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

Research Article Open Access

Estimating the Unobserved Signals in the Mean and Dispersion in the Number of People who Tested Positive of COVID-19 in the UK

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) Estimating the Unobserved Signals in the Mean and Dispersion in the Number of People who Tested Positive of COVID-19 in the UK. J Biostat Biometric App 8(1): 102

Received Date: September 21, 2023 **Accepted Date:** October 21, 2023 **Published Date:** October 24, 2023

Abstract

COVID-19 pandemic is a global threat, where the rate of infection with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), without prevention strategies, increases exponentially, and the spread of the virus from person to person is very fast. Fitting for COVID-19 infectious disease countshas received a great deal of attention, and modelling the dispersion of COVID-19infectious disease counts can help 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 peoplewho tested positive of COVID-19 per day in the UK, and estimates the impact of the 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 people who tested positive of COVID-19 in the UK are fitted withthe generalized structural time series (GEST) model.

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

Background

Time series of infectious diseases represent the counts of event of infection per unit of time (daily, weekly, monthly or quarterly). The most commonly used distribution for analysis of counts is a Poisson where the sample space of the Poisson distribution is the set of non-negative integers. However, the Poisson model requires that the mean equalsthe variance, an assumption that is inadequate in the analysis of infectious disease countsas it is shown that infectious diseases exhibit spikes or outbreaks over time. To measure these outbreaks, a more flexible distribution is needed. Negative binomial distribution ismore flexible than the Poisson as it has a dispersion parameter that can vary over time and measures infectious disease outbreaks. The dispersion parameter is interpreted by [1] as the degree of clumping in the population, as the population becomes more clumped, the variance is expected to be higher than the mean. [2-5] among others, proposed time series models for counts and applied them for Poliomyelitis cases in the USA from the year 1970 to 1983 using negative binomial and Poisson distributions and fitted the conditional mean to investigate whether or not the incidence of Poliomyelitis has been decreasing since 1970. [6], used the negative binomial distribution to model the dispersion parameter as a measurement of a superspreading effect. [7] proposed a non-stationary negative binomial model with time-dependent covariates for time series of Enterococcus counts in Boston Harbour to evaluate the effects of court-mandated improvements in sewage treatment.

The rational for modelling the time trend with a random walk of order two (RW2) goes back to [11,12] in the analysis of mortality table in actuaries. Prior to Whittaker-Henderson method, actuaries used Moving Weighted Average (MWA) technique for dealing with data near the extremities as de- scribed by [13]. Whittaker-Henderson graduation (smoothing) technique was a new and a better method to smooth the data near extremities and has been widely employed by actuaries since then. [14] analysed seasonal ad- justed macroeconomic time series using RW2 Whittaker-Henderson method to model the business cycle and growth. [15] developed a flexible smoothing with B-splines based on Whittaker graduation (smoothing) technique. Rue and Held (2005) used RW2 to model the trend in Gaussian Markov field, The second-order random walk (RW2) model is now commonly used for smoothing data and modelling response functions and it is computationally efficient due to the Markov properties of the joint (intrinsic) Gaussian density as described by [16].

If the seasonality is present in the data but ignored by the model-building process, the result is likely to be a misspecified model which is likely to lead to residual autocorrela-tion of the order of the seasonality. To estimate stochastic seasonality in the mean and overdispersion, random seasonality model uses unity factors with levels of data frequency. To allow for a disturbance noise in seasonality, the sum of the unit levels follows a random walk process with mean zero and variance σ^2_W, [17,18].

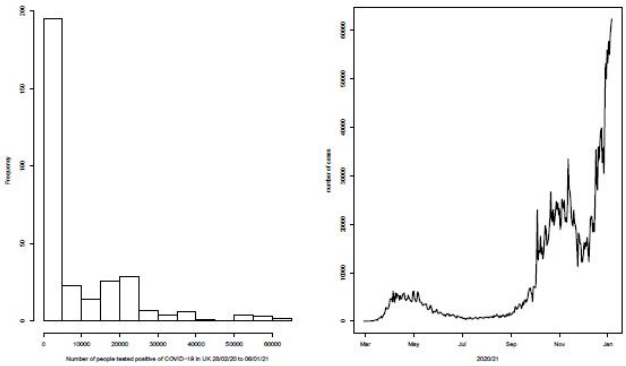

Figure 1: Histogram and time series of daily counts of people who tested COVID-19 in the UK from 28 Feb 2020 to 6 Jan 2021

Statistical Modeling

The usual decomposition of time series into four components: baseline, trend, seasonalityand white noise, can be thought of as a one-dimensional decomposition of the mean parameter of the assumed probability distribution for the observations. The proposed model in this article, which is more flexible, uses a two-dimensional decomposition of theparameters of the probability distribution function for serially correlated observations, where time series of the mean and dispersion parameters of the probability distribution function are decomposed jointly and simultaneously into baseline, trend and seasonality.The goal of the two-dimensional decomposition of the observations over time is to extracta signal of information for unobserved overdispersion of the observations from the noise, if there is a significant time-varying overdispersion in the data, then the residuals from thefitted mean model will have a higher variance than the standard normal distribution if theoverdispersion is fitted with a constant. A four-dimensional decomposition was developedin 2015 by Djennad et al.

The United Kingdom imposed a national lockdown with coercive measures on 23 March 2020 to bring the rate of transmission and infection under control. This prevention strategy brought the rate of infection down in June, July and August 2020 but since the mid of September to 6 January 2021 the numberof COVID-19 cases were rocketing. The daily number of people who tested positive of COVID-19 in the UK from day 28 February 2020 until 06 January 2021 are shown in Figure 1. The of Figure 1 presents the frequency distribution of the daily counts which looks highly positively skewed and the left hand of Figure 1 presents a time series of the daily counts. The average value of the observations is 9,135.48 people and the variance is 154,711,307people, which shows evidence of a higher variance in the data. The data is available at https://coronavirus.data.gov.uk/details/cases

2.1 Generalized Structural Time Series Model

The time series of number of people who tested positive of COVID-19 across the United Kingdom from day 28 February to 6th January 2021 were examined using the generalizedstructural time series model with the Poisson and the negative binomial distributions. Thenatural logarithm of the expected number of people who tested positive over time and the natural logarithm of the dispersion vector over time were jointly and simultaneouslydecomposed into baseline, trend and seasonality.

Let Y_t be the daily number of people who tested positive of COVID-19 in the UK from 28 February 2020 to 06 January 2021, and $D = N\,BI(\mu_t, \sigma_t)$ where N BI represents the negative binomial type I distribution of the response variable,

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

where n is the size of United Kingdom population in 2020 estimated at 67,886,011 peopleat 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 overdispersion parameters, the $\gamma_{k,t}$ represent the time-varying trends in the mean and overdispersion of the observations where the trends are extracted with a random walk of order two (RW2), $s_{k,t}$ represent time-varying day-of-the-week effects in the mean and overdispersion parameters, and $b_{k,t}$ and $w_{k,t}$ are independently distributed disturbance terms with mean zero and variances σ^2_{bk} , σ^2_{wk} where $b_k \sim N_{T-J}\ 0, \sigma^2\ I_{T-J}$ and $\omega_k \sim N_{T-Jk}\ 0, \sigma^2_{bk}\ I_{T-M+1}$.

