ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

A data-driven approach to measuring epidemiological susceptibility risk around the world

Alessandro Bitetto
(Università di Pavia)

Paola Cerchiello
(Università di Pavia)

Charilaos Mertzanis
(Abu Dhabi University)

# 200 (05-21)

Via San Felice, 5
I-27100 Pavia

economiaweb.unipv.it

A data-driven approach to measuring epidemiological susceptibility risk around the world

Abstract

Epidemic outbreaks are extreme events that become less rare and more severe, associated with large social and economic costs. It is therefore important to assess whether countries are prepared to manage epidemiological risks. We use a fully data-driven approach to measure epidemiological susceptibility risk at the country level using time-varying information. We apply both principal component analysis (PCA) and dynamic factor model (DFM) to deal with the presence of strong cross-section dependence in the data. We conduct extensive in-sample model evaluations of 168 countries covering 17 indicators for the 2010-2019 period. The results show that the robust PCA method accounts for about 90% of total variability, whilst the DFM accounts for about 76% of the total variability. Our index could therefore provide the basis for developing risk assessments of epidemiological risk contagion. It could be also used by firms to assess likely economic consequences of epidemics with useful managerial implications.

Keywords: Risk analysis, Epidemiological risk, Data-driven, Policy framework, Machine learning

JEL: I18, C55, C38, F68

During the past year, the Covid-19 pandemic has infected more than 100 million people and caused more than 2 million deaths in more than 200 countries around the world. The associated economic and social costs are huge. Some estimates raise the global economic cost of the Covid-19 pandemic for the next few years to several USD trillion (The International Monetary Fund, 2020). A great concern has been the virus' spread to countries with weaker epidemics management systems. Thus, knowing how countries with different degrees of preparedness have responded to the pandemic is important for assessing cross-country epidemiological risk and optimally deploying resources in support of this global health emergency. This is critical knowledge of globally susceptible populations, with several countries reporting infection levels exceeding their average historical levels. These policy concerns have remained valid during all phases of the Covid-19 pandemic and especially during the process of gradual removal of the lockdown restrictions. The question of country preparedness has surfaced again following the recent mutation of the virus (The World Health Organization, 2020d).

The question of countries' preparedness to manage epidemiological risk must be addressed from a long-term perspective. It is likely that the world will continue to face epidemic risks, which many countries are still ill positioned to manage. In addition to climate change and urbanization, global population displacement and migration—now happening in nearly every corner of the world—create favorable conditions for the emergence and spread of new pathogens. Countries also face an increasing potential threat of accidental or deliberate release of deadly engineered pathogens, which could cause even greater harm than a naturally occurring pandemic. Scientific advances that help in fighting epidemic diseases have also allowed pathogens to be engineered or recreated in laboratories. Meanwhile, cross-country disparities in capacity and inattention to biological threats have exacerbated preparedness gaps. Measuring country preparedness emerges as a key economic policy challenge for both countries and firms.

We contribute to addressing this policy challenge by creating an index of epidemiological susceptibility risk (ESR) for 168 countries. Various economic and non-economic factors affect the extent to which a country is susceptible to epidemiological risk. We produce a new epidemiological preparedness measure that relies on objective information that facilitates policy choices. We build on previous studies and our index information accounts for the role of environmental, health, transport and communications infrastructures; economic activity; demographics; and governance institutions. We deal with the complexity of these factors by implementing a fully data-driven approach to measuring their influence on epidemiological risk. In contrast to previous studies (Rivers et al., 2019, Polonsky et al., 2019, Mertzanis and Papastathopoulos, 2021), our fully data-driven approach produces results that provide a better evidence basis to support reasoning and decision. While there are no data-driven algorithms that can lead to fully optimal assessments of risk, our approach has considerable advantages, such as avoiding the subjective weight determination and the need for post-hoc rationalization. Evidence shows that data-driven models offer better predictive accuracy in epidemiological research than knowledge-based ones (Rajabi et al., 2014). Given the complexity of the problem, we choose different versions of principal component analysis (PCA) as well as dynamic factor models (DFM) to deal with the presence of strong cross-section dependence in the data due to unobserved common factors. We conduct extensive in-sample model evaluations of 168 countries covering 17 indicators during the period 2010-2019. Our results show that the robust PCA method explains more than 90% of total variability, whilst the DFM explains about 76% of the total variability.

Our paper contributes to the literature in the following ways: it builds on previous studies by proposing a substantially improved index of epidemiological susceptibility risk that is fully model-based and data-driven, tested and validated according to advanced statistical techniques. We use alternative versions of an unsupervised statistical learning techniques, which make neither a priori 14 assumptions on the relationship among the input variables nor a subjective decision on the variables to be possibly dropped. Further, our data-driven approach does not need to define a target variable, thereby avoiding a further risk of subjectivity. The only model assumption lays on the number of components built on the original variable space reflecting the desired level of captured variability and predictive ability. Moreover, the new coordinates must, by construction, lie on a linear space and be orthogonal (i.e., uncorrelated). No correlation ensures that each new principal component or dynamic factor describes a specific and unknown in advance latent phenomenon through the linear combination of the initial variables. We produce the index values with different methods, which allow policy makers to assess country preparedness according to specific needs and objectives.

