Early detection of variants of concern via funnel plots of regional reproduction numbers

Simone Milanesi
University of Pavia
Francesca Rosset
University of Udine https://orcid.org/0000-0002-1202-2712
Marta Colaneri
Fondazione IRCCS Policlinico San Matteo
Giulia Giordano (giulia.giordano@unitn.it)
University of Trento https://orcid.org/0000-0002-8600-1738
Kenneth Pesenti
University of Trieste
Franco Blanchini
University of Udine https://orcid.org/0000-0002-4109-5531
Paolo Bolzern
Politecnico di Milano
Patrizio Colaneri
Polytechnic University of Milan
Paolo Sacchi
Division of Infectious and Tropical Diseases, Fondazione IRCCS Policlinico San Matteo, University of Pavia
Giuseppe De Nicolao
University of Pavia
Raffaele Bruno
University of Pavia

Article

Keywords:

Posted Date: April 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1538799/v1

Early detection of variants of concern via funnel plots of regional reproduction numbers

Department of Mathematics, University of Pavia; Pavia, Italy
Simone Milanesi (ORCID: 0000-0002-6314-1965)
Department of Mathematics, Computer Science and Physics, University of Udine; Udine, Italy
Francesca Rosset (ORCID: 0000-0002-1202-2712)
Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo, Pavia
Marta Colaneri
Department of Industrial Engineering, University of Trento; Trento, Italy
Giulia Giordano (ORCID: 0000-0002-8600-1738)
Department of Medicine, University of Trieste; Trieste, Italy
Kenneth Pesenti
Department of Mathematics, Computer Science and Physics, University of Udine; Udine, Italy
Franco Blanchini
Department of Electronics, Information and Bioengineering, Politecnico di Milano; Milan, Italy
Paolo Bolzern
Department of Electronics, Information and Bioengineering, Politecnico di Milano; Milan, Italy
Institute of Electronics, Information Engineering and Telecommunication (IEIIT) of the Italian National Research Council (CNR), Turin, Italy.
Patrizio Colaneri
Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo; Pavia, Italy
Paolo Sacchi
Department of Electrical, Computer and Biomedical Engineering, University of Pavia; Pavia, Italy
Giuseppe De Nicolao (ORCID: 0000-0002-3712-9911)
Department of Clinical, Surgical, Diagnostic, and Pediatric Sciences, University of Pavia
Division of Infectious Diseases I, Fondazione IRCCS Policlinico San Matteo; Pavia, Italy
Raffaele Bruno

Contributions

S.M.: Conceptualisation, Data curation, Formal analysis, Methodology, Validation, Visualisation, Writing—Original draft (Abstract, Introduction, Results, Methods), Writing—Review and Editing.

M.C. & G.G. & K.P.: Methodology, Validation, Writing—Original draft (part of Results and Discussion), Writing—Review and Editing.

F.R.: Data curation, Methodology, Validation, Visualisation, Writing—Original draft (part of Results and Discussion), Writing—Review and Editing.

F.B. & P.B. & P.C. & R.P. & P.S.: Writing—Review and Editing

R.B.: Supervision, Funding acquisition, Writing—Review and Editing

G.D.N.: Conceptualisation, Data curation, Formal analysis, Methodology, Supervision, Validation, Visualisation, Writing—Original draft (Abstract, Introduction, Results, Methods), Writing—Review and Editing.

Corresponding author

Giulia Giordano (ORCID: 0000-0002-8600-1738)

giulia.giordano@unitn.it

Abstract

Tools to early detect the emergence of a new variant of concern are essential to develop strategies that contain epidemic outbreaks and their health-economic-social consequences. For example, knowing in which region a variant of concern appears or starts spreading enables prompt actions to circumscribe the diffusion area. This paper presents 'funnel plots' as a statistical method that can quickly identify regions of a country where the reproduction number is anomalous with respect to the national one, thus triggering cross-cutting research, while keeping false alarms under control. COVID-19 data demonstrate the efficacy of the method in the early detection of Delta and Omicron variants in India, South Africa, England, and Italy, as well as a malfunctioning episode of the diagnostic infrastructure in England.

Introduction

All viruses, including SARS-CoV-2, evolve over time. Mutations happen frequently and, in most cases, have little or no impact on the viral function. However, a group of mutations with similar genetic lineage, denoted by public health organizations as Variants of Concern (VoC), have gained global attention because of their faster spread and evidence for higher transmissibility and possibly higher virulence [1].

Surveillance aimed at the early detection of a new VoC is fundamental. The World Health Organization (WHO) and its international networks of experts closely monitor SARS-CoV-2 variants [2], but a surveillance system at a national and sub-national level is crucial to identify the emergence of new variants with the potential to spread worldwide, as well as the spread of already detected variants. Local authorities are thereby currently encouraged to strengthen surveillance and sequencing capacities, to early detect unusual epidemiological events. However, several countries still have limited capacity, despite the enormous efforts to facilitate the access to existing international networks [3], and the implementation of low-cost whole genome sequencing (WGS) methods [4].

As suggested by SARS-CoV outbreaks [5], we can expect that new SARS-CoV-2 variants with unforeseen mutations will continue to emerge [6,32,33], also with the potential risk of immune evasion [7,8].

The Omicron variant (B.1.1.529 lineage), which contains over 30 mutations in the spike protein, including the same mutations of pre-existing VoC, will definitely not be the last, and possibly not the most challenging we will ever face.

To support monitoring based on epidemiological data, we propose a methodology that is easy to apply and can allow the early detection of anomalous events, consequently triggering further inquiries. With respect to massive genomic sequencing, statistical methods based on epidemiological data are faster and reduce costs and needed resources; of course, they do not replace sequencing, but integrate it and may defer the genomic sequencing methods to a more targeted and purpose-driven framework, to effectively detect potential VoC, and prevent their spread.

The keystone of the approach investigated in the present paper is the use of statistical quality control to monitor the homogeneity of the reproduction numbers estimated in different regions of a country. In particular, a methodology is proposed that is able to account for different sample sizes. In the context of healthcare monitoring, this issue had come under the spotlight in the early 2000, in a series of works [9,10,11]. An example was the detection of abnormal mortality rates in cardiac surgery wards [9]: through the characterization of the baseline variability, it was possible to build control charts with statistical limits which, if exceeded, suggested the existence of an abnormal cause

explaining the anomalous data. When the key performance indicators were affected by the sample size, it was shown that their monitoring could rely on so-called funnel plots [12,13].

In the case of epidemics, anomalies can be detected by comparing the regional effective reproduction number, R_t, whose uncertainty depends on the number of infected subjects in the given region. It is therefore worth investigating if funnel plots could distinguish whether a large regional R_t is due to some special cause, e.g., the emergence of a new VoC, a breakdown of testing infrastructures, or widespread reckless behaviours, or is just caused by statistical fluctuations due to sampling noise.

This work proposes a framework to monitor the onset of anomalies in regional R_t distribution, deriving suitable funnel plots with control limits able to reveal abnormal trends while preventing false alarms. We validate our proposed methodology based on publicly available epidemiological data from Italy, the England, India and South Africa and show how control limits being exceeded promptly reveal the emergence of new more transmissible variants or the malfunctioning of the diagnostic infrastructure.