Model Estimation

The GEST model defined by equation (2) 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 in trend and seasonality, where $b \sim N_{T-J}(0, \sigma^2 I_{T-J})$, $\omega \sim N_{T-M+1}(0, \sigma^2 I_{T-M+1})$.

5.1 Maximum Likelihood Estimation

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

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma_b^2, \sigma_w^2) &= \int\int f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}|\sigma_b^2, \sigma_w^2) d\boldsymbol{\gamma} d\mathbf{s} \\ &= \int\int f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}) f(\boldsymbol{\gamma}, \mathbf{s}|\sigma_b^2, \sigma_w^2) d\boldsymbol{\gamma} d\mathbf{s} \end{aligned} \tag{3}$$

where

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}) = \prod_{t=1}^{T} f(y_t|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}) \tag{4}$$

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

$$l = \log f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}) = \sum_{t=1}^{T} \log f(y_t|\mu_t, \sigma_t) \tag{5}$$

denotes the log-likelihood function. The extended log-likelihood function in Lee, [19] p227-279, is defined as:

$$l_e = l + \log f\left(\boldsymbol{\gamma}, \mathbf{s}|\sigma_b^2, \sigma_w^2\right) \tag{6}$$

where $f(\gamma, s|\sigma^2_b, \sigma^2_w)$ is the joint density function of γ_t and s_t given σ^2 and σ^2. 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 hierar- chical generalized linear model [20]. However, the integration of (3) isintractable for a non-Gaussian response variable and becomes more difficult when thereis more than one random effect component, here we have two random effect componentsγ and s. This integral can be approximated using Laplace approximation which gives the following approximative marginal log likelihood:

$$l(\boldsymbol{\beta}, \sigma_b^2, \sigma_w^2) = \log f(\mathbf{y}|\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}}) + \log(\hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}}|\sigma_b^2, \sigma_w^2) - \frac{1}{2} \log \left| \frac{\hat{\mathbf{D}}_{\boldsymbol{\gamma},\mathbf{s}}}{2\pi} \right| \tag{7}$$

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 (β, σ^2_b, σ^2_W), and $\hat{D}_{\gamma,s}$ is the second derivative of the ex- tended likelihood with respect to (γ, s) evaluated at $\gamma = \hat{\gamma}$ and $s = \hat{s}$

Following Appendix B2 and C in [21], the maximizationof the extended likelihood in (6), can be achieved by using the GEST algorithm describedbelow, which provides posterior mode estimates of the sets of parameters of β, γ, s and by maximizing the extended log likelihood for hyperparameters σ^2_{bk} , σ^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 fiteach of the distribution parameter vectors, $\theta_k = (\mu_t, \sigma_t)$

(i) evaluate the current *iterative response variable* $\mathbf{z}_k$ and current *iterative weights* $\mathbf{W}_k$, where $\mathbf{z}_k = \boldsymbol{\eta}_k + \mathbf{W}_k^{-1}\mathbf{u}_k$, $\mathbf{W}_k = -\frac{\partial^2 \ell}{\partial \eta \partial \eta^\top}$, or $-E\left[\frac{\partial^2 \ell}{\partial \eta \partial \eta^\top}\right]$ or $\left(\frac{\partial \ell}{\partial \eta}\right)^2$, 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 $s_k = z_k - \gamma_k - s_k$ 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 like-lihood function Q, and then estimate γ_k and s_k using the equation $(\gamma_k, s_k)^T = A_k + D_k^T M_k^{-1} D_k^{-1} A (s_k, 0_k)^T$, 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 $\eta = g(\theta_k)$.

(C) end the inner cycle on convergence of θ_k. 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 obtainedbefore 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 season-ality model, is given by:

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

where $s = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)^T$, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_T)^T$, $s = (s_1, s_2, \ldots, s_T)^T$, $\Sigma^{-1} = \sigma^{-2}W$,

Σ is a [T x T] matrix, A is a [2T × 2T] matrix of Σ^{-1}, M is a matrix diagonal$[\sigma^2 \mathbf{I}_{T-J}, \sigma^2 \mathbf{I}_{T-M+1}]$ 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) × T]. TheQ 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 fittingalgorithm, is given by solving the following equation:

$$\boldsymbol{\gamma} = \left[\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right]^{-1}\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\epsilon}. \tag{8}$$

5.2 Effective Degrees of Freedom

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

$$df = df_{\mu_t} + df_{\sigma_t} = p_k + d_k \tag{9}$$

for k = 1, 2, where p_k is the length of β_k and the d_k is the effective degrees of freedomfor the fitted random walk trend and random seasonality. Let B =A + $D^TM^{-1}D$ and 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, $(\hat{\gamma},\hat{s}) = \hat{B}s$. 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:

$$d_k = trace\left[\hat{\mathbf{B}}_k\right] = trace\left\{\left[\hat{\mathbf{A}}_k + \hat{\mathbf{D}}_k^\top\hat{\mathbf{M}}_k^{-1}\hat{\mathbf{D}}_k\right]^{-1}\hat{\mathbf{A}}_k\right\}. \tag{10}$$

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

$$d_k = trace\left[\hat{\mathbf{B}}_k\right] = trace\left\{\left[\hat{\boldsymbol{\Sigma}}_k^{-1} + \hat{\sigma}_{b_k}^{-2}\hat{\mathbf{D}}_k{}^\top\hat{\mathbf{D}}_k\right]^{-1}\hat{\boldsymbol{\Sigma}}_k^{-1}\right\}. \tag{11}$$

As d_k is difficult to calculate directly for large T, it can be calculated by setting $\partial Q/\partial\sigma^2 = 0$ giving on convergence $d = J + \hat{\sigma}^{-2}\hat{\gamma}^TD^TD\hat{\gamma}$, 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 $d = J + M - 1 + \hat{\sigma}^{-2}\hat{\gamma}^TD_\gamma^T D_\gamma\hat{\gamma} + \hat{\sigma}_W^{-2}\hat{s}^TD_s^T D_s\hat{s}$, for the random walk trend and random seasonality model, using the result [2] $\log|xC + F| = \mathrm{tr}\,[(xC + F)^{-1}C]$, 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 $\hat{\sigma}_W^{-2}$. The estimates of these variances are often very small (see Table 3), so the inverse of these variances is very

large with very large effective degrees of freedom. In therandom walk of order 2, the fitted trend is a smooth curve with no much variability and the variance is very small and inverse is very large, but because the smoothing matrix D_γ is a second order differencing matrix, the $\hat{\gamma}^T D_\gamma^T D_\gamma \hat{\gamma}$ become smaller and it compensates the very large value of the inverse of the variance

Model testing

6.1 Testing for Overdispersion

In the GEST-NBI, σ_t is modelled with a random walk of order two and random seasonalityusing a log link, testing the overdispersion is performed by testing the null hypothesis$H_0 : \sigma_t = 0$ against the alternative $H_1 : \sigma_t > 0$, at 1% significance level. 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 see [10].

