scientific reports

Check for updates

OPEN

A global perspective on the intrinsic dimensionality of COVID-19 data

Abhishek Varghese[1,2], Edgar Santos-Fernandez[1,2]✉, Francesco Denti[3], Antonietta Mira[4,5]✉ & Kerrie Mengersen[1,2]

We develop a novel global perspective of the complexity of the relationships between three COVID-19 datasets, the standardised per-capita growth rate of COVID-19 cases and deaths, and the Oxford Coronavirus Government Response Tracker COVID-19 Stringency Index (CSI) which is a measure describing a country's stringency of lockdown policies. We use a state-of-the-art heterogeneous intrinsic dimension estimator implemented as a Bayesian mixture model, called Hidalgo. Our findings suggest that these highly popular COVID-19 statistics may project onto two low-dimensional manifolds without significant information loss, suggesting that COVID-19 data dynamics are generated from a latent mechanism characterised by a few important variables. The low dimensionality imply a strong dependency among the standardised growth rates of cases and deaths per capita and the CSI for countries over 2020–2021. Importantly, we identify spatial autocorrelation in the intrinsic dimension distribution worldwide. The results show how high-income countries are more prone to lie on low-dimensional manifolds, likely arising from aging populations, comorbidities, and increased per capita mortality burden from COVID-19. Finally, the temporal stratification of the dataset allows the examination of the intrinsic dimension at a more granular level throughout the pandemic.

High-dimensional datasets are generally challenging for statistical inference and data analysis. Their analysis is made even more challenging by longitudinal measurements and temporal autocorrelation. Fortunately, these kinds of data have a high degree of redundancy typically and, therefore, may project onto low-dimensional manifolds without losing substantive information[1,2]. The dimensionality of these manifolds is called the *intrinsic dimension* (ID) of the data, and it can provide important information about the properties of datasets.

Data science methods for high-dimensionality datasets have been utilised and explored in multiple contexts to aid decision-making and analysis during the COVID-19 pandemic. For example, citywide smart card travel data has been utilised in Sydney, Australia, to cluster passenger types along multiple mobility dimensions and develop intervention strategies for disease spread[3]. Similarly, manifold learning techniques have been applied to cell-phone mobility data in the United States during the COVID-19 pandemic, distinguishing mobility trends in multiple geographic regions and demographics[4]. Others have leveraged dimensionality reduction techniques to cluster and analyse highly dimensional genome sequence data of COVID-19[5] and identified essential features in predicting the mode of instruction in American universities during the COVID-19 pandemic[6].

Additionally, these statistical techniques have enabled decision-makers to parse the large body of communication transmitted online during the COVID-19 pandemic to glean new insights. Uniform manifold approximation and projection and latent Dirichlet allocation have been used to parse Twitter data during the COVID-19 pandemic and distinguish topics, identify trends and patterns in social network behaviours[7,8]. In another example, Doanvo et al.[9] utilise similar techniques to analyse a large body of open access COVID-19 research studies and classify research output to identify existing knowledge gaps in research.

However, there has been little work done to explore the latent dynamics of the pandemic spread across continents and countries. Sivakumar and Deepthi[10] examined the temporal dynamics of COVID-19 daily cases and deaths in 40 countries, using a False Nearest Neighbour method to identify the relevant embedding dimension (ED) for each country. The authors recognise that new COVID-19 cases and deaths exhibit a low- to medium-level ED. However, it is essential to note that ED does not account for points in a dataset lying on low-dimensional manifolds. Thus, identifying the ID is generally more valuable as it accounts for inherent structures in the data and remains a more accurate representation of underlying structural complexity in a dataset[11,12]. This research

[1]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia. [2]Centre for Data Science (CDS), Queensland University of Technology (QUT), Brisbane, Australia. [3]Department of Statistics, Università Cattolica del Sacro Cuore, Milan, Italy. [4]Data Science Lab, Università della Svizzera italiana, Lugano, Switzerland. [5]Department of Science and High Technology, Università degli Studi dell'Insubria, Como, Italy. ✉email: santosfe@qut.edu.au; antonietta.mira@usi.ch

work will seek to bridge an important gap and provide valuable information towards understanding the complexity and dimensionality of the COVID-19 pandemic in different countries, and develop a deeper understanding of the spread of the pandemic.

This paper provides an application of the recent heterogeneous ID algorithm (Hidalgo). Hidalgo is a Bayesian mixture model capable of clustering the observations into groups characterised by similar IDs. The ID can be considered an indicator of the complexity of the data: the higher its value, the larger the number of relevant directions are required to represent the data points faithfully. More information about ID may be found in the next section, and more formal definitions of ID can be found elsewhere[2,13].

The vast majority of statistical methods assume and estimate a unique value for the ID. However, this assumption is often too strong for datasets containing information generated by intricated systems with complex dynamics, such as a global pandemic. Hidalgo extends this framework, allowing the presence of multiple manifolds characterised by different ID values in the same dataset. The Bayesian local ID estimator has been applied successfully to a diverse range of datasets for scenarios such as financial markets, neuroimaging, proteomics[14], genomics[15], and high-resolution player tracking data[11]. Here, we seek to organise the pandemic dynamics of different countries into groups with similar ID to help us unveil complex patterns related to the dynamics of the COVID-19 pandemic. Finally, we temporally stratify the dataset to examine the ID at a more granular level throughout the COVID-19 pandemic.

Methods

Likelihood-based intrinsic dimension estimation. A large number of ID estimators are currently available in the literature. Likelihood-based estimators are particularly appealing because of their theoretical foundation and the immediate ability to provide estimates for uncertainty quantification.

Recently, building on previous work[1,16], the 'Two Nearest Neighbours' (TWO-NN) estimator was introduced[17], based on the following distributional result. Assume we observe n units in a dataset of nominal dimension D (intuitively, the number of recorded columns in a tall dataset), where the data lie on a manifold of smaller dimension d, the ID. In other words, some dimensions are irrelevant, or there may be a functional relationship between two or more coordinates. From a modeling perspective, we consider the dataset as a configuration obtained from a Poisson point process over $\mathbb{R}^D$ characterised by a homogeneous intensity function ρ. In that case, one can prove that the ratio of the distances between a given point and its second and first nearest neighbours (NN) is Pareto distributed with shape parameter d and scale parameter identically equal to 1. Algebraically, denoting with $r_{i,j}$ the distance between the i-th point and its j-th NN, we have:

$$\mu_i = \frac{r_{i,2}}{r_{i,1}} \sim Pareto(1, d), \qquad \mu_i \in (1, +\infty), \qquad i = 1, \ldots, n. \tag{1}$$

Although the theoretical derivation requires a uniform intensity of the point process, the result in Eq. (1) is empirically valid as long as the homogeneity assumption holds up to the second NN for every point.

As previously mentioned, methods that return a unique ID value to describe the entire dataset can often be limiting and unrealistic since data may lie on multiple latent ID manifolds. This shortcoming has been addressed[14] by partitioning the data in subgroups characterised by locally homogenous ID via a Bayesian mixture model[18]. We now suppose that the ratios μ_i, for $i = 1, \ldots, n$, are potentially generated from L different Pareto distributions, obtaining:

$$f(\mu_i | \boldsymbol{d}, \boldsymbol{\pi}) = \sum_{l=1}^{L} \pi_l \, d_l \mu_i^{-(d_l+1)}, \qquad i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$ is the vector of mixture weights and $\boldsymbol{d} = (d_1, \ldots, d_L)$ is a vector containing the ID parameters. The Bayesian model is completed with prior distribution specifications. In particular, the authors chose identically distributed and independent conjugate Gamma priors for each element of $\boldsymbol{d}$, with shape and rate parameters $a_d > 0$ and $b_d > 0$, respectively: $d_l \sim Gamma(a_d, b_d) \ \forall \ l$. Moreover, a Dirichlet prior for the mixture weights is adopted: $\boldsymbol{\pi} \sim Dirichlet(\alpha_1, \ldots, \alpha_L)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$ is a vector of positive concentration parameters.

Here, we adopt a sparse mixture specification[19,20] which permits, similarly to a nonparametric approach, a distinction between the number of fitted components L and the number of estimated clusters, L^*, which coincide with the populated components. To this aim, a careful choice of the vector $\boldsymbol{\alpha}$ (e.g., by setting all its entries to small values, say $\alpha_l \leq 0.05, \ \forall l$) allows the method to automatically select the necessary number of mixture components $L^* \leq L$, preventing the need to fit multiple models with different values of L and then rely on post-hoc solutions, such as the comparison of information criteria (e.g., AIC, BIC) or marginal likelihood to select the best model. Indeed, within this context, the value L in Eq. (2) is interpreted as an upper bound on the number of populated clusters, and the actual number of manifolds is directly estimated by the data.