Results

The funnel plot methodology has been applied to five case studies, corresponding to different stages of the COVID-19 pandemic, chosen in view of their relevance with respect to the spread of VoC's or flaws of the diagnostic infrastructure. Two cases studies refer to England (initial spread of Delta and large failure of a diagnostic lab), and the other three to Italy (initial spread of Omicron), India (first emergence of the Delta variant), and South Africa (first emergence of the Omicron variant). In all cases, the early detection capabilities enabled by statistical process control tools are illustrated and discussed.

Omicron spread in Italy.

We first apply the funnel-plot methodology to the Italian regional data in the period going from 6 December to 31 December 2021, based on epidemiological indicators daily released by the Civil Protection Department, which provides 21 regional time series (for 19 regions and the 2 autonomous provinces of Trento and Bolzano). The Delta variant was dominant in Italy until December 2021, when the Omicron variant started to spread across the country. It is therefore interesting to observe whether and how funnel plots can detect this spread.

The results are summarized in Fig. 1. In the Panels a-d, the estimates of Italian regional R_t's are plotted against the infectious cases on four selected dates. When a single variant, i.e., Delta, is homogeneously present in the country and contact rates do not vary much across regions, differences between estimated R_t's are due to natural variability alone and the 21 points are expected to lie within

the funnel, centered around the national R_t (see Methods), as indeed observed on 7 December 2021 (see Panel a). On 18 December 2021, Lombardy (dark red) crossed the alarm limit (see Panel b) and on 25 December 2021 (see Panel c) it was definitely outside the upper alarm limit. In fact, as confirmed by a survey by the Italian National Institute of Health published on 31 December 2021 [26], Lombardy was the first Italian region to be colonized by the Omicron variant. As other regions became increasingly colonized by the Omicron variant, their R_t's rose as well and, by 31 December 2021, Lombardy was absorbed again within a funnel, now with a higher mean with respect to early December's (see Panel d).

The trend can be monitored by plotting the standardized R_t's on a control chart with ±3 sigma limits (Panel e), where the arrival of the Omicron variant in Lombardy in mid-December is clear. Panel e also shows that a few regions (Bolzano, Friuli Venezia Giulia and Veneto) exhibit an undershoot that exceeds the lower alarm limit. Although further investigations are required, this phenomenon might be attributed to an earlier recovery from the previous Delta pandemic wave in these regions.

Delta emergence in India.

We apply our methodology to epidemic data from India, where, from 10 February to 5 March 2021, the Delta variant emerged and started spreading from the state of Maharashtra. In figure 2, in Panels a-d, funnel plots on four selected days are shown, with colour-coded circles representing the R_t's of the 36 Indian states. While on February 13 all circles fell within the funnel, on February 16 the state of Maharashtra (dark red) exceeded the alert threshold (thus suggesting when the Delta variant started spreading), further departing from the mean on February 22. Lastly, on 4 March 2021, the R_t dispersion chart of all the regions but Kerala (orange) shaped a new funnel, with a higher mean, which incorporated Maharashtra back in. The peculiar dropping of Kerala's R_t below the lower alert threshold, despite the very high number of infectious cases, might be explained by the overlapping of Alpha and Delta variants during the same period, resulting in a lower R_t value in Kerala than in the areas predominantly hit by the Delta variant.

The trend can be monitored by plotting the standardized R_t's on a control chart with ±3 sigma limits (see Panel e), where the rise of the Delta variant in Maharashtra is clearly visible since mid-February 2021. One month later, on March 17, it was made public that a 10-lab research consortium had alerted the Union health ministry about a new variant spreading in Maharashtra [27], and a week later the Indian Ministry of Health and Family Welfare issued a press release on the new VoC [28]. This case study demonstrates that the use of funnel plots could have allowed an earlier detection of the variant.

Omicron emergence in South Africa.

From 25 October to 3 December 2021, the Omicron variant colonized South Africa, starting with the province of Gauteng. In Fig. 2 (Panels f-i) four funnel plots are shown, with colour-coded circles representing the R_t's of the South African provinces. Until the very beginning of November 2021, the Delta variant prevailed and the differences in R_t levels across provinces were merely a result of natural fluctuations (see Panel f). By mid-November the Gauteng province crossed the upper alert threshold (see Panel g) and then further diverged (see Panel h). This is precisely the timing when the Omicron variant was first identified, as declared by the WHO [29], and became a threat [30]. By 2 December 2021, Gauteng was reabsorbed within the funnel, now with a definitely higher mean, following the spread of Omicron in the other provinces and the consequent rise of their R_t (see Panel i). On the control chart with ±3 sigma limits (Panel j), the out-of-control trajectory of the Gauteng province (red) is plainly evident.

Omicron spread in England.

From 1 December 2021 to 6 January 2022, the Omicron variant massively spread in England. The four funnel plots show colour-coded circles corresponding to the R_t's of the English regions, see Panels k-n of Fig. 2. On December 3, all the regions were within the alarm limits (Panel k). By 10 December 2021, the London region was out of the funnel (Panel l), further diverging from the upper limit on 16 December (Panel m). This suggests that Omicron was more prevalent in London than in the rest of England and, indeed, on 13 December 2021 it was reported that 20% of the cases in England and over 44% in London were due to Omicron [31]. As the other regions were colonized, the distribution of their R_t's moved upward and, on 24 December 2021, the London region was again inside the funnel (Panel n). An early detection would have been allowed by the funnel-plot control chart, where London first crossed the alarm limit in early December (Panel o).

Immensa scandal in England.

Our last case study concerns England in the period from 27 August to 25 September 2021. In Panels a-d of Fig. 3, four funnel plots in selected dates are displayed, with colour-coded circles corresponding to the R_t's of the English regions. On 28 August 2021, all English regions were within the funnel (Panel a). By 2 September 2021, the South West (red) had moved below the lower alarm limit (Panel b) and remained below the lower limit for about two weeks (Panel e). The timing of this swing coincides with the period during which the Immensa lab in Wolverhampton gave some 43,000 incorrect negative tests relative to South West and West Midlands territories [14,15]. While the

suspension of lab operations came in mid-October, the control chart indicated an out-of-control condition as early as late August and would have allowed a much earlier detection of the anomaly.

Discussion

We proposed funnel plots as a valuable framework for the early detection of a new emerging or imported VoC and showed its effectiveness in five real-life scenarios based on epidemiological data from Italy, India, South Africa and England. These case studies demonstrate that the proposed methodology, besides being direct and inexpensive, allows the early detection of anomalies with various possible causes, ranging from the emergence of a new VoC, and its colonization of new countries, to flaws in the diagnostic system, e.g., the Immensa COVID-19 testing scandal. Once the method identifies anomalous patterns, further inquiries are needed to assess its cause.

