

## Research



**Cite this article:** Swallow B, Xiang W, Panovska-Griffiths J. 2022 Tracking the national and regional COVID-19 epidemic status in the UK using weighted principal component analysis. *Phil. Trans. R. Soc. A* **380**: 20210302.

<https://doi.org/10.1098/rsta.2021.0302>

Received: 1 October 2021

Accepted: 10 March 2022

One contribution of 18 to a theme issue 'Technical challenges of modelling real-life epidemics and examples of overcoming these'.

### Subject Areas:

statistics, bioinformatics, computational biology

### Keywords:

principal component analysis, COVID-19, multivariate statistics, dimension reduction, spatial epidemiology

### Author for correspondence:

Ben Swallow

e-mail: [ben.swallow@glasgow.ac.uk](mailto:ben.swallow@glasgow.ac.uk)

# Tracking the national and regional COVID-19 epidemic status in the UK using weighted principal component analysis

Ben Swallow<sup>1</sup>, Wen Xiang<sup>2</sup> and

Jasmina Panovska-Griffiths<sup>3,4</sup>

<sup>1</sup>School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK

<sup>2</sup>Department of Statistics, London School of Economics and Political Science, London WC2B 4RR, UK

<sup>3</sup>The Big Data Institute and the Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK

<sup>4</sup>The Queen's College, University of Oxford, Oxford OX1 4AW, UK

BS, 0000-0002-0227-2160; WX, 0000-0003-4077-4791; JP-G, 0000-0002-7720-1121

One of the difficulties in monitoring an ongoing pandemic is deciding on the metric that best describes its status when multiple intercorrelated measurements are available. Having a single measure, such as the effective reproduction number  $R$ , has been a simple and useful metric for tracking the epidemic and for imposing policy interventions to curb the increase when  $R > 1$ . While  $R$  is easy to interpret in a fully susceptible population, it is more difficult to interpret for a population with heterogeneous prior immunity, e.g. from vaccination and prior infection. We propose an additional metric for tracking the UK epidemic that can capture the different spatial scales. These are the principal scores from a weighted principal component analysis. In this paper, we have used the methodology across the four UK nations and across the first two epidemic waves

(January 2020–March 2021) to show that first principal score across nations and epidemic waves is a representative indicator of the state of the pandemic and is correlated with the trend in  $R$ . Hospitalizations are shown to be consistently representative; however, the precise dominant indicator, i.e. the principal loading(s) of the analysis, can vary geographically and across epidemic waves.

This article is part of the theme issue ‘Technical challenges of modelling real-life epidemics and examples of overcoming these’.

## 1. Introduction

Throughout the SARS-CoV-2 pandemic, there have been consistent requirements from modellers to ‘nowcast’ the pandemic, i.e. provide a quantitative assessment of the current state of the pandemic and to ‘forecast’ the epidemic, i.e. predict the future trajectory at varying spatial and temporal scales. There are inherent difficulties in modelling the current state of a pandemic, with challenges associated with data censoring, reporting lags, uncertainty in cause of death, asymptomatic cases and testing errors all contributing to possible biases in informed measures of incidence [1,2]. While several metrics can be used to provide a quantitative evaluation of the current state of an epidemic, the effective reproduction number,  $R$ , has been the most commonly and widely used metric (see Pellis *et al.* [3] for further discussion of this).  $R$  is a measure of the number of secondary infections stemming from a single infection and reflects the transmissibility or infectiousness of the viral variant during an epidemic. It allows tracking of the status of the epidemic with  $R < 1$ , suggesting that the epidemic is in the declining phase, whereas  $R > 1$  describes increased transmission and a growing epidemic. In the UK, it is derived as a consensus range from a number of different models and reported weekly by the UK Health Security Agency [4]. Since the onset of the pandemic in the UK, this  $R$  consensus range has been used to track the epidemic status and to inform and guide policy decision-makers in imposing and removing interventions such as imposing reduced social interactions intervention, i.e. lockdowns. However, while  $R$  is a useful measure, it is sensitive to, for example, the choice of data being used (e.g. the number of tests being carried out and any delays in reporting of cases) or the method for calculating it (e.g. the combination and the type of models used or the length of the time-slice being used for the calculation). It can also be computationally demanding to provide accurate estimates. Furthermore, once the population is partially or fully vaccinated defining population-wide  $R$  may not suffice. There are also difficulties associated with the fact that the onset of symptoms often happen a few days after the onset of infection, and hence, the current value of  $R$  really represents the state of the pandemic at some point in the recent past; in the UK, the consensus is that the current value of  $R$  is lagged by 2–3 weeks [4].

Other metrics such as the rate of daily hospital admissions have been suggested as alternative metrics that can be used alongside  $R$ , especially in a very heterogeneous population with mixed immunity from vaccination or from prior infection with a specific viral variant. In this article, we develop and apply a statistical method to derive a different metric related to the daily cases, hospitalizations, mechanical ventilation bed (MVB) admissions or deaths related to COVID-19 to track the national and regional epidemic status. For this purpose, we analyse the time series of these metrics to determine the important temporal, spatial and mechanistic dimensions of the data and highlight potential outliers in the time series.

We utilize multivariate projection methods, specifically methods of dimension reduction, aiming to find lower-dimensional representations of multiple (correlated) measurements to provide new measurement axes that are weighted as linear combinations of the original measurements. Principal components analysis (PCA) is one such method, which also has the added advantage of representing a geometric rotation of the data into orthogonal, or mutually independent, axes that are maximal in terms of variance retention. Given there would *a priori* be expected to be a high degree of correlation between the time series measuring varying dimensions

of the pandemic status, dimension reduction techniques would be expected to project the data matrix down into a much smaller number of uncorrelated bases. Larger numbers of infected cases generally lead to higher numbers of those requiring medical intervention or mortality rates. However, within a horizon of emerging variants that may be more transmissible but possibly less severe, this relationship may not be linear or consistent across measurements and settings.

There appears to have been little attempt to account for these inherent correlations between measurements thus far. The formal structures inherent in existing models are not always directly applicable to simple statistical approaches and can often be computationally complex when conducting inference. Methods that enable practitioners to obtain indicators of the current or recent state of the pandemic in a quick and efficient manner are therefore highly desirable.

We aim to do this by developing an approach similar to Xiang & Swallow [5] in analysing data from the UK COVID-19 dashboard [6] to study simpler representations of the multivariate output of cases, deaths, hospitalizations and MVB occupancy. We use S-Mode and T-mode PCA, with a temporal weight matrix calculated from median correlation in residuals following a generalized additive model fitted to a smooth of date index. We conduct the PCA on deaths, cases, hospitalizations and MVBs both at a UK level, and separately for the four UK nations. We also conduct additional analyses looking for differences between the dynamics of the first two principal COVID-19 waves. The analyses aim to explore whether a single epidemic metric or a combination of metrics can be a useful indicator of the status of the epidemic across nations and waves and hence be potentially useful to track in future alongside  $R$  and growth rate.