As customary in Bayesian mixture models, we can augment the parameter space to enhance inference and ease posterior computation adding the auxiliary parameters $c_i \in \{1, \ldots, L\}$, for $i = 1, \ldots, n$.

These latent membership labels link each observation to a cluster. In other words, $c_i = l$ implies that the i-th unit is assigned to the l-th mixture component. Unfortunately, even given this expansion to the model space, fitting the model presented in Eq. (2) is exceptionally challenging: the overlaying support of the Pareto distributions jeopardises the clustering assignment, which in turn prevents the derivation of reliable estimates of the ID. To address this issue[14], enhanced the model by introducing a local homogeneity assumption, postulating that data points close to each other are more likely to lie on the same latent manifold and, therefore, should be clustered together. This way, the clustering is aided by spatial information about the data points, which was previously

ignored. In particular, the authors make use of the $n \times n$ binary similarity matrix $\mathcal{N}^{(q)}$, with a generic entry defined as $\mathcal{N}_{ij}^{(q)} = 1$ if the j-th observation is among the first q NNs of the i-th observation. To enforce local homogeneity, $\mathbb{P}\left[\mathcal{N}_{i,j}^{(q)} = 1 \mid c_i = c_j\right] = \zeta$ and $\mathbb{P}\left[\mathcal{N}_{i,j}^{(q)} = 1 \mid c_i \neq c_j\right] = 1 - \zeta$. This model extension leads to the following specification:

$$\mathscr{L}\left(\{\mu_i\}_{=1}^{n}, \mathcal{N}^{(q)} | \boldsymbol{d}, \boldsymbol{c}, \zeta\right) = \prod_{i=1}^{n} d_{c_i} \mu_i^{-(d_{c_i}+1)} \times \frac{\zeta^{\sum_{j=1}^{n} \mathcal{N}_{ij}^{(q)} \mathbb{1}_{c_i = c_j}} (1-\zeta)^{\sum_{j=1}^{n} \mathcal{N}_{ij}^{(q)} \mathbb{1}_{c_i \neq c_j}}}{\mathscr{Z}_i}, \qquad c_i | \boldsymbol{\pi} \sim Cat_L(\boldsymbol{\pi}), \tag{3}$$

where $\mathscr{Z}_i$ is a normalizing constant and Cat_L denotes a Categorical distribution over the set $\{1, \ldots, L\}$. A closed-form expression for the posterior distribution is not available, so we rely on Markov Chain Monte Carlo (MCMC) techniques to simulate a posterior sample. The interested reader can find more technical discussions of this model specification and the validity of the underlying hypothesis in the Supplementary Material of related papers[14,15]. In these references, one can also find more details about the Gibbs sampler algorithm used for fitting the model and the post-processing tools adopted to deal with computational issues such as label-switching. In this work, we apply the model defined by Eq. (2) and the corresponding Hidalgo algorithm[14] to assess global COVID-19 disease dynamics. More details are provided in the following subsection.

Data description. This work utilises three datasets to explore the disease and spreading dynamics of COVID-19 in countries: COVID-19 new cases, deaths per million population (pmp)[21], and the COVID-19 Stringency Index (CSI) from the Oxford Coronavirus Government Response Tracker (OxCGRT)[22] (now referred to as CSI). The CSI describes the stringency of government measures by recording the number of government policies in each country and their strictness. The index is a composite measure based on nine response indicators, including school and workplace closures, travel bans, etc. These indicators are rescaled to a value from 0 to 100 (100 = strictest response). Together, these three datasets represent the health and social representation of the effects of COVID-19 on each country. The CSI has informed studies in the health sciences, such as estimating the impact of various physical distancing measures on disease incidence[23] and relating different levels of healthcare resources to the associated transmission risk[24,25]. Political scientists have employed the CSI to consider whether stringency measures vary by regime type[26,27], and whether upcoming elections influenced the strength of responses[28].

We source the datasets from the *Our World in Data* 'Data Explorer', which formats and aggregates a variety of datasets from academic and public institutions globally[21]. *Our World in Data* sources data on worldwide COVID-19 cases and deaths from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University[29]. Given the open nature of these datasets, no ethical approval or specific permissions were required for this study.

Each row of the final dataset contains a country index and the corresponding concatenated relevant time-series (datasets of cases, deaths, and stringency index). Figure 1 provides an excerpt of the combined dataset.

The dataset spans a period of 454 days from 1st Mar 2020 to 29th May 2021 and initially included 190 countries. Given that this analysis includes three datasets each containing 454 temporal measurements, the nominal dimension of the dataset is $D = 454 \times 3 = 1362$.

To improve the robustness of the study, we included countries if they meet certain data availability and size requirements. Countries with more than 20% of missing data in the given period for any of the three time-series were excluded. Any remaining missing values were imputed with a linear regression using the `imputeTS` package in R[30,31]. Additionally, countries with smaller populations (< 1 million) can display higher volatility in new cases and deaths pmp. Thus, such countries were excluded from the dataset to limit the influence of outliers in the analysis. These data pre-processing procedures leave $n = 115$ countries.

All data manipulation or transformation tasks performed as part of the pre-processing methodology were undertaken using through the tools available in the `tidyverse` package in R[31,32]. The completed pre-processed

Code	GSI* (01/03/2020)	GSI* (02/03/2020)	GSI* (03/03/2020)	...	New cases ppm* (01/03/2020)	New cases ppm* (02/03/2020)	New cases ppm* (03/03/2020)	...	New deaths ppm* (01/03/2020)	New deaths ppm* (02/03/2020)	New deaths ppm* (03/03/2020)	...
AFG	-0.34	-0.34	-0.34		-0.81	-0.81	-0.81		-1.00	-1.00	-1.00	
ALB	-4.03	-4.03	-4.03		-0.95	-0.95	-0.95		-1.03	-1.03	-1.03	
ARE	-2.69	-2.69	-2.69		-1.19	-1.19	-1.19		-1.00	-1.00	-1.00	
ARG	-4.59	-4.59	-4.59		-1.25	-1.25	-1.25		-1.05	-1.05	-1.05	
AUS	-3.19	-3.19	-3.19		-0.58	-0.58	-0.57		-0.45	-0.45	-0.45	
AUT	-2.68	-2.68	-2.68		-0.85	-0.85	-0.85		-0.78	-0.78	-0.78	
AZE	-4.29	-4.29	-3.46	⋮	-0.69	-0.69	-0.69	⋮	-0.80	-0.80	-0.80	⋮
BEL	-3.84	-3.84	-3.84		-0.74	-0.74	-0.74		-0.82	-0.82	-0.82	
BEN	-2.01	-2.01	-2.01		-0.78	-0.78	-0.78		-0.64	-0.64	-0.64	
BFA	-1.65	-1.65	-1.65		-0.71	-0.71	-0.71		-0.76	-0.76	-0.76	
BGD	-4.50	-4.50	-4.50		-1.22	-1.22	-1.22		-1.45	-1.45	-1.45	
BGR	-2.28	-2.28	-2.28		-0.80	-0.80	-0.80		-0.82	-0.82	-0.82	
BHR	-2.87	-2.87	-2.87		-1.46	-1.46	-1.47		-1.19	-1.19	-1.19	
BIH	-2.40	-2.40	-2.40		-0.96	-0.96	-0.96		-0.87	-0.87	-0.87	
...		...				...				...		

Figure 1. Input dataset excerpt. Data format utilised in this analysis. The asterisk (*) denotes a dataset standardised to a z-score via Eq. (4). Here, ppm refers to 'per million population'.

dataset and the corresponding code to replicate the data preparation methodology, results, and figures are available on Github.

Additionally, the original dataset may be temporally stratified into four equally separate pandemic stages to reveal the ID at a more granular level in the dataset. The original dataset described in Table 1 documents the date range of each stage.

Hidalgo requires unique observations in a dataset[14] and thus performs optimally on datasets with continuous data or discrete numbers within a broad range. Therefore, this analysis scales new cases and deaths by the country population to satisfy this assumption and enable disease dynamics to be compared across countries with different populations. Second, Hidalgo assumes identically and independently distributed observations in a dataset. To limit temporal autocorrelation, two pre-processing steps are applied to the chosen datasets. Firstly, 'new' cases, and deaths pmp are selected, as opposed to their 'active' or 'cumulative' counterparts. Additionally, each of the three datasets are normalised to z-scores across all countries, given by Eq. (4):

$$z_k = \frac{x_k - \bar{x}_k}{s_k}, \tag{4}$$

where x represents a dataset, $k \in [1, 2, 3]$ denotes each of the three datasets used in this analysis, and $\bar{x}_k$ and s_k represent the mean and standard deviation of a dataset respectively.

