Journal of Biostatistics and Biometric Applications
Volume 8 | Issue 1
ISSN:2455-765X

Research Article Open Access

Trends and Day-of-the-Week Effects Decomposition in the Mean and Dispersion in the Number of Subjects Testing Positive for COVID-19 in the UK in 2020

Abdelmajid Djennad*

Statistics, Operational Research and Mathematics (STORM) Research Centre, London Metropolitan University, UK

***Corresponding Author:** Djennad Abdelmajid, Statistics, Operational Research and Mathematics (STORM) Research Centre, London Metropolitan University, E-mail: a-djennad@hotmail.co.uk

Citation: Djennad Abdelmajid (2023) Trends and Day-of-the-Week Effects Decomposition in the Mean and Dispersion in the Number of Subjects Testing Positive for COVID-19 in the UK in 2020. J Biostat Biometric App 8(1): 102

Received Date: November 01, 2023 **Accepted Date:** December 01, 2023 **Published Date:** December 05, 2023

Abstract

COVID-19 pandemic is a global threat, where the rate of infection with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) increases exponentially, and the spread of the virus from person to person is very fast. Fitting models for COVID-19 counts receive a great attention, and modelling the dispersion of COVID-19 counts helps measuring the spread of the disease in a population and evaluating the intervention. This paper examines the presence and persistence of day-of-the-week effects in both the mean and dispersion in the number of subjects testing positive for COVID-19 in the United Kingdom in 2020, and estimates the impact of first national lockdown on the spread of the disease in the population. The conditional mean and dispersion parameters of the probability distribution for the daily number of subjects testing positive for COVID-19 in the UK in 2020 are examined with the generalized structural time series (GEST) model.

Keywords: COVID-19, lockdown, day-of-the-week effect, negative binomial distribution.

Background

The SARS-CoV-2 spread globally in 2020 causing significant morbidity and mortality [1]. Between 12 and 15 million individuals are estimated to be infected with SARS-CoV-2 across all eleven European countries combined, up to 4 May 2020, representing between 3.2% and 4.0% of the population [2].

Through the examination of COVID-19 surveillance data in Middlesex County, Massachusetts, from 2 March 2020 through 7 November 2021, a new research [3] finds a strong evidence of day- of-the-week (DOW) effect in COVID-19 positive cases, where cases are twice as likely to be reported on Tuesdays-Fridays in the most stringent phases of non-pharmaceutical intervention, and half as likely to be reported on Mondays and Tuesdays in less stringent non-pharmaceutical intervention phases compared to Sundays.

Preliminary analysis of COVID-19 surveillance data for hospital admissions and deaths in Italy indicates a day-of-the-week pattern, and this pattern is confirmed for COVID-19 surveillance data in the United Kingdom [4]. For COVID-19 cases in the United Kingdom, the DOW indicates a daily pattern in cases as laboratory confirmation for the virus tends to slow down during weekends [4].

The United Kingdom imposes a national lockdown on 23 March 2020 by implementing non- pharmaceutical interventions to bring down the rate of infection. The daily number of subjects testing positive for COVID-19 in the UK from day 28 February 2020 to 06 January 2021 are plotted in Figure 1. In June, July and August 2020, the number of cases drop significantly but since mid September to 6 January 2021, the number of cases increase exponentially. The average value of the counts is 9,135 cases and the variance is 154,711,307 cases, suggesting overdispersion in the data. The data is available at https://coronavirus.data.gov.uk/details/cases

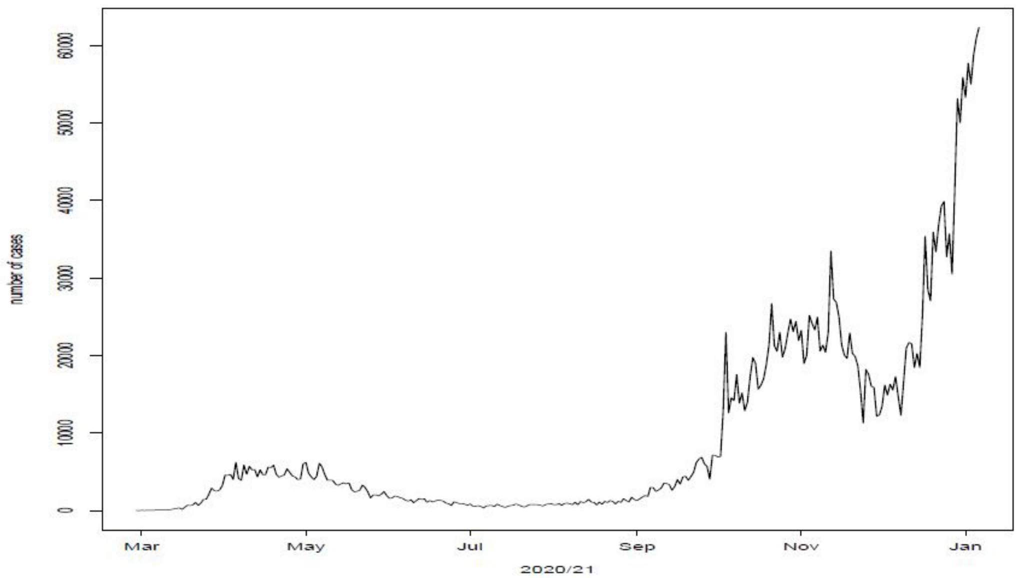

Figure 1: Daily counts of subjects testing positive for COVID-19 in the UK from 28 Feb 2020 to 6 Jan 2021.

The decomposition of time series into trend and seasonality has proven to be extremely useful in the analysis of infectious diseases. Seasonal-trend decomposition procedure [5] can be thought of one-dimensional decomposition of the mean parameter of the assumed probability distribution function of the outcome. A good example of seasonal-trend decomposition of Campylobacter cases in England and Wales and the impact of weather parameters on Campylobacter is illustrated in [6].

However, some data are better fitted with probability distributions with more than one random parameter, see for example generalized additive models for location, scale and shape (GAMLSS) [7]. The generalized structural time series (GEST) model [8] extends the GAMLSS model to serially dependent data.

The advantage of high-dimensional decomposition of the parameters of the probability distribution function of time series is the extraction of trend and seasonality in location, scale, skewness, and kurtosis, when estimating the impact of intervention or the effects of explanatory variables. This new methodology becomes more complex when the data is non-Gaussian [8].

The rationale for fitting a smooth trend with random walk of order two (RW2) originates in the analysis of mortality table in actuaries [9,10]. Prior to Whittaker-Henderson method, actuaries use moving weighted average (MWA) technique for dealing with data near the extremities [11]. The Whittaker-Henderson graduation (smoothing) technique becomes a better method used by actuaries to smooth the data near extremities, and it is also used to model the business cycle and growth in seasonal-adjusted macroeconomic data [12]. Flexible smoothing with B-splines is RW2 Whittaker-Henderson graduation (smoothing) technique [13]. The second-order random walk (RW2) model is commonly used for smoothing data and modelling response functions as it is computationally efficient due to the Markov properties of the joint (intrinsic) Gaussian density [14].

In addition, if seasonal or day-of-the-week effects are present in the data but ignored by model-building process, the result is likely to be a misspecified model that leads to residual autocorrelation of the order of the seasonality or DOW effects. Stochastic seasonality can be fitted using state space methods [8,15,16].

Statistical Modeling

Time series of number of subjects testing positive for COVID-19 in the United Kingdom from day 28 February to 6 January 2021 are examined using the generalized structural time series model with the Poisson and the negative binomial distributions. The natural logarithm of the expected number of daily counts and the natural logarithm of the dispersion are jointly and simultaneously decomposed into baseline, trend and DOW effects.