Funnel plots provide an innovative and rigorous tool for monitoring the distribution of regional R_t's. Our method can be seen as an extension to epidemiology of the funnel charts advocated by Spiegelhalter in the assessment and comparison of institutional performances in the healthcare sector [12]. Before then, funnel plots were mainly known as a standard visual tool for investigating publication and other bias in meta-analysis studies. As such, they have also been employed in the context of COVID-19 meta-analyses, see e.g., [34].

Prompt identification of a VoC before its unavoidable large-scale spread, leading to impactful public health implications, is a key goal in the control of the SARS-CoV-2 pandemic and in preventing and controlling future pandemics. However, as the relentless and flashy worldwide dissemination of the Omicron variant has largely proven, some doubts remain about the most effective way to achieve this goal.

Although some rRT-PCR–based algorithms and/or NAAT-based screening assays have been proposed for the early identification of VoCs [16,17] and might be implemented in routine laboratories [18], the WGS, or at least the complete or partial sequencing of the spike (S) protein-gene remains the unique tool able to both effectively identifying the different variants and follow the evolution of SARS-CoV-2 [19,20]. However, WGS is time consuming, expensive, and needs dedicated structures with technical experience to be timely implemented. Furthermore, it is challenging to be applied on low viral loads samples [21].

It is exactly in this breach that the potential support of surveillance based on funnel plots might accelerate the detection of a new VoC, without requiring, at least initially, the backup of a specialized microbiology laboratory. The value of WGS is undisputed, but, in a resource-limited setting, the combined use of easy and inexpensive data-driven statistical methodologies for surveillance may support a more targeted and focused adoption of genomic sequencing, guided by the detection of a

suspicious epidemiological pattern through funnel plots. Therefore, funnel plots are not only extremely useful where sequencing is lacking due to scarce resources but are also precious to inform and suggest where sequencing efforts should be concentrated. Furthermore, funnel plots allow the detection of anomalies that cannot be revealed by sequencing, such as failures of the testing infrastructure, as shown by the Immensa case study.

Thanks to the statistical underpinning of the methodology, the natural variability of the phenomenon is taken into account, which helps to prevent false alarms even in the presence of "dirty" data (as long as they are statistically stable), e.g., due to late registration of new cases. Although polished data may arrive with even weeks of delay, funnel plots can be used successfully in real-time based on dirty data, which can notably anticipate the detection of anomalies and allow prompt interventions. For instance, the Italian funnel plots of the first case study were drawn from daily published data.

Other authors have proposed the application of statistical process control methods for monitoring the evolution of the COVID-19 pandemic, see [35], where hybrid control charts were proposed to detect within a geographic area the start and end of exponential growth in reported deaths. An interesting use of hybrid control charts was investigated in [36], keeping under control exponential and non-exponential growth and decline of cases, disaggregated at regional and subregional level, to inform local mitigation and containment strategies. Compared to these studies, our approach leverages the characterization of collective distribution of regional R_t's: we do not monitor each region individually, but rather the homogeneity of the distribution is surveilled.

In view of its nature, the proposed method reveals the loss of statistical stability, but cannot of course reveal its cause. Consistently with common quality control practices, it can be used as a trigger to start an inspection. Out of our five case studies, in four cases the regions going out-of-control turned out to be affected by the emergence, or early colonization, of a new VoC, while in the fifth case (Immensa scandal) the anomaly was due to a completely different cause: the failure of testing infrastructures. Therefore, the funnel plot cannot be strictly seen as a VoC-detector, but more precisely as an anomaly detector: early detection enables focused inquiries aimed at discovering what triggered the anomaly.

Methods

Data

Data regarding new positive cases can be obtained from publicly available sources: https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni for Italian data, https://data.covid19india.org/ for Indian data, https://mediahack.co.za/datastories/coronavirus/data/# for South African data, https://coronavirus.data.gov.uk/details/download for English data.

The distribution for the serial interval is obtained from [22].

All data were filtered using a double seven-day moving average, which was necessary because of systematic errors in the data, partly due to the weekly periodicity, partly due to other delays and bureaucratic errors. The outliers were corrected as follows.

Let $p(t)$ be the number of new positive cases on a region at time t. On $t = 23$ November 2021, the South African data presents an irregular data item. It is replaced by an imputation of the following form

$$log\,(p(t)) : (log\,(p(t-1)) + log\,(p(t+1)) = log\,(p(t-7)) : (log\,(p(t-8)) + log\,(p(t-6))$$

When the Indian data on new positives are negative, they are replaced by an imputation of the following form

$$p(t) = \frac{p(t-1) + p(t+1)}{2}$$

A further imputation of the same type is performed for the Chhattisgarh region on 28 January 2021.

Funnel plots and control charts

A funnel plot is a graphical tool for comparing the characteristics of units of analysis. In particular, a measured or estimated quantity is plotted against an interpretable measure of its precision. A funnel plot is composed of four elements [12]: (i) an indicator Y that represents the quantity to be monitored, (ii) a reference value θ that specifies the expectation of the indicator, (iii) a precision parameter ρ that determines the accuracy with which the indicator is measured, (iv) the control limits y_{lower}, y_{upper} that specify whether a point is in or out-of-control. An example of funnel plot can be seen in Fig 1. Each dot (ρ_i, y_i) corresponds to the i-th region, where ρ_i is the number of infectious cases and y_i is the region's reproduction number R_t for a given day. The horizontal line θ shows the national average R_t and the funnel-shaped pair of control limits y_{lower} and y_{upper} show where we expect the Italian regions to lie if their R_t's were statistically indistinguishable from each other, see Panel d in Fig 1.

In several circumstances an exact or approximate normal distribution of the indicator Y can be assumed

$$Y|\theta, \rho \sim N[\theta, g(\theta)/\rho] \qquad (1)$$

where g is a function of θ. Under this null hypothesis, with probability $1-\alpha$,

$$\theta - z_{\alpha/2}\sqrt{g(\theta)/\rho} \leq Y \leq \theta + z_{\alpha/2}\sqrt{g(\theta)/\rho}$$

where $z_{\alpha/2}$ is such that $P(Z \leq z_{\alpha/2}) = 1-\alpha/2$ for a standard normal variable Z. For instance, $z_{\alpha/2} = 1.96$, when $\alpha = 5\%$, and $z_{\alpha/2} = 3$, when $\alpha = 0.27\%$. This means that in $100(1-\alpha)\%$ of cases Y is expected to lie within the lower and upper control limits defined as

$$y_{lower} = \theta - z_{\alpha/2}\sqrt{g(\theta)/\rho}$$

$$y_{upper} = \theta + z_{\alpha/2}\sqrt{g(\theta)/\rho}$$

By introducing the Z-score

$$z_i = \frac{y_i - \theta}{\sqrt{g(\theta)/\rho}}$$

we have that $P(|z_i| \leq z_{\alpha/2}) = 1-\alpha$. In Statistical Process Control, the common practice is to select a false alarm probability as small as $\alpha = 0.27\%$., corresponding to $z_{\alpha/2} = 3$. A point lying outside the funnel or, equivalently, a point whose Z-score is either less than -3 or greater than 3, is said to be *out of (statistical) control* and is deemed worthy of study to find a special cause of variation that explains its departure from the mean. According to the terminology of statistical decision theory, α is the false positive rate, i.e., the probability of rejecting the null assumption when it is not actually violated. In other terms, when ± 3 control limits are used, there is a 0.27% probability of reporting an out-of-control point when no special cause of variation is actually perturbing the process and the outlier arises by pure chance under common causes of variation.