Computational details. Hidalgo was run on this dataset for 25,000 MCMC iterations, after a burn-in of 1000. The fast convergence was confirmed by running a secondary analysis with 10000 iterations and obtaining the same results. A sparse mixture modelling approach[19,20] is employed in this analysis, with $L = 6$ mixture components, and $\alpha = 0.05$ for the Dirichlet priors of the mixture weights. Three matrices are produced as the output[15]:

1. Membership labels (dim: $nsim \times n$) where each column contains the MCMC sample of the membership labels for every observation;
2. Cluster probabilities (dim: $nsim \times L$) where each column contains the MCMC sample of the mixing weights for each mixture component;
3. Intrinsic dimensions (dim: $nsim \times L$) where each column contains an MCMC sample for every ID parameter estimated in each cluster.

The MCMC chains produced by Hidalgo may exhibit label-switching issues, which prevents direct extraction of inference from the MCMC output. Indeed, label-switching arises whenever a mixture model with a-priori symmetric components is adopted. Due to label-switching, mixture components can be discarded, emptied, or repopulated across iterations.

To obtain a reliable clustering estimate, one can inspect the posterior co-clustering matrix $PCM = \{p_{ij}\}$ computed across the n countries, where each entry p_{ij} is defined as the proportion of times that countries i and j have been clustered together across the $nsim$ MCMC iterations. Once the PCM is estimated, one can recover a clustering estimation by minimising a loss function over the space of the possible partitions. A widely used method is the minimisation of the Variation of Information (VI) loss function[33,34]. This way, we can estimate the number of latent ID manifolds in the dataset. Moreover, we can also obtain more specific results by following a post-processing procedure[15], devised to address the label-switching issue. The algorithm that is used maps the L different parameter-specific chains – one for each mixture component parameter $\{d_l\}_{l=1}^{L}$ – to n observation-specific chains $\{d_{c_i}\}_{i=1}^{n}$. This way, not only are we able to draw inferences about the clusters characterised by heterogeneous ID present in the data, but we can also focus on the observation-specific ID estimates. Thus, in our application, we can compare the different country-specific ID estimates in addition to ID estimates of latent manifolds in the dataset.

Results and discussion

Global COVID-19 data is characterised by low complexity. A summary of the ID analysis of global data is shown in Fig. 2. In particular, Fig. 2E highlights the posterior distribution of IDs in each cluster group, from which we can obtain a visual estimate of the variability of the ID estimates in each cluster. Hidalgo automatically identifies two manifolds ($L^* = 2$) of posterior mean IDs $d_1 = 12$ and $d_2 = 9$, indicating the COVID-19 disease dynamics and corresponding government-established non-pharmaceutical interventions (NPIs) display higher redundancy in some countries than others. Countries assigned a higher ID indicate complex dynamics,

Stage	Beginning	Ending
1	1st Mar 2020	23rd June 2020
2	24th June 2020	15th October 2020
3	16th October 2020	6th February 2021
4	7th February 2021	29th May 2021

Table 1. Date range for data subsets.

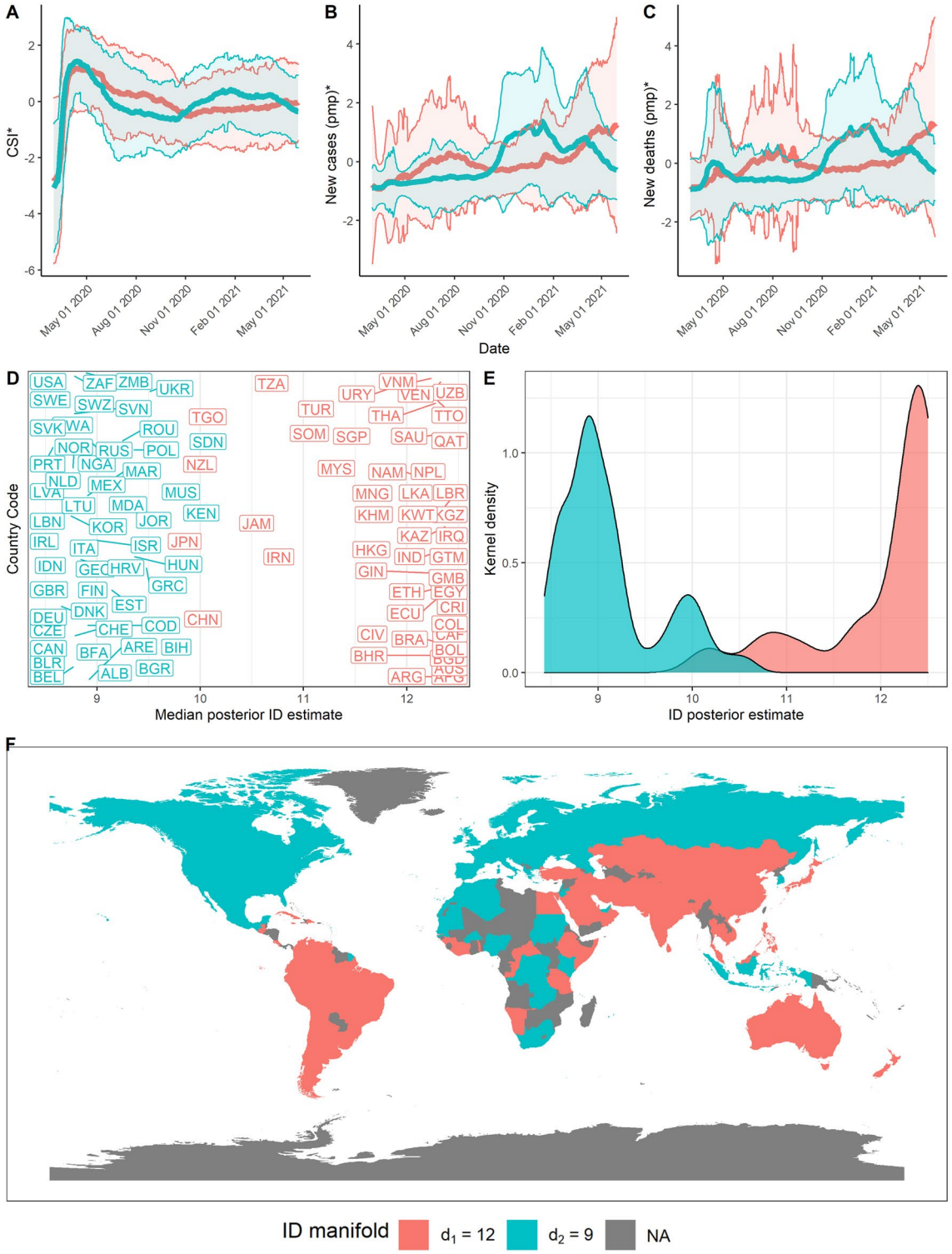

Figure 2. Summary of results over the time period from 1st Mar 2020 to 29th May 2021. (**A**) Mean and standard deviation of standardised CSI aggregated by ID manifold, (**B**) standardised new cases pmp, and (**C**) standardised new deaths pmp. (**D**) Median posterior ID estimate by country, (**E**) posterior ID density estimated by manifold, and (**F**) world map of countries, coloured by ID manifold.

as Hidalgo identifies these points project onto a high-dimensional manifold. Conversely, countries with a lower ID suggest simpler dynamics, as Hidalgo identifies these points project onto a low-dimensional manifold.

Given the high dimensionality of the dataset, IDs of 12 and 9 represent a dimensionality reduction of 99.34 and 99.11% respectively, suggesting strong dependence on the standardised new cases per million population

(pmp), new deaths pmp, and the CSI for a country over the given period. Notably, these results indicate that a small set of parameters govern the COVID-19 dynamics, which has important implications for practitioners seeking to model these dynamics or apply dimensionality reduction techniques. For example, authors such as[35] have identified that lower IDs lower the sample complexity of learning, enabling more accessible learning for neural networks and better model generalisation from training to test data.

Despite the overall low dimensionality of the dataset, the two ID manifolds identified differ by at least three dimensions. This result warrants further inspection to examine potential explanations for the dimensionality of each ID manifold.

Global distribution of COVID-19 data complexity demonstrates spatial autocorrelation. Notably, we identify evidence of spatial autocorrelation in the ID of global COVID-19 data, supported by Fig. 2D,F. Figure 2D highlights the individual ID of each country included in the dataset. The colour of each country code corresponds to the ID manifold to which each country belongs. The ID manifold of each country may also be presented geographically on a map, as displayed in Fig. 2F. Upon visual examination, Fig. 2D,F demonstrate that countries geographically close together tend to belong to the same ID manifold.