Generalized Structural Time Series Model

Let Y_t be the daily number of subjects testing positive for COVID-19 in the UK from 28 February 2020 to 06 January 2021, and $D = NBI(\mu_t, \sigma_t)$ where NBI represents the negative binomial type I distribution of the response variable, then the GEST model of two parameters (μ_t, σ_t), is defined as:

$$
\begin{aligned}
Y_t|\mu_t, \sigma_t &\sim \mathcal{NBI}(\mu_t, \sigma_t) \\
\log(\mu_t) &= \log(n) + \beta_1 + \gamma_{1,t} + s_{1,t} \\
\gamma_{1,t} &= 2\gamma_{1,t-1} - \gamma_{1,t-2} + b_{1,t} \\
s_{1,t} &= -\sum_{m=1}^{M-1} s_{1,t-m} + w_{1,t} \\
\log(\sigma_t) &= \beta_2 + \gamma_{2,t} + s_{2,t} \\
\gamma_{2,t} &= 2\gamma_{2,t-1} - \gamma_{2,t-2} + b_{2,t} \\
s_{2,t} &= -\sum_{m=1}^{M-1} s_{2,t-m} + w_{2,t}
\end{aligned}
\qquad (1)
$$

where n is the size of United Kingdom population in 2020 estimated at 67,886,011 people at mid year according to UN data [https://www.worldometers.info/world-population/uk-population/] and the log(n) is an offset variable, β_k is a constant vector in the mean and dispersion parameters, the $\gamma_{k,t}$ represent RW2 trends in the mean and dispersion, $s_{k,t}$ represent random day-of-the-week effects in the mean and dispersion, $b_{k,t}$ and $w_{k,t}$ are independently distributed disturbances with zero mean and variances σ^2_{bk}, σ^2_{wk} where $b_k \sim N_{T-Jk}(0, \sigma^2_{bk} I_{T-Jk})$ and $\omega_k \sim N_{T-J_k}(0, \sigma^2_{wk} I_{T-M+1})$.

The generalized structural time series (GEST) model [8] assumes that, conditional on the past, the response variable Y_t comes from a parametric distribution with probability (density) function $fY_t(y_t|\theta_t)$, where θ_t is a vector of unknown distribution parameters. Here θ_t is restricted to two parameters: $\theta_t = (\mu_t, \sigma_t)$, where μ_t is in general a location parameter, σ_t a scale parameter. Each parameter (μ_t, σ_t) is modelled by a structural time series model and/or linear, non-linear or smooth non-parametric models to account for explanatory variables. Each structural model is a random walk or autoregressive model, and a random seasonality.

Model Estimation

The GEST model defined by equation (1) has distinct sets of parameters: β, γ, s, σ^2_e, σ^2_b, σ^2_w where σ^2_b *and* σ^2_w are referred to as hyperparameters and represent the variances of the normal disturbance vectors $b_{k,t}$ and $\omega_{k,t}$ for $k = 1, 2$ where $b \sim N_{T-J}(0, \sigma^2_b I_{T-J})$, $\omega \sim N_{T-M+1}(0, \sigma^2_w I_{T-M+1})$.

Maximum likelihood estimation of β, σ^2_b and σ^2_w is defined by:

$$
\begin{aligned}
L(\boldsymbol{\beta}, \sigma_b^2, \sigma_w^2) &= \int\int f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{s}|\sigma_b^2, \sigma_w^2) d\boldsymbol{\gamma} d\mathrm{s} \\
&= \int\int f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{s}) f(\boldsymbol{\gamma}, \mathrm{s}|\sigma_b^2, \sigma_w^2) d\boldsymbol{\gamma} d\mathrm{s} \qquad (2)
\end{aligned}
$$

where

$$
f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{s}) = \prod_{t=1}^{T} f(y_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{s}) \qquad (3)
$$

denotes the conditional density function of the response vector y_t given β, γ and s, and

$$
l = \log f\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathrm{s}\right) = \sum_{t=1}^{T} \log f(y_t|\mu_t, \sigma_t) \qquad (4)
$$

denotes the log-likelihood function. The *extended* log-likelihood function in [17] p227-279, is defined as:

$$
l_e = l + \log f\left(\boldsymbol{\gamma}, \mathrm{s}|\sigma_b^2, \sigma_w^2\right) \qquad (5)
$$

where $f(\gamma, s|\sigma^2_b, \sigma^2_w)$ is the joint density function of γ_t and s_t given σ^2_b and σ^2_w. The likelihood $f(y|\beta, \gamma, s)f(\gamma, s|\sigma^2_b, \sigma^2_w)$ is knows as the joint or extended likelihood in hierarchical generalized linear model [18]. However, the integration of (2) is intractable for a non-Gaussian response variable and becomes more difficult when there is more than one random effect component, here there are four random effect components γ and s. This integral can be approximated using Laplace approximation which gives the following approximative marginal log likelihood:

$$
l(\boldsymbol{\beta}, \sigma_b^2, \sigma_w^2) = \log f(\mathbf{y}|\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}, \hat{\mathrm{s}}) + \log(\hat{\boldsymbol{\gamma}}, \hat{\mathrm{s}}|\sigma_b^2, \sigma_w^2) - \frac{1}{2}\log\left|\frac{\hat{\mathbf{D}}_{\boldsymbol{\gamma},\mathrm{s}}}{2\pi}\right| \qquad (6)
$$

where $\hat{\gamma}$ and $\hat{s}$ are the fitted value of γ and s estimated by maximising the extended likelihood over (γ, s) for given $(\beta, \sigma^2_b, \sigma^2_w)$, and $\hat{D}_{\gamma,s}$ is the second derivative of the extended likelihood with respect to (γ, s) evaluated at $\gamma = \hat{\gamma}$ and $s = \hat{s}$

Following Appendix B2 and C in [7], the maximization of the extended likelihood in (5), can be achieved by using the GEST algorithm described below, which provides posterior mode estimates of the sets of parameters of β, γ, s and by maximizing the extended log likelihood for hyperparameters $\sigma^2_{bk}, \sigma^2_{wk}$ for $k = 1, 2$.

GEST Algorithm for Estimating β, γ, s Given Fitted σ^2_b, σ^2_w

(A) initialise $(\theta_1, \theta_2) = (\mu_t, \sigma_t)$, and set initial $\gamma_k = 0$ and $s_k = 0$ for k =1,2.

(B) start the *outer cycle* to fit each of the distribution parameter vectors θ_k sequentially until convergence where, $\theta_1 = \mu_t = (\mu_1, \mu_2, \ldots, \mu_T)^T$, and $\theta_2 = \sigma_t = (\sigma_1, \sigma_2, \ldots, \sigma_T)^T$,

(a) start the *inner cycle (*or local scoring) for each iteration of the outer cycle to fit each of the distribution parameter vectors, $\theta_k = (\mu_t, \sigma_t)$

(i) evaluate the current *iterative response variable* z_k and current *iterative weights*

$$\mathbf{W}_k, \text{ where } \mathbf{z}_k = \boldsymbol{\eta}_k + \mathbf{W}_k^{-1}\mathbf{u}_k,\ \mathbf{W}_k = -\frac{\partial^2 \ell}{\partial\eta\partial\eta^\top}, \text{ or } -E\left[\frac{\partial^2 \ell}{\partial\eta\partial\eta^\top}\right] \text{ or } \left(\frac{\partial \ell}{\partial \eta}\right)^2, \text{ and } \mathbf{u}^{(r)} = \frac{\partial \ell}{\partial \eta}$$

(ii) Start the Gauss-Seidel (or backfitting) algorithm

(I) estimate β_k by regressing the current partial residuals $\epsilon_k = z_k - \gamma_k - s_k$ against design matrix X_k using current iterative weights W_k.
(II) estimate the hyperparameters σ^2_{bk} and σ^2_{wk} by maximising their likelihood function Q, and then estimate γ_k and s_k using the equation

$(\gamma_k, s_k)^T = [A_k + D^T{}_K M_k^{-1} D_k]^{-1} A_k(\epsilon_k, 0_k)^T$, where 0 is a vector of zeros of length T,

(iii) end the Gauss-Seidel algorithm on convergence of β_k, γ_k and s_k

(iv) update θ_k and $\eta = g(\theta_k)$.

(b)end the *inner cycle* on convergence of θ_k.

(c) end the *outer cycle* when the global deviance ($= -2 \times l$) of the estimated model converges.

GEST Algorithm for Estimating the Hyperparameters $\alpha = (\sigma^2_b, \sigma^2_w)$

1. Select starting values for $\alpha = (\sigma^2_b, \sigma^2_w)$.

2. Maximize Q over α using a numerical algorithm, where γ, s given α are obtained before calculating Q in the function evaluating Q.