## 2. Methods

### (a) Principal components analysis

Multivariate projection and decomposition methods, such as PCA, enable the extraction of structures in multivariate data through an eigen-decomposition of the correlation or covariance matrix. The eigenvectors form a rotated basis of the data, or equivalently a new set of uncorrelated axes that are ordered by magnitude of their corresponding eigenvalues. This corresponds directly to the proportion of variation in the original measurements that they explain.

For a data matrix  $\mathbf{X}$  of dimension  $n \times p$ , a PCA can be conducted through a singular value decomposition (SVD) of the column mean-centred matrix. The SVD decomposes  $\mathbf{X}$  into

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where  $\mathbf{U}$  and  $\mathbf{V}$  are the column matrices of left and right singular vectors, respectively, and  $\mathbf{D}$  a diagonal matrix of singular values. The objective is to find a linear transformation  $\mathbf{Y} = \mathbf{U}^T\mathbf{X}$ , where  $\mathbf{U}^T = (U_{11}, \dots, U_{p1})^T$  is a matrix of constants such that the  $\mathbb{V}ar(\mathbf{Y}_1)$  is maximized, subject to the normalizing constraint  $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ . It can therefore be seen both as a variance-maximization projection of the covariance matrix or a linear transformation into an orthogonal set of bases. As the singular values are ordered by magnitude, or hence proportion of variation explained, for highly collinear data, lower order approximations can be produced by setting these singular values to zero.

Extensions to standard PCA relax the *a priori* assumption of unknown structure and allow users to account for existing spatial and/or temporal structures inherent in the data through the use of spatial and/or temporal weighting matrices [7]. Accounting for existent temporal structures in the data allows the extraction of important residual joint structures that can be more readily interpreted than if these known structures are not accounted for. In conventional PCA, it is common practice to standardize the data matrix by column standard deviations. Standardizing by a corresponding weight matrix ensures that the temporal structures are more comparable and are not dominated by measurements that are in larger units or are leading lags.

Alternative rotations, often referred to as S-Mode or T-Mode PCA of the data, can lead to different bases. Assuming the rows of the matrix correspond to the time points of the data and the columns to the individual time series, the S-Mode PCA aims to find dominant temporal trends

across the four data streams. Conversely, T-Mode PCA is conducted on the transpose of the matrix and aims to find different patterns across the time series and the associated time points at which they occur.

## (b) Flow-directed PCA

### (i) Weighted PCA

Adjusting PCA to account for known spatial and/or temporal structure in the data can support the extraction of novel trends in the data, as well as account for spatial variability in units across space or temporal lags in the time series. Assuming the data matrix is a  $n \times p$  matrix, a  $p \times p$  column weight matrix  $\Omega$  and  $n \times n$  row weight matrix  $\Phi$  can be constructed so that PCA is applied to a transformed matrix  $\tilde{X} = \Phi X \Omega$ .

In our analyses here, we only consider the temporal column matrix,  $\Omega$ , as the spatial aspect is less evident when considering measurements of different patient status; however, we describe the full process for completeness. Weighting or supervising the PCA by a temporal matrix should transform the data matrix towards a standard multivariate normal, so that the subsequent PCA selects eigenvectors with a high ratio of spatial or inter-measurement variability relative to temporal variation. This reduces the potential of overfitting by prioritizing variation between measurements, while still ordering the eigenvalues by maximal variance retention.

Let  $\tilde{X}$  denote the scaled data matrix, weighted by spatial and/or temporal weight matrices as follows.

$$\tilde{X} = \Phi X \Omega = \tilde{U} \tilde{D} \tilde{V}^T,$$

where the decomposed matrices generated on the scaled variables are denoted with a tilde. In this instance,  $\Phi$  is the spatial weight matrix and  $\Omega$  is the temporal weight matrix. As only a temporal matrix is used here, we remove dependence on  $\Phi$ , concentrating only on the temporal matrix  $\Omega$ . Hence, the principal scores (PCs) of the new weighted variables become  $X \Omega \tilde{V}$ , and the loadings are  $\Omega^{-T} \tilde{V}$ .

The temporal weight matrix is constructed similar to that by Gallacher *et al.* [8] using independent generalized additive models [9]. For measurement  $Y$ , we fit the model

$$\mathbb{E}[Y] = g^{-1} \left( \beta_0 + \sum_{k=1}^K f_k(x) \right),$$

where the  $f_k(\cdot)$  are smooth functions, often represented as splines and  $g(\cdot)$  is the link function mapping to the scale of the response. These models are fitted by restricted maximum likelihood in the mgcv package [10] to each of the time series with an intercept and an univariate smooth function of date as predictor, i.e. a univariate  $x$ . As in the study by Xiang & Swallow [5], this aims to remove trends specific to each stream, with residual variation used to determine the PCs. Remaining correlation in the model residuals between time  $[1, \dots, (n-1)]$  and  $[2, \dots, n]$  is calculated for each stream and then the median value is used as an estimate of  $\rho$ , the average global temporal correlation. The median residual temporal correlations  $\rho$  are estimated for each of the output time series (these will vary in each analysis depending on the spatial resolution/measurements used). The  $i$ th row and  $j$ th column element of the temporal weight matrix  $T_{i,j}$  is then specified as  $\rho^{|i-j|}$  for all time indices in the original data matrix. The weight matrix  $\Omega$  is then taken as the matrix square-root of  $T$ , i.e.  $\Omega = T^{1/2}$ . The temporal weight matrix is applied to the data matrix, and dimension reduction is then conducted to create a new uncorrelated set of bases. The method is identifiable up to a change in sign, so in some cases, similar trends are apparent, only inverted.

## (ii) S-mode and T-mode PCA

Next we describe how these methods can be applied to explore important global spatial and temporal trends in cases and deaths from COVID-19. To extract the important trends, PCA and similar dimension reduction techniques are an obvious choice. PCA conducts an eigen-decomposition of the covariance (or correlation) of a data matrix, with eigenvalues ordered by magnitude to reduce a set of  $p$  correlated variables to a smaller set of  $k < p$  orthogonal variables. Versions of PCA for spatio-temporal data were referred to by Richman [11] as S-mode and T-mode PCA, the particular mode depending on whether the columns of are time points (T-mode) or time series index (S-mode).

S-Mode PCA aims to find dominant temporal trends across the spatial locations, highlighting a small number of dominant temporal trends across all countries and/or time series. Conversely, T-Mode PCA aims to find different spatial patterns in the data and the associated time points at which they occur. In general, however, PCA finds unsupervised structures in the data by conducting an eigen-decomposition of the correlation of covariance matrix. While this can often be useful in visualizing data in lower dimensions, it is not possible to guide the structure of the new axes using prior information or independent data. To account for known spatio-temporal correlations inherent in the data, we use spatio-temporally weighted S-mode and T-mode PCA, which aim to find dominant temporal and spatial patterns, respectively. Gallacher *et al.* [8] extended these approaches to account for known spatio-temporal structures in river flow systems through the use of spatial and/or temporal weight matrices to inform spatio-temporal structure. Analyses were conducted using the *stpca* package in R.