We confirmed these results using the Moran's `I` test, which is a widely used spatial statistic for detecting spatial autocorrelation[36,37]. Moran's `I` ranges from -1 to 1, and is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{5}$$

where N is the number of spatial units indexed by i and j; x is the individual median posterior ID estimate of a country; $\bar{x}$ is the mean of x ; w_{ij} is a neighbour adjacency matrix with zeroes on the diagonal (i.e., $w_{ii} = 0$); and W is the sum of all w_{ij}[36]. In line with common approaches, we assign a weight of 1 for neighbouring zones and 0 otherwise[37]. A neighbourhood is defined such that every country has at least one neighbour in the spatial weights matrix.

Applying this test to the geographical distribution of median posteriors of ID produces an `I` value of 0.85 ($p < 0.001$) using the `spdep` package in R[31,38], indicating significant positive spatial autocorrelation.

This result is a compelling finding, as the input dataset does not include any information about the countries geographical locations. Neighbouring countries may share the complex dynamics of COVID-19 as the pandemic spreads worldwide, resulting in positive spatial autocorrelation in the distribution of median posterior IDs of countries included in the analysis[39] suggests that geographically close countries are likely to share spatio-temporal dynamics due to human spatial dynamics and similar demographic factors across geographic regions. In reviewing the available literature[40], highlights that a country's interconnectedness influences the spreading dynamics of COVID-19. This literature suggests that geographical closeness and interconnectivity have substantial implications for the spreading dynamics of COVID-19, allowing this to be a potential explanation for the spatial autocorrelation identified in the complexity of spreading dynamics observed in the analysis.

High-income countries are characterised by lower complexity data. Our analysis reveals that countries with higher income level groups are more likely to lie in low-dimensional manifolds. Figure 3 presents the distribution of income levels across the two ID manifolds. The World Bank assigns one of four income levels to each country, ranging from low- to high-income[41]. For the 2022 fiscal year, low-income countries fall under a Gross National Income (GNI) per capita of $1,045 (USD) or less in 2020; lower-middle-income between $1,046 and $4,095; upper-middle-income between $4,096 and $12,695; and high-income from $12,696 or more. GNI per capita represents the value produced by each person in a country's economy in a given year, regardless of whether the source of the value created is domestic production or receipts from overseas. While the GNI per capita does not entirely summarise a country's level of development or welfare, it has proved to be a useful and readily available indicator that closely correlates with other, non-monetary measures of the quality of life, such as life expectancy at birth, mortality rates of children, and enrollment rates in school[42].

A range of factors may explain the skewed distribution of income levels towards the low-ID manifold. High-income countries usually have aging populations, arising from declining fertility and improving mortality due to income growth, changes in health behaviours, and higher education levels[43,44]. Aging populations in many high-income countries have played a role in creating a greater mortality burden during the COVID-19 pandemic over 2020 to 2021 due to increased vulnerability to serious infections in population groups aged over 65[45]. Additionally, underlying diseases such as diabetes, cardiovascular disease, and other diseases significantly contribute to increased severity risk from COVID-19[46]. Importantly, chronic medical conditions are widely prevalent in aging populations in high-income countries[47]. These factors have significantly impacted the mortality burden per capita over the COVID-19 pandemic. Research from the World Bank[45,48] estimate that high-income countries have had 2 to 3 times the COVID-19 mortality burden per capita compared to other countries over 2020. The age distribution disparity across the two ID manifolds is evident in Fig. 4A,B.

Figure 4A reveals that countries assigned to a high-ID manifold had less than 7% of the population aged over 65 on average. In comparison, countries assigned to a high-ID manifold have 13% of the population aged over 65 on average, despite displaying bimodality due to some low-income countries in the low-ID manifold. A Kolgomorov-Smirnov test may be applied to evaluate the null hypothesis that the distributions are sampled from a population with the same distribution, which is subsequently rejected at the $p < 0.001$ significance level[49]. The mean age distribution of countries in each ID manifold presented in Fig. 4B corroborates that countries assigned to a low-ID manifold host a higher proportion of the population aged over 65, while countries assigned to a high-ID manifold host a higher proportion of the population aged between 0 and 14.

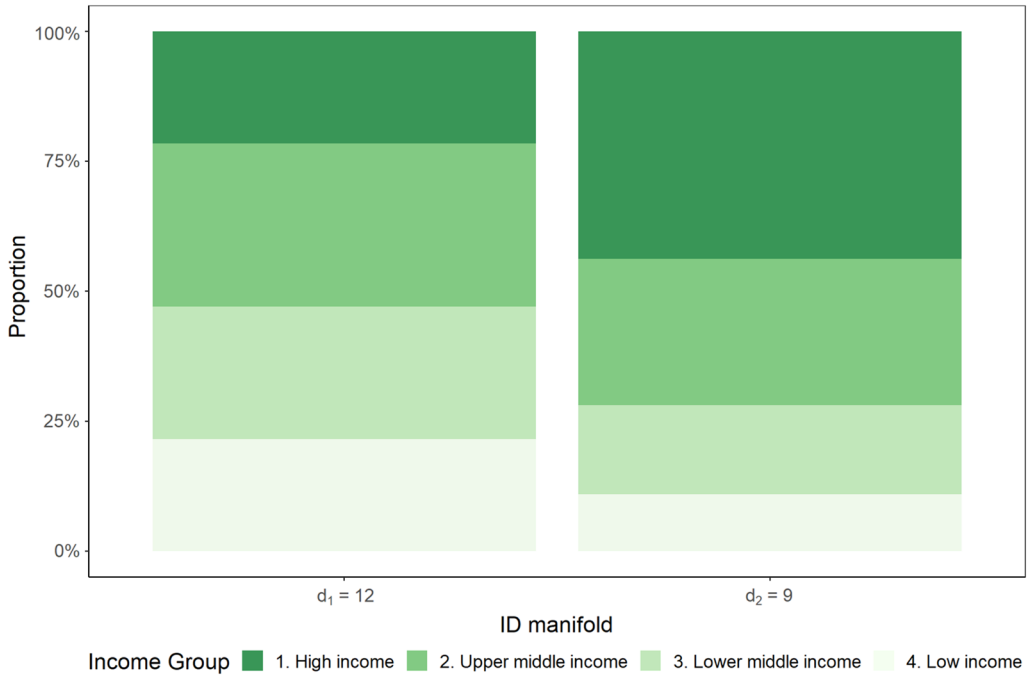

Figure 3. Distribution of countries by income in each ID manifold. Income group classifications are retrieved from the World Bank database[41].

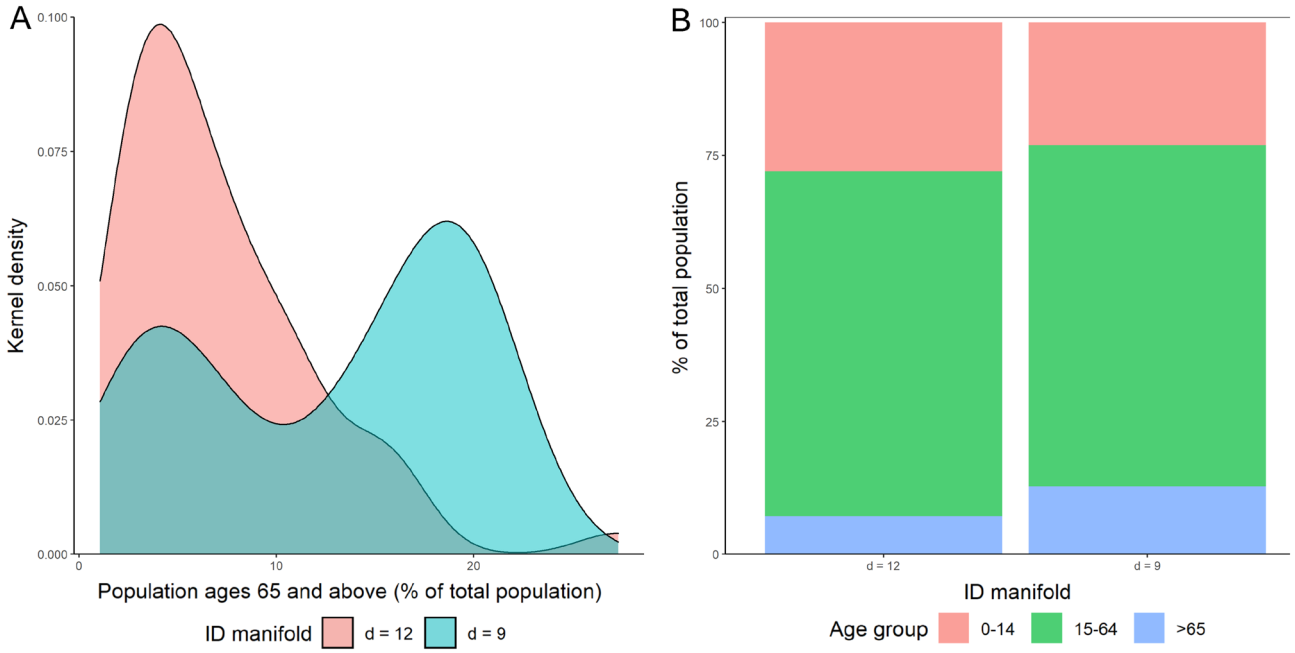

Figure 4. Population distribution of countries in each ID manifold. (**A**) Kernel density estimation of population ages 65 and above (% of total), by ID manifold. (**B**) Mean population distribution of countries by ID manifold.