When the indicators y_i measure a frequency of occurrence, e.g., the mortality rates in heart surgery units, it is reasonable to assume a binomial model, with θ representing the probability of the event and ρ_i the number of surgeries in the i-th unit. For the binomial model, the variance of y_i is $\theta(1-\theta)/\rho_i$ so that, given θ, the variance of y_i is completely specified. For a large enough ρ, the binomial converges to a normal random variable that follows distribution (1) with $g(\theta) = \theta(1-\theta)$. An analogous case is when the products $\rho_i y_i$ are Poisson distributed with expectation $\rho_i\theta$. If $\rho_i\theta$ is large enough, the indicators y_i are normally distributed as (1) with $g(\theta) = \theta$. Therefore, for both the binomial and Poisson model it appears that estimating the mean of y_i suffices to specify both the centerline and the alarm limits of the funnel plot.

However, as discussed in [13], the use of funnels based on the ideal variance implies that the number of units of analysis lying outside the alarm limits exceeds by large the theoretical false positive rate.

This phenomenon, well known in the statistical literature, goes under the name of *overdispersion:* "*This typically arises when there is insufficient risk adjustment; there are many small institutional factors that contribute to excess variability, and these may not be particularly important nor indicate poor quality. The consequence is that, if one is not careful, the majority of institutions can be labelled as abnormal, and this appears a contradiction in terms"* [13]. This can be dealt with by modifying (1) with the introduction of a suitable overdispersion parameter ϕ to be estimated from data:

$$Y|\theta,\rho \sim N[\theta, \phi g(\theta)/\rho] \tag{2}$$

The control limits and the Z-scores are redefined accordingly as

$$y_{lower} = \theta - z_{\alpha/2}\sqrt{\phi g(\theta)/\rho}$$

$$y_{upper} = \theta + z_{\alpha/2}\sqrt{\phi g(\theta)/\rho}$$

$$z_i = \frac{y_i - \theta}{\sqrt{\phi g(\theta)/\rho}}$$

When the indicators to be monitored are time series depending on time index t, i.e., $y_i = y_i(t)$, a distinct funnel plot can be drawn for each time instant. However, it might be desirable to have a graphical tool that displays the trend of the Z-scores, highlighting the out-of-control episodes. Under (2), we have that $z_i \sim N[0,1]$, so that the Z-scores can be plotted on a control chart with zero centerline and alarm limits equal to $\pm z_{\alpha/2}$, see panel e in Fig. 1.

Distribution of regional R_t's

The reproduction number at time t, named R_t, captures the number of secondary infections from a population including both susceptible and immune individuals. For its estimation, a range of model frameworks and estimation procedures have been proposed [23]. Herein we adopt the approach of Cori et al. [24] that makes minimal assumptions about the mathematical model of the epidemic process. Cori's formula uses the time series of the new cases and estimates of the distribution of the generation time. i.e., the time between infections.

According to [24] the estimate $\hat{R}_t$ of the instantaneous reproduction number R_t is obtained as

$$\hat{R}_t = \frac{I_t}{\sum_{s=1}^{t} w_s I_{t-s}} \tag{3}$$

where I_t denotes the daily number of new infected cases and w_s are the coefficients, summing to one, of the infectivity profile, often approximated by the distribution of the serial interval. The denominator

$$\Lambda_t = \sum_{s=1}^{t} w_s I_{t-s}$$

can be interpreted as the total infectiousness of individuals that are currently infected at time t. In view of the typical models of the infectivity profile, e.g., lognormal or gamma density functions, Λ_t is a smoothed version of the time series I_t of the daily cases. If seven-day moving averages are used to filter out weekly oscillations, I_t is already smooth by itself, so that the resulting Λ_t will be insensitive to the precise shape of the infectivity profile. This feature may prove helpful when a new VoC arises whose infectivity profile is unknown or known only approximately.

To derive the distribution of $\hat{R}_t$, we model disease transmission as a Poisson process with mean $R_t\Lambda_t$:

$$I_t|R_t, \Lambda_t \sim Pois[R_t\Lambda_t],$$

When its mean is large enough, a normal approximation can be used:

$$I_t|R_t, \Lambda_t \sim N[R_t\Lambda_t, R_t\Lambda_t]$$

In view of (3), it follows that $\hat{R}_t|R_t, \Lambda_t \sim N[R_t, R_t/\Lambda_t]$. For the sake of interpretability, rather than using the notion of total infectiousness Λ_t, it is more intuitive to refer to the number of infectious individuals. For this purpose, we introduce the parameter

$$\gamma = \frac{1}{\sum_{s=1}^{\infty} s w_s}$$

i.e., the inverse of the mean serial interval, which, within the SIR model, can be interpreted as the removal rate [25]. Then, $\rho_t = \Lambda_t/\gamma$ represents the number of individuals that are infected at time t. Letting $\theta = R_t, g(\theta) = R_t/\gamma$, it follows that

$$\hat{R}_t|\theta, \rho_t \sim N[\theta, g(\theta)/\rho_t]$$

Comparing the above distribution with (1), it follows that, for any given t, the scatter plot of $\hat{R}_t$ against ρ_t is indeed a funnel plot. Also in this case, it is convenient to allow for an overdispersion parameter ϕ, so that, in accordance with (2), the final model becomes

$$\hat{R}_t|\theta, \rho_t \sim N[\theta, \phi g(\theta)/\rho_t] \tag{4}$$

A useful byproduct of accounting for overdispersion is that ϕ accommodates the effects on the variance of $\hat{R}_t$ of possible errors or uncertainties in the estimated mean serial interval. Indeed, the variance of $\hat{R}_t$ is inversely proportional to γ, but the effect of a wrong γ is automatically compensated when estimating ϕ from the data.

Parameter estimation

To estimate the model parameters, observe that, under (2), the estimated reproduction number $\hat{R}_t^i$, of the i-th region can be written as

$$Y_i = \theta + \frac{\varepsilon_i}{\sqrt{x_i}} \tag{5}$$

where $Y_i = \hat{R}_t^i$, $x_i = \rho_t^i$ is the number of infectious individuals, and $\varepsilon_i \sim N[0, \sigma^2]$, $i = 1, \ldots, n$, are mutually independent. Letting $v_i = \varepsilon_i/\sqrt{x_i}$, and converting in matrix form, the model becomes

$$Y = \Phi\theta + v$$

where $Y = [\ldots\ Y_i\ \ldots\ ]'$, $\Phi = [\ldots 1 \ldots]'$, $v \sim N[0, \sigma^2\Sigma]$, and

$$\Sigma = \begin{pmatrix} \frac{1}{x_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{x_n} \end{pmatrix}$$

According to the Gauss-Markov theorem, the estimated parameters are

$$\hat{\theta} = (\Phi'\Sigma^{-1}\Phi)^{-1}\Phi'\Sigma^{-1}Y$$

$$\widehat{\sigma^2} = \frac{1}{n}\, e' \Sigma^{-1} e \qquad (6)$$

where $e = Y - \Phi\, \hat{\theta}$ is the vector of the residuals. Recalling that $g(\theta) = \theta/\gamma$, the overdispersion parameter ϕ is estimated as $\hat{\phi} = \gamma\widehat{\sigma^2}/\hat{\theta}$. In order to reduce the effect of possibly spurious outliers, data winsorization can be performed as detailed in [12].