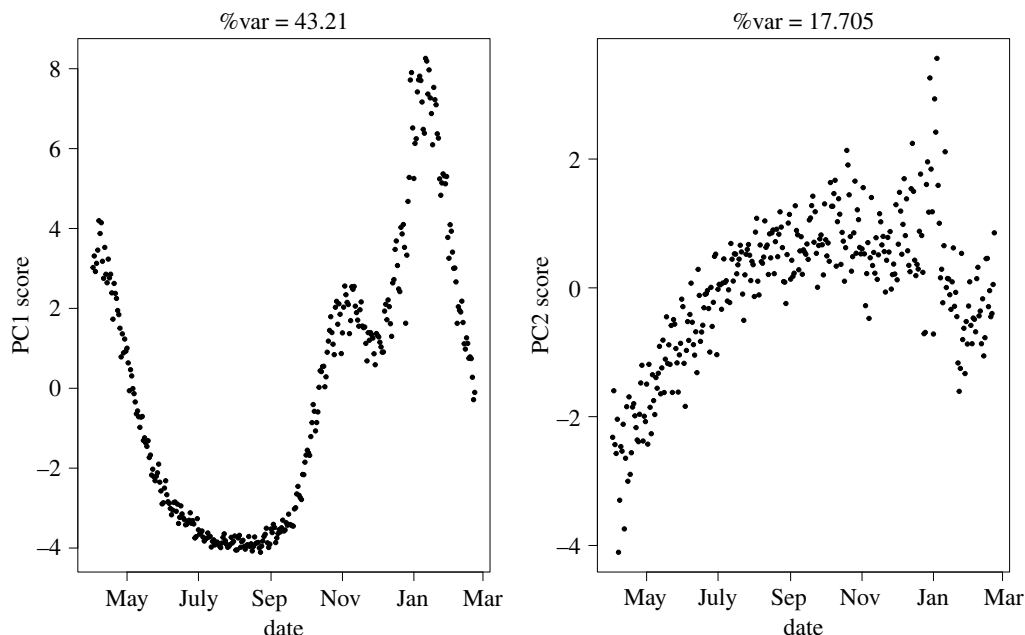
S-Mode PCA will be conducted with the columns of matrix in each formulation being the particular data stream and the rows corresponding to time points (day). For T-Mode PCA, the columns correspond to the time points and the rows are the data streams. In S-Mode, the output of interest will be a series of time-indexed points projected into the dimensions of principal variance, denoted the PCs. In T-mode, the scores will be a projection of the time points into a single score for each of the data streams. In both cases, the PCs will be a linear combination of the original higher-dimensional data with loadings representing the contribution of each of the original measurements to the corresponding score. It would be expected that these scores are centred around zero, with approximately symmetric variation around this. Any deviations from this would suggest outliers or patterns that warrant further consideration.

## (c) Data

Data used in this study are extracted from the UK COVID-19 Dashboard [6] and consist of daily measurements of reported cases, deaths, hospitalizations and MVBs occupied (representing hospital occupancy where ventilation support was required) from 2 April 2020 to 22 February 2021 (327 days of 16 measurements). This time period spans the first two waves of the pandemic in the UK and terminates at the point that the vaccination programme was rolled out on a wide scale. Numbers were available at the level of the four individual UK nations, as well as aggregated across the UK as a whole. The nationally segregated data were used for all analyses, with some analyses run both jointly across all nations and independently for each nation, to determine differences in results depending on spatial scale. Data were checked for inconsistencies and outliers that may have impacted results, although none were found.

For additional analyses, the data were also stratified into nation-specific data matrices and matrices corresponding to the two individual waves, namely, March to May 2020 (59 days) representing the first epidemic wave and September 2020 to April 2021 (175 days) representing the second epidemic wave (as per ONS, [12]), were also constructed. All matrices are column mean-centred before analysis to ensure easier comparison between dimensions.

The data and computer code used to run the analyses and generate figures reported in this article are openly available at <https://doi.org/10.5281/zenodo.6078749>.



**Figure 1.** First two PC scores for temporally weighted S-Mode PCA for pooled UK data, with the corresponding proportion of variance explained above each. Data correspond to April 2020 to March 2021.