A possible explanation for high-income countries being assigned a low-ID manifold arises after identifying a link between high-income countries, aging populations, and increased per capita mortality burden. Namely, since new deaths pmp are a subset of new cases pmp, increased new deaths pmp in high-income countries may provide explanatory power to new cases pmp, resulting in greater dependency between the two time-series datasets and thus requiring a lower ID. Conversely, lower rates of reported new deaths pmp in low-income countries would lower the dependence in the entire dataset included in the analysis requiring a higher number of dimensions to represent the data accurately.

Furthermore, issues in data quality in COVID-19 data in low- and middle-income countries are widely researched. They may be another factor contributing towards the existing distribution of countries and

corresponding designations to ID manifolds[45,50] identifies that under-reporting in deaths varies globally but is highest in low-income and fragile settings. Such data artifacts could lead to a higher number of unexplained values, thereby lowering the dependence in the dataset and requiring more dimensions to describe the data effectively.

Changes in ID over stages of the COVID-19 pandemic. Stratifying the datasets has provided a granular view of the ID over the course of the pandemic, and a summary of the results for each stage is presented in Figs. 5, 6, 7 and 8.

We can observe that countries lie between 2 to 3 ID manifolds throughout the pandemic. From March to June 2020 (Stage 1), manifolds have a similar ID which could reflect a generally united global response to the pandemic (Fig. 5, $d_1 = 9, d_2 = 8.6$). In June 2020 to October 2020 however, we find that the data lies on 3 different manifolds (Fig. 6, $d_1 = 10, d_2 = 9.2, d_3 = 9.75$). These 3 ID manifolds continue from October 2020 to February 2021, with all 3 manifolds lying on an ID between 9 and 10 (Fig. 7, $d_1 = 10, d_2 = 9.2, d_3 = 9.75$). Countries belonging to the manifold with the ID of 5.9 are mostly European (e.g., France, Italy). These countries experience a rise in the growth rate of cases and deaths, which precedes a corresponding rise in countries lying on the manifold with an ID of 6.9 (e.g., US, Spain, UK, Russia). Meanwhile, other countries with an ID of 9.2 continue to experience the average growth rates in cases and deaths (e.g., Australia, China, India, and much of South America). Finally, from February to May 2021, some countries lie on one clear manifold, with an ID of 7.5 (Fig. 8, $d_1 = 7.5, d_2 = 6.4$).

Implications and future research. We have successfully identified heterogenous ID manifolds for a dataset incorporating publicly available COVID-19 data such as government stringency levels, cases, and death rates per capita utilising Hidalgo, a Bayesian mixture model. Applying this model to the dataset reveals low intrinsic dimensionality, highlighting a potential for significant dimensionality reduction in the dataset. These findings suggest that few independent dimensions are required to effectively describe the dataset, enabling practitioners to discern better the level of model complexity required when describing or forecasting such data.

Furthermore, we demonstrate how heterogenous ID estimators like Hidalgo may be employed to partition and simplify high-dimensional datasets. We reveal interesting spatial and demographic patterns in data that capture the unfolding of the global pandemic. It may be valuable for practitioners to consider these tools as part of their arsenal, to quantify data complexity and heterogeneity meaningfully, as part of a quest to effectively extract useful information contained in high-dimensional data.

Ultimately, the results of this analysis are subject to the quality of data available. While every effort has been made to correct for issues in the data, the inherent discrepancy in data quality across countries inevitably affects the results of this analysis. As previously ascertained[45,51], it is currently infeasible to account for all under-reporting and data quality issues for specific countries and therefore remain an artifact in this dataset. It is also important to note that inherent assumptions in the Hidalgo algorithm require a careful choice of datasets in addition to some scaling transformations to limit temporal autocorrelation. These requirements limit the immediate applicability of Hidalgo for time series datasets as analysis must be conducted on standardised first-order differences with continuous values (e.g. new cases pmp, new deaths pmp), precluding practitioners from considering more intuitive datasets like cumulative cases or deaths.

In our analysis, we emphasized the interpretation of the spatial characteristics of the results. Nonetheless, we acknowledge that further studies can be conducted to deepen the understanding of our findings from a temporal point of view. One option is to temporally align the data by considering the first date a COVID-19 case was reported. This temporal restructuring of the data would provide valuable insights into the temporal evolution of the pandemic and its impact on different regions over time. By examining the dynamics of the outbreak from a dynamic system perspective, we can gain a deeper understanding of how the pandemic unfolded and its varying effects across regions. We are mindful of the importance of investigating these temporal features, and we plan to pursue this line of research in the future.

Moreover, it would be valuable to conduct further examination on other factors contributing to the complexity (ID) of the COVID-19 data dynamics of a country to better understand drivers for complexity in pandemics. While we have identified that income level, age distribution, disease burden, and data quality all play a role in determining the ID of a country, developing a more nuanced understanding of these contributing factors would provide utility to the broader scientific community. For example, this could encompass additional significant socio-economic and environmental covariates[52,53].

Finally, from a methodological perspective, we recall that Hidalgo is based on ratios of distances between a point and its first and second NNs. In principle, one could rely on ratios of distances from NNs of *generic* order as a suitable estimator for homogeneous ID has been recently proposed by[54]. Future work is needed to extend this methodology to a mixture framework to account for the presence of heterogeneous IDs. Although considering larger neighbourhoods leads to a reduction of the estimator variance, we remark that considering more generic ratios would imply more stringent assumptions. These assumptions, such as a broader local homogeneity, may be violated when working with real-world data.

Conclusions

This work evaluated the complexity of a dataset consisting of the standardised per-capita growth rate of COVID-19 cases, deaths, and an index describing a country's stringency of NPI measures (CSI), using a heterogenous intrinsic dimension estimator implemented as a Bayesian mixture model (Hidalgo). We identify that the COVID-19 dataset may be projected onto two low-dimensional manifolds ($d_1 = 12, d_2 = 9$). Lower dimensionality suggests stronger dependence in the standardised growth rates of cases and deaths per capita and the CSI for a

