHU CSSE data: least noisy 28-day sequences – clustering parameters for the countries with the 10 lowest ϕ_i^{28} values; extended table: zenodo.4432080/phi_N_28days_jhu.dat.

country	N_i	$\langle n_i^{28} \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^{28}	starting date
TR	246861	1014.5	0.03	1.00	0.14	2020-06-30
DZ	37664	154.1	0.10	0.75	0.17	2020-05-13
BY	69308	921.9	0.14	0.89	0.21	2020-05-08
AE	63819	512.8	0.08	0.23	0.23	2020-04-14
RU	910778	5456.3	0.56	0.62	0.28	2020-07-18
SA	295902	1182.2	0.47	0.54	1.11	2020-04-12
AL	7117	77.7	0.20	0.46	1.33	2020-06-23
IR	338825	1863.3	0.20	0.97	2.85	2020-03-30
HR	6258	60.2	0.27	0.89	3.24	2020-03-28
NE	64292	128.8	0.10	0.99	3.39	2020-06-04

Table A3. As in Table 3, for the JHU CSSE data: least noisy 14-day sequences – clustering parameters for the countries with the 10 lowest ϕ_i^{14} values; extended version of table: zenodo.4432080/phi_N_14days_jhu.dat.

country	N_i	$\langle n_i^{14} \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^{14}	starting date
AE	63819	521.2	0.11	0.56	0.09	2020-04-19
DZ	37664	144.1	0.11	0.48	0.09	2020-05-23
TR	246861	971.6	0.12	0.86	0.11	2020-07-08
BY	69308	945.6	0.22	1.00	0.13	2020-05-12
RU	910778	5165.5	0.47	0.51	0.28	2020-08-01
SA	295902	1227.5	0.38	0.96	0.30	2020-04-19
AL	7117	131.5	0.78	1.00	0.53	2020-08-01
PL	55319	299.9	0.55	0.68	0.53	2020-06-17
KE	29334	126.2	0.54	0.91	0.57	2020-06-03
CA	123605	1181.5	0.52	0.71	1.22	2020-05-08

Table A4. As for Table 4, for the JHU CSSE data: least noisy 7-day sequences – clustering parameters for the countries with the 10 lowest ϕ values; extended table: zenodo.4432080/phi_N_07days_jhu.dat.

country	N_i	$\langle n_i^{7} \rangle$	P_i^{Poiss}	P_i^{KS}	ϕ_i^{7}	starting date
AE	63819	544.9	0.24	0.99	0.05	2020-04-27
BY	69308	947.9	0.60	0.94	0.05	2020-05-13
TR	246861	929.6	0.22	0.93	0.05	2020-07-15
DZ	37664	188.6	0.20	0.99	0.06	2020-05-20
PL	55319	297.0	0.51	0.99	0.10	2020-06-20
PA	79402	171.1	0.82	0.96	0.17	2020-05-09
HN	49487	160.7	0.89	0.99	0.18	2020-06-01
RU	910778	5873.9	0.68	0.87	0.21	2020-07-19
DK	15758	71.1	0.48	0.94	0.28	2020-05-11
AL	7117	78.9	0.40	0.79	0.32	2020-07-05