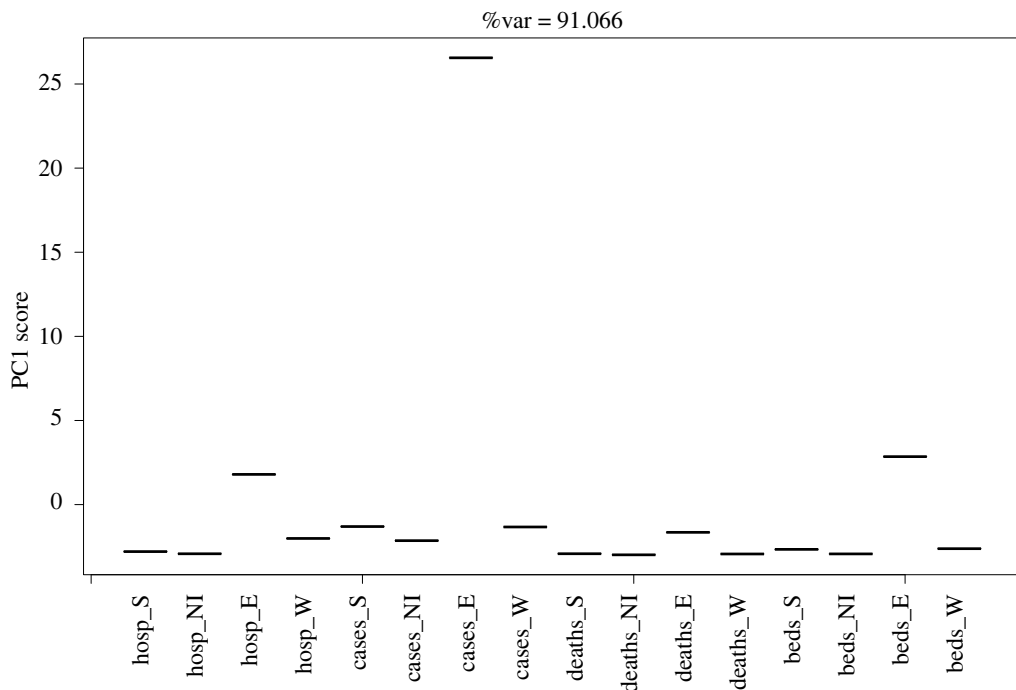
### 3. Results

#### (a) Combined UK analysis

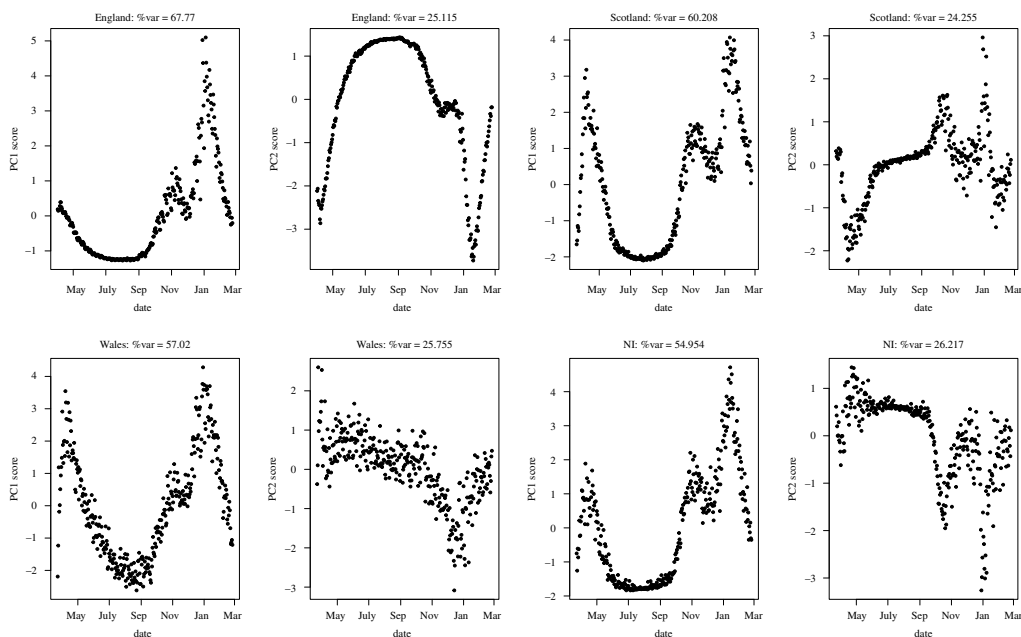
As a result of the PCA analysis, we derived the first and second PCs. The initial results obtained at the UK-wide scale suggest that a single combined index of the four different measurements from each of the four nations is able to explain approximately 42% of variation, with the second PC accounting for a further 18% (figure 1), with loadings being roughly equal. The observed trend in PC1 is incredibly smooth, with an initial peak very early in the studied period before a steep decline in the state of the pandemic towards the summer months. The overall state then worsened from early September through to early November, when a lockdown was implemented. The release of restrictions over the Christmas period from late December leads to a drastic increase in the overall observed trend until early January, when further interventions were introduced as well as the initialization of the vaccination roll-out. Following this point, there is a steep linear reduction in the observed overall state to the end of the modelled period in March.

In addition, there is a further trend observed in the second PC, which shows a general increase in the pandemic state from April through to September, after which it remains relatively stable. Given the relatively smooth nature of this second score and its asymptotic behaviour, this would suggest a more subtle and smooth change in the observed behaviour. The change in testing capacity over this time frame increased significantly, particularly up until the autumn/winter period. The Spearman correlation coefficient was calculated as 0.36 between the second score and the reported number of new tests [6], giving some support to the idea this second score could be associated with an underlying change in the data collection, rather than necessarily a change in the pandemic.

Figure 2 shows the first PC for the T-mode analysis. The aim of this approach is to detect time series that deviate significantly from each other. In this analysis, only a single PC is required to explain over 90% of the variation in the data, suggesting that further PCs are not informative. Following temporal standardization, it would be expected that these follow approximately a



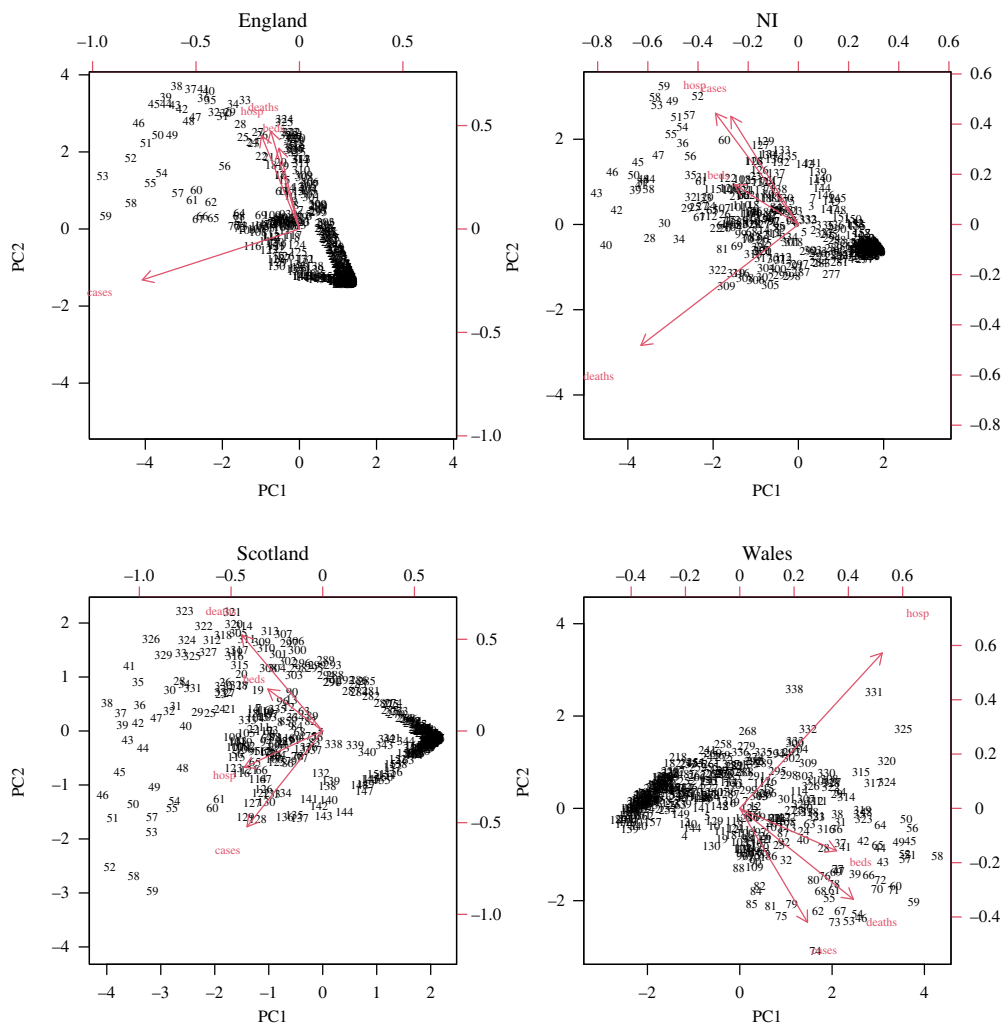
**Figure 2.** First PC for temporally weighted T-Mode PCA in a combined UK analysis. The approach highlights measures that deviate markedly from the others/average.



**Figure 3.** First two PCs for temporally weighted S-Mode PCA across the different UK nations.

multivariate normal distribution. Most of the scores lie around zero, with some small deviations for hospitalizations and MVBs in England. What is particularly noteworthy, however, is the significant deviation for cases away from the central trend, suggesting that case numbers in





**Figure 4.** Biplots for the temporally weighted S-Mode PCA across the different UK nations. Length of the arrows corresponds to the loading contribution of that measure to each of the first two PCs. (Online version in colour.)