6.11 Logarithmic Transformation

Anscombe (1950) pointed out that the logarithm transformation of the observations makes the variance independent of the mean and Preston (1948) recommended that in a situation where the zero counts are not recorded, transforming the data by logarithm transformation it appears approximately normal and make the method of analysis of variance appropriate. If the variance of Y is of the form $Var(y) = \varphi\mu^p$, where $\varphi > 0$ is a scale parameter, the standard deviation is proportional to the mean and hence the log transformation stabilizes the variance and makes analysis of variance more appropriate. For negative binomial distribution, $p = 2$, so the log (Y) stabilizes the variance. In the data, there were none zero observations recorded in any day from 28 February 2020 to 06 January 2021, so time series of number of people who tested positive of COVID-19 in the UK was logarithmical transformed for the analysis of variance. The goodness of fit of GEST- NO model with a fixed standard deviation, $\log(\sigma_t) = \beta_2$, and GEST-NO model with a time-varying standard deviation, $\log(\sigma_t) = \beta_2 + \gamma_{2,t} + s_{2,t}$, were compared using Akaike information criterion (AIC). The histogram, time series of COVID-19 positive countsand fitted values of GEST-NO model for mean and standard deviation are shown in Supplementary file.

6.2 Testing Normality of Fitted NBI and Poisson Models

Normalized (randomized) quantile residuals of the fitted GEST model are checked usingthe detrended transformed Owen's plot DTOP [15] to check the adequacy of the fitted probability distribution function. If the fitted Poisson distribution is shown to be adequate, then the model for overdisperisonis 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 originalmodel distribution. Hence the fitted normalized (randomized) quantile residuals are al- ways compared to a standard normal distribution making interpretation and comparisonof the resulting plots (for different original model distributions) easier. Significant depar-tures from the model (resulting in significant differences between the residuals and theirstandard 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. Similarlythe DTOP of the fitted residuals provides a guide to model adequacy, if the fitted dis- tribution 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 [21] Table 1.

6.3 Model Selection Criteria and Comparison

The GEST with the Poisson and negative binomial were fitted and compared (see Table 2) using the Akaike information criterion (AIC) [where AIC=−2l + 2df ; l is the log likelihood of the model, (−2l) is the global deviance of the model, and df is the total effective degrees of freedom of the fitted GEST model (equation 9); so a good fit of the model yields to a low value of AIC]. The AIC of parsimonious GEST models m1 and m2were compared with model m3, where m1 [GEST-Poisson model] was fitted with a RW2 trend and random seasonality for μ_t, model m2 was fitted with a RW2 trend and random seasonality for μ_t and a constant for σ_t, and m3 was fitted with a RW2 trend and randomseasonality for μ_t and a RW2 trend and random seasonality for σ_t. A further comparison was performed using the centile curves of the fitted models. The 100p centile of a randomvariable Y is the value y_p such that $p(Y \leq y_p) = p$, i.e. $y_p = F^{-1}(p)$, so y_p is the inverse cumulative distribution

function of Y applied to p. The conditional centile of Y given explanatory variable X = x, i.e. $y_p(x) = F^{-1}{}_{Y|X=x}(p)$. By varying x a 100p centile curve of $y_p(x)$ against x is obtained. Centile curves can be obtained for different values of p. Note that a z-score given y and x is defined by $z_p = \Phi^{-1}[F_{Y|X=x}(y)]$, where Φ^{-1} is the inverse cumulative distribution function of a standard normal variable. The z-scores are used as residuals in fitted GEST and GAMLSS [23]. The World Health Organization (2007) uses 100p=(3,15,50,85,97) in its charts and 100p=(1,3, 5, 15, 25, 50, 75, 85, 95, 97, 99) in its tables.

Table 2: The fitted GEST models with Poisson and NBI, random walk of order 2 (RW2)trend over time and random seasonality (Rseas) with frequency = 7 days.

Models	Probability distribution	$\mu_t(\gamma_{1,t}, s_{1,t}) =$	$\sigma_t(\gamma_{2,t}, s_{2,t}) =$
m1	Poisson	RW2 + Rseas	-
m2	Negative binomial type I	RW2 + Rseas	constant RW2
m3	Negative binomial type I	RW2 + Rseas	+ Rseas

Results

7.1 Centile Curves of the Fitted GEST-Poisson and GEST-NBI

The percentiles 100p = (0.5, 2.5, 25, 50, 75, 97.5, 99.5) of number of people tested positiveof COVID-19 in the UK from day 28 February 2020 to 06 January 2021 and the percentilesestimates of GEST Poisson and NBI models were given in Table 4. Below 2.5% percentile of number of people tested positive, the estimate percentile of GEST-Poisson model wasoverestimated at 23.89%, with GEST-NBI with a fixed overdispersion was a bit higher at 3.82% and GEST-NBI with a stochastic over dispersion was very close at 2.23%. Below 50% percentile of number of people tested positive, the estimate percentile of GEST- Poisson model m1 was identical 50.00%, with GEST-NBI m2 was quite low at 46.50% andGEST-NBI m3 was very close at 48.09%, and below 97.5% percentile of number of peoplewho tested positive, the estimate percentile of GEST-Poisson m1 was underestimatedat 80.25%, GEST-NBI model m2 and GEST-NBI model m3 were very close at 97.77%.The percentile curves for 100p = (2.5, 50, 97.5) were plotted in Figure 3 and FiguresS1, S2. The fitted centile curves for GEST-Poisson were tight and overlapping closer to 50% (see Figure S1). The percentiles below 50, 100p = (0.5, 2.5, 25), the Poisson model overestimated percentiles at $100\hat{p} = (19.43, 23.89, 41.72)$ respectively, and for 100p above 50%, 100p = (75, 97.5, 99.5), the Poisson model underestimated the percentiles at $100\hat{p} = (61.46, 80.25, 84.39)$ respectively. In GEST-NBI, the fitted centile curves of NBI models were wider and dispersed unlike the Poisson. For percentiles below 50%, 100p = (0.5, 2.5, 25), GEST-NBI model m2 fairly overestimated the percentiles at $100\hat{p} = (1.59, 3.82, 23.25)$ respectively, and for 100p above 50%, 100p = (75, 97.5, 99.5) the NBI model m2 fairly underestimated the percentiles at $100\hat{p} = (77.39, 97.77, 99.68)$ respectively. For GEST-NBI model m3, has very good prediction of the percentiles with $100\hat{p}$ =(0.00, 2.23, 28.03, 48.09, 73.89, 97.77, 99.68) respectively. The panels (a), (b), (c), (d), and (e) of Figure 2 shows the number of people who tested positive below each percentile 100p = (0.5, 2.5, 25, 50, 75, 97.5, 99.5) together with the predicted percentiles of the fitted modelm3.