3. Use the maximizing values for α to calculate the maximizing values for γ.

In step [(B).(a).(ii).(II)], the Q function, for a random walk trend and random seasonality model, is given by:

$$\begin{aligned} Q &= \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}, \mathrm{s}) + \log f(\boldsymbol{\gamma}, \mathrm{s}) - \frac{1}{2}\log\left|\mathbf{A} + \mathbf{D}^\top\mathbf{M}^{-1}\mathbf{D}\right| + T\log 2\pi \\ \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}, \mathrm{s}) &= -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\epsilon} - \boldsymbol{\gamma} - \mathrm{s})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\epsilon} - \boldsymbol{\gamma} - \mathrm{s}) \\ \log f(\boldsymbol{\gamma}, \mathrm{s}) &= -\frac{1}{2}\log|2\pi\mathbf{M}| - \frac{1}{2}(\boldsymbol{\gamma}^\top\mathrm{s}^\top)\mathbf{D}^\top\mathbf{M}^{-1}\mathbf{D}(\boldsymbol{\gamma}^\top\mathrm{s}^\top)^\top \end{aligned}$$

where $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)^T$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_T)^T$, $s = (s_1, s_2, \ldots, s_T)^T$, $\Sigma^{-1} = \sigma_e^{-2}W$, Σ is a [T x T] matrix, A is a [2T × 2T] matrix of Σ^{-1}, M is a matrix diagonal $[\sigma^2_b I_{T-J}, \sigma^2_w I_{T-M+1}]$ and D is a matrix diagonal $[D_\gamma, D_s]$, where D_γ is a second order differencing matrix and D_s is a unit matrix of M levels and size $[(T - M + 1) \times T]$. The Q function for a random walk trend model, is given by:

$$\begin{aligned} Q &= \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}) + \log f(\boldsymbol{\gamma}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right| + \frac{T}{2}\log 2\pi \\ \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}) &= -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\epsilon}-\boldsymbol{\gamma})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\epsilon}-\boldsymbol{\gamma}) \\ \log f(\boldsymbol{\gamma}) &= -\frac{1}{2}\log\left|2\pi\sigma_b^2\right| - \frac{1}{2}\sigma_b^{-2}\boldsymbol{\gamma}^\top\mathbf{D}^\top\mathbf{D}\boldsymbol{\gamma}^\top \end{aligned}$$

The maximization of Q over γ given α, where α is estimated by the GEST fitting algorithm, is given by solving the following equation:

$$\boldsymbol{\gamma} = \left[\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right]^{-1}\boldsymbol{\Sigma}^{-1}\ \boldsymbol{\epsilon}. \tag{7}$$

Effective Degrees of Freedom

The total effective degrees of freedom of the fitted GEST model, df, combines those of the model for μ_t and σ_t, given by $df_{\mu t}$ and $df_{\sigma t}$ respectively. Hence,

$$df = df_{\mu_t} + df_{\sigma_t} = p_k + d_k \tag{8}$$

for k = 1, 2, where p_k is the length of β_k and the d_k is the effective degrees of freedom for the fitted random walk trend and random seasonality. Let $\hat{B}, \hat{A}, \hat{M}, \hat{\gamma}, \hat{s}$ be the values of B, A, M, γ , s on convergence of the GEST fitting algorithm. On convergence, $(\hat{\gamma}, \hat{s}) = \hat{B}\epsilon$. Hence d_k, the effective degrees of freedom used for random walk trend and random seasonality model in μ_t and σ_t, is given by:

$$d_k = trace\left[\hat{\mathbf{B}}_k\right] = trace\left\{\left[\hat{\mathbf{A}}_k + \hat{\mathbf{D}}_k^\top\hat{\mathbf{M}}_k^{-1}\hat{\mathbf{D}}_k\right]^{-1}\hat{\mathbf{A}}_k\right\}. \tag{9}$$

In addition, let $B = [\Sigma^{-1} + \sigma^{-2}_b D^T D]^{-1}\Sigma^{-1}$ and let $\hat{B}, \hat{\Sigma}, \hat{D}, \hat{\gamma}, \hat{\sigma}_b^{-2}$ be the values of **B, Σ, D, γ** and σ_b^{-2} on convergence of the GEST fitting algorithm. On convergence, $\hat{\gamma} = \hat{B}\epsilon$. Hence d_k, the effective degrees of freedom used for random walk trend model in μ_t and σ_t, is

$$d_k = trace\left[\hat{\mathbf{B}}_k\right] = trace\left\{\left[\hat{\boldsymbol{\Sigma}}_k^{-1} + \hat{\sigma}_{b_k}^{-2}\hat{\mathbf{D}}_{\mathbf{k}}^\top\hat{\mathbf{D}}_k\right]^{-1}\hat{\boldsymbol{\Sigma}}_k^{-1}\right\}. \tag{10}$$

As d_k is difficult to calculate directly for large T, it can be calculated by setting $\partial Q/\partial\sigma^2_b = 0$ giving on convergence $d = J + \hat{\sigma}^{-2}_b\hat{\gamma}^T D^T D\hat{\gamma}$, for the random walk trend model, and by setting $\partial Q/\partial\sigma^2_b = 0$ and $\partial Q/\partial\sigma^2_w = 0$ giving on convergence $d = J + M - 1 + \hat{\sigma}_b^{-2}\hat{\gamma}^T D^T D_\gamma\hat{\gamma} + \hat{\sigma}_w^{-2}\hat{s}^T D_s^T D_s\hat{s}$, for the random walk trend and random seasonality model, using the result $\partial/\partial x \log|xC + F| = tr[(xC + F)^{-1}C]$, where x is a scalar and C and F are r x r matrices (provided $|xC + F| \neq 0$), Hence, for each distribution parameter, d_k is calculated using the values of $\hat{\gamma}_K$, $\hat{s}_k$, $\hat{\sigma}^2_{bk}$, $\hat{\sigma}^2_{wk}$ on convergence of the GEST fitting algorithm. Note that the formula for the effective degrees of freedom has reciprocals $\hat{\sigma}^{-2}_b$ and $\hat{\sigma}^{-2}_w$. The estimates of these variances are often very small ,so the inverse of these variances are very large with very large effective degrees of freedom. In the random walk of order 2, the fitted trend is a smooth curve with no much variability the variance is very small and inverse is very large, but because the smoothing matrix D_γ is a second order differencing matrix, the $\hat{\gamma}^T D^T D_\gamma\hat{\gamma}$ become smaller and it compensates the very large value of the inverse of the variance.

Model Selection Procedure

GEST Poisson and negative binomial models are fitted (see Table 1) and compared using the Akaike information criterion (AIC) with the total effective degrees of freedom given by equation 9. The AIC of parsimonious GEST models (m1 and m2) are compared with model m3, where m1 is fitted with a RW2 trend and random DOW for μ_t, model m2 is fitted with a RW2 trend and random DOW for μ_t and a constant dispersion σ_t, and m3 is fitted with RW2 trend and random DOW for both μ_t and σ_t. A further comparison is performed using the centile estimates of the fitted models [19], where the 100p centile of a random variable Y is the value y_p such that $p(Y \le y_p) = p$, i.e. $y_p = F^{-1}(p)$, so y_p is the inverse cumulative distribution function of Y applied to p, see [19] for more details. The z-scores are used as residuals in fitted GEST models.

Table 1: The fitted GEST models with Poisson and NBI, random walk of order 2 (RW2) trend over time and random day-of-the-week (RDOW) effects.

Models	Probability distribution	$\mu t\ (\gamma_{1,t}, s_{1,t}) =$	$\sigma t\ (\gamma_{2,t}, s_{2,t}) =$
m1	Poisson	RW2 + RDOW	–
m2	Negative binomial type I	RW2 + RDOW	constant
m3	Negative binomial type I	RW2 + RDOW	RW2 + RDOW

Testing Normality of the Residuals

Normalized (randomized) quantile residuals of the fitted GEST models are examined using the detrended transformed Owen's plot (DTOP) [20]. If the fitted Poisson distribution is shown to be adequate, then the model for dispersion is not needed and $\sigma_t = 0$. The true normalized (randomized) quantile residuals always have a standard normal distribution for any regression type model whatever the original model distribution. Hence the fitted normalized (randomized) quantile residuals are always compared to a standard normal distribution making interpretation and comparison of the resulting plots (for different original model distributions) easier. Significant departures from the model (resulting in significant differences between the residuals and their standard normal distributions, if the model is correct) are indicated by the confidence bands not including the horizontal zero line for some value(s) in the DTOP. Similarly the DTOP of the fitted residuals provides a guide to model adequacy, if the fitted distribution of the model is adequate, then the normalized quantile residuals has exactly a standard normal distribution and they should not cross the horizontal line, see Table 1 in [20].