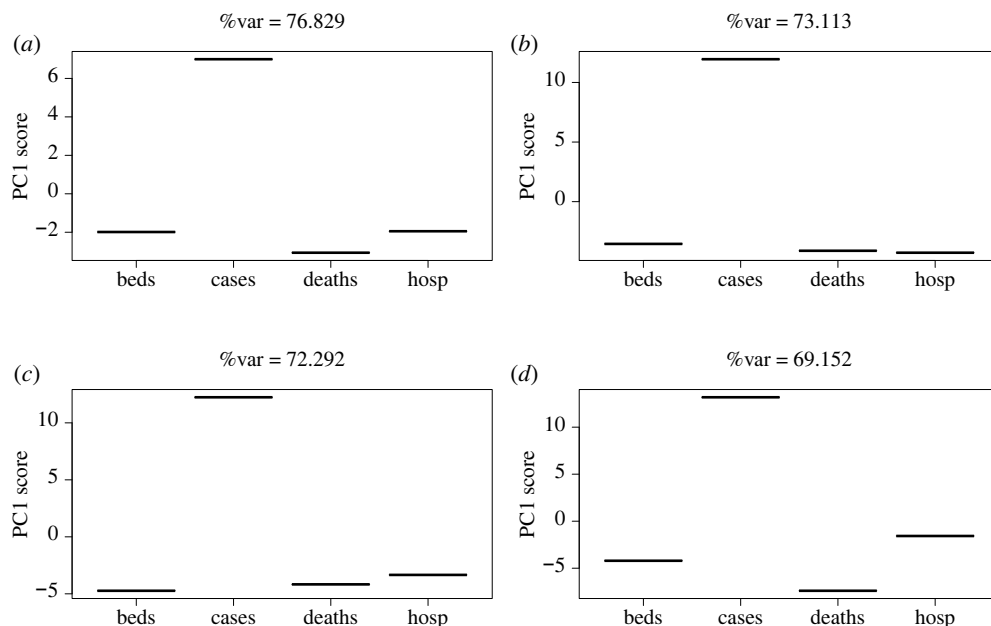
England, and to a lesser extent hospitalization and MVBs, deviated markedly away from trends in the other nations and other time series.

### (b) Nation-specific analysis

For the nation-specific analyses, the first S-mode PC was able to explain between 65% and 70% of the total variation in the data, with around 25% explained by PC2. This suggests that there is a single dominant trend, with a second less-dominant independent trajectory occurring (figure 3). The corresponding loadings show that the dominant contributor to the principal axis varies across nations (figure 4). For England, it was cases; Northern Ireland deaths; Scotland showed roughly equal contributions; and Wales hospitalizations.

All analyses showed peaks in April 2020 and January 2021, with a smaller peak in November 2021 that showed initial signs of reduction before increasing again to the major peak in January 2021. The lockdown that was introduced in November 2020 seems to have been particularly beneficial in Scotland, which showed a reduction down to similar low levels as over the summer months in 2020. The subsequent significant growth up to January 2021 follows temporary





**Figure 5.** First PC for temporally weighted T-Mode PCA—(a) England, (b) Northern Ireland, (c) Scotland and (d) Wales.

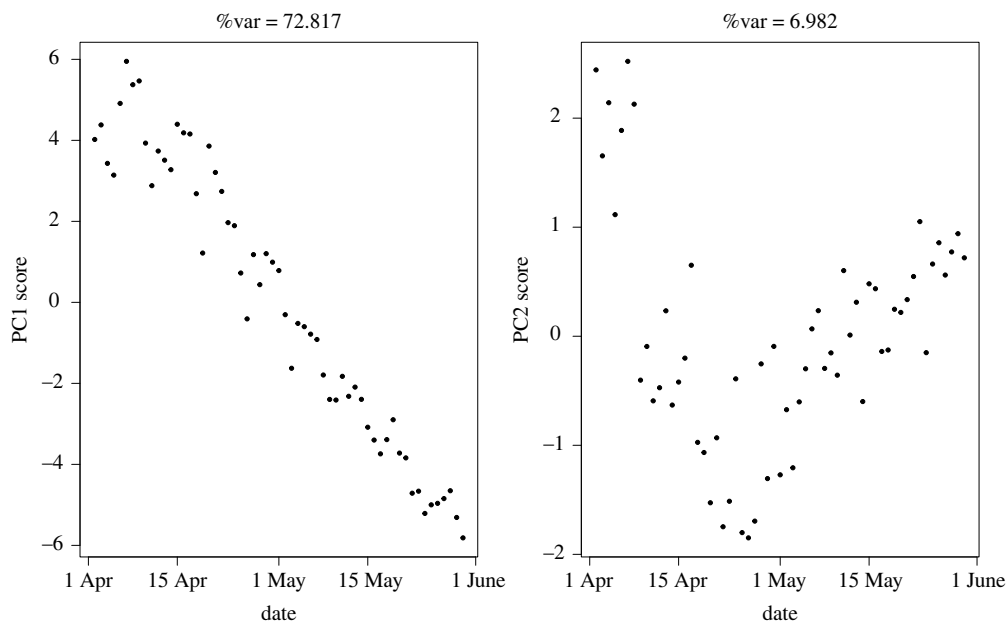
Christmas reductions in distancing measures and may be a result of this. A combined impact of strict intervention measures after the Christmas holidays and in the New Year, combined with the rolling out of the vaccination programme for susceptible individuals, led to a reduction across all trends, and this is highlighted in the first PC.

While Scotland, Wales and Northern Ireland show a short sharp peak in April 2020 that drops sharply down to September, the first PC in England showed a much shallower change over the initial months in 2020 before similar dynamics over the winter 2020 to the other nations. The change in the early months observed in the other three nations is relegated to PC2 in England, suggesting that overall changes were less dramatic than in other areas of the UK. This pattern in PC2 for England relative to the other nations suggests that the earlier wave was distinctly different from the first wave.