Table 4: The estimates of percentiles 100p = (0.5, 2.5, 25, 50, 75, 97.5, 99.5) from the fitted models m1 (Poisson), m2 (NBI with constant dispersion) and m3 (NBI with randomdispersion).

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

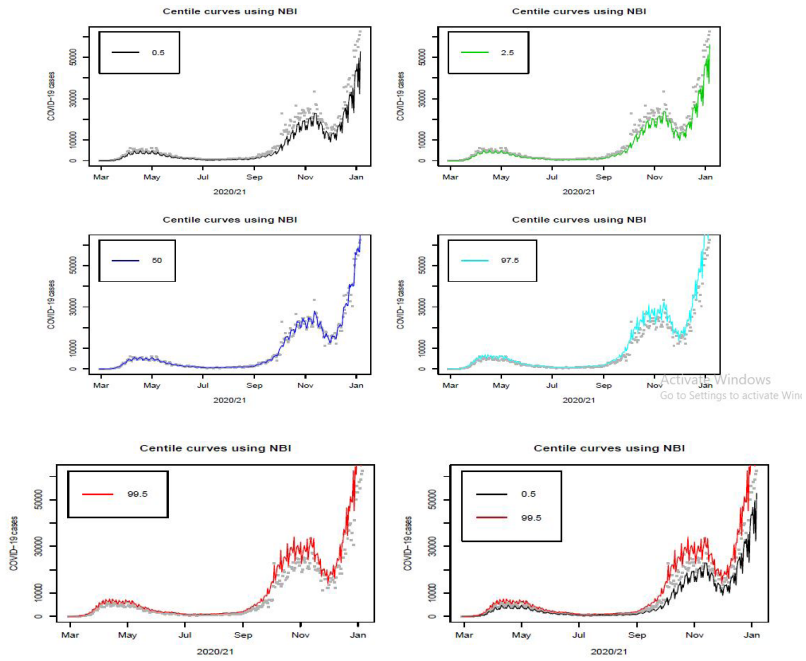

Figure 2: The number of people tested positive of COVID-19 from 28 February 2020 to 06 January 2021 in UK and the centile curves of fitted model. The graph shows thesample percentages below each centilecurve for comparison with the model predicted percentages. Cases below 0.5% model centile is 0%, cases below 2.5% the model centile is 2.23%, cases below 50% model centile is 48.09%, cases below 97.5% model centile is 97.77 and cases below 99.5% model centile is 99.68%

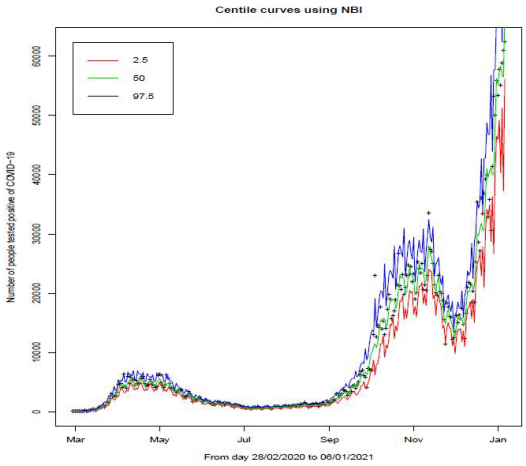

Figure 3: Daily counts of people who tested positive of COVID-19 from day 28 February 2020 to 06 January 2021 in the United Kingdom together with centile curves of the fitted GEST-NBI m3 with a time-varying overdispersion. The graph shows the sample percentages below each centile curve for comparison with the model predicted percentages.Cases below 2.5% model centile is 2.23%, cases below 50% model centile is 48.09%, casesbelow 97.5% model centile is 97.77%

7.2 Model Comparison

Table 2 reports the GEST-Poisson and GEST-NBI which were fitted and Table 3 reportsthe estimates of their parameters: baselines, hyperparameters for the trend and seasonality and Akaiki Information Criteria. The AIC of m2 [for a fixed $\hat{\sigma}_t$] was 4831, it dropped by 52 in m3 [for a RW2 trend and seasonality for $\hat{\sigma}_t$], where AIC(m3) = 4779 (see Table 3), therefore,

model m3 has a better fit, it implies that a RW2 trend and seasonality for overdispersion improves the predictive values of the GEST-NBI in comparison with the model of a constant overdispersion. The AIC of GEST-Poisson m1 [RW2 trend and a random seasonality for $\hat{\mu}_t$], was 6169, it dropped by 1390 in GEST-NBI of m3 where AIC(m3) = 4779 (see Table 3). The fitted RW2 trend over time in m1 and m3 look similar with a constant but if the constant is dropped, then trend of the NBI is much higher thanthe trend of the Poisson model, where the estimate of the baseline of number of infectionper day was 9323.94 in Poisson and the estimate of the baseline of number of infection per pay was 3223.88 in NBI (see Table 3). The fitted random seasonality removed the autocorrelation of the residuals in all models, in m1 the seasonality was more stochastic than the fitted seasonality in m3 which was constant over time. The fitted value of the disturbance variance in m1 seasonality was $\hat{\sigma}^2_{w1,t}$= 0.007344, which is higher than the disturbance variance in m3 seasonality $\hat{\sigma}^2_{w1,t}$= 4.47×10^{-5}. In GEST-Poisson, the fitted RW2 trend and random seasonality model in $\hat{\mu}_t$ overfitted the observations, whereas in GEST-NBI the fitted $\hat{\mu}_t$ was smoother with no much variability but the seasonality for overdispersion σ_t was more stochastic with disturbance variance $\hat{\sigma}^2_{w2,t}$= 0.0532.

Table 3: The estimates of models hyperparameters, baselines, -Log Lik. , effective degreesof freedom and AIC for models m1 (Poisson), m2 (NBI) and m3 (NBI)

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	4830.99
m3	0.000173	4.4723e-05	9.0509e-05	0.05321	3223.88	0.01676	2301.24	88.14	4778.8

7.3 The Chosen Model

Model m3 is considered to be better than models m1 and m2 as it has the smallest AIC, where the maximum likelihood estimates of the mean and dispersion μ_t and σ_t, denotedby $\hat{\mu}_t$ and $\hat{\sigma}_t$ respectively, for the number of people who tested positive of COVID-19 in the UK per day from 28 February 2020 to 06 January 2021 were given by:

$$\begin{aligned} Y_t|\mu_t,\sigma_t &\sim \mathcal{NBI}(\hat{\mu}_t,\hat{\sigma}_t) \\ \hat{\mu}_t &= 67,886,011 \times \exp\left[-9.955+\hat{\gamma}_{1,t}+\hat{s}_{1,t}\right] \\ \hat{\sigma}_t &= \exp\left[-4.089+\hat{\gamma}_{2,t}+\hat{s}_{2,t}\right] \end{aligned} \tag{12}$$

where

$$\begin{aligned} \hat{\gamma}_{1,t} &= 2\hat{\gamma}_{1,t-1}-\hat{\gamma}_{1,t-2}+\hat{b}_{1,t} \\ \hat{s}_{1,t} &= -\sum_{m=1}^{M-1}\hat{s}_{1,t-m}+\hat{w}_{1,t} \\ \hat{\gamma}_{2,t} &= 2\hat{\gamma}_{2,t-1}-\hat{\gamma}_{2,t-2}+\hat{b}_{2,t} \\ \hat{s}_{2,t} &= -\sum_{m=1}^{M-1}\hat{s}_{2,t-m}+\hat{w}_{2,t} \end{aligned} \tag{13}$$