Testing Overdispersion

In GEST-NBI, the dispersion parameter σ_t is fitted with RW2 trend and random DOW effects, testing for overdispersion can be achieved by testing the null hypothesis $H_0 : \sigma_t = 0$ against the alternative $H_1 : \sigma_t > 0$, at 1% significance level. A sound practice to test overdispersion, is to estimate both Poisson and negative binomial models and test the null hypothesis $H_0 : \sigma_t = 0$ against the alternative $\sigma_t > 0$ [21]. The likelihood ratio test for overdispersion is – 2 time the difference in the fitted log-likelihood of the two models and the asymptotic distribution of the LR test statistic has probability mass of 0.5 at zero and a half-$\chi^2(1)$ distribution above zero [21] p89-91.

Results

The estimated parameters of GEST-Poisson and GEST-NBI are reported in Table 2. With regard to fitted distribution, the negative binomial distribution fits better the Poisson, since m2 has smaller AIC than model m1, and the detrended transformed Owen's plot of GEST-Poisson model has a negative slope which indicates that the variance in the residuals is too large (Table 3), the variance in the model response variable is too small, and the confidence bands cross the zero horizontal line. Therefore, the Poisson distribution is inadequate to fit the count of COVID-19 cases. For negative binomial distribution, the confidence bands in models m2 and m3 do not cross the zero horizontal line, and the variance in the residuals are too small (see Table 3), indicating that the negative binomial distribution is more adequate to fit the counts of COVID- 19 cases. In the matter of overdispersion, model m2 fits better than m1, since the AIC of m2 is smaller than the AIC of m1 (see Table 2), and the likelihood ratio test statistic is 1019.36 [= -2 × (2366.28 - 2875.96)] which exceeds the 1% critical value of $\chi^2_{.98}$= 5.41, therefore, the test results strongly reject the null hypothesis of a Poisson model and indicates the existence of overdispersion in the data. In terms of random overdispersion, the model with a random dispersion fits better than the model with a constant dispersion, since the AIC of m3 is smaller than the AIC of m2 (see Table 2), and the likelihood ratio test statistic is 130.08 [= -2 × (2301.24 -2366.28)] also exceeds the 1% critical value of $\chi^2_{.98}$= 5.41.

Table 2: Fitted parameters and AIC for models m1, m2 and m3.

Mod.	$\hat{\sigma}^2_{b1,t}$	$\hat{\sigma}^2_{w1,t}$	$\hat{\sigma}^2_{b2,t}$	$\hat{\sigma}^2_{w2,t}$	$ne^{(\hat{\beta}1)}$	$e^{(\hat{\beta}2)}$	-logLik	df	AIC
m1	0.001102	0.007344	-	-	9323.94	-	2875.96	208.34	6168.6
m2	0.000186	6.0065e-05	-	-	3211.01	0.01824	2366.28	49.22	4830.9
m3	0.000173	4.4723e-05	9.0509e-05	0.05321	3223.88	0.01676	2301.24	88.14	4778.8

Table 3: Randomised quantile residuals of models m1, m2 and m3.

Models	mean	variance	skewness	kurtosis
m1	-0.1324988	8.088289	-0.0382294	3.310703
m2	-0.0040586	1.024121	0.0205053	5.083053
m3	-0.0186246	0.996569	0.0063311	2.528352

Table 4: Centiles estimates of models m1, m2 and m3.

Models	0.5%	2.5%	25%	50%	75%	97.5%	99.5%
m1	19.43	23.89	41.72	50.00	61.46	80.25	84.39
m2	1.59	3.82	23.25	46.50	77.39	97.77	99.68
m3	0.00	2.23	28.03	48.09	73.89	97.77	99.68

Table 4 reports the centiles 100p = (0.5, 2.5, 25, 50, 75, 97.5, 99.5) of models m1, m2 and m3. Below 2.5%, the centile estimate of model m1 is far too high at 23.89%, m2 fits a bit higher at 3.82%, but m3 fits better at 2.23%. Below 50%, the centile estimate of m1 is identical 50.00%, m2 is quite low at 46.50%, but m3 fits better at 48.09%, Below 97.5%, the centile of m1 is low at 80.25%, m2 and m3 are much better at 97.77%. Figure 2 shows the number of subjects testing positive and centile estimates of model m1, m2, and m3 for 100p = (2.5, 50, 97.5). The Poisson model appears to overestimate all centiles below 50%, for example for 100p = (0.5, 2.5, 25), the Poisson estimates are $100\hat{p}$ = (19.43, 23.89, 41.72) respectively. On the other hand, centiles above 50%, 100p = (75, 97.5, 99.5), Poisson model underestimates the centiles $100\hat{p}$ = (61.46, 80.25, 84.39) respectively. This is due to equidispersion feature of the Poisson model; i.e., the mean equals the variance. In negative binomial, the centile curves of models m2 and m3 are wider and dispersed as GEST-NBI model allows for extra variation beyond that expected in a Poisson model. Clearly, model m3 provides better centiles estimates. Hence, m3 is considered to be better than models m1 and m2.

The Chosen Model

The maximum likelihood estimates of the mean and dispersion in model m3, denoted by $\hat{\mu}_t$ and $\hat{\sigma}_t$ respectively, are given by:

$$\begin{aligned} Y_t|\mu_t,\sigma_t &\sim \mathcal{NBI}(\hat{\mu}_t,\hat{\sigma}_t) \\ \hat{\mu}_t &= 67,886,011 \times \exp\left[-9.955+\hat{\gamma}_{1,t}+\hat{s}_{1,t}\right] \\ \hat{\sigma}_t &= \exp\left[-4.089+\hat{\gamma}_{2,t}+\hat{s}_{2,t}\right] \end{aligned} \tag{11}$$

where

$$\begin{aligned} \hat{\gamma}_{1,t} &= 2\hat{\gamma}_{1,t-1}-\hat{\gamma}_{1,t-2}+\hat{b}_{1,t} \\ \hat{s}_{1,t} &= -\sum_{m=1}^{M-1}\hat{s}_{1,t-m}+\hat{w}_{1,t} \\ \hat{\gamma}_{2,t} &= 2\hat{\gamma}_{2,t-1}-\hat{\gamma}_{2,t-2}+\hat{b}_{2,t} \\ \hat{s}_{2,t} &= -\sum_{m=1}^{M-1}\hat{s}_{2,t-m}+\hat{w}_{2,t} \end{aligned} \tag{12}$$

$\hat{\gamma}_{1,t}$ is the fitted RW2 trend and $\hat{s}_{1,t}$ is the fitted random DOW effects in the mean parameter $\hat{\mu}_t$ of NBI; $\hat{b}_{1,t}$ and $\hat{w}_{1,t}$ represent their white noise disturbances. The estimates of the disturbances variances are $\hat{\sigma}^2_{b1,t} = 1.73 \times 10^{-4}$ and $\hat{\sigma}^2_{w1,t} = 4.47 \times 10^{-5}$ respectively. $\hat{\gamma}_{2,t}$ is the fitted RW2 trend and $\hat{s}_{2,t}$ is the fitted random DOW effects in the dispersion parameter $\hat{\sigma}_t$ of NBI; $\hat{b}_{2,t}$ and $\hat{w}_{2,t}$ represent their white noise disturbances. The estimates ofthe disturbances variances are $\hat{\sigma}^2_{b2,t} = 9.05 \times 10^{-5}$ and $\hat{\sigma}^2_{w2,t} = 0.05321$ respectively. The estimate of the mean baseline is 3224. Hence, the baseline rate of infection over time is estimated at 3224 cases per day.The fitted conditional variance is given by $\hat{\mu}_t + \hat{\sigma}_t \hat{\mu}_t^2 = n \times \exp(-9.955 + \hat{\gamma}_{1,t} + \hat{s}_{1,t}) + \exp(-4.089 + \hat{\gamma}_{2,t} + \hat{s}_{2,t}) \times [n \times \exp(-9.955 + \hat{\gamma}_{1,t} + \hat{s}_{1,t})]^2$, where n=67,886,011. The fitted mean number of cases over time is plotted together with the observations in Figure 3 (panel a), and fitted dispersion over time is shown in Figure 4 (panel d).