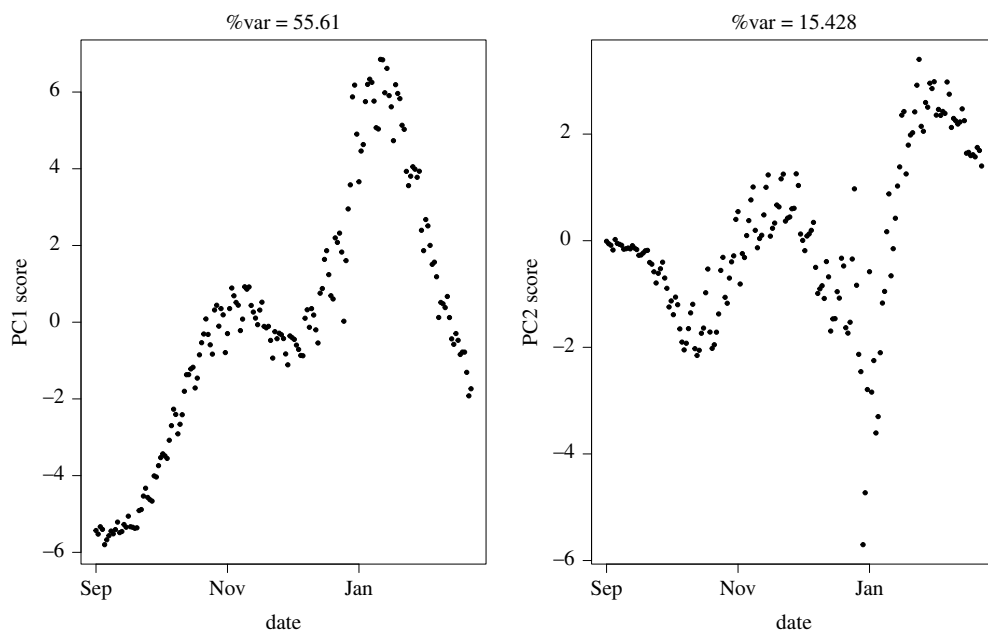
The T-mode analysis by nation (figure 5) suggests that hospitalizations are consistently close to zero when projected into the new axes. Cases vary markedly from this, suggesting that they are very different and therefore caution should be taken when considering them as representative of the epidemic state.

### (c) Comparison of waves

We also rerun the unified S-Mode analysis to compare dynamics of trends separately for the two principal waves of COVID-19 over the studied period (figures 6 and 7). We denote wave 1 to consist of data between 20 January 2020 and 31 May 2020 and wave 2 to correspond to data from 1 September 2020 to the end of March 2021 [12]. The results were consistent with the full analysis in terms of which variables dominated the first principal component; however, there were clearly different general temporal trends in each of these periods. Particularly noteworthy is the principal linearly increasing trend seen in PC1 for wave 1 dynamics. PC2 for this analysis explains little of the variance, so is not worth focusing on. Wave 2 shows more variation through time, including an interrupted linear growth that temporarily decreases at the time of a national lockdown in November. A similar linear reduction is seen at the end of wave 2. Both of these time periods correspond to strict lockdown periods in the UK.

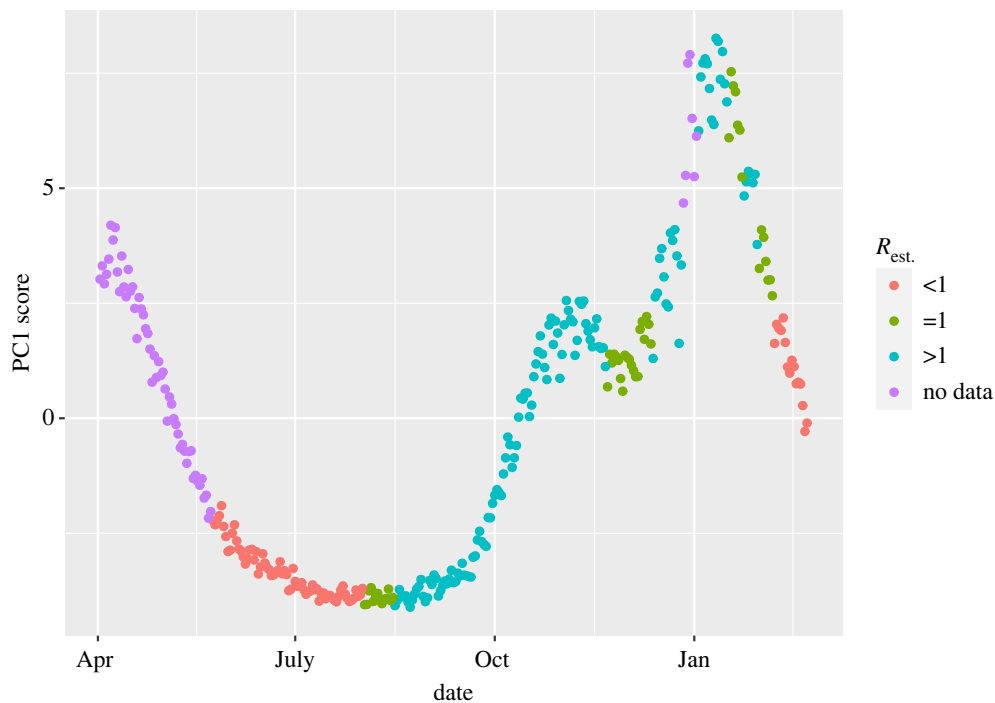


**Figure 6.** First two PCs for temporally weighted S-Mode PCA covering dates in COVID-19 Wave 1, April to June 2020.

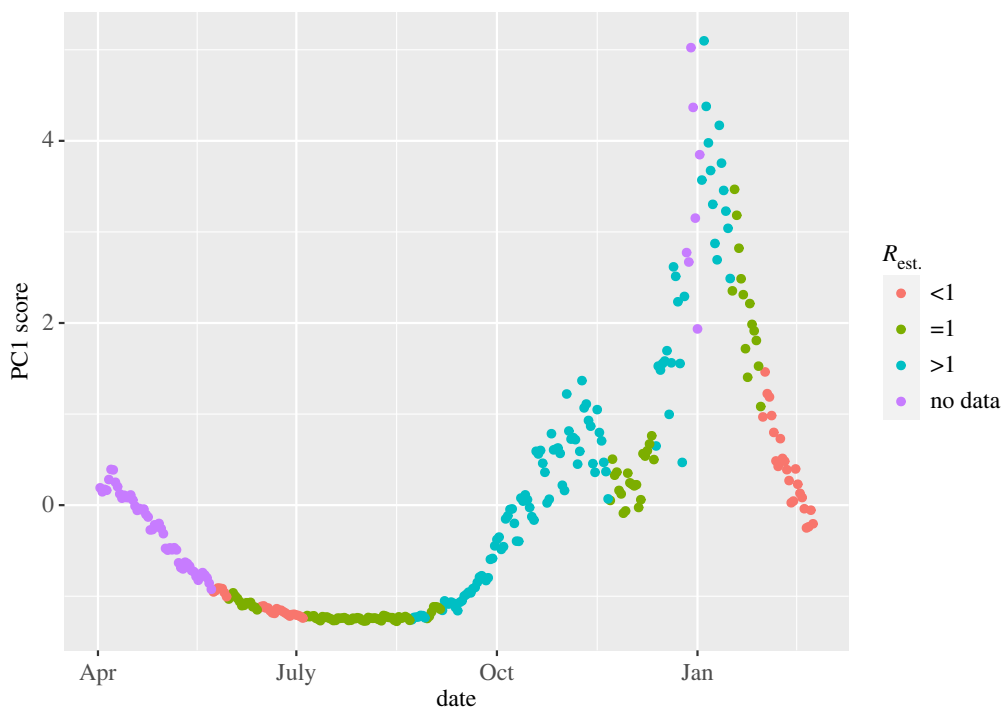


**Figure 7.** First two PCs for temporally weighted S-Mode PCA covering dates in COVID-19 Wave 2 spanning September 2020 to March 2021.