Moreover, our paper contributes to the limited literature on the conceptualization and measurement of epidemiological risk (Gupta et al., 2018). However, most studies focus on epidemics forecasting and they do not explicitly consider the preparedness question. The key novelty of our ESR measure is the consideration of long-term, policy-relevant conditions, and not merely of the temporary incidence of diseases, affecting the contagion of epidemics. Our ESR index is not meant to predict an epidemic outbreak itself but rather the post-outbreak risk of contagion, largely reflecting the effect of policy. Finally, our analysis complements recent risk assessments based on the use of machine learning methods (Lin et al., 2012). Indeed, the authors stress that, beside the efficiency of the learning algorithm (often ensemble models do the job), the dataset, the selection of leading variables and the preprocessing phase in general play a key role in producing accurate assessments. We have placed special emphasis on these aspects in our analysis.

We organize the paper in the following parts: section **??** analyzes the related literature; section describes the methodological framework, the sources of data and the dimensionality reduction strategy; section presents the analytical results and introduces some robustness checks and finally section concludes the analysis of the paper.

Most efforts to contain the spread and effects of epidemics use the results of prediction models (Rivers et al., 2019, Polonsky et al., 2019). The prediction of the Covid-19 behavior has deployed sophisticated methods that include big data, social media information, stochastic models and data science/machine learning techniques along with medical (symptomatic and asymptomatic) parameters (Shinde et al., 2020, Nikolopoulos et al., 2021). However, prediction accuracy is limited due to the short period of data availability, data suitability, lockdown policies, difficulties in tracking the movement of people, changes in the incubation period and mutation of the virus, but also inappropriate algorithms and models.

The prediction of an epidemic establishes an alarm, which calls for a decision on what policy measures to undertake. The decision must be based on appropriate optimization of the prediction parameters, the likelihood of epidemic spread and its potential impact. Thus, it can be very complex and difficult, especially for locations with large and dense populations or critical infrastructure. Epidemics managers must factor prediction uncertainty into their decision-making models. However, while prediction methods have improved considerably and can handle increasing levels of complexity (Reich et al., 2019, Spreco et al., 2018, Debellut et al., 2018), prediction is essentially a short-term research enterprise. Instead, the overall preparedness of a country is a crucial long-term factor that guides the making of optimal decisions in response to an epidemic prediction.

The emergence of various epidemic outbreaks in the recent years led to the formulation of various country preparedness approaches that use different information and data aggregation methods. We briefly survey the most important ones. The Global Health Security Index (GHSI) represents a comprehensive assessment and benchmarking of health security and related capabilities of the countries that participate in the WHO's International Health Regulations. The GHSI is a joint project of the Nuclear Threat Initiative, the Johns Hopkins Center for Health Security, and The Economist Intelligence Unit (Johns Hopkins University Centre for Health Security, 2019). The GHSI provides a measure of a country's preparedness based on the capacity gaps of countries in their potential response to epidemics (T.Craig et al., 2020). However, the GHSI has been first published in 2019 and therefore it does not provide historical data to be used in thorough economic research. Further, the GHSI is too broad and includes global catastrophic and biological hazards, which on the one hand endows it with a broad coverage capacity but, on the other hand, make it less flexible and less suitable for a tool of prediction of epidemic-driven economic outcomes. Najmul (2020) find insignificant correlation between the GHSI and the incidence of Covid-19. After multiple testing, they suggest the inclusion of information on demographics and the reappraisal of its aggregation methodology. Razavi et al. (2020) argue that, while very comprehensive, the GHSI scoring may not be suitable for determining priorities and comparing countries with one another, calling for a further refinement of the index process that rationalizes the index's extensive focus on developed countries and health-related variables and its weighting methodology.

A related effort to assess country preparedness is the Joint External Evaluation (JEE) assessment tool. The latter is an externally validated, voluntary and collaborative assessment of 19 technical blocs of information necessary to validate the countries' capacity to detect and respond to public health risks (The World Health Organization, 2017). Unlike the GHSI, which allows inter-country comparisons, the JEE is a formal component of the WHO's Monitoring and Evaluation Framework, which all UN member states must implement. The JEE is not designed for making inter-country comparisons, but instead it is a technical tool for providing support to WHO member countries in setting quantified baseline thresholds for assessing progress. Shahpar (2019) use the average of the JEE's 19 technical areas for benchmark/comparison and argue that the JEE represents an initial effort at policy coordination that requires more global collaboration and prioritization of intervention. Garfield et al. (2019) tested the effectiveness of the JEE tool in a few African countries and found a high level of correspondence between score and policy text at the country level but also considerable differences in actual country responses relative to the benchmark JEE scores. They propose a better alignment of the JEE measures with the timing and depth of the country responses, which also reflect the contribution of international assistance in these areas.