In practice, there is no guarantee that the regional estimates Y_i satisfy the assumption (5), because, due to some special cause of variation, homogeneity could have been disrupted resulting in $E[Y_i] \neq \theta$ for some i. According to the statistical process control methodology, such situation can be dealt with as follows. First the parameters θ and ϕ are estimated and the Z-scores z_i, $i = 1, \ldots, n$, for all the n regions are computed. When $|z_i| > z_{\alpha/2}$, the i-th region is labeled as out-of-control, and the parameters θ and ϕ are re-estimated after removing the i-th observation from the dataset. The procedure is iteratively repeated until no Z-score exceeds the alarm limits.

Setting the alarm limits at $z_{\alpha/2}$ guarantees that the false alarm rate is α if the Z-scores are normally distributed. Even when σ^2 is estimated according to (6), $\pm z_{\alpha/2}$ control limits are often maintained as done in [12], in which case the actual false alarm rate is going to exceed α. For the false alarm rate to remain exactly equal to α, wider alarm limits derived from the Student's t distribution with $n - 1$ degrees of freedom might be used. For large values of n, obviously, there is no practical difference.

Data availability statement

Data regarding new positive cases can be obtained from publicly available sources:
https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni for Italian data;
https://data.covid19india.org/ for Indian data;
https://mediahack.co.za/datastories/coronavirus/data/# for South African data;
https://coronavirus.data.gov.uk/details/download for English data.

Code availability statement

Code will be made available online.

References

[1] https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html
[2] Guidance for surveillance of SARS-CoV-2 variants: Interim guidance, 9 August 2021 (who.int)
[3] https://www.who.int/initiatives/global-influenza-surveillance-and-response-system/virus-sharing/shipping-and-logistics-activities

[4] Gohl DM, A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genomics*. 2020 Dec 4;21(1):863. doi: 10.1186/s12864-020-07283-6. PMID: 33276717; PMCID: PMC7716288.

[5] Zhao Z et al. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol*. 2004 Jun 28; 4:21. doi: 10.1186/1471-2148-4-21. PMID: 15222897; PMCID: PMC446188

[6] Harvey et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* 19, 409–424 (2021), doi: 10.1038/s41579-021-00573-0.

[7] Callaway E. Fast-spreading COVID variant can elude immune responses. *Nature*. 2021 Jan;589(7843):500-501. doi: 10.1038/d41586-021-00121-z. PMID: 33479534.

[8] Eguia RT et al. A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog*. 2021 Apr 8;17(4):e1009453. doi: 10.1371/journal.ppat.1009453. PMID: 33831132; PMCID: PMC8031418.

[9] Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet*. 2001 Feb 10;357(9254):463-7. doi: 10.1016/s0140-6736(00)04019-8. PMID: 11273083.

[10] Goldstein, Harvey, David J. Spiegelhalter. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159.3 (1996): 385-409.

[11] Aylin, Paul, et al. Following Shipman: a pilot system for monitoring mortality rates in primary care. *The Lancet* 362.9382 (2003): 485-491.

[12] Spiegelhalter, David J. Funnel plots for comparing institutional performance. *Statistics in medicine* 24.8 (2005): 1185-1202.

[13] Spiegelhalter, David J. Handling over-dispersion of performance indicators. *BMJ Quality & Safety* 14.5 (2005): 347-351.

[14] https://www.gov.uk/government/news/testing-at-private-lab-suspended-following-nhs-test-and-trace-investigation

[15] Fetzer, Thiemo, et al. Measuring the Epidemiological Impact of a False Negative: Evidence from a Natural Experiment. University of Warwick, Department of Economics, 2021.

[16] Matic, Nancy, et al. Early Release-Rapid Detection of SARS-CoV-2 Variants of Concern, Including B. 1.1. 28/P. 1, in British Columbia, Canada.

[17] Neopane P, Nypaver J, Shrestha R, Beqaj SS. SARS-CoV-2 Variants Detection Using TaqMan SARS-CoV-2 Mutation Panel Molecular Genotyping Assays. *Infect Drug Resist*. 2021;14:4471-4479. Published 2021 Oct 27. doi:10.2147/IDR.S335583

[18] Ong, David SY, et al. Rapid screening method for the detection of SARS-CoV-2 variants of concern. *Journal of Clinical Virology*, 2021, 141: 104903.

[19] World Health Organization. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. 2021. Jan 8 https://www.who.int/publications/i/item/9789240018440

[20] Boudet, Agathe, et al. Limitation of screening of different variants of SARS-CoV-2 by rt-pcr. *Diagnostics*, 2021, 11.7: 1241.

[21] Charre et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus evolution*, 2020, 6.2: veaa075.

[22] Geismar, Cyril, et al. Serial interval of COVID-19 and the effect of Variant B. 1.1. 7: analyses from prospective community cohort study (Virus Watch). *Wellcome open research*, 2021, 6.

[23] Anderson, et al. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. *The Royal Society*, 2020.

[24] Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, *178*(9), 1505-1512.

[25] BAR-ON, Yinon M., et al. A quantitative compendium of COVID-19 epidemiology. *arXiv preprint arXiv:2006.01283*, 2020.

[26] Survey ISS (Istituto Superiore di Sanità) https://www.iss.it/en/cov19-cosa-fa-iss-varianti

[27] Singh, Jasdeep, et al. SARS-CoV-2 variants of concern are emerging in India. *Nature medicine*, 2021, 27.7: 1131-1133.

[28] Genome Sequencing by INSACOG shows variants of concern and a Novel variant in India https://pib.gov.in/PressReleasePage.aspx?PRID=1707177

[29] Bull World Health Organ 2022;100:4–5. doi: 10.2471/BLT.22.010122 https://apps.who.int/iris/bitstream/handle/10665/351061/PMC8722635.pdf

[30] WHO et al. Weekly Bulletin on Outbreak and other Emergencies: Week 48: 22-28 November 2021. 2021.