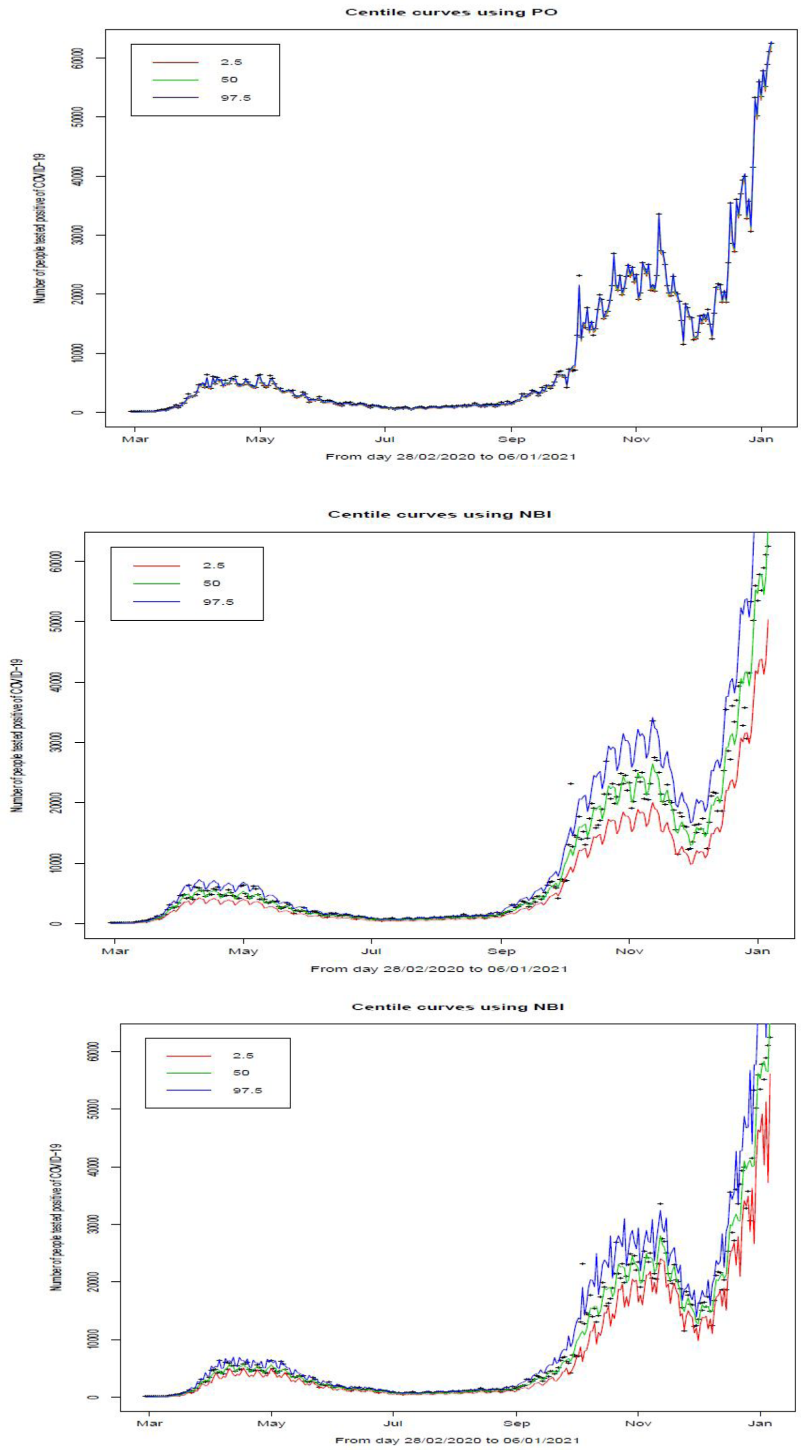

Figure 2: Daily counts of subjects testing positive of COVID-19 with three centile curves of models m1 (top), m2 (middle), m3 (bottom).

The Decomposition of the Fitted Mean and Dispersion

The decomposition of the fitted values of the mean ($\hat{\mu}_t$) and dispersion ($\hat{\sigma}_t$) into RW2 trend and DOW effects are shown in Figure 3 and Figure 4 respectively.

Panel (b) shows the fitted smooth trend over time in the mean number of cases without DOW effects. From 28 February 2020 to 01 April 2020 - the first 34 days - the trend of infection increases exponentially. The mean number of cases without DOW effects is estimated at 6 cases on 28 February 2020 and changes within 47 days to 5,161 cases on 14 April 2020. It slows down between 01 and 14 April 2020 when it reaches the peak of infection on 14 April 2020. The trend continues to drop significantly to 555 cases on 7 July 2020. Between 14 and 20 August 2020 the trend is stable. From 24 August to 12 November 2020, the trend accelerates very fast from 1,145 cases to 25,409 cases. From 1 December to 6 January 2021, the trend increases exponentially from 14,280 to 67,112.

The fast increase and the shift in the peak number of infection, from estimated 5,161 cases on average on 14 April 2020 to 25,409 cases on average on 12 November 2020, are due to an increase in testing volume and laboratory capacities. Therefore, the numbers of subjects testing positive for COVID-19 appear low in March and April 2020, because testing is happening in hospitals only and a large scale of population testing starts in May 2020.

In addition, the fast increase in COVID-19 cases is due a new variant of SARS-CoV-2 emerges in South East England which spreads rapidly in the population as pointed out by the European Centre for Disease Prevention and Control (ECDC) on 20 December 2020 report:

Over the last few weeks, the United Kingdom (UK) has faced a rapid increase in COVID-19 cases in South East England, leading to enhanced epidemiological and virological investigations. Analysis of viral genome sequence data identified a large proportion of cases belonged to a new single phylogenetic cluster [22].

Panel (c) shows the fitted DOW effects in the mean number of cases. The DOW reflects a daily testing cycle over time. This cycle is likely to be dominated by reporting imperfections. There is a small change in DOW effects in the mean number of cases. The fitted DOW effects represent the relative risk of the day-of-the-week on mean number of infection. The relative risk vary between 0.8651 to 1.1246.

Panel (e) shows the fitted smooth trend over time in dispersion of cases without DOW effects. The trend in the spread of COVID-19 declines sharply from estimated dispersion of 3.76 on 28 February 2020, before lockdown, to estimated dispersion of 1.18 on 23 March 2020, when the lockdow starts. It continues to decline to lower level of dispersion of 0.5, between 16 April and 27 May 2020. It starts to increase again but very slowly, from 0.52 on 1 June 2020 to 0.95 on 7 July 2020. On first of June 2020, schools are re-opened in England, and on twenty-third of June 2020, the first national lockdown ends.

Panel (f) shows the fitted DOW effects in dispersion of COVID-19 cases. The pattern exhibits significant changes over time in DOW spread (see more details in next Section).

Day-of-the-Week Effects

The estimates of DOW effects in the mean and dispersion are plotted in Figure 5 in blue and red respectively. DOW effects in dispersion represent the daily spikes in COVID-19 infection. These spikes in COVID-19 can be observed in the daily counts of COVID-19 cases in gray in Figure 3 Panel (a), above the fitted mean in blue. From 28 February 2020 to 21 March 2020, the time before lockdown, the spikes are occurring on Sundays, Mondays and Wednesdays with estimated effects of $\hat{s}_t$ = 1.90; 1.98; 3.23 respectively. Fridays, Saturdays, Tuesdays, and Thursdays do not have spikes, with $\hat{s}_t$ = 0.5. After lockdown the pattern of daily spikes starts to change. This can easily be seen in Figure 6.

Spikes of Sundays are significant in March-April, they decline in May-August, then they increase dramatically in September-December. Spikes of Wednesdays are the biggest in March-July but they drop later on. Spikes of Tuesdays increase rapidly between October-December. Table 5 and Table 6 report the average DOW effects in mean and dispersion per month.