The maximum likelihood estimate of time-varying mean of number of people who tested positive of COVID-19 in the UK per day from 28 February 2020 to 06 January2021 is given by $\hat{\mu}_t = 67,886,011 \times \exp[-9.955 + \hat{\gamma}_{1,t} + \hat{s}_{1,t}]$ and plotted together with time series of observations in Figure 4 (top right panel). The maximum likelihood estimate of time-varying overdispersion is given by $\hat{\sigma}_t = \exp(-4.089 + \hat{\gamma}_{2,t} + \hat{s}_{2,t})$ and plotted in Figure 4 (top left panel). Hence, the estimate of time-varying variance of number of people who tested positive of COVID-19 in the UK per day from 28 February 2020 to 06 January 2021 is given by $\hat{\mu}_t + \hat{\sigma}_t\hat{\mu}^2 = n \times \exp(-9.955 + \hat{\gamma}_{1,t} + \hat{s}_{1,t}) + \exp(-4.089 + \hat{\gamma}_{2,t} + \hat{s}_{2,t}) \times [n \times \exp(-9.955 + \hat{\gamma}_{1,t} + \hat{s}_{1,t})]^2$, where n=67,886,011, and plotted in Figure S3.

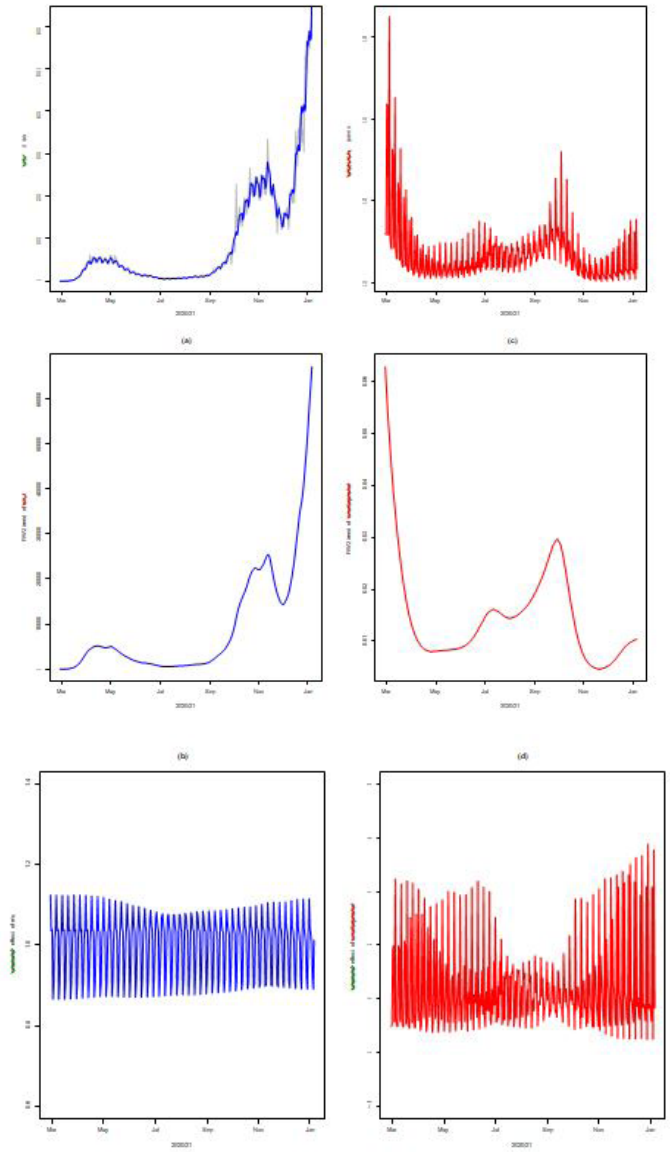

Figure 4: Time series of number of people tested positive of COVID-19 from 28 Feb 2020 to 06 Jan 2021 in UK together with the fitted $\hat{\mu}_t$ in blue and below is decomposition of the fitted $\hat{\mu}_t$ into RW2 trend and seasonality in blue. Fitted $\hat{\sigma}_t$ in red and below is decomposition of the fitted $\hat{\sigma}_t$ into RW2 trend and seasonality in red.

The estimates of the baselines for the mean and overdispersion were 3223.88 and 0.02 respectively [i.e. 67,886,011 × exp (−9.955) = 3223.88 and exp (−4.089) = 0.01676]. Hence, the baseline for the rate of infection with COVID-19 in the United Kingdom from day 28 February 2020 to 06 January 2021 is estimated at 3224 people per day.

The maximum likelihood estimates of $\hat{\gamma}_{1,t}$, $\hat{s}_{1,t}$, $\hat{b}_{1,t}$, and $\hat{w}_{1,t}$ represent the fitted values of RW2 trend over time, fitted stochastic seasonality, and disturbances in RW2 trend, andseasonality in the fitted mean level $\hat{\mu}_t$.

The maximum likelihood estimates of $\hat{\gamma}_{2,t}$, $\hat{s}_{2,t}$, $\hat{b}_{2,t}$, and $\hat{w}_{2,t}$ represent the fitted values of RW2 trend over time, fitted stochastic seasonality, and disturbances in RW2 trend, andseasonality in the fitted overdispersion level $\hat{\sigma}_t$

The maximum likelihood estimates of the variances of white noises $\hat{b}_{1,t}$ and $\hat{w}_{1,t}$ in the mean were $\hat{\sigma}^2_{b1,t} = 1.73 \times 10^{-4}$ and , $\hat{\sigma}^2_{w1,t} = 4.47 \times 10^{-5}$ respectively, where $\hat{b}_{1,t} \sim N_{T-J}(0, \hat{\sigma}^2_{b1,} I_{T-J})$, $\hat{w}1,t \sim N_{T-M+1}(0, \hat{\sigma}^2_{w1,t} I_{T-M+1})$, $t = (1, \ldots, T)$, $T = 314$, J

$= 2$, $M = 7$, I_{T-J} and I_{T-M+1} are unit matrices of size $T - J$ and $T - M + 1$ respectively. The maximum likelihood estimates of the variances of white noises $\hat{b}_{2,t}$ and $\hat{w}_{2,t}$ in the overdispersion were $\hat{\sigma}^2_{b1,t} = 9.05 \times 10^{-5}$ and $\hat{\sigma}^2_{w2,t} = 0.05321$ respectively, where $\hat{b}_{2,t} \sim N_{T-J}(0, \hat{\sigma}^2_{b2,t} I_{T-J})$, $\hat{w}2,t \sim N_{T-M+1}(0, \hat{\sigma}^2_{w1,t} I_{T-M+1})$, $t = (1, \ldots, T)$, $T = 314$, $J = 2$, $M = 7$, I_{T-J} and I_{T-M+1} are unit matrices of size $T - J$ and $T - M + 1$ respectively.