Moreover, the Joint Research Centre (JRC), the European Commission's science and knowledge service, has cooperated with the World Health Organization to produce the Index for Risk Management (INFORM) (Doherty et al., 2018). The latter is a composite indicator that identifies countries at risk of humanitarian crisis and disaster that would overwhelm national response capacity and would be more likely to require international assistance. The INFORM model is based on risk concepts published in scientific literature and envisages three dimensions of risk: hazards and exposure, vulnerability, and lack of coping capacity. Risk components factored into the analysis include natural disasters, socioeconomic factors, such as inequality and aid dependency, and institutional capacity, such as built environment and access to health care. However, the INFORM framework does not adequately capture the effect of biological hazards (i.e., epidemic outbreaks). The INFORM Annual meeting 2017 in Rome agreed to proceed by incorporating ancillary information from the WHO epidemiological risk initiative relating to health components to improve the overall INFORM index (The INFORM Annual Meeting Report, 2017). The index measures a wide variety of hazard risks and less so epidemiological ones and its multi-level and complex construction also makes it less flexible and less suitable for use as a policy tool.

Another comprehensive effort to develop a preparedness index was expended by the U.S. Center for Disease Control and Prevention (CDCP). Following the emergence of various national hazards, the CDCP produced the National Health Security Preparedness Index at the U.S. state level to measure the preparedness (NHSPI, 2015). The NHSPI uses information from six broad domains of national health security (NHSPI, 2015, CDCP, 2014). The domains are the management of incident and information, the delivery of health-care services, the improvement of occupational and environmental health conditions, the management of countermeasures, community engagement and planning conditions, and the surveillance of health security conditions. After reviewing these occupational and environmental health domains, we observe no inclusion of indicators of occupational health and safety but only measures of environmental health. Overall, while the NHSPI is comprehensive, it covers only one country (the U.S.) for only a few years. Moreover, we do not find evidence of using the NHSPI to predict economic outcomes in the US economy.

Furthermore, E.Marcozzi et al. (2020) present a Hospital Medical Surge Preparedness Index (HMSPI) that can be used to systematically evaluate health care facilities across the U.S. states regarding their capacity to handle patient surges during disasters. The index aims to ensure that the US health care delivery system is poised to respond to mass casualty events by assessing the ability of victims to access health care (Kaji et al., 2008) as well as resolving weaknesses and reinforcing strengths in hospital and emergency management planning and capacity (Simiyu et al., 2014). The HMSPI uses four domains of surge capacity: staff, supplies, space, and integrated systems, and their subcomponents. However, the HMSPI is a static measure and of interest mainly to the US researchers.

Finally, Mertzanis and Papastathopoulos (2021) propose a composite index of epidemiological susceptibility risk, which they use to predict tourist flows around the world. They use information on time-varying, policy-relevant factors, such as infrastructure; demographics, economic activity and institutions, which they standardize and combine based on a standard PCA method to produce a continuous value index, using equal weights. While their index proves a significant predictor of tourist flows, their methodological approach is a rather simple one depriving their index from its full predictive potential. The authors acknowledge the need for using more sophisticated dimensionality reduction methods to achieve better results.

A common characteristic of the above preparedness measures is that they are composite indicators (CIs). Their construction involves stages where subjective judgments need to be made on the selection of indicators, the treatment of missing values, the choice of aggregation process and the weights of the indicators, etc. The unavoidable subjectivity involved in their construction may undermine their credibility and therefore it is important to identify the sources of subjectivity. However, the absence of an objective way to determine weights and the aggregation methods should not compromise their validity provided that the overall construction process is transparent (Nardo et al., 2005). This paper proposes a data-driven approach, which overcomes potential subjectivity bias in weight selection, takes into consideration dynamic effects and provides a better understanding of the complexity in approximating epidemic effects. After all, evidence-based evaluation of national epidemic management programs is critical to their future success (Koplan et al., 1999).

The conception of our ESR index originated in our observation that the spread of COVID-19 differed among countries. We observed that some countries fared better than others in containing the spread, regardless of their level of economic or institutional development, which was mainly the result of policy choices. Our index construction reflects our effort to include relevant policy variables. To this end, it reflects the importance of infrastructure, demographics, economic activity and governance (Najmul, 2020, Razavi et al., 2020, Mertzanis and Papastathopoulos, 2021).

The literature on epidemiological risk provides justification for these factors. First, quality health care infrastructure facilitates the timely detection and monitoring of infectious people in time and space, and therefore the successful containment of the epidemic (Morse, 2007). Global coordination increases monitoring efficiency. Moreover, quality health care infrastructure helps improve productivity and employment and hence production resilience, economic stability and social inclusion (Boyce and Brown, 2019). Adequate financing of health care infrastructure contributes decisively to its effectiveness (Kruk and Freedman, 2008).

Second, an effective communications infrastructure improves market surveillance, raises public awareness of epidemics risks and facilitates the swift private and public responses by assembling and broadcasting suitable information (Rainwater-Lovett et al., 2016). A new survey finds that about 53 percent of adults in the U.S. say that the internet has been essential for dealing with the pandemic, whilst 34 percent describe it as “important, but not essential” (Pew Research Center, 2020).

Third, an effective transportation infrastructure facilitates the monitoring and control of infectious population but also the response and timely provision of necessary care (Meyer and Elrahman, 2019). This is especially important with respect to passenger aviation that unavoidably contributes to the spread of an epidemic. Hufnagel et al. (2004) found a significant association between heterogeneity in airline connectivity networks and epidemic predictability.