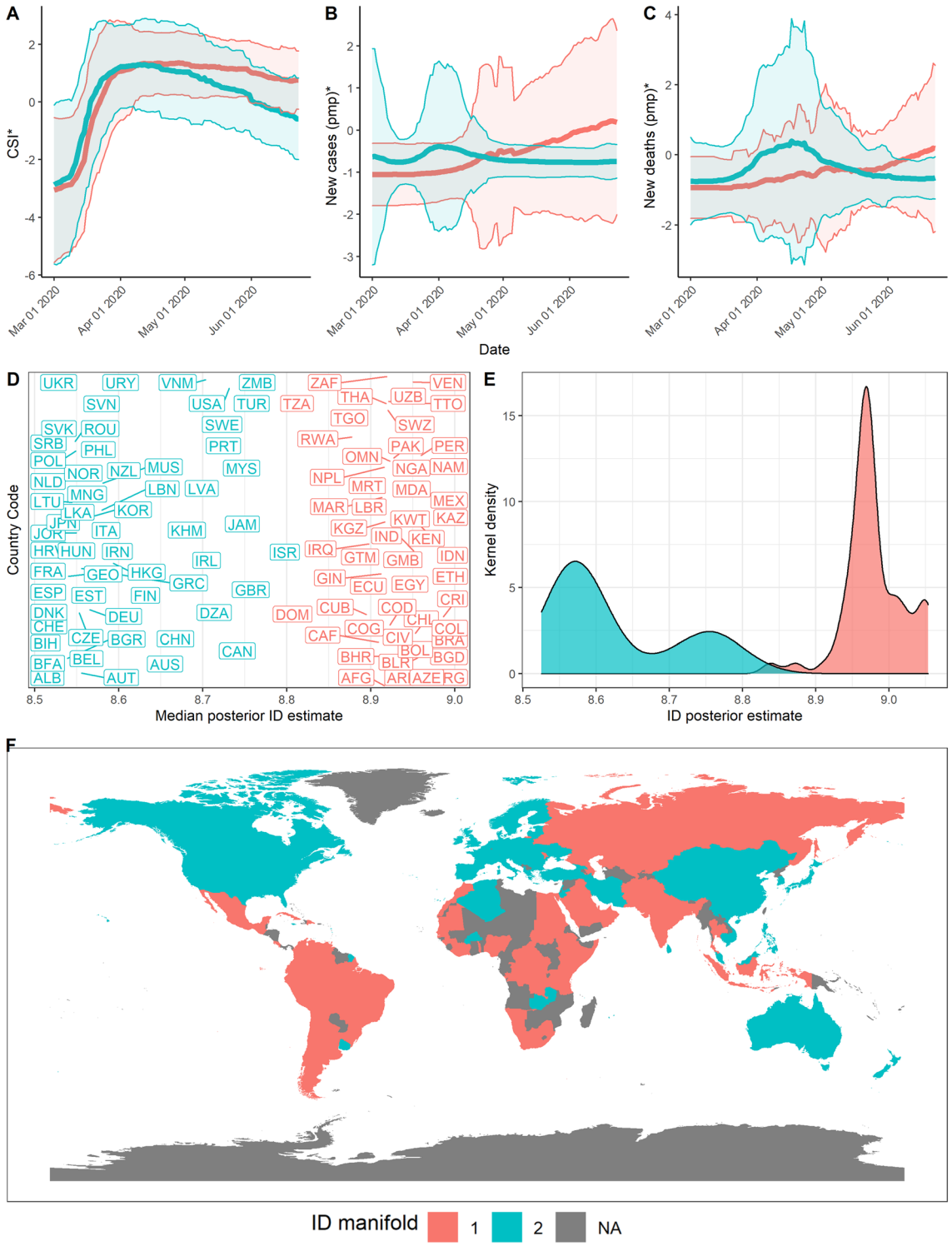

Figure 5. Stage 1—Summary of results over the time period from 1st Mar 2020 to 23rd June 2020. (**A**) Mean and standard deviation of standardised CSI aggregated by ID manifold, (**B**) standardised new cases pmp, and (**C**) standardised new deaths pmp. (**D**) Median posterior ID estimate by country, (**E**) posterior ID density estimated by manifold, and (**F**) world map of countries, coloured by ID manifold.

country over the given period. Notably, it indicates that COVID-19 data dynamics are governed by a small set of parameters, which has important implications for practitioners seeking to model these dynamics or apply dimensionality reduction techniques on this data.

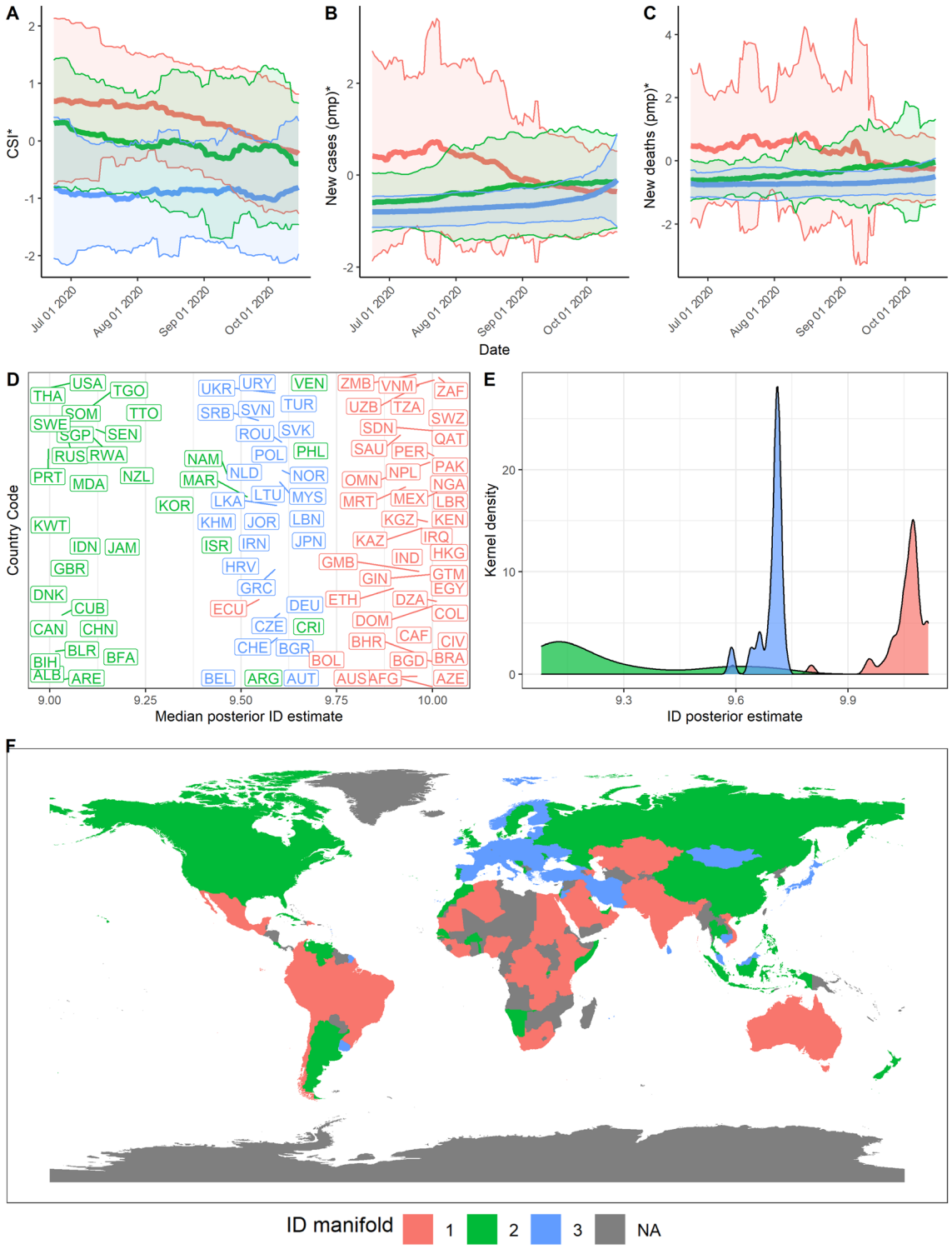

Figure 6. Stage 2—Summary of results over the time period from 24th June 2020 to 15th October 2020. (**A**) Mean and standard deviation of standardised CSI aggregated by ID manifold, (**B**) standardised new cases pmp, and (**C**) standardised new deaths pmp. (**D**) Median posterior ID estimate by country, (**E**) posterior ID density estimated by manifold, and (**F**) world map of countries, coloured by ID manifold.

This work has demonstrated how the intrinsic dimension can help extract novel insights across multiple complex datasets and identify engaging ways to effectively segregate data. For example, we identify spatial auto-correlation in the distribution of ID estimates for countries. Furthermore, we highlight a skewed distribution of high-income countries projected on a low-dimensional ID manifold due to the increased per capita mortality