7.4 The Fitted Mean and Overdispersion of COVID-19 Cases

The GEST-NBI decomposition of the fitted values of the time-varying mean ($\hat{\mu}_t$) into base- line × RW2 trend [i.e. 67,886,011 × exp (−9.955) × exp ($\hat{\gamma}_{1,t}$)] and stochastic seasonality [i.e. exp ($\hat{s}_{1,t}$)] were plotted in Figure 4 panels (a) and (b) on antilog scale, respectively. Panel (a) shows that the mean number of people who tested positive of COVID-19 was increasing exponetially in the first 34 days [from day 28 February 2020 to 01 April 2020] ofCOVID-19 pandemic, it slowed down from 01 April 2020 to 14 April 2020 when it reachedthe peak of infection in 14 April 2020. The mean number was estimated at 5.59 people inday 28 February 2020 and soared to 5160.70 people in 14 April 2020. From 14 April to 25April 2020 the mean number of people test positive dropped from 5160.70 to 4770.23, then it increased again from 4770.23 to 4999.31 [from 25 April to 1 May 2020], then decreasedsignificantly to a lowest level in 7 July 2020 since April, from 4999.31 to 555.40 people on average as shown in Figure 4 panel (a) The mean trend decelerate to 4770.23 people in 25 April 2020 then increased again to 4999.31 in day 1 May 2020. From 1 May 2020, the trend was plummeting continuously until 7 July 2020 from 4999.31 to 555.40 people.From 14 to 20 August 2020 the mean was stable, from 1050.201 to 1074.61 people. From 24 August to 12 November 2020, the mean accelerated from 1144.64 to 25408.88, then decelerate to 14279.66 in 1 December 2020. From 1 December to 6 January 2021, the mean was increasing very fast from 14279.66 to 67111.62. This rapid increase in COVID- 19 cases arose as a result of emergence of a new variant of SARS-CoV-2 in South EastEngland which was spreading rapidly in the population as pointed out by the European Centre for Disease Prevention and Control (ECDC)[24] in 20 December 2020 assessment re- port: 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 virolog-ical investigations. Analysis of viral genome sequence data identified a large proportion of cases belonged to a new single phylogenetic cluster. In addition, the fast increase or ashift in the peak of mean number of cases from 5160.70 people in 14 April 2020 to higher peak of 25408.88 in 12 November 2020, was due to an increase in testing volume overthat time. While the numbers of reported cases in March and April appear low, this was because only testing in hospitals was occurring and large scale population testing only started in May. Panel (b) shows a constant daily seasonality in the mean number of cases, reflecting a daily testing cycle throughout the year in number of people tested positive,there was a small change on daily seasonality from March to July, then a slight increasein the cycle. The coefficients of daily seasonality in the mean number of cases varied from 0.8651 to 1.1246, where [(exp ($\hat{s}_{1,t}$) − 1) × 100%] can be interpreted as the proportionate change in the mean number of people tested positive due to a one-unit change in the day- of-the-week effects ,this percentage change varied from -13.49% to 12.46%, where some days-of-the-week increased the mean number of people tested positive by [0% to 12.46%] and other days-of-the-week decreased the mean number of people tested positive by [0%to -13.49%], with more details in the next Section.

Figure 4 panels (c) and (d) plots the GEST-NBI decomposition of the fitted valuesof time-varying overdispersion ($\hat{\sigma}_t$) into baseline × RW2 trend [i.e. 0.01676 × exp ($\hat{\gamma}_{2,t}$)] and stochastic seasonality [i.e. exp ($\hat{s}_{2,t}$)] on antilog scale respectively. Panel (c) shows the fitted values of baseline × RW2 trend of time for overdispersion in number of people who tested positive of COVID-19 in the UK from day 28 February to 6 January 2021. This trend without baseline varied from 0.28 to 3.76 and with baseline varied from 0.01 to 0.06. It started with a higher estimate in 28 February with and decreased gradually over time. Panel (d) shows a time-varying daily seasonality for overdispersion in COVID-19 cases. The overdispersion in number of people who tested positive of COVID-19 in the UK from day 28 February to 6 January 2021 exhibited a significant time-varying cyclical pattern reflecting a considerable change in the degree of infection in the population, themagnitude of the cycle dropped significantly between July and Sept, then increased again between Oct to Jan. The significant variation in overdispersion seasonality arose as theresult of the wider spread in number of people who tested positive of COVID-19 and the impact of lockdown prevention strategy to minimise its wider spread.

The time-varying day-of-the-week effects [$\hat{s}_{1,t}$, $\hat{s}_{2,t}$] in mean and overdispersion of num- ber of people who tested positive with COVID-19 were plotted per month in Figure 5 (in blue and red respectively), where [($e^{\hat{s}k,t} - 1$) *100] can be interpreted as

the propor- tionate change in the mean and overdispersion number of people tested positive due to a one-unit change in the seasonality of mean and overdispersion respectively. The fitted daily seasonality in the mean in Figure 5 (in blue) did not show significant variation, it was stable with a higher peak detected on Fridays, this peak plateaued and decreased slightly from July to Dec. The fitted daily seasonality in overdispersion in Figure 5 (inred) showed significant variation and changes in March to June and October to Decem- ber. This provides strong evidence that the national lockdown dropped the seasonality inoverdispersion significantly.