Fourth, an effective infrastructure securing clean water and sanitation services is necessary for containing the speed and spread of epidemics and induces the health care sector’s response to adhere to high sanitary standards (D.Phelps et al., 2017). During epidemic outbreaks, the transmission of diseases occurs through both access to local water distribution facilities and the availability of man-made or natural water resources and sanitation systems. The OECD (2020) argues that enhancing environmental health through better air quality, water and sanitation, waste management, along with efforts to safeguard biodiversity, will reduce the vulnerability of communities to the effects of epidemics. KWR (2020) found that screening for Covid-19 at municipal wastewater plants in the Netherlands contributed to a better monitoring of its spread.

Fifth, demographics is also important. The increasing life expectancy and decreasing fertility rates change the patterns of consumption thereby affecting the dynamic of epidemics. For instance, Geard et al. (2015) argue that declining fertility rates are associated with an older mean age of disease infection that affects the spread of epidemics, depending on vaccination and other policy measures. Further, the rising urbanization rate globally affects epidemics in two ways (Neiderud, 2014): it causes improvements in health infrastructure in urban areas, but also provides a fertile ground for the emergence of new pathogens due to tighter human encounter. Population density is generally associated with a faster and wider spread of epidemics Tarwater and Martin (2001), Li et al. (2018).

Sixth, economic activity also affects the spread of epidemics. Relman et al. (2020) report the views of different experts on how travel, trade and conflict move people, animals and plants globally affecting the transmission of diseases. Adda (2015) finds that economic booms increase people's mobility among different transmission venues (ports, airports, etc.) and interpersonal economic interaction thereby contributing to a wider and faster spread of epidemics. Suhrcke et al. (2011) argue that economic downturns cause higher urbanization and congestion of people seeking jobs, worsening living and health care access conditions of living, which in turn lead to adverse epidemic effects. Kafertein (1997) argued that the rapid concentration of global food trade in a few multinational corporations increased the transmission of foodborne diseases. Lang (2001) stressed the effects of mass production and logistics procedures on the spread of infectious diseases.

Finally, institutional governance matters. Quah (2007) and Pritchett et al. (2013) document from different perspectives how institutional governance, exerted through various social interactions, social coordination and risk management policies, affect the spread of epidemics. However, the capacity of governance institutions develops differently among countries, subject to political influence, uncertainty or conflict (Gayer et al., 2007). The OECD (2010) argues that higher human capital improves governance and health outcomes through stronger social capital networks, employment prospects and psychological responses.

Methods

Sources of data. The preceding literature provides the broad directions and information for constructing the epidemiological susceptibility risk index (ESR). The index broadly captures the effects of the above-described building blocks of epidemiological risk. Following previous studies, we select objective and periodically reproducible variables that, given the relevant literature, best capture the extent to which a country may be susceptible to epidemiological risk and for which there is adequate and ongoing country coverage. The index does not model restrictions per se, but the objective outcome of restrictions in terms of people and products. Our initial dataset includes the values of 17 time-varying variables for 206 countries during the 2010-2019 period, classified in seven groups to construct the ESR index. To capture health infrastructure effects, we use (1) the value of health expenditure per capita (current USD); (2) the index value of health care access and quality; (3) the response rate to public health hazards; (4) the number of physicians per 1,000 people; and (5) the number of hospital beds per 1,000 people. To capture transport infrastructure effects, we use (6) the (inverse of the) number of air passengers as a percent of total population. To capture demographic effects, we use (7) the number of urban populations as a percent of total population; (8) the number of people per Km2 of land (population density); and (9) the population of 65+ years of age as a percent of total population. To capture environmental safety infrastructure effects, we use (10) the number of people using safely managed drinking water services as a percent of total population; (11) the number of people using safely managed sanitation services as a percent of total population. To capture relevant economic activity effects, we use (12) the value of trade in services as a percent of total trade and (13) the value of trade in goods as a percent of total trade. To capture communications infrastructure, we use (14) the number of individuals using the internet as a percent of total population. Finally, to capture institutional effectiveness, we use (15) the extent of human capital development; (16) the value of government effectiveness indicator and (17) the value of the rule of law indicator. The World Health Organization (WHO)[1] database provides the data for variables (1) to (4); the World Development Indicators (WDI)[2] database provides the data for variables (5) to (15); the Penn Tables (PT)[3] database provides the data for variable (16) and the Worldwide Governance Indicators (WGI)[4] database provides the data for variables (17) to (18). Tables A1 and A2 in the Appendix present the summary statistics of the index's constituent variables Var1 to Var17 and their pairwise correlations. In order to ensure the adequate sample size suitable for the presented methodologies we run the Kaiser–Meyer–Olkin test (Kaiser, 1970) resulting in the large score of 84.5%. Moreover, we run the Im-Pesaran-Shin test (Im et al., 2003) obtaining p-values $p \ll 0.01$ for both model specifications, i.e. "individual intercepts" and "individual intercepts and trends" for the underlying Augmented Dickey-Fuller test, implying the acceptance of alternative hypothesis of stationarity for the input variables time-series.