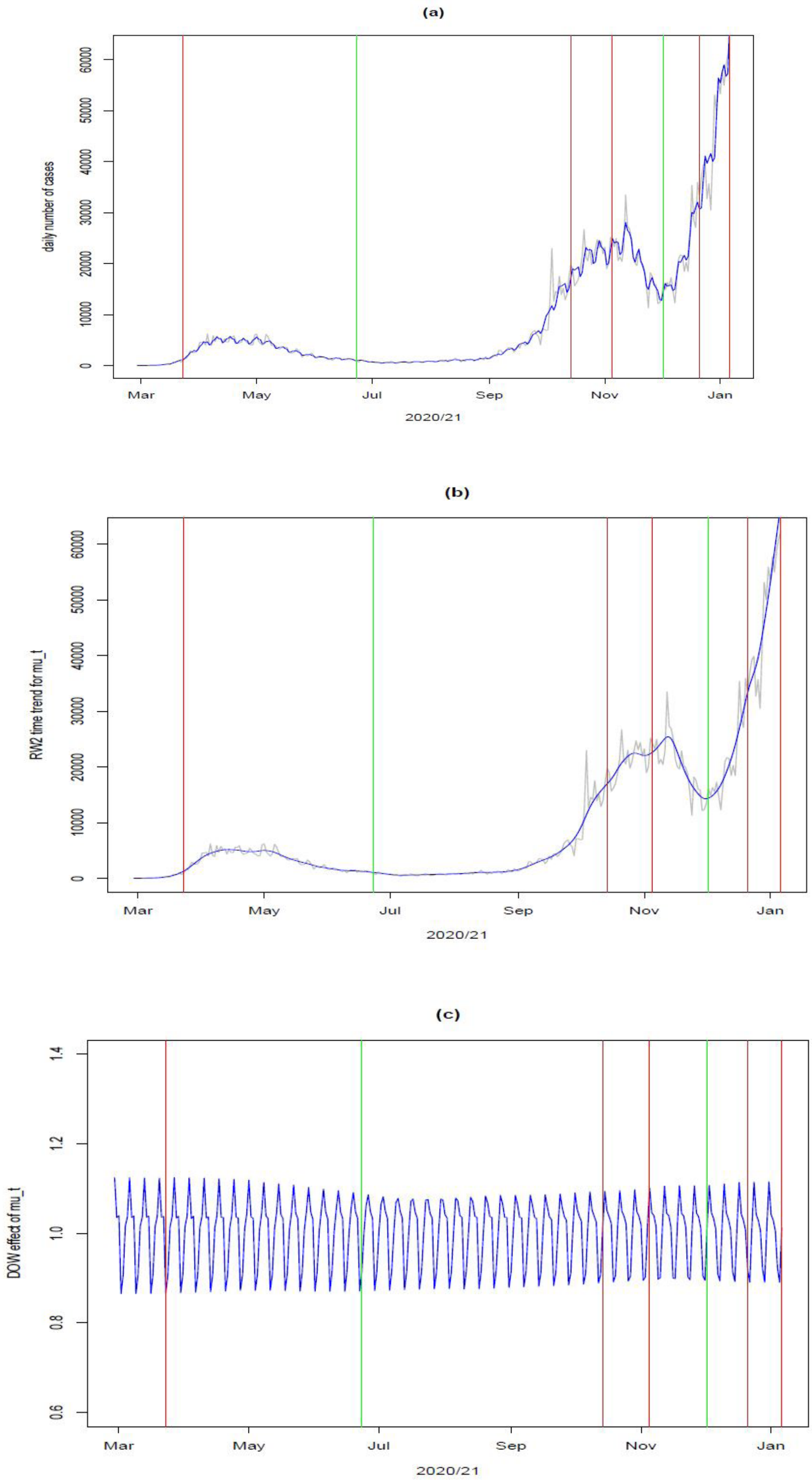

Figure 3: Panel (a): daily counts of subjects testing positive for COVID-19 in the UK from 28 Feb 2020 to 6 Jan 2021 in gray with fitted mean in blue. Panel (b): fitted RW2 trend. Panel (c): fitted random DOW effects. Panels (b) and (c) show Trend-DOW effects decomposition of the mean.

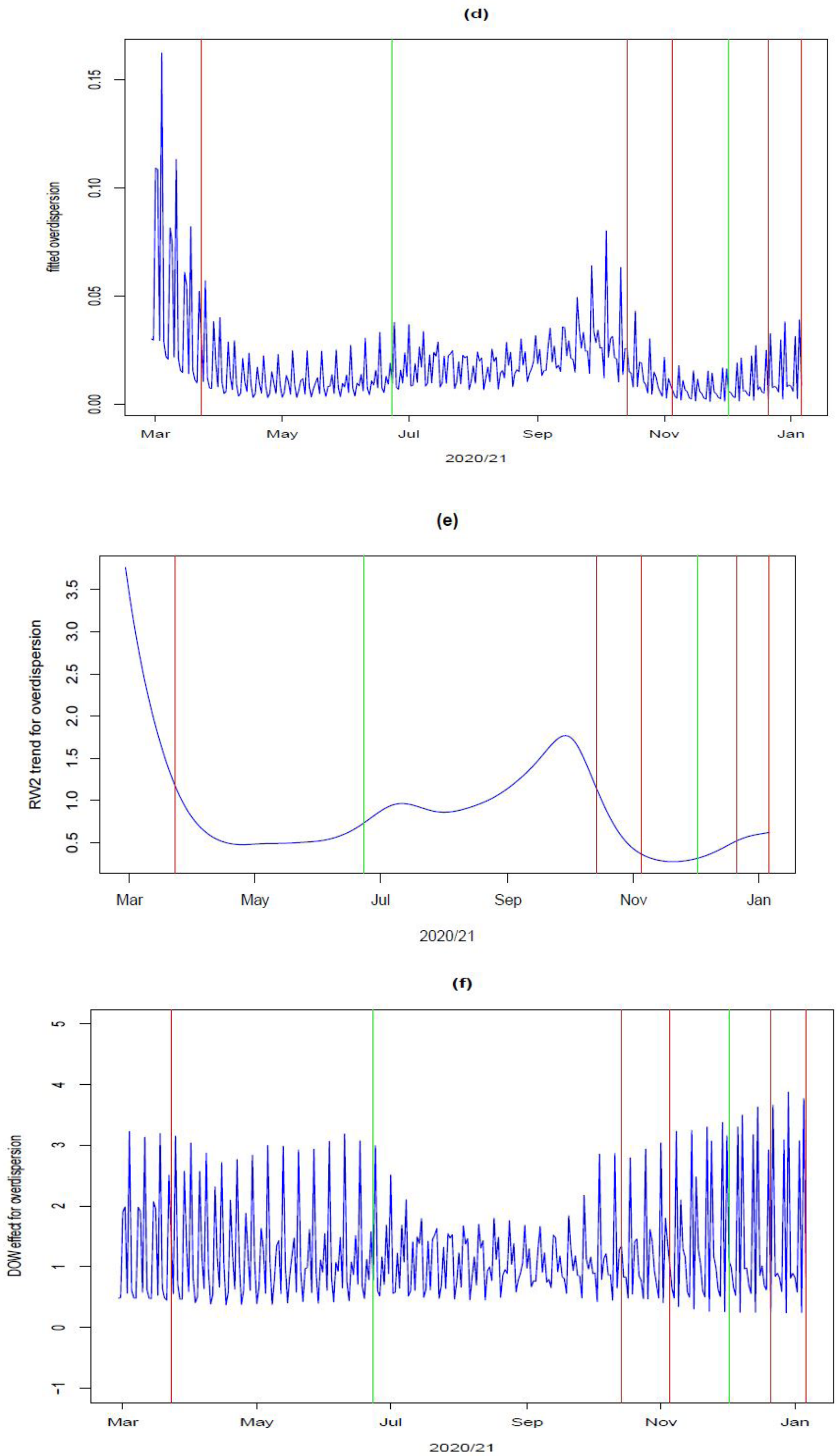

Figure 4: Panel (d): fitted overdispersion. Panel (e): fitted RW2 trend. Panel (f): fitted random DOW effects. Panels (e) and (f) show Trend-DOW effects decomposition of overdispersion.

Table 5: Estimates of average day-of-the-week effects per month on the fitted mean of GEST- NBI model m3, $\hat{s}_{\mu}t \mid day$.