[31] Oral statement on COVID-19 by the Health and Social Care Secretary, 2021, 13 December https://www.gov.uk/government/speeches/health-and-social-care-secretary-oral-statement-on-covid-19

[32] E Callaway. Beyond Omicron: what's next for Sars-Cov-2 evolution. *Nature.* 2022 https://www.nature.com/articles/d41586-021-03619-8,

[33] Gov UK, Long term evolution of SARS-CoV-2, 26 July 2021 https://www.gov.uk/government/publications/long-term-evolution-of-sars-cov-2-26-july-2021/long-term-evolution-of-sars-cov-2-26-july-2021

[34] Zhang A, Leng Y, Zhang Y, Wu K, Ji Y, Lei S, Xia Z. Meta-analysis of coagulation parameters associated with disease severity and poor prognosis of COVID-19. Int J Infect Dis. 2020 Nov;100:441-448. doi: 10.1016/j.ijid.2020.09.021. Epub 2020 Sep 15. PMID: 32947052; PMCID: PMC7490635.

[35] Perla RJ, Provost SM, Parry GJ, Little K, Provost LP. Understanding variation in reported covid-19 deaths with a novel Shewhart chart application. *Int J Qual Health Care*. 2021 Mar 5;33(1):mzaa069. doi: 10.1093/intqhc/mzaa069. PMID: 32589224; PMCID: PMC7337871.

[36] Inkelas M et al. Using control charts to understand community variation in COVID-19. *PLoS One*. 2021 Apr 30;16(4):e0248500. doi: 10.1371/journal.pone.0248500. PMID: 33930013; PMCID: PMC8087083.

Acknowledgements

This research received funding from the European Union's Horizon 2020 Research and Innovation Program 'PERISCOPE: Pan European Response to the ImpactS of COvid 19 and future Pandemics and Epidemics' under grant agreement no. 101016233, H2020-SC1-PHE-CORONAVIRUS-2020-2-RTD.

Ethics declarations

No ethics review and informed consent/animal welfare protocols were needed because the research made use of epidemiological time series that are publicly accessible.

Competing interests

All the authors declare the absence of competing interests.

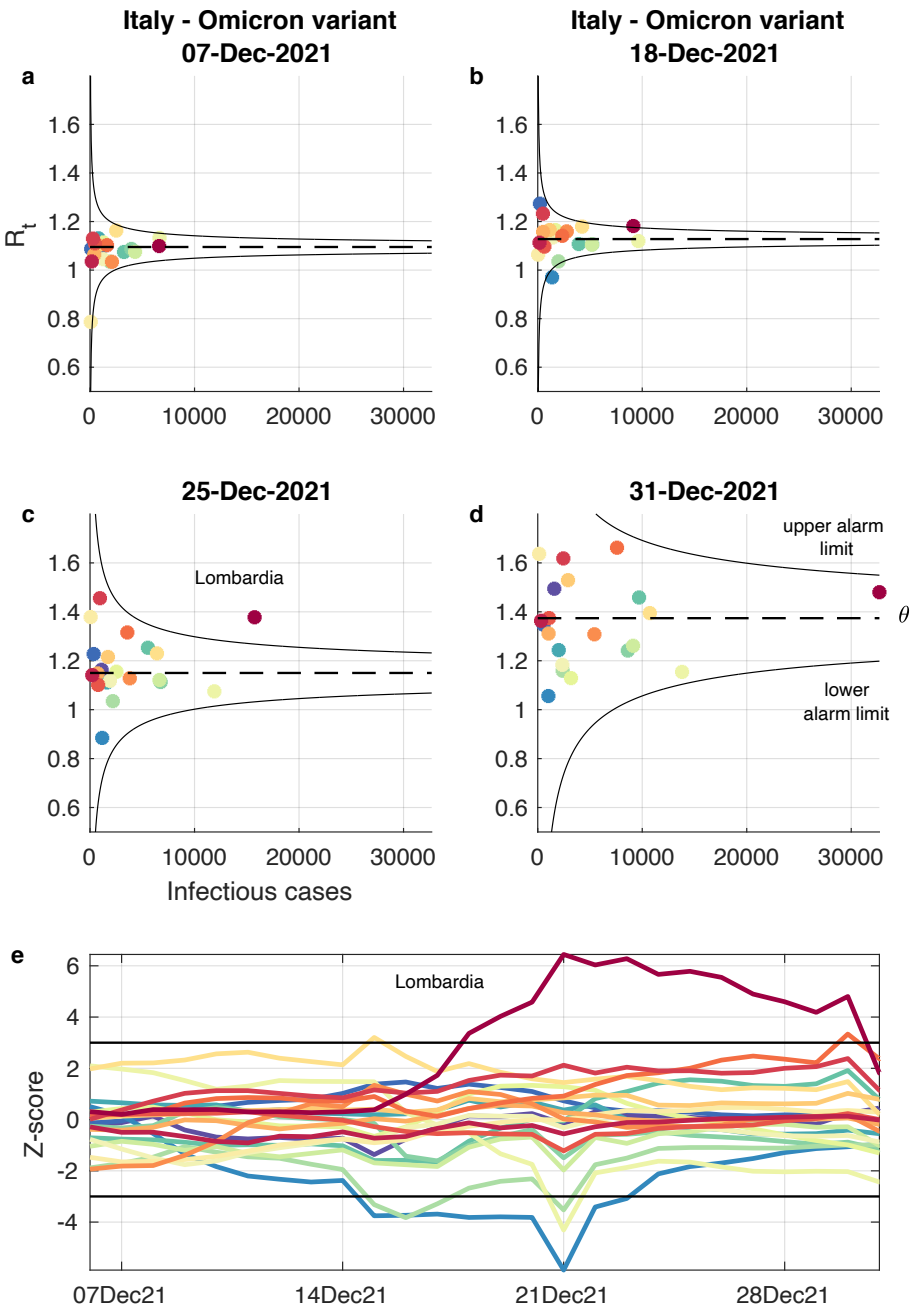

Fig. 1 Monitoring regional reproduction numbers (R_t's): funnel plots and control chart. In Panels a-d, Italian regional R_t's (colour-coded circles) are plotted against the infectious cases on four selected dates. When the epidemic evolution is homogenous across regions, differences between R_t's are exclusively due to natural statistical variability and the circles are expected to lie outside the black alarm limits only in the 0.27% of cases. The alarm limits have the shape of a funnel because the variance of the estimated R_t is inversely proportional to the number of infectious cases. The central dashed line represents the average R_t. A circle is out of statistical control if it lies outside the black funnel. Out-of-control circles might therefore reveal anomalies that disrupt the homogeneity between regions. In Panels a-d, the majority of the points, lying in the funnel, are essentially indistinguishable and therefore not even named. On 18 December 2021, Lombardy (dark red) crossed the alarm limit and on 25 December 2021 it was completely outside the upper alarm limit. As confirmed by a survey by the Italian National Institute of Health, Lombardy was the first Italian region being colonized by the Omicron variant. As the other regions are colonized too, the distribution of their R_t's moves upward and, on 31 December 2021, Lombardy is again inside the funnel. The trend can be monitored by plotting the standardized R_t's on a control chart with ± 3 sigma limits (Panel e), where the arrival of the Omicron variant in Lombardy in mid-December is clearly visible.