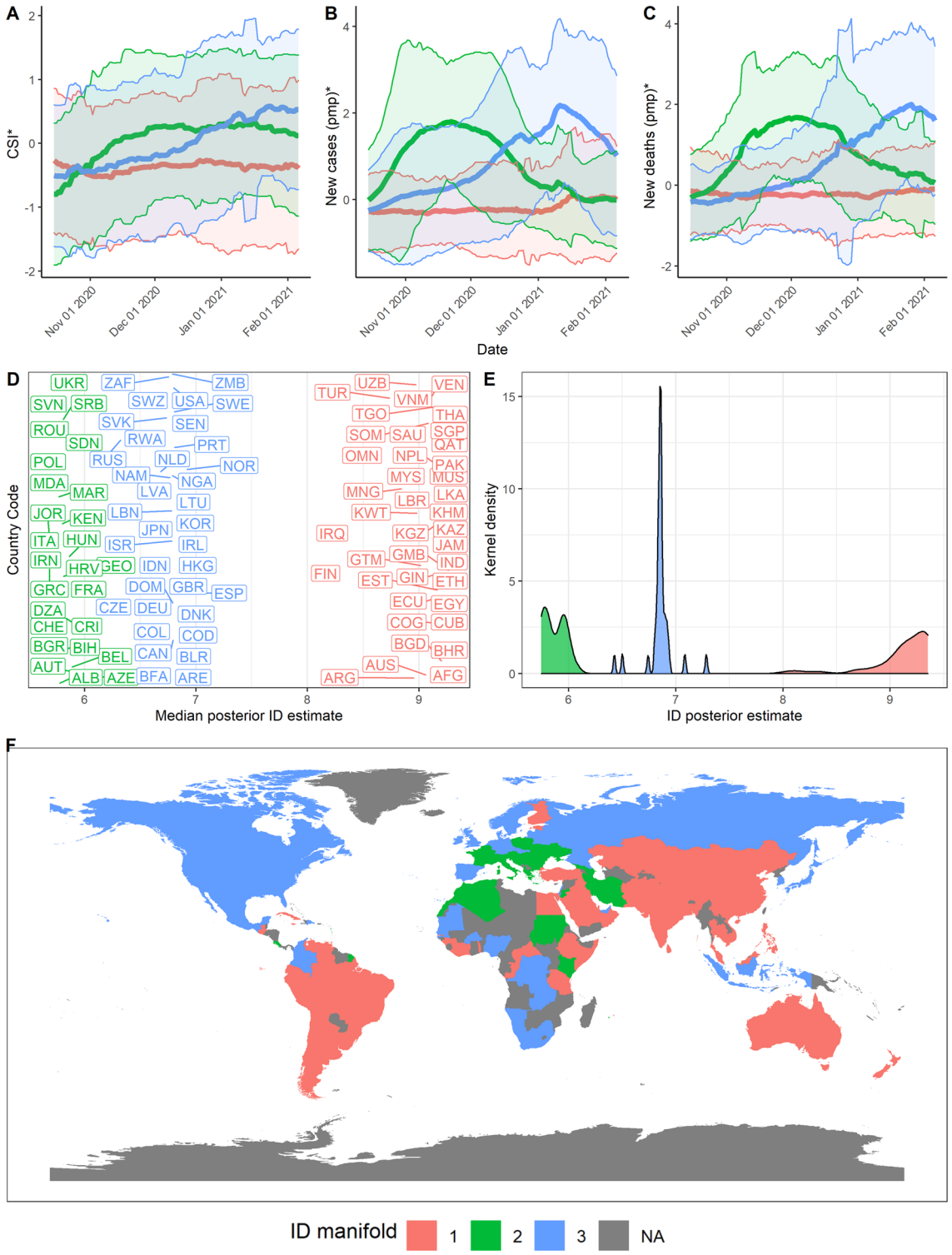

Figure 7. Stage 3—Summary of results over the time period from 16th October 2020 to 6th February 2021. (**A**) Mean and standard deviation of standardised CSI aggregated by ID manifold, (**B**) standardised new cases pmp, and (**C**) standardised new deaths pmp. (**D**) Median posterior ID estimate by country, (**E**) posterior ID density estimated by manifold, and (**F**) world map of countries, coloured by ID manifold.

burden from COVID-19 arising from aging populations and the increased prevalence of comorbidities. While we make significant progress towards understanding drivers for complexity in the included COVID-19 datasets, developing a more nuanced understanding of these contributing factors would enable decision-makers to better account for complexity in pandemics and is identified as an area of future research.

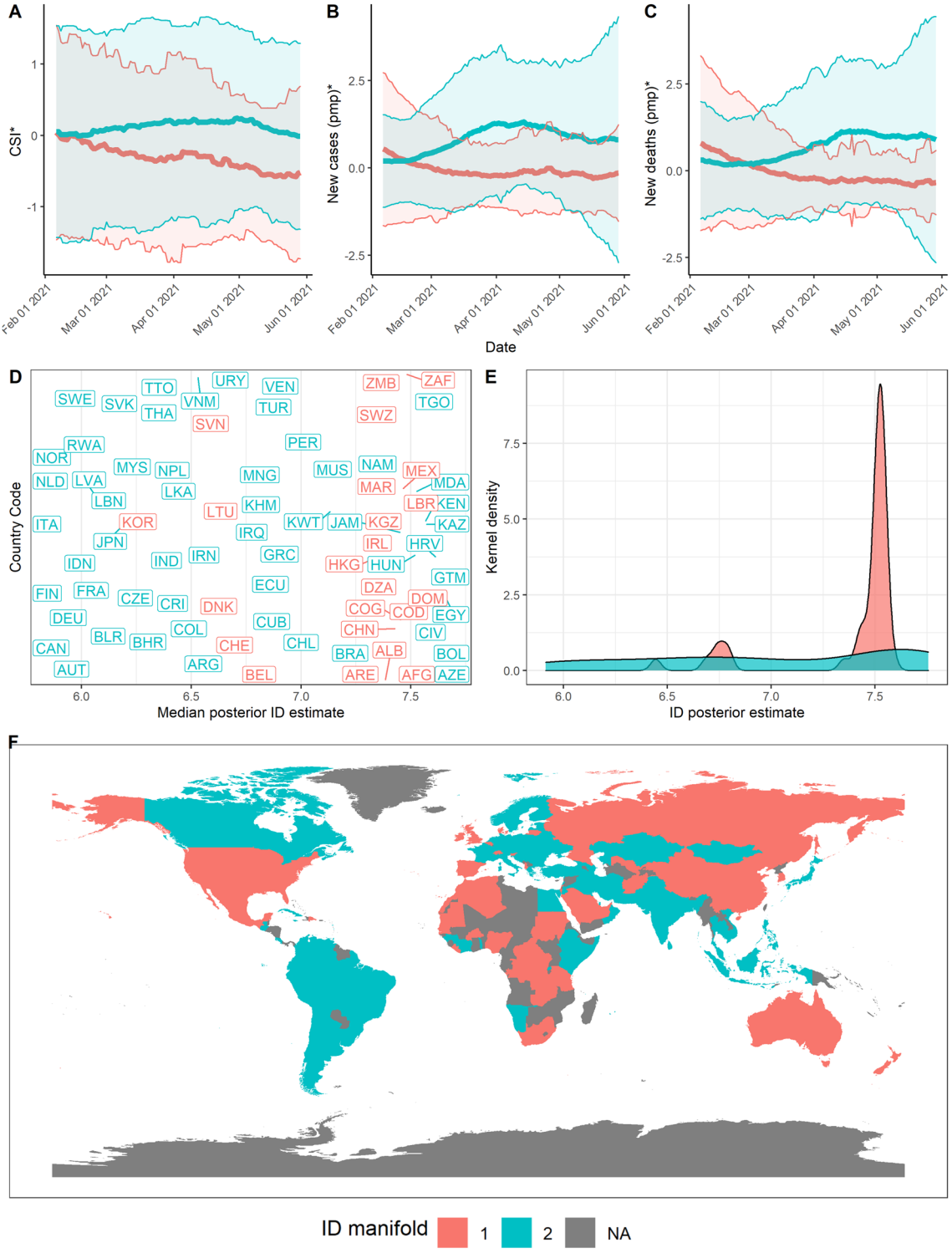

Figure 8. Stage 4—Summary of results over the time period from 7th February 2021 to 29th May 2021. (**A**) Mean and standard deviation of standardised CSI aggregated by ID manifold, (**B**) standardised new cases pmp, and (**C**) standardised new deaths pmp. (**D**) Median posterior ID estimate by country, (**E**) posterior ID density estimated by manifold, and (**F**) world map of countries, coloured by ID manifold.

Data availability

The datasets used in this paper are publicly accessible and are sourced from the *Our World in Data* website (ourworldindata.org). No request for access and ethics approvals were required to retrieve the data used in this

paper. The aggregated dataset used in this analysis is available at https://github.com/EdgarSantos-Fernandez/covid19_ID.

Code accessibility

R code to reproduce the outcomes and plots presented in this paper can be found at https://github.com/EdgarSantos-Fernandez/covid19_ID.

Received: 13 September 2022; Accepted: 30 May 2023
Published online: 16 June 2023