$\hat{s}\mu t \mid day$	Mar	Apr	May	June	Jul	Aug	Sept	Oct	Nov	Dec	Jan
$\hat{s}\mu t \mid Fri$	1.124	1.123	1.112	1.093	1.077	1.070	1.062	1.051	1.045	1.044	1.042
$\hat{s}\mu t \mid Sat$	1.035	1.035	1.042	1.048	1.048	1.039	1.037	1.038	1.039	1.031	1.029
$\hat{s}\mu t \mid Sun$	1.039	1.036	1.033	1.033	1.034	1.035	1.032	1.024	1.015	1.008	1.005
$\hat{s}\mu t \mid Mon$	0.866	0.870	0.872	0.871	0.874	0.877	0.882	0.890	0.900	0.916	0.922
$\hat{s}\mu t \mid Tue$	0.908	0.908	0.909	0.915	0.913	0.916	0.913	0.907	0.899	0.893	0.891
$\hat{s}\mu t \mid Wed$	1.015	1.012	1.008	1.004	1.002	1.001	1.007	1.014	1.015	1.015	1.012
$\hat{s}\mu t \mid Thu$	1.037	1.041	1.047	1.056	1.071	1.082	1.086	1.094	1.105	1.112	

Table 6: Estimates of average day-of-the-week effects per month on the fitted dispersion of GEST-NBI model m3, $\hat{s}_{\sigma}t \mid day$.

$\hat{s}\sigma t \mid day$	Mar	Apr	May	June	Jul	Aug	Sept	Oct	Nov	Dec	Jan
$\hat{s}\sigma t \mid Fri$	0.479	0.385	0.398	0.487	0.648	0.832	0.823	0.817	0.625	0.736	0.820
$\hat{s}\sigma t \mid Sat$	0.472	0.569	0.877	1.133	1.323	1.047	0.577	0.462	0.504	0.574	0.583
$\hat{s}\sigma t \mid Sun$	2.209	2.218	1.210	0.786	0.636	0.818	1.936	2.901	3.240	3.115	3.075
$\hat{s}\sigma t \mid Mon$	1.746	1.171	1.471	1.562	1.569	1.724	1.268	0.573	0.315	0.248	0.244
$\hat{s}\sigma t \mid Tue$	0.564	0.636	0.577	0.725	1.366	1.163	0.939	1.448	2.524	3.567	3.771
$\hat{s}\sigma t \mid Wed$	3.179	2.799	2.983	2.968	1.915	1.413	1.210	1.365	1.248	0.914	0.825
$\hat{s}\sigma t \mid Thu$	0.648	1.038	0.945	0.685	0.504	0.531	0.799	0.872	1.043	0.953	

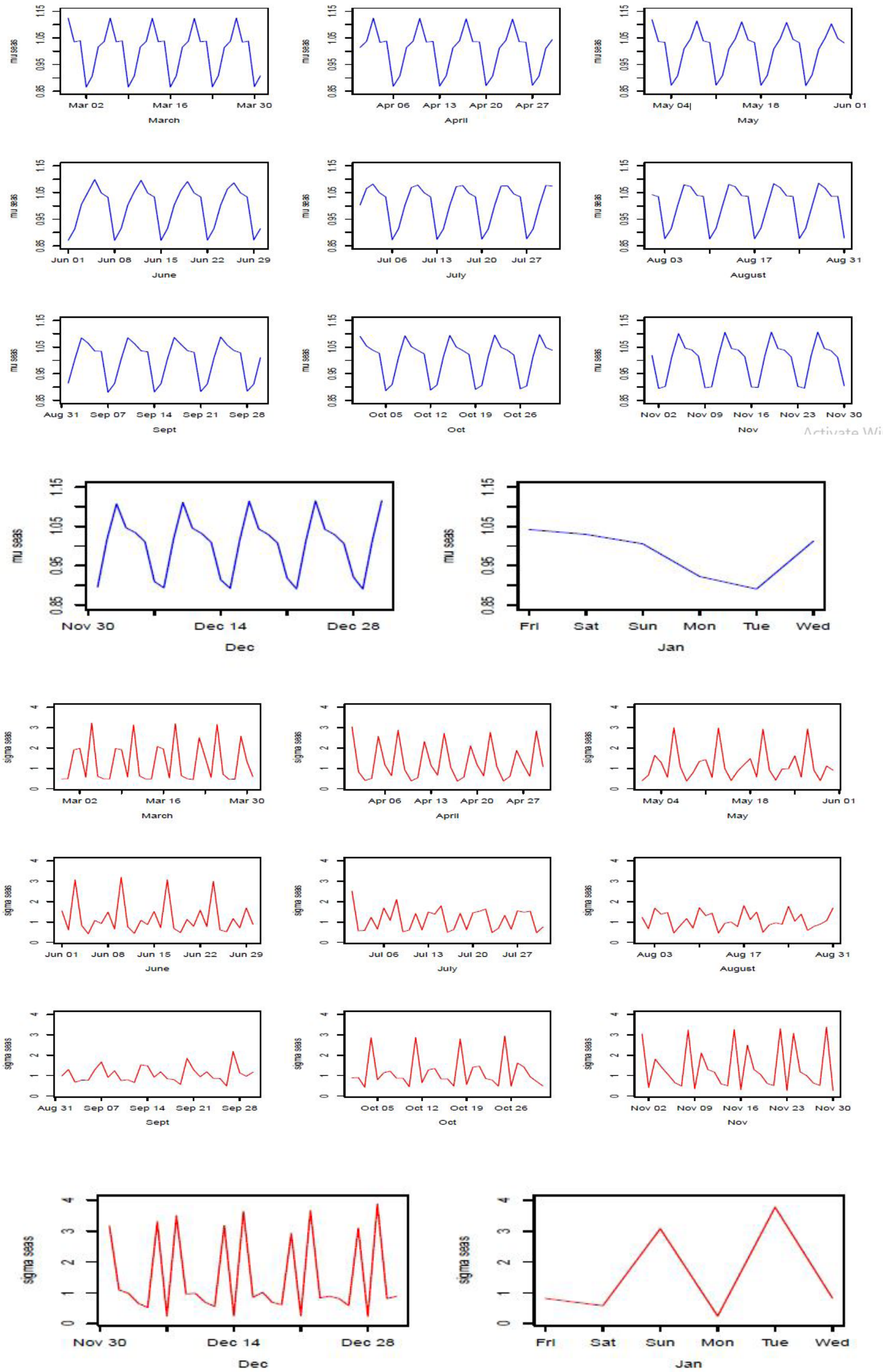

Figure 5: Fitted random DOW effects in the mean on separate months (in blue) and the fitted random DOW effects in the overdispersion on separate months (in red).