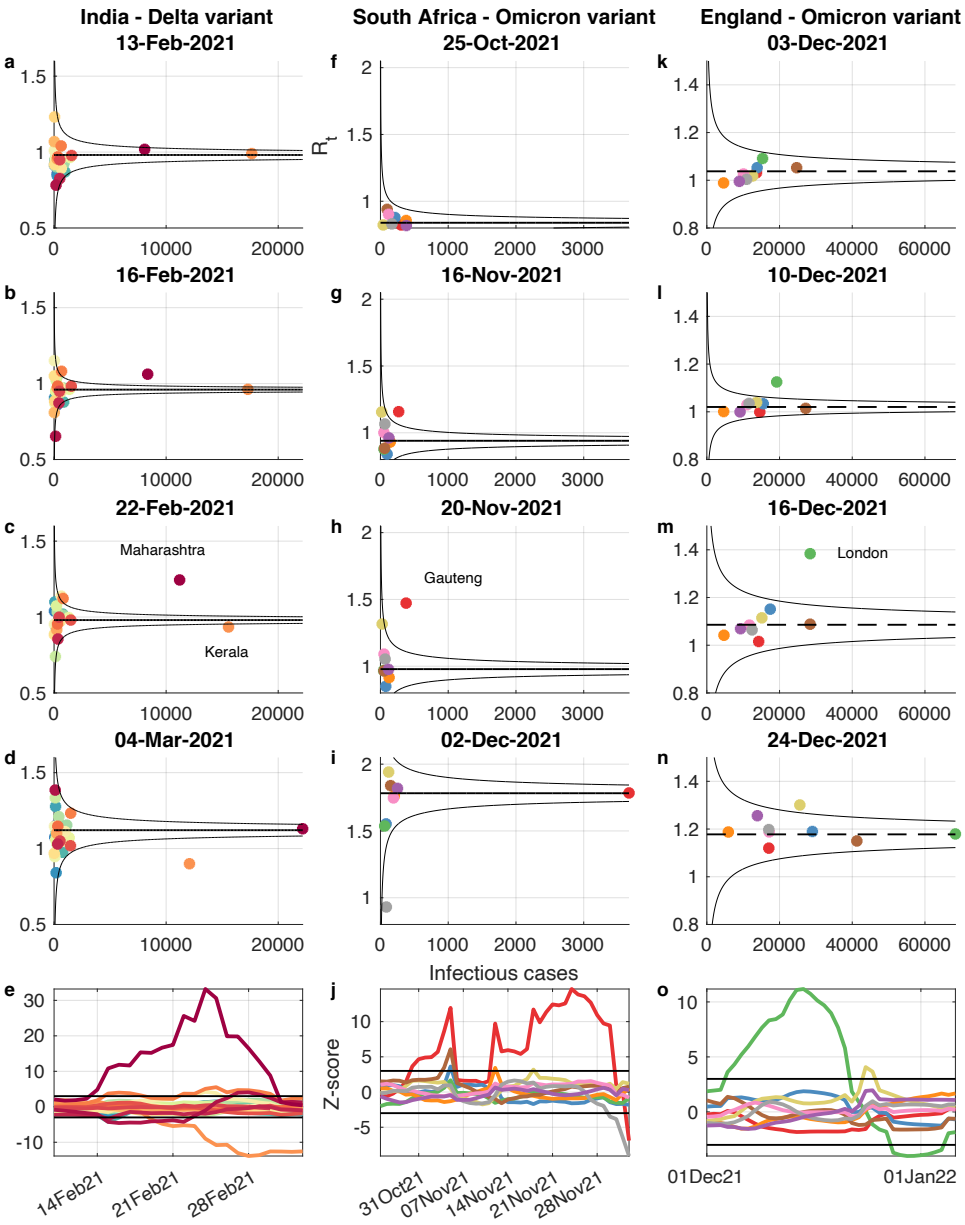

Fig. 2 Funnel plots help detect anomalies: spread of Delta in India and of Omicron in South Africa and England. India: in Panels a-d, the funnel plots in four selected days are displayed, with colour-coded circles corresponding to the R_t's of the Indian states. On 13 February 2021, all points are within the funnel, but on 16 February 2021, when Delta variant starts spreading, there is an out-of-control point corresponding to Maharashtra (dark red), which on 22 February 2021 is further apart from the mean. Finally, on 4 March 2021 the R_t's of all regions except Kerala (orange) converge to a new distribution characterized by a higher R_t. The trend can be monitored by plotting the standardized R_t's on a control chart with ± 3 sigma limits, see Panel e, where the rise of the Delta variant in Maharashtra is clearly visible. South Africa: in Panels f-i, the funnel plots in four selected days are displayed, with colour-coded circles corresponding to the R_t's of the South African provinces. The rise of the Omicron variant in the Gauteng province (red) is well visible both in the funnel plots and in the control chart reported in Panel j. England: in Panels k-n, the funnel plots in four selected days are displayed, with colour-coded circles corresponding to the R_t's of English regions. The spread of Omicron in England started from the London region (green) whose R_t on 10 December had already crossed the alarm limit. As the other regions are colonized, the distribution of their R_t's moves upward and, on 24 December, the London region is again inside the funnel, as also seen in the control chart reported in Panel o.