References

1. Levina, E. & Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, 777–784 (2005).
2. Camastra, F. & Staiano, A. Intrinsic dimension estimation: Advances and open problems. *Inf. Sci.* **328**, 26–41. https://doi.org/10.1016/j.ins.2015.08.029 (2016).
3. Shoghri, A. E., Liebig, J., Jurdak, R., Gardner, L. & Kanhere, S. S. Identifying highly influential travellers for spreading disease on a public transport system. *IEEE Comput. Soc.*https://doi.org/10.1109/WoWMoM49955.2020.00020 *(2020).*
4. Levin, R., Chao, D. L., Wenger, E. A. & Proctor, J. L. Cell phone mobility data and manifold learning: Insights into population behavior during the COVID-19 pandemic. preprint, Epidemiology (2020). https://doi.org/10.1101/2020.10.31.20223776.
5. Wisesty, U. N. & Mengko, T. R. Comparison of dimensionality reduction and clustering methods for SARS-CoV-2 genome. *Bull. Electr. Eng. Inform.* **10**(4), 2170–2180. https://doi.org/10.11591/eei.v10i4.2803 (2021).
6. Hearn, A. The Higher-Ed Coronavirus Response in the Public Sector: The Usage of Dimensionality Reduction Techniques and Feature Importance Algorithms to Analyze Fall Re-Opening Plans. *online manuscript* (2020).
7. Pierri, F. *et al.* Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Sci. Rep.* **12**, 5966 (2022).
8. Ordun, C., Purushotham, S. & Raff, E. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. arXiv:2005.03082 [cs] (2020).
9. Doanvo, A. *et al.* Machine Learning Maps Research Needs in COVID-19 Literature. *Patterns***1**, 100123–100123, https://doi.org/10.1016/j.patter.2020.100123 (2020). Place: United States Publisher: Elsevier Inc.
10. Sivakumar, B. & Deepthi, B. Complexity of COVID-19 Dynamics. *Entropy* **24**, 50. https://doi.org/10.3390/e24010050 (2021).
11. Santos-Fernandez, E., Denti, F., Mengersen, K. & Mira, A. The role of intrinsic dimension in high-resolution player tracking data—Insights in basketball. *Annals of Applied Statistics* **16**(1), 326–348 (2022).
12. Eneva, E., Kumaraswami, K. & Matteucci, M. Wekkem: A study in fractal dimension and dimensionality reduction. In *Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches* (2002).
13. Bishop, C. M. *Neural Networks for Pattern Recognition* (Clarendon Press, 1995).
14. Allegra, M., Facco, E., Denti, F., Laio, A. & Mira, A. Data segmentation based on the local intrinsic dimension. *Sci. Rep.* 1–27. https://doi.org/10.1038/s41598-020-72222-0 (2019).
15. Denti, F. intRinsic: An R package for model-based estimation of the intrinsic dimension of a dataset. *J. Stat. Softw.* **106**, 1–45. https://doi.org/10.18637/jss.v106.i09 (2023) arXiv:2102.11425.
16. MacKay, D. & Ghahramani, Z. Comments on 'Maximum Likelihood Estimation of Intrinsic Dimension' by E. Levina and P. Bickel (2004). *Comment on personal webpage* (2005).
17. Facco, E., d'Errico, M., Rodriguez, A. & Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **7**, 12140 (2017).
18. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* (Chapman and Hall/CRC, 1995).
19. Rousseau, J. & Mengersen, K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x (2011).
20. Malsiner-Walli, G., Frühwirth-Schnatter, S. & Grün, B. Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26**, 303–324. https://doi.org/10.1007/s11222-014-9500-2 (2016).
21. Ritchie, H. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
22. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.*https://doi.org/10.1038/s41562-021-01079-8 *(2021).*
23. Edejer, T.T.-T. *et al.* Projected health-care resource needs for an effective response to COVID-19 in 73 low-income and middle-income countries: a modelling study. *Lancet Glob. Health* **8**, e1372–e1379. https://doi.org/10.1016/S2214-109X(20)30383-1 (2020).
24. Islam, N. *et al.* Physical natural experiment in 149 countries. *BMJ* **370**, m2743. https://doi.org/10.1136/bmj.m2743 (2020).
25. Hale, T. *et al.* Government responses and COVID-19 deaths: Global evidence across multiple pandemic waves. *PLOS ONE* **16**, e0253116. https://doi.org/10.1371/journal.pone.0253116 (2021).
26. Hale, T. *et al.* Pandemic Governance Requires Understanding Socioeconomic Variation in Government and Citizen Responses to COVID-19. SSRN Scholarly Paper ID 3641927, Social Science Research Network, Rochester, NY (2020).
27. Frey, C. & Chen, C. Democracy, Culture, and Contagion: Political Regimes and Countries Responsiveness to Covid-19* (2020).
28. Pulejo, M. & Querubín, P. Electoral Concerns Reduce Restrictive Measures During the COVID-19 Pandemic. Working Paper 27498, National Bureau of Economic Research (2020). https://doi.org/10.3386/w27498. Series: Working Paper Series.
29. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).
30. Moritz, S. & Bartz-Beielstein, T. imputeTS: Time series missing value imputation in R. *R J.* **9**, 207–218 (2017).
31. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021).
32. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686. https://doi.org/10.21105/joss.01686 (2019).
33. Meilă, M. Comparing clusterings-an information based distance. *J. Multivar. Anal.* **98**, 873–895. https://doi.org/10.1016/j.jmva.2006.11.013 (2007).
34. Wade, S. & Ghahramani, Z. Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Anal.* **13**, 559–626. https://doi.org/10.1214/17-BA1073 (2018) arXiv:1505.03339.
35. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M. & Goldstein, T. The Intrinsic Dimension of Images and Its Impact on Learning. *ICLR 2021 Conference* (2021).
36. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23. https://doi.org/10.2307/2332142 (1950).
37. Jackson, M. C., Huang, L., Xie, Q. & Tiwari, R. C. A modified version of Moran's I. *Int. J. Health Geogr.* **9**, 33. https://doi.org/10.1186/1476-072X-9-33 (2010).
38. Bivand, R. S., Pebesma, E. & Gomez-Rubio, V. *Applied Spatial Data Analysis with R* 2nd edn. (Springer, 2013).
39. McLafferty, S. Placing pandemics: Geographical dimensions of vulnerability and spread. *Eurasian Geogr. Econ.* **51**, 143–161. https://doi.org/10.2747/1539-7216.51.2.143 (2010).

40. McMahon, T., Chan, A., Havlin, S. & Gallos, L. K. Spatial correlations in geographical spreading of COVID-19 in the United States. *Sci. Rep.* **12**, 699. https://doi.org/10.1038/s41598-021-04653-2 (2022).
41. Hamadeh, N., van Rompaey, C. & Metreau, E. New World Bank country classifications by income level: 2021-2022 (2021).
42. International Economics Department. Per Capita Income: Estimating Internationally Comparable Numbers (1989).
43. Ho, J. Y. & Hendi, A. S. Recent trends in life expectancy across high income countries: Retrospective observational study. *BMJ* **362**, k2562. https://doi.org/10.1136/bmj.k2562 (2018).
44. Lee, J. & Smith, J. P. *Health, Economic Status, and Aging in High-Income Countries* (National Academies Press (US), 2018). Publication Title: Future Directions for the Demography of Aging: Proceedings of a Workshop.
45. Schellekens, P. & Sourrouille, D. *COVID-19 Mortality in Rich and Poor Countries: A Tale of Two Pandemics?* (World Bank, 2020).
46. Booth, A. *et al.* Population risk factors for severe disease and mortality in COVID-19: A global systematic review and meta-analysis. *PLOS ONE* **16**, e0247461. https://doi.org/10.1371/journal.pone.0247461 (2021).
47. Ofori-Asenso, R. *et al.* Recent patterns of multimorbidity among older adults in high-income countries. *Popul. Health Manag.* **22**, 127–137. https://doi.org/10.1089/pop.2018.0069 (2019).
48. Bayati, M. Why is COVID-19 more concentrated in countries with high economic status?. *Iran. J. Public Health* **50**, 1926–1929. https://doi.org/10.18502/ijph.v50i9.7081 (2021).
49. Kolmogorov, A. N. & Smirnov, V. I. Kolmogorov-Smirnov Test. In *The Concise Encyclopedia of Statistics*, 283–287, https://doi.org/10.1007/978-0-387-32833-1_214 (Springer, New York, NY, 2008).
50. Whittaker, C. *et al.* Under-reporting of deaths limits our understanding of true burden of covid-19. *BMJ* **375**, n2239. https://doi.org/10.1136/bmj.n2239 (2021).
51. Lloyd-Sherlock, P., Sempe, L., McKee, M. & Guntupalli, A. Problems of data availability and quality for COVID-19 and older people in low- and middle-income countries. *Gerontol.*https://doi.org/10.1093/geront/gnaa153 *(2020).*
52. Weaver, A. K., Head, J. R., Gould, C. F., Carlton, E. J. & Remais, J. V. Environmental factors influencing covid-19 incidence and severity. *Annu. Rev. Public Health* **43**, 271–291 (2022).
53. Kong, J. D., Tekwa, E. W. & Gignoux-Wolfsohn, S. A. Social, economic, and environmental factors influencing the basic reproduction number of covid-19 across countries. *PLoS ONE* **16**, e0252373 (2021).
54. Denti, F., Doimo, D., Laio, A. & Mira, A. The generalized ratios intrinsic dimension estimator. *Sci. Rep.*https://doi.org/10.1038/s41598-022-20991-1 *(2022).*
55. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

Acknowledgements

AV received funding from the Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers for Big Data, Big Models and New Insights (ACEMS) under Grant Number CE140100049 and the First Byte Grant through the Centre for Data Science at the Queensland University of Technology. KM was supported by an ARC Laureate Fellowship under Grant Number FL150100150. AM was supported by FISR 2020 COVID No. FISR2020IP_03843 and by European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101016233. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. Data visualizations, including the maps, were generated using the R package ggplot2 (3.4.1)[55].

Author contributions

A.V. contributed to data aggregation, application of the methodology, securing funding, analysing the results, and writing the paper and code. E.S.F. contributed in writing the paper, securing funding and providing important feedback and guidance. F.D. contributed in writing the paper, developing the methodology, and providing important feedback. A.M. and K.M. contributed towards the ideation and methodology development, in providing supervisory guidance throughout the paper, securing and providing funding, and providing important feedback in the editing process.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.S.-F. or A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.