Higher values of these variables are associated with a lower risk of a country being susceptible to epidemiological contagion or, alternatively, they indicate better preparedness to manage these risks. While there are other relevant variables, the selected variables reflect factors and conditions that the literature has highlighted; they are objectively (not perceived) measured across countries, exhibit a low incidence of missing values and they are reproducible on a periodic basis. We did not include time-invariant factors (e.g., culture, religion, genetics) for we intend the index to capture mainly policy-relevant dynamic influences. For the same reason, we did not include time-varying factors relating to the environment conditions (e.g., temperature, rainfall) and slowly changing institutional factors (e.g., legal systems). We believe these factors should act as external controls mediating the predictive ef-

[1] https://www.who.int/data/collections

[2] https://databank.worldbank.org/source/world-development-indicators

[3] https://www.rug.nl/ggdc/productivity/pwt/?lang=en

[4] https://databank.worldbank.org/source/worldwide-governance-indicators

fectiveness of the ESR index on economic behavior rather than being constituent elements of the index itself. We do acknowledge the limitation of choosing certain variables than others or many more, but we had to draw the line somewhere. We do believe there is room for future improvements in the index's conceptualization and construction. An advantage of this construction is that our ESR index is mainly a policy-based and not a perceptions-based measure, which allows us to explore its effects on economic behavior largely devoid of perceptions, which would make it more severely prone to endogeneity.

Imputation of missing data. We next assess the data quality and completeness and address the problem of missing data. There are various imputation methods that are suitable for different data sets and conditions (Johnson and Young, 2011). There is a trade-off between data availability and the construction of a comprehensive composite index. We stress that the goal of our index construction is not to create artificial data series. We do not use the individual series as standalone predictors. Instead, we combine them to produce a single composite index that reflects consistently epidemiological susceptibility risk, rather than data availability. King et al. (2001) argue that imputation of missing data and their combination into aggregate indices is highly common in social sciences research because the nature of measured phenomena is associated with incomplete records. Typically, more data is available for larger countries and in recent years. We therefore restricted our sample to the 2010-2019 period and 168 countries in which the missing data tolerance rate does not exceed 40%, which gives a total of 28,560 country-year observations. Table A3 in the Appendix presents the full list of the sample countries and their rate of missing data. Table A4 in the Appendix presents the missing data incidence by year.

Since the presence of many missing values can extremely impact the quality and the reliability of results, we set an operational protocol of missing values treatment and imputation. In our final sample, 114 out of 168 countries show a rate of missing data between 20-39%. To address the missing values problem that would make possible the application of robust data aggregation methods, we test two different data imputation techniques: Matrix Completion with Low Rank SVD (MC-SVD) proposed by Hastie et al. (2015) and Bayesian Tensor Factorization (BTF) proposed by Khan and Ammad-ud-din (2016).

Briefly, MC-SVD solves the minimization problem $\frac{1}{2}\|X - AB^T\|_F^2 + \frac{\lambda}{2}(\|A\|_F^2 + \|B\|_F^2)$ for A and B where $\|\cdot\|_F$ is the Frobenius norm by setting to 0 the missing values. Once estimated, AB^T can approximate the original matrix X, including the missing values. This is applied on the 2-dimensional "slice" of countries-variables for each year. Subsequently, we apply the BTF method, which in addition uses a tensorial decomposition of the 3-dimensional tensors that stack all the annual “slices” together so that the imputation process involves information coming from a temporal dimension as well.

We assess comparative imputation performance by testing the imputation algorithm in three settings. In the first setting (named *Original*) we consider the whole dataset made of 168 countries and the 17 constituents variables over 10 years for a total of 28,560 entries. The full sample has 25% of missing values, thus we randomly remove some additional values, representing 10%, 20% and 30% of the initial dataset. In the second setting (named *No missing*) we drop all entries with missing values and apply the same incremental sampling procedure on the remaining subset. In the last setting (named *Some missing*) we drop all countries with at least 3 missing values for any year and apply again the incremental sampling procedure on the remaining subset. Furthermore, we fit the two methods, MC-SVD and BTF, on the previous 3 cases with different sampling percentages and we evaluate the Mean Absolute Reconstruction Error (MARE) on the excluded observations as follows:

$$MARE = \frac{1}{M}\sum_{i}^{M}|x_{excluded} - x_{reconstructed}|$$

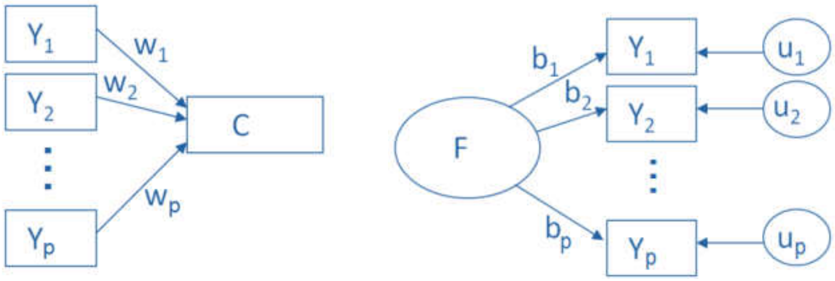
where M is the total number of excluded values. Moreover, we check the sensitivity to the original percentage of missing values by comparing the MARE on *No missing* and *Some missing* with the one on *Original*. Figure B.6 in the Appendix presents the imputation results, which show that the BTF method is the most efficient one for dealing with the missing value problem (for an exhaustive explanation, the reader can refer to Appendix B) .