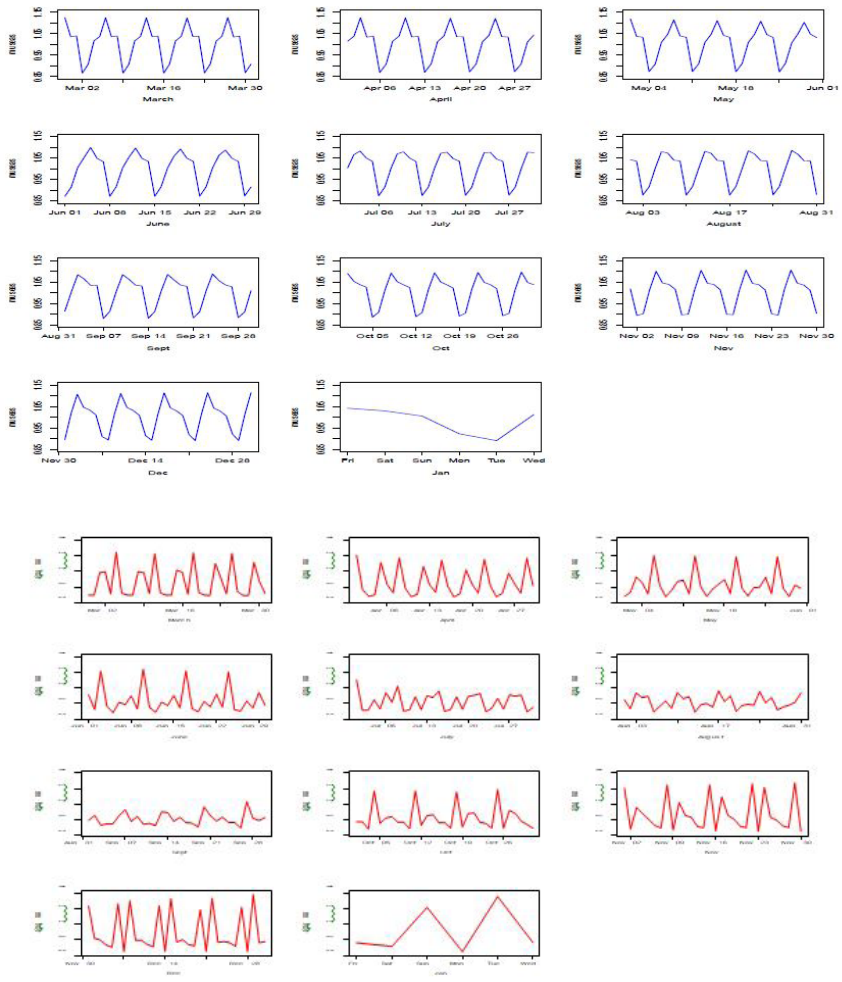

Figure 5: Estimation of fitted day-of-the-week effects $\hat{s}_{1,t}$ in the mean per month (in blue) and the fitted day-of-the-week effects $\hat{s}_{2,t}$ in the overdispersion per month (in red) of GEST-NBI model m3

7.5 Day-of-the-Week Effects

The maximum likelihood estimates of the coefficients of the day-of-the-week effects in $\hat{\mu}_t$ and $\hat{\sigma}_t$ on average per month [i.e. $(\hat{s}_{\mu}t|\text{day})$ and $(\hat{s}_{\sigma}t|\text{day})$] were given in Tables 6 and 7 respectively and were plotted in Figure 6. The estimated day-of-the-week effects could be interpreted as a positive or a negative effect of a calendar on the mean and overdispersionin number of subjects who tested positive, where the coefficients above 1 represent a posi- tive calendar effect, and the coefficients below 1 represent a negative calendar effect. The maximum likelihood estimates of the coefficients for Friday-Thursday effects in March were estimated at $\hat{s}_{\mu}t|\text{days} = (1.1241, 1.0352, 1.0386, 0.8651, 0.9076, 1.0151, 1.0374)$ re- spectively. Hence, there was an increase of 12.41% in the mean number of cases on Friday, an increase in the mean by 3.52% on Saturday, an increase of 3.86% on Sunday, a decrease in the mean by - 13.42% on Monday, a decrease of - 9.24% on Tuesday, an increase of1.54% on Wednesday and 3.73% on Thursday. The estimate of average Sundays effectsin overdispersion in March was $\hat{s}_{\sigma}t|\text{Sun} = 2.209$ [an increase of 120.9%]

and dropped to $\hat{s}_{\sigma t|\mathrm{Sun}} = 0.786$ in June [-21.4%], then decreased to $\hat{s}_{\sigma t|\mathrm{Sun}} = 0.636$ in July [-36.4%] and to $\hat{s}_{\sigma t|\mathrm{Sun}} = 0.818$ in August [-18.2%] (see Table 7). The average Wednesdays effects in overdispersion dropped significantly from $\hat{s}_{\sigma t|\mathrm{Wed}} = 3.179$ in March [increase of 217.9%] to $\hat{s}_{\sigma t|\mathrm{Wed}} = 0.825$ in 1-6 Jan 2021 [-17.5%], (see Table 7), where $[(\hat{s}_{\sigma t|\mathrm{day}} - 1) \times 100\%]$ can be interpreted as the proportionate change in the day-of-the-week effects of overdispersion in number of people tested positive due to a one-unit change in the seasonality. The averageSundays effects in overdispersion increased rapidly from $\hat{s}_{\sigma t|\mathrm{Sun}} = 0.818$ in August [-18.2%] to $\hat{s}_{\sigma t|\mathrm{Sun}} = 1.936$ in September [93.6%], to $\hat{s}_{\sigma t|\mathrm{Sun}} = 2.901$ in October [190.1%], to $\hat{s}_{\sigma t|\mathrm{Sun}} = 3.240$ in November [224%], to $\hat{s}_{\sigma t|\mathrm{Sun}} = 3.115$ in December [211.5%], to $\hat{s}_{\sigma t|\mathrm{Sun}} = 3.075$ in 1-6 January 2021 [207.5%]. The average Tuesdays effects in overdispersion increased from $\hat{s}_{\sigma t|\mathrm{Tue}} = 0.725$ in June [-27.5%] to $\hat{s}_{\sigma t|\mathrm{Tue}} = 3.771$ in 1-6 January 2021 [277.1%] (see Table 7).

7.6 Normality Test: Adequacy of the Fitted Distribution

The DTOP in Figure S4 for GEST-Poisson model panel shows a negative slope which indicates that the variance in the residuals is too large and the variance in the model response variable is too small. Hence, the Poisson distribution is inadequate to fit the number of people who tested positive with COVID-19 in the UK since the bands cross thezero horizontal line, whereas in GEST-NBI m3 the bands did not cross the zero horizontalline which indicates that the negative binomial distribution is more adequate. Table 5 shows higher variance in the residuals of the fitted Poisson model.

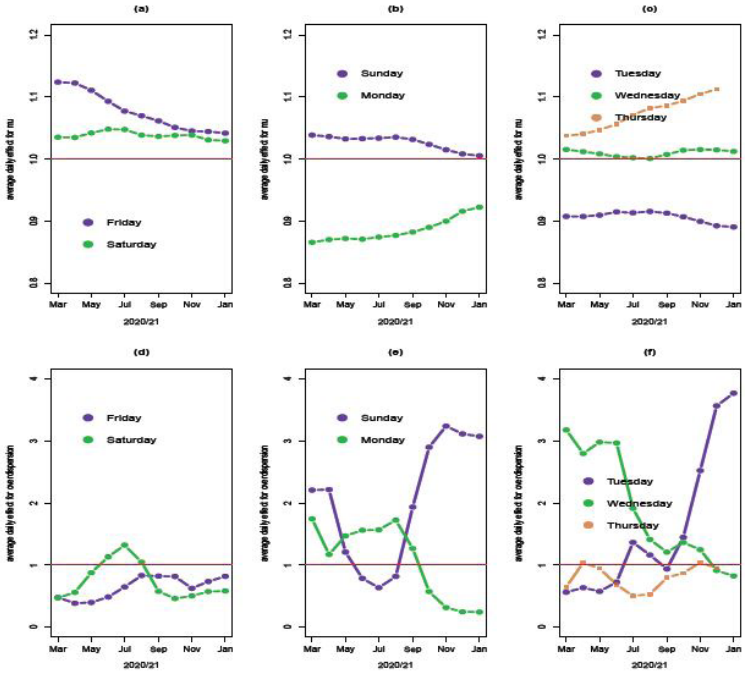

Figure 6: Average day-of-the-week effects per month: panels (a), (b) and (c) show averageday-of-the-week effects over time per month in the mean number of people tested positive of COVID-19 in the UK from 28 Feb 2020 to 6 Jan 2021, and panels (d), (e) and (f)show average day-of-the-week effects over time per month on the time-varying dispersion of number of people tested positive of COVID-19 in the UK from 28 Feb 2020 to 6 Jan 2021.