The loadings between the two analyses, which represent the contribution of an individual measurement to each linear transformation, showed some interesting trends. Specifically, in wave 2, the loadings decreased for hospitalizations in all nations; deaths in England and cases in Wales and Northern Ireland. All other measurements were more dominant in wave 1.



**Figure 8.** First PC from figure 1 coloured by strata of the upper bound of the estimated UK reproductive number upper bound from [4] in the same week. Red is  $R < 1$ ; green is  $R = 1$  and teal is  $R > 1$ . Purple corresponds to weeks in which no estimate was provided. (Online version in colour.)



**Figure 9.** First PC from S-Mode PCA in England coloured by strata of the upper bound of the estimated UK reproductive number upper bound from [4] in the same week. Red is  $R < 1$ ; green is  $R = 1$  and teal is  $R > 1$ . Purple corresponds to weeks in which no estimate was provided. (Online version in colour.)

### (d) Relationship with $R$

Finally, we compare the results from the full S-mode analysis in figure 1 and the England-specific analysis with the corresponding estimates of UK-wide reproductive number from [13]. Comparable data from the other nations were not publicly available. Both upper and lower bounds of the reproduction number  $R$  were available, generated from an ensemble of epidemiological models with the upper bound being used in figure 1. During periods of increased  $R$ , there is very good agreement in the trend between the time series of  $R$  and the corresponding observed increase in the first PC (figures 8 and 9). Similarly, when  $R$  drops and remains below one, we also observe a inflection in the first PC. This agreement suggests that the PC is strongly correlated with  $R$  (Spearman's coefficient  $\rho = 0.84$  and  $\rho = 0.72$  for the UK and England, respectively), and hence, PC may feasibly provide a viable additional quantity for fast nowcasting of the pandemic.

## 4. Conclusion

Overall across nations, our results suggest that hospitalizations are consistently important as a representative indicator of the state of the pandemic at any given time point, being highlighted as outside the standard linear combinations of the other variables in most analyses. This matches the fact that these are likely to be relatively unbiased compared with cases [14] and less age- or health status-specific than MVBs. MVBs are also expected to be limited by capacity, unlike the overall hospital admissions, the former of which policy has changed significantly over the course of the pandemic [15]. The combination of results from S-mode and T-mode analyses at different scales allow for a greater understanding of the pandemic at different levels.

Elsewhere in this issue, alternative approaches have been proposed for calculation of  $R$  from cases only [16,17]; however, our approach utilizes all available data streams and is also able to find additional second-order structures in the data.

In some analyses, the epidemic trends can also be driven by the number of reported cases, notably in Scotland, where non-pharmaceutical interventions (NPIs), particularly in the populated central belt incorporating Glasgow and Edinburgh, were maintained much longer over the study period so that the dynamics would be expected to be different guided by these NPIs' impact. For example, Greater Glasgow and neighbouring authorities remained under varying levels of restrictions between early September 2020 and May 2021, which were particularly severe between November and December. However, reported numbers of cases can be highly variable and subject to possible underreporting and are strongly dependent on testing, contact tracing and isolation behaviour and policies.

For the nation-specific analyses, variation was observed in the dominant contributor to the principal axis, suggesting that simple averages of different measures are not appropriate. Numbers of hospitalized patients appear to be consistently highlighted as outliers in the data, however. This is aligned with hospitalizations and especially occupancy levels, being a trend that, together with  $R$ , has been important in driving policy decisions over the first and second epidemic waves, i.e. over the study period here.

Deaths and MVBs as a proxy for hospital occupancy have also been used as epidemic indicators by policy decision-makers. In our analysis, however, they do not appear to be a significant indicator of the epidemic status. We note that this could be due to possible bias in these as reported death figures have been challenging during the pandemic due to delays in reporting and uncertainty over the exact cause of death. For instance, there have been several different figures adopted, e.g. mortality within 28 days of a positive test, deaths with COVID-19 on the death certificate and deaths with rather than from COVID-19. All of these metrics represent deaths related to COVID-19, but all have aspects of the data that can cause biases and give uncertainty in reporting deaths related to COVID-19. The number of MVBs, while perhaps slightly less biased than the reported numbers of cases and deaths, is likely to be susceptible to changes in policy and is severely limited by the carrying capacity at each hospital and movement of patients in and

out of these wards depending on need. Hospitalization figures indicating admission to hospital with COVID-19 and are less susceptible to either of these biases [14]. Hence, our suggestion that overall hospitalizations may be a good epidemic indicator and an additional potential metric to track alongside  $R$  and growth rate is sensible.

The good agreement between our approach and independent estimates of the reproduction number in England gives support to our approach as a complementary fast, real-time method for determining the state of the pandemic from multivariate noisy data. As we had mentioned previously, tracking  $R$  in a highly heterogeneous population with an imprinted prior immunity from vaccination and prior infection with different variants is very different to  $R$  in a unvaccinated population or fully susceptible one at the onset of the epidemic or after immunity has fully waned. Furthermore, the current tracking of the epidemic status via the nowcasting process by the modelling team at the UKHSA is reliant on a combination of mathematical models that are calibrated to the vast amount of currently available data from the UK COVID-19 dashboard [4]. Looking towards a situation where less data may be produced daily, e.g. when potentially only hospitalizations may be tracked rather than also cases and deaths related to COVID-19, an alternative method to assess epidemic status such as the one proposed here will be of great value.

Via our step-by-step statistical analysis, we highlight the care that must be taken when choosing the level of aggregation of data for statistical analysis [18,19]. For example, our results highlight that aggregating the data to UK level changes the dynamics of the data streams and masks important local dynamics.

There were limited differences in general trends between the first and second wave of COVID-19 across the UK, with both waves showing linear reductions post-lockdown interventions. MVBs generally became less dominant while hospitalizations emerged as more dominant in the second wave. Deaths in England were also more dominant in wave 1. This could relate to a change in policy of only intervening in the most serious cases, or an improvement in outcomes due to an improved understanding of the disease and the impact from the large-scale vaccination programme from December 2020.

The fact that most analyses of spatial trends show two major principal components, while temporal analyses tend to show only one dominant trend in space, is worth highlighting. The principal spatial trend will generally show the dominant trend over larger regions, while the second spatial trend could relate to an increase in testing capacity over the time period, showing an overall increase in numbers detected over the whole, whereas early in the emergence of SARS-CoV-2, only those with severe symptoms would likely be detected. It could also highlight second-order structure in the data, corresponding to higher variability in times of policy changes (e.g. lockdowns, reopening of schools, reopening of public services) or local variations in trends not picked up by the main principal component.

One of the drawbacks of conventional PCA is that as an unsupervised method, it is often challenging to interpret the often abstract outputs that it produces. Also, in many scenarios, known structures in the data can be readily available and are not themselves of interest. One of the interests in these scenarios is finding latent structures in the data that explain hidden correlations between measured quantities. The study of these structures can then inform further on possible important mechanisms within the system under study.