Normalization of data. We remove differences in magnitude among the input variables by standardising the values, i.e. we subtract the mean and divide by the standard deviation. Having all variables on the same reference scale is crucial for unbiased estimation when applying dimensionality reduction techniques. Standardisation relates country performance of a variable as a bounded (by unitary standard deviation) variation from an average value (set to zero by definition) across all countries and years, which facilitates variable aggregation expressed in different measurement units. Further, when applying dimensionality reduction methods, component weights can have a significant effect on the overall composite indicator and country rankings. Several weighting techniques exist (Nardo et al., 2005). Some are based on statistical models (e.g., factor analysis), whilst others are based on participatory methods (e.g., analytical hierarchy process). Regardless of the method used, weights are essentially value judgments. However, our data-driven approach overcomes the problem of arbitrary and subjective choice of weights that could constrain the index's predictive efficacy.

Dimensionality reduction The aim of our analysis is to extract a synthetic indicator that summarizes at best the relationship among variables in a lower dimensional space. We apply two alternative but complementary statistical methodologies to reduce dimensionality and construct the index: Principal Component Analysis (*PCA*) and Factor Analysis (*FA*). *PCA* aims at creating new variables from a larger set of observed covariates, where each one is a linear combination of the Y original variables (see Figure 1a). The model is represented by the equation $C = w_1Y_1 + \ldots + w_iY_p$, where C is the new principal component, Y_i are the original variables and w_i are the weights of the linear combination for $i = 1, \ldots, p$.

FA, on the other hand, models the measurement of latent variables, seen through the relationships they cause in a set of Y variables (see Figure 1b). The model is represented by a set of equations $Y_i = b_iF_i + u_i, i = 1, \ldots, p$, where Y_i are the original variables, F_i are the latent factors and b_i, u_i are the parameters of the combination.

Recalling that our dataset has three dimensions, *Country*, *Variable* and *Time*, we



(a) *PCA:* the model is an equation $C = w_1 Y_1 + \ldots + w_p Y_p$

(b) *FA:* the model is a set of equations $Y_i = b_i F_1 + u_i, i = 1, \ldots, p$

Fig. 1. Principal Component Analysis and Factor Analysis

use PCA to model country/variable interaction for each year whereas FA to model country/time interaction, for all variables. Thus, using PCA, we create a low dimensional (1 way) indicator, explaining the maximum variance of the data and considering each year separately. Whereas, using FA, we estimate a single latent component able to capture the temporal interactions among the original variables. We describe the application of each dimensionality reduction method below in more detail.

We evaluate PCA on each year separately, producing T models. To ensure the stability and robustness of results, we apply and compare three different PCA techniques: regular PCA, Robust PCA and Robust Sparse PCA. PCA aims at finding new and wise linear combinations of the original data, in a way that the amount of explained variance of the data is maximised. Those combinations are mathematically constrained to be mutually orthogonal (that is uncorrelated) and are called Principal Components (PC) or loadings. Given a $n \times p$ data matrix $\mathbf{X}$, where n is the number of observations and p is the number of variables, we want to find the $k \times p$ Principal Component matrix C, with usually $k << p$ such that the projected data matrix $W = XC^T$, also called scores matrix, will have dimension $n \times k$. The maximization problem is stated as follows:

$$\begin{aligned} \underset{\mathbf{C}}{\text{minimize}} \quad & \|\mathbf{X} - \mathbf{X}\mathbf{C}\mathbf{C}^T\|_F^2 \\ \text{subject to} \quad & \mathbf{C}^T\mathbf{C} = \mathbf{I} \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. We implement the model using *R* package `prcomp`. Since we do not rely on the classical PCA but, rather, we seek for a robust estimation of the Principal Components, we can decompose the data matrix *X* into a low rank component *L* that represents the intrinsic low dimensional features and an outlier component *S* that captures anomalies in the data. The maximization problem is stated as follows:

$$\begin{aligned} \underset{\mathbf{L},\mathbf{S}}{\text{minimize}} \quad & ||\mathbf{L}||_* + \lambda ||\mathbf{S}||_1 \\ \text{subject to} \quad & \mathbf{L} + \mathbf{S} = \mathbf{X} \end{aligned}$$

where $||L||_*$ is the nuclear norm and λ is a penalization term. Following the procedure of Candes et al. (2009), once fitted, **L** can be used as a proxy for **X** with the extreme values excluded. Finally, following Erichson et al. (2018), we produce both a robust estimation and a sparse representation of the principal components by adding a sparsity constraint on the matrix *C*. The associated maximization problem is stated as follows:

$$\begin{aligned} \underset{\mathbf{C},\mathbf{W}}{\text{minimize}} \quad & ||\mathbf{X} - \mathbf{W}\mathbf{C}^T - \mathbf{S}||_F^2 + \psi(\mathbf{C}) + \phi(\mathbf{W}) + \lambda ||\mathbf{S}||_1 \\ \text{subject to} \quad & \mathbf{C}^T\mathbf{C} = \mathbf{I} \end{aligned}$$

ψ and ϕ are regularizing functions (i.e. LASSO or Elastic Net).