Table 5: Summary of the randomised quantile residuals NBI and Poisson models

Models	mean	variance	skewness	kurtosis	Filliben corr.
m1	-0.1324988	8.088289	-0.0382294	3.310703	0.997329
m2	-0.0040586	1.024121	0.0205053	5.083053	0.987658
m3	-0.0186246	0.996569	0.0063311	2.528352	0.997337

Table 6: Estimates of average day-of-the-week effects per month on the fitted mean of GEST-NBI model m3, $\hat{s}_{\mu t}|\text{day}$, where $[(\hat{s}_{\mu t}|\text{day} - 1) \quad 100]$ can be interpreted as the propor- tionate change in the seasonal effect in the mean number of people tested positive due toa one-unit change in the seasonality

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

Table 7: Estimates of average day-of-the-week effects per month on the fitted random dispersion of GEST-NBI model m3, $\hat{s}_{\sigma t}|\text{day}$, where $[(\hat{s}_{\sigma t}|\text{day} \_\_ 1) \quad 100]$ can be interpreted as the proportionate change in the seasonal effect of overdispersion in number of peopletested positive due to a one-unit change in the seasonality.

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

7.7 Overdispersion Test

The fitted values of the time-varying overdispersion $\hat{\sigma}_t$ were positive but close to zero, $\hat{\sigma}_t \in [0.001 - 0.16]$, suggesting an equidispersion Poisson model. However, the likelihood ratio test statistic was 1149.44 [-2 x (2301.24-2875.96) = 1149.44] which exceeded the 1% critical value of $\chi^2_{.98} = 5.41$, therefore, the tests results strongly reject the null hypothesis of the equidispersion GEST-Poisson model, indicating the presence of (modest) overdispersionin the data. The fitted conditional variance exceeded the mean $\text{V ar}(Y_t|\mu_t, \sigma_t) >> \mu_t$ as shown in Figure S3. In GEST-NO model for log transformation of number of peopletested positive with COVID-19 in the UK, model with random standard deviation has lower AIC than the model with a constant standard deviation. log-likelihood was 224.61 and AIC = -271.06 compared to log-likelihood of 162.74 and AIC = -207.05. The fitted values of the time-varying standard deviation in GEST-NO $\hat{\sigma}_t$ varied between 0.04 to 0.53 [i.e. $\hat{\sigma}_t \in [0.04 - 0.53]$], as shown in Figure S7. The log transformation reduced the disturbance in the fitted standard deviation seasonality but the RW2 trend for standard deviation was higher in GEST-NO as shown in Figure S8.

Conclusion

This article provides strong evidence that the first national lockdown suppressed the epidemic of COVID-19 significantly, and suppressed the day-of-the-week effect of overdis-persion of COVID-19 infection significantly. By quantifying the unobserved signal of the trend over time with a random walk of order 2, the first national lockdown decreased the trend of COVID-19 infection significantly to a lowest level in 7 July 2020 from the peak of infection in 14 April 2020. In first 34 days [from day 28 February 2020 to 01 April 2020], the fitted RW2 trend in the mean number of people who tested positive withCOVID-19 increased exponentially. This article shows strong evidence that the fitted variance exceeded the fitted mean over time in the number of people tested positive of COVID-19 in the UK suggesting that there is an overdispersion which arose because of extra source of variation among the number of people who tested positive with COVID-19and that the relationship between the variance of the mean is not linear. By quantifying the unobserved signal of the seasonality in the overdispersion, there was a strong evidence that the national lockdown decreased the day-of-the-week effects of overdispersion signif- icantly. This article also shows strong evidence of a fast increase of overdispersion on Sundays and Tuesdays from September to 6 January 2021, in post-lockdown, as shown inFigure 5 and Figure 6. In order to decelerate this rapid increase in the number of people who tested positive with COVID-19, as detected by the fitted RW2 time trend and higheroverdispersion seasonality, the UK government announced a second national lockdown on 4 January 2021 and started a vaccination programme against COVID-19 from December2020 in order to bring the number of infections down.

References

1. White GC, Bennetts RE (1996) Analysis of frequency count data using the neg- ative binomial distribution. Ecology 67.

2. Zeger SL (1988). A regression model for time series of counts. Biometrika 75: 621-9.

3. Davis RA, Rodriguez-Yam G (2005) Estimation for state-space models: an approximate likelihood approach. Statist. Sinica 15: 381-406.

4. Benjamin MA, Rigby RA, stasinopoulos DM (2003) Generalized au- toregressive moving average models. J. Am. Statist. Assoc 98: 214-23

5. Davis RA, Wu R (2009) A negative binomial model for time series of counts. Biometrika 96: 735-49

6. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Super- spreading and the effect of individual variation on disease emergence. Nature 438: 355-9.

7. Houseman EA, Coull BA, Shine JP (2006) A nonstationary negative binomial time series with time-dependent covariates: Enterococcus counts in Boston Harbor. J. Am. Statist. Assoc 101: 1365-76

8. Anscombe FJ (1950) Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37: 358-82.

9. Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distribu- tions, 3nd edn. Wiley, New York.

10. Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans. Amer. Soc. Mech. Eng., J. Basic Engineering 82: 35-45.

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

12. Whittaker ET (1923) On a new method of graduation. Proceedings of the Edin- burgh Mathematical Society 41: 63-75.

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

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

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

16. 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

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

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

19. Harvey AC (1989) Forecasting, Structural Time Series Models and the Kalman Fil- ter. Cambridge University Press, UK.

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

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

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

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

24. Stasinopoulos DM, Rigby RA, Akantziliotou C (2008) Instructions on how to use the GAMLSS package in R, Second Edition, STORM Research Centre, Lon- don Metropolitan University, London.

25. ECDC (2020) Rapid increase of a SARS-CoV-2 variant with multiple spike protein mu- tations observed in the United Kingdom.

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

27. Henderson CR (1975) Best linear unbiased estimation and prediction under a se- lection model. Biometrics 31: 423-47.

28. Kitagawa G (1987) Non-Gaussian state space modeling of nonstationary time series, (with discussion). J. Amer. Statist. Assoc 82: 1032-63.

29. 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.

30. Shumway RH, Stoffer DS (2011) Time series analysis and its applications with R examples, 3rd edi, Springer.

31. WHO-Growth-Reference-Study-Group (2007) WHO Child Growth Standards: head circumference-for-age, arm circumference-for-age, triceps circumference-for-age and sub-scapular skinfor-for-age: methods and development. Geneva: World Health Organization.

32. Durbin J, Koopman SJ (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84: 669-84.