In the study by Xiang & Swallow [5], the authors analysed data from global trends relating to reported cases of, and deaths resulting from, SARS-CoV-2. Their analyses found that a single temporal trend dominated the global spread of the disease. This suggests that there are multi-scale processes occurring, namely, a general spread across countries that happens smoothly and consistently, and more local dynamics within a country that occur at a finer resolution. This again highlights the importance of considering appropriate scales when looking at dynamics of infectious diseases (e.g. Garabed *et al.* [20]). Aggregation of data may be beneficial computationally or in terms of reducing the impact of biases and errors in the data; however, it will inevitably change results.

Removing common temporal correlation from the data through a temporal weight matrix reduced the amount of variation explained in all analyses, but had little impact on the overall

conclusions. This supports the idea that there is a great deal of similarity across nations within the UK in terms of seasonal dynamics and principal trends; however it is not sufficient to treat the streams as either independent or entirely analogous.

There appear to have been few attempts to look at the structures inherent in multiple measurements of the pandemic and how to reduce these down to single measures, other than the  $R$  value widely applied. Rahman *et al.* [21] constructed a structural equation model, in which dimension reduction is included, and a variable relating to ‘pandemic severity’ is generated as one of their dimensions of interest. However, this is not the main focus of their research, and they only include cases and deaths in their analysis. Their analyses support the idea that treating the measurements as independent is clearly erroneous, as they clearly have large amounts of correlation. Also, errors and biases in each stream will be compounded if independent models are fitted. Using dimension-reduction techniques, such as directed PCA, allows for the correlation to be incorporated into the methodology and reduces the impact of any single stream bias by integrating over them all.

In summary, in this study, we have used an established statistical technique to generate an alternative, yet complementary, indicator of the epidemic status to  $R$  and growth rate. We have used the PCA methodology described here to show that overall, across nations and epidemic waves, the level of hospitalizations with COVID-19 is a good indicator of the epidemic status. However, the precise best indicator, i.e. the PC(s) of the PCA, vary geographically and across epidemic waves.

**Data accessibility.** Data and code to conduct the analyses and generate the figures in this manuscript can be found at <https://doi.org/10.5281/zenodo.6078749>. The code and data are provided openly and there are no restrictions to their access.

**Authors' contributions.** B.S.: conceptualization, data curation, formal analysis, methodology, project administration, software, supervision, validation, visualization and writing—original draft; W.X.: data curation, investigation and software; J.P.: investigation, methodology, validation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

**Conflict of interest declaration.** J.P.G.'s work was supported by funding from the UK Health Security Agency and the UK Department of Health and Social Care. This funder had no role in the study design, data analysis, data interpretation, or writing of the report. The views expressed in this article are those of the authors and not necessarily those of the UK Health Security Agency or the UK Department of Health and Social Care.

**Funding.** J.P.G.'s work was supported by funding from the UK Health Security Agency and the UK Department of Health and Social Care (DHSC).

**Acknowledgements.** B.S. is a member of the Scottish COVID-19 Response Consortium, which was undertaken in part as a contribution to the Rapid Assistance in Modelling the Pandemic (RAMP) initiative, coordinated by the Royal Society.

## References

1. Shadbolt N, Brett A, Chen M, Mario G, McKendrick IJ, Panovska-Griffiths J, Pellis L, Reeve R, Swallow B. 2021 The challenges of data in future pandemics. See <https://www.newton.ac.uk/documents/preprints/>.
2. Swallow B *et al.* 2022 Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. *Epidemics* **38**, 100547. (doi:10.1016/j.epidem.2022.100547)
3. Pellis L *et al.*, JUNIPER Consortium. 2021 Estimation of reproduction numbers in real time: conceptual and statistical challenges. See <https://rss.org.uk/RSS/media/File-library/News/2021/PellisBirrel.pdf>.
4. UKHSA. 2020 The  $R$  value and growth rate. See <https://www.gov.uk/guidance/the-r-value-and-growth-rate>.
5. Xiang W, Swallow B. 2021 Multivariate spatio-temporal analysis of the global covid-19 pandemic. *medRxiv* (doi:10.1101/2021.02.08.21251339)
6. Coronavirus (COVID-19) in the UK. 2021. See <https://coronavirus.data.gov.uk/details/download>.

7. Baldwin MP, Stephenson DB, Jolliffe IT. 2009 Spatial weighting and iterative projection methods for EOFs. *J. Clim.* **22**, 234–243. (doi:10.1175/2008JCLI2147.1)
8. Gallacher K, Miller C, Scott EM, Willows R, Pope L, Douglass J. 2017 Flow-directed PCA for monitoring networks. *Environmetrics* **28**, e2434. (doi:10.1002/env.2434)
9. Wood SN. 2017 *Generalized additive models: an introduction with R*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC Texts in Statistical Science.
10. Wood SN. 2012 MgcV: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. See <https://CRAN.R-project.org/package=mgcv>.
11. Richman MB. 1986 Rotation of principal components. *J. Climatol.* **6**, 293–335. (doi:10.1002/joc.3370060305)
12. ONS. 2021 Coronavirus (COVID-19) infection survey technical article: waves and lags of COVID-19 in England, June 2021.
13. UKHSA. 2020 Reproduction number (*R*) and growth rate: methodology. See <https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology>.
14. Vekaria B *et al.* 2021 Hospital length of stay for covid-19 patients: data-driven methods for forward planning. *BMC Infect. Dis.* **21**, 700. (doi:10.1186/s12879-021-06371-6)
15. NICE. 2021 COVID-19 rapid guideline: managing COVID-19. See <https://www.nice.org.uk/guidance/ng191>.
16. Ackland G, Ackland J, Antonioletti M, Wallace D. 2022 Fitting the reproduction number from UK coronavirus case data and why it is close to 1. *PTRSA* **380**, 20210301. (doi:10.1098/rsta.2021.0301)
17. Li W, Buliekova K, Gregor B, White L, Kolaczyk E. 2022 Estimation of local time-varying reproduction numbers in noisy surveillance data. *PTRSA* **380**, 20210303. (doi:10.1098/rsta.2021.0303)
18. Jeffery C, Ozonoff A, Pagano M. 2014 The effect of spatial aggregation on performance when mapping a risk of disease. *Int. J. Health Geogr.* **13**, 9. (doi:10.1186/1476-072X-13-9)
19. Jeffery C, Ozonoff A, White LF, Nuño M, Pagano M. 2009 Power to detect spatial disturbances under different levels of geographic aggregation. *J. Am. Med. Inform. Assoc.* **16**, 847–854. (doi:10.1197/jamia.M2788)
20. Garabed RB, Jolles A, Garira W, Lanzas C, Gutierrez J, Rempala G. 2020 Multi-scale dynamics of infectious diseases. *Interface Focus* **10**, 20190118. (doi:10.1098/rsfs.2019.0118)
21. Rahman MM, Thill J-C, Paul KC. 2020 COVID-19 pandemic severity, lockdown regimes, and people's mobility: early evidence from 88 countries. *Sustainability* **12**, 9101. (doi:10.3390/su12219101)