Dynamic Factor Model. Moreover, we evaluate a temporal dependent version of FA called Dynamic Factor Model (DFM), using all the available years within the same model. Given the $p \times n$ matrix **X**, the model assumes that there exist some $k \times n$ factors **F** such that their mutual interaction over time can be expressed by a $k \times k$ interaction matrix **A** and the observed variable can be expressed as a linear function of the factors themselves through a $p \times k$ loading matrix **C**. The problem can be solved as a system of equations:

$$\begin{cases} \mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \mathcal{N}(0, \mathbf{Q}) \\ \mathbf{X}_t = \mathbf{C}\mathbf{F}_t + \mathcal{N}(0, \mathbf{R}) \end{cases} \tag{1}$$

where $\mathscr{N}$ is the normal probability distribution and $\mathbf{Q}$ and $\mathbf{R}$ are the covariance matrix of the residuals of each equation in (1), respectively. Due to the short time series of the input variables, this model cannot be fitted considering all countries together as the resulting system of equations (1) is under-determined. Thus, we deal with the problem as follows: first, following Holmes et al. (2018), we fit DFM for each country, obtaining the factor matrices $\mathbf{F}^i$, the factor interactions $\mathbf{A}^i$ and the factor loadings $\mathbf{C}^i$, $i = 1, \ldots, n$. Second, we fit a Vector Auto Regressive (VAR) model in order to get $\hat{\boldsymbol{A}}$ 1-year lag matrix that incorporates cross-countries interactions of $\mathbf{A}^i$. We implement the model using *R* package `sparsevar` because this calibration problem has too many parameters to estimate relative to the number of observations, thus requiring a sparse approach. Finally, we use Kalman Filter to get smoothed factors $\widehat{\mathbf{F}^i}$ using $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{C}} = diag(\boldsymbol{C}^i)$, that is to get latent factors that incorporate cross-countries interactions. Briefly, Kalman filter re-estimates the factor matrix $\mathbf{F}$ iterating the two equations in (1) until the error between the predicted observed variables $\hat{\boldsymbol{X}}$ and the true one is minimized. We implement the model using *R* package `FKF`. We assume $\hat{\boldsymbol{C}}$ to be diagonal in order not to double-count correlations within the observed variables and because cross-country interactions are already modelled through the VAR.

In both cases (PCA and DFM), the final index ESR will be represented by the scores matrix W and the factor matrix F respectively, both k-dimensional. One of the goal is to select the optimal number of components k as a trade-off between the maximal explained variance and the smallest value of components k.

Validation. Applying a dimensionality reduction technique by merely maximising the amount of explained variance with the smallest set of components, could be misleading and conduct to hardly interpretable results. Thus, once identified the most reliable results, we compare the fitting power of the produced indexes to a baseline benchmark. We accordingly estimate several parametric and non-parametric regression models to produce comparisons of the produced ESR index with the original set of variables. We use, as target variable, the following macro-economic variables: real GDP per capita, government consumption (percent of total), price level of capital formation, trade volume, unemployment rate, outstanding loans of commercial banks. Our validation process aims at demonstrating the relevance of the new index in representing the information conveyed by the original component variables. If the modeling ability of the composite ESR index, measured by the root mean square error (RMSE), is comparable to the original one based on the initial variables, we can conclude that the produced indicator is not only satisfactory according to the chosen dimension reduction technique but also effective in terms of predictive power within a simplified framework.

Results

We standardize the dataset for each year and then we apply first the PCA method in all different versions, as previously described. Table 1 reports the results of the different PCA versions. We report the average variance explained by loadings across years, as well as the average R^2 on both the whole dataset and subsets with values trimmed for the $95th$ and $99th$ percentiles in order to check for outliers impact. In our context R^2 means the ratio of the amount of variance explained by our retained components over the total variance contained in the original variables. Moreover, we run the Im-Pesaran-Shin test on the PCA index and $p - values \ll 0.01$ for all model specifications ensure its stationarity. The stationarity is important because we can infer that the changes over time, which the index is expected to capture, can be statistically robust and not caused by any trend in the data or mean-reversion effects. The results show that the robust PCA method performed best regardless the employed data (full data set, 1% trimmed and %5 trimmed). Accordingly, we retain only the first principal component, which explains at its minimum a remarkable 87% of the total variance and therefore renders the resulting ESR index visually interpretable. Figure 2 shows the scree plots of the variance explained by the loadings using the robust PCA method only. Figures C.7 through C.9 in the Appendix report the full comparison among all PCA methods as well as the relative importance of the loadings. This includes the percent of variance explained by the first principal component of each PCA method per year.

Table 1
Results from Robust PCA. Mean is evaluated over years. Mean Explained Variance is evaluated from the eigenvalues of PCA, R^2 is reported for the full dataset and for the $99th$ and $95th$ percentiles. Im-Pesaran-Shin test for stationarity on the ESR index as well.