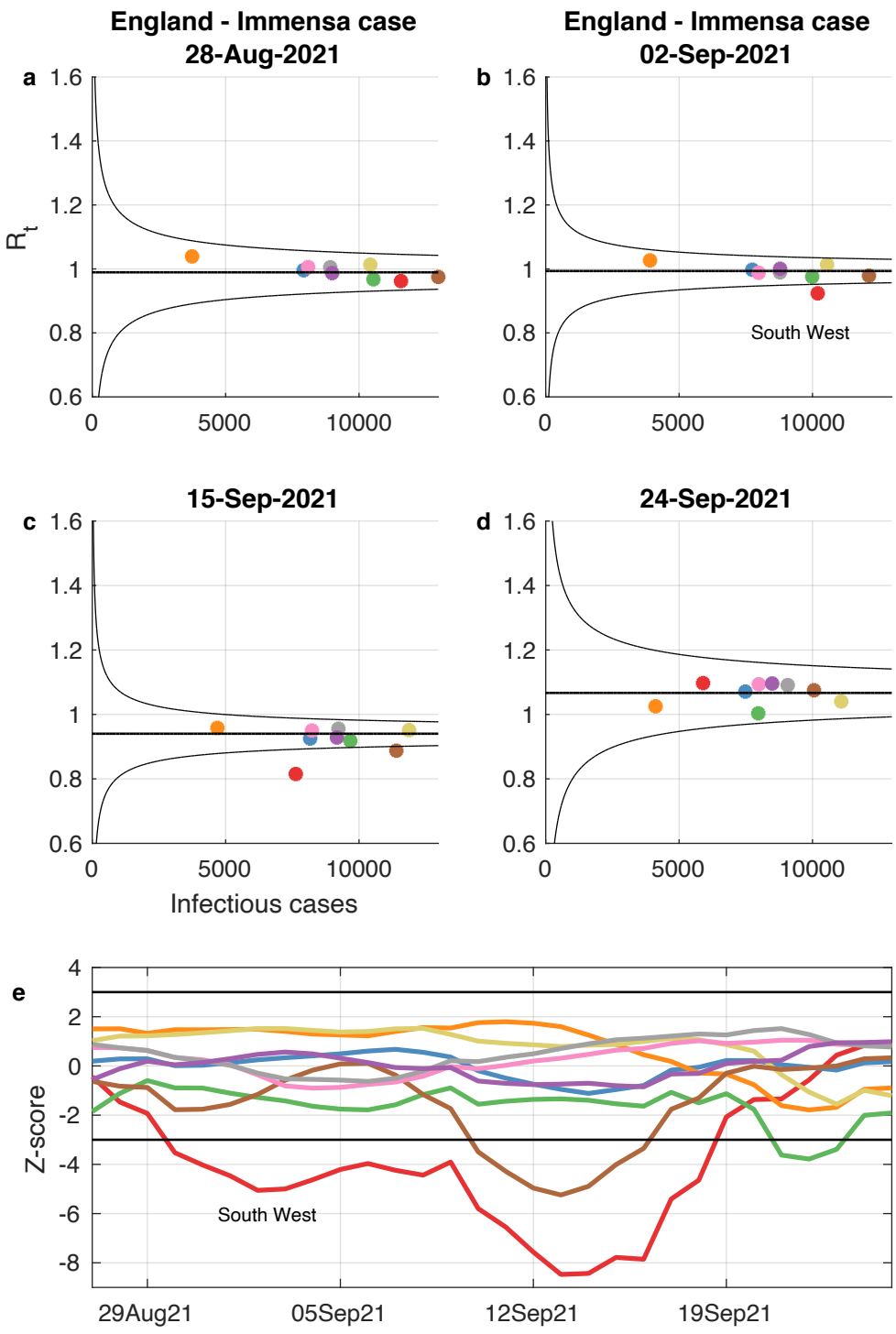

Fig. 3 Funnel plots help detect anomalies: the incorrect negative tests of the Immensa lab in England. In Panels a-d, the funnel plots in four selected days are displayed, with colored circles corresponding to the R_t's of the England regions. On 28 August 2021, all circles are inside the funnel, but on 2 September 2021 there is an out-of-control point below the lower alarm limit corresponding to South West (red), which is further apart from the mean on 15 September 2021, when also West Midlands (brown) is below the lower limit. Finally, on 24 September 2021 the R_t's of all regions return within the limits. The anomalous decrease of South West's R_t corresponds to the period during which the Immensa lab (Wolverhampton) gave some 43,000 incorrect negative tests relative to South West and West Midlands. The whole trend can be monitored by plotting the standardized R_t's on a control chart with ± 3 sigma limits, see Panel e. The suspension of lab operations came in mid-October, while the control chart indicated an out-of-control condition as early as late August.

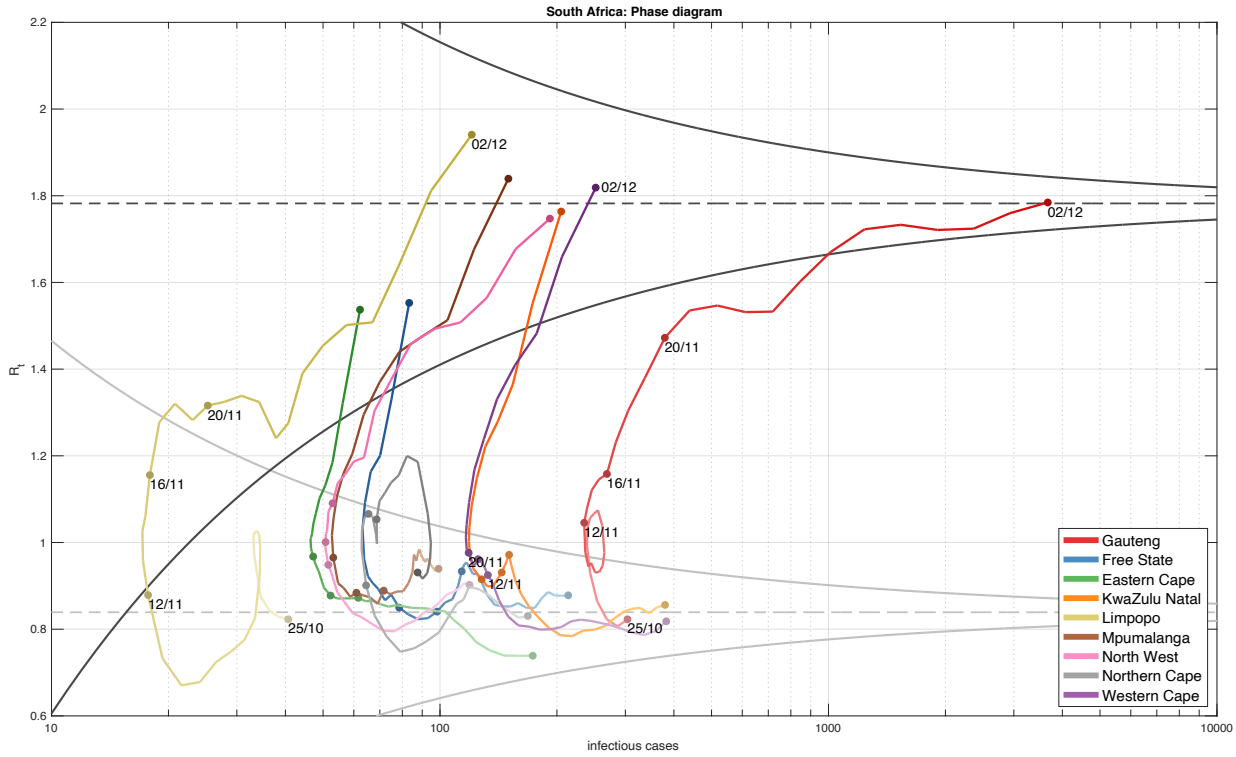

Fig. 4 Moving funnels: the effect of Omicron on the R_t distribution in South Africa. The figure displays the trajectories of infectious cases and R_t's of the South African provinces from 25 October 2021 to 2 December 2021, with the colours getting darker over time. The funnel plots of 25 October 2021 (grey) and 2 December 2021 (black) are plotted with their mean (dashed) and alarm limits (continuous). On 25 October 2021, before the spread of Omicron, the average R_t is below 1 and all points lie inside the grey funnel plot. In the subsequent days, Gauteng's R_t (red) moves upwards, followed by the other provinces. Eventually, on 2 December 2021, the R_t's of all provinces, except Northern Cape, lie inside the black funnel plot, whose mean is about 1.8. Overall, Omicron causes an upward escape of the province where it first becomes dominant (Gauteng, red line), followed by a collective drift of the R_t's of other provinces, until a new funnel, i.e., the black one, is established at a higher level. The trajectories go leftwards when R_t is less than one, because the infectious cases tend to decrease, while the trajectories go rightwards when R_t is greater than one. Therefore, the trajectories exhibit a characteristic clockwise trend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nreditorialpolicychecklistfilled.pdf
- nrreportingsummaryfilled.pdf