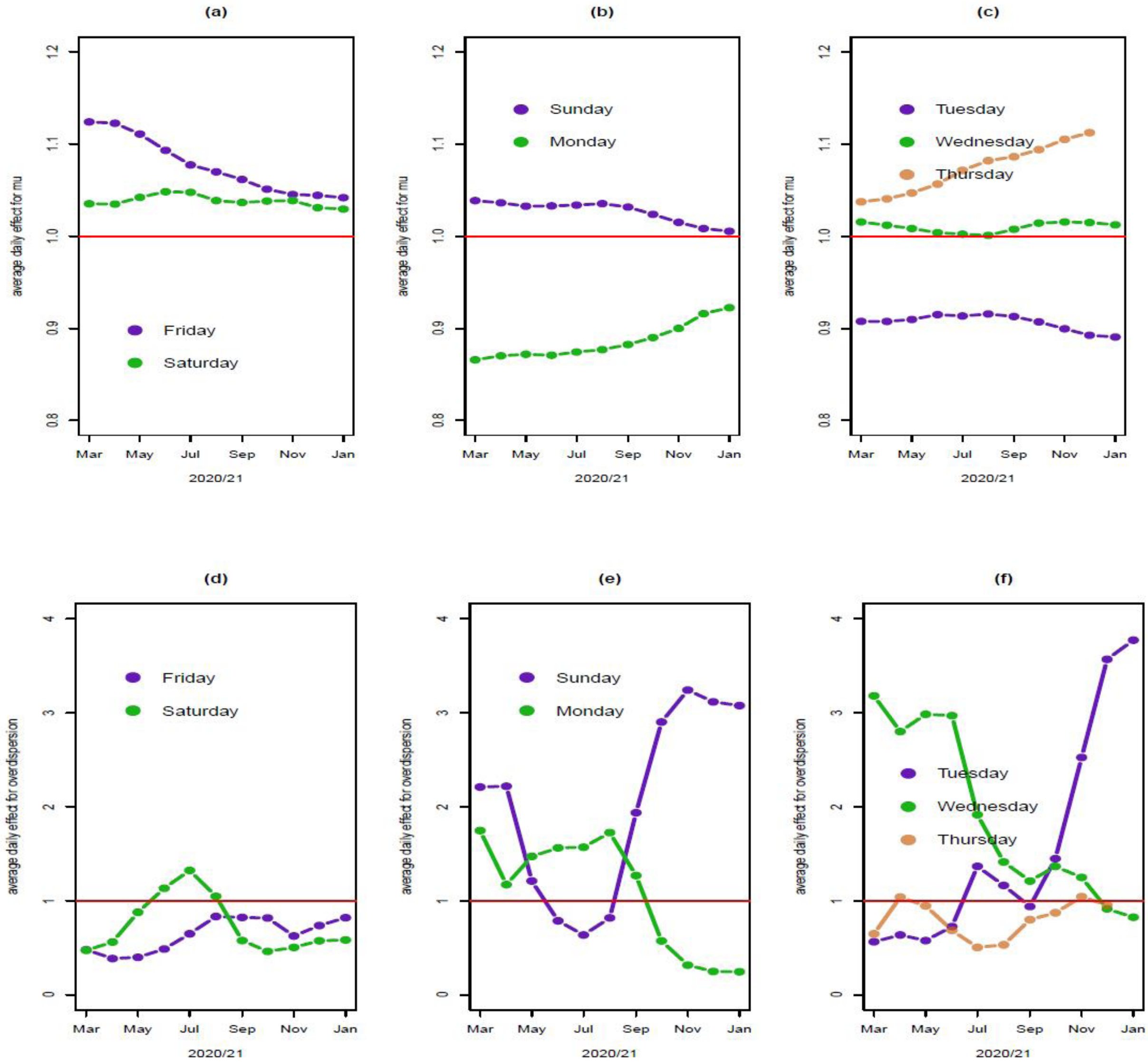

Figure 6: Average random DOW effects per month: panels (a), (b) and (c) show average random DOW effects in the mean per month, and panels (d), (e) and (f) show average random DOW effects in overdispersion per month.

Discussion

This article introduces a new methodology for time series analysis of infectious diseases.

The high-dimensional seasonal-trend decomposition performs better than one-dimensional decomposition. GEST models m1 and m2 are regarded as one-dimensional decomposition for the mean parameter, using Poisson and negative binomial distributions respectively. However, GEST model m3 is a novel methodology, where the dispersion parameter is decomposed into a random walk trend and random DOW effects jointly and simultaneously with the mean parameter.

One of the reasons for fitting the three models is to test the overdispersion, if model m3 does not improve the fitted values of the mean, then the simple model of one-dimensional decomposition of m2 is adequate, given that the NBI fits better than the Poisson, therefore there is no need to decompose the second parameter of the assumed probability distribution function.

The currents time series models assume that the overdispersion of the data is constant through time; this is not the case in COVID-19 data as the GEST model m3 provides strong evidence of presence and persistence of DOW effects and RW2 trend in the overdispersion of COVID-19 positive counts in the UK.

Day-of-the week effects in COVID-19 data in USA are estimated using segmented negative binomial regression models and Kolmogorov-Zurbenko adaptive filters [3]. The extended Kolmogorov- Zurbenko filter is developed to detect outbreaks/spikes in cases in the presence of noisy variance [23]. However, the GEST model detects outbreak signals in the mean and dispersion for non-Gaussian observations, therefore it is a more flexible parameter-driven model.

The small changes in DOW effects in the mean number of cases could be due to minimum access to health care services through non-pharmaceutical intervention phases. The estimated DOW effects in the mean is a useful tool for health professionals and health care services planning. The significant changes in DOW effects in dispersion reflects the temporal change in the spread of COVID-19 during lockdown and after. More examinations are needed to understand the driving force of the DOW effects in COVID-19 spread and the methods can be applied to other respiratory diseases.

Conclusion

This article provides strong evidence that the first national lockdown suppresses the pandemic of COVID-19 significantly. The non-pharmaceutical interventions during the first national lockdown decrease the trend of COVID-19 infection significantly, from the peak of infection on 14 April 2020 to a lower level of infection on 7 July 2020. Also, they decrease the trend of COVID-19 spread significantly. The estimated pattern of DOW effects in mean number of cases is stable through the year, the differences in DOW effects per day is likely to be dominated by reporting imperfection and the volume of testing per day. On the other hand, the estimated pattern of DOW effects in spread of cases is not stable and varies through the year. This stochastic pattern is possibly driven by social interactions, indoor activities, seasonal travels after lockdown and weather.

References

1. Khoury DS, Cromer D, Reynaldi A et al. (2021) Neutralizing antibody levels are highly predictive of immune protection from symptomatic SARS-CoV-2 infection. Nature Med., 27: 1205-11.

2. Flaxman S, Mishra S, Gandy A, Unwin HJT et al. (2020) Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature, 584: 257-61.

3. Simpson RB, Lauren BN, Schipper KH, McCann JC et al. (2022) Critical periods, critical time points and day-of-the-week effects in COVID-19 surveillance data: an example in Middlesex County, Massachusetts, USA. Int. J. Environ. Res. Public Health, 19: 1321.

4. Harvey A, Kattuman P (2020) Time series models based on growth curves with applications to forecasting coronavirus. Harvard Data Sci Rev, 4.

5. Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: A seasonal- trend decomposition procedure based on Loess (with discussion). J. Official Stats. 6: 3-33.

6. Djennad A, Lo Iacono G, Sarran C, Lane C, et al (2019) Seasonality and the effects of weather on Campylobacter infections. BMC Infectious Diseases. 19: 255.

7. Rigby RA and Stasinopoulos DM (2005). Generalized additive models for location, scale and shape (with discussion). J. R. Statist. Soc., C, 54: 507-54.

8. Djennad A, Rigby RA, Stasinopoulos D, Voudouris V, Eilers PHC (2015) Beyond location and dispersion models: the generalized structural time series model with applications. Munich Pers REPIC Repos. 1-33.

9. Whittaker ET (1923) On a new method of graduation. Proceedings of the Edinburgh Mathematical Society, 41: 63-75.

10. Whittaker ET and Robinson G (1924) The calculus of observations: a treatise on numerical mathematics, 3rd., Blackie and Son Limited, UK.

11. Greville TNE (1977) Moving-weighted-average smoothing extended to the extremities of the data. Technical Summary Report, University of Wisconsin-Madison, Mathematics Research Center.

12. Hodrick RJ and Prescott EC (1997) Postwar U.S. business cycles: an empirical investigation. J. of Money, Credit and Banking, 29: 1-16.

13. Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Science, 11: 89-121.

14. Lindgren F and Rue H (2008) On the second-order random walk model for irregular locations. Scand. Jour. of Statist, 35: 691-700.

15. Kitagawa G (1989) Non-Gaussian seasonal adjustment. Computers Math. Applic, 18: 503-14.

16. Durbin J and Koopman SJ (2012) Time series analysis by state space methods. 2nd ed. Oxford University Press, Oxford.

17. Lee Y, Nelder JA and Pawitan Y (2006) Generalized linear models with random effects: unified analysis via h-likelihood. Chapman & Hall-CRC.

18. Lee Y and Nelder JA (1996) Hierarchical generalized linear models (with discussion). J. R. Statist. Soc., B, 58: 619-78.

19. Rigby RA and Stasinopoulos DM (2014) Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. Stat Methods Med Res, 32: 318-32.

20. Djennad A, Rigby RA, Stasinopoulos D, Voudouris V (2012) Detrended transformed Owen's plot: a diagnostic tool for checking the adequacy of a fitted model distribution. STORM Research Centre, London Metropolitan University, London.

21. Cameron AC and Trivedi PK (2013) Regression analysis of count data. 2nd ed.Cambridge University Press, USA.

22. ECDC (2020). Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. https://www.ecdc.europa.eu/sites/default/ files/documents/SARS- CoV-2-variant-multiple-spike-protein-mutations-United-Kingdom.pdf

23. Zurbenko IG, Sun M (2017) Applying Kolmogorov-Zurbenko Adaptive R-Software. International Journal of Statistics and Probability; 6: 110-8.

24. R Development Core Team (2020) R: A Language and Environment for Statistical Computing. Austria: R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, URL http://www.R-project.org.