Method	Number of PC	Mean Explained Variance	Mean R^2	Mean R^2 on 99th	Mean R^2 on 95th	Im-Pesaran-Shin test
PCA	1	$49.9 \pm 0.9\%$	$49.9 \pm 0.9\%$	$57.3 \pm 1.1\%$	$65.3 \pm 0.9\%$	$\ll 0.01$
RobPCA	1	$87 \pm 0.9\%$	$94.8 \pm 0.3\%$	$95.4 \pm 0.2\%$	$96.5 \pm 0.2\%$	$\ll 0.01$
RobSparPCA	1	$50.2 \pm 0.9\%$	$28.5 \pm 3\%$	$33.6 \pm 3.6\%$	$38.2 \pm 4.5\%$	$\ll 0.01$

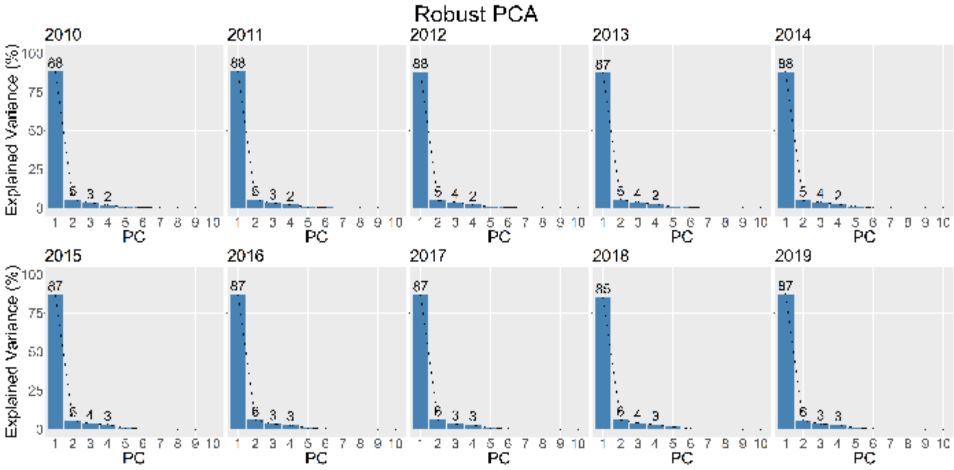


Fig. 2. Scree plot for Robust PCA method.

Figure 2 clearly shows how important is the first component whatever year we take into account. Such result has several important implications: PCA proves that there exists a strong latent component which is highly connected to almost all the variables. Moreover, the possibility of building up our ESR index considering just one component eases the interpretation, the relative employment and the subsequent monitoring.

Then we apply the DFM method, as previously described, which depends upon two hyper-parameters: the sparsity coefficient α of the VAR and the correlation structure of the residuals for Kalman filter. Thus, we simulate synthetic factors $\widetilde{\mathbf{F}}$ with different combinations of number of observed variables, countries, years, latent factors $\mathbf{F}$, and we generate the corresponding $\mathbf{X}_t$ given different combination of $\mathbf{A}$, defined by α, and $\mathbf{C}$, randomly generated, using equation (1). Then, for each of the previous combination and correlation structure of residuals $\mathbf{Q}$, we apply the described algorithm and assess the reconstruction error on the fitted factors $\widetilde{\mathbf{F}}$ with the simulated factors $\mathbf{F}$. The optimal parameters found are $\alpha = 0.2$ and a diagonal structure. Afterwards, we evaluate the R^2 on DFM model. Table 2 reports the DFM results. In this case, the poorer performance is due to the small size of the dataset compared to the number of parameters, despite mitigated with sparseness. Moreover, the estimated interactions factor in $\hat{\boldsymbol{A}}$ turns out to be very small (values range in $[-0.06, 0.05]$), so we assume to be valid the no interactions setting, which has produced the highest R^2 (73.6%). We run the Im-Pesaran-Shin test also on the DFM based index obtaining p-values $\ll 0.01$ for both model specifications and ensuring its stationarity as for the PCA case. Figure D.10 in the Appendix shows the relative importance of the loadings for the DFM model with interpretation.

Table 2
Results for DFM. R^2 is reported for the full dataset and for the $99th$ and $95th$ percentiles. We also report Im-Pesaran-Shin test for stationarity on the ESR index.

Method	Number of Factors	R^2	R^2 on 99th	R^2 on 95th	Im-Pesaran-Shin test
DFM with interactions	1	−204.5%	−43.8%	7.7%	$\ll 0.01$
DFM without interactions	1	−405.4%	38.6%	73.6%	$\ll 0.01$

As robustness check, we compare the two ESR index values generated by the competing methods in terms of predictive power within a supervised analysis setting. To this end, we use the following macro-economic variables: real GDP per capita, government consumption (percent of total), price level of capital formation, trade volume, unemployment rate and outstanding loans of commercial banks. We standardize the target variables before fitting the algorithms to make the results comparable. We use both linear and non-linear data-driven learning algorithms to capture potential non-linearity effects in the data. We use alternatively the learning techniques of Random Forest, Regularized OLS (Elastic-Net), Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Multivariate Adaptive Regression Spline (MARS) and a single layer Neural Network (NN). All the hyper-parameters are tuned with Bayesian Optimization and a 5-fold cross-validation. When fitting Elastic-Net with a single regressor, we use the OLS regression. Final performances are evaluated using a further 5-folds cross-validation and the average test set Root Mean Square Error (RMSE) is considered. The seed used to select the cross-validation folds has been kept fixed for all algorithms in order to ensure reproducible results.

We provide examples of the comparison results. Table 3 shows the RMSE percent increase in predicting Unemployment rate with the single index as regressor compared to the RMSE obtained with all 17 original variables. RMSE of models which are fitted considering ESR index solely tends to increase as we would reasonably expect. However, RMSE increases are always within one standard deviation bound suggesting that a much simplified analysis based on 1 unique index is significant and largely satisfies the parsimony principle. Table 3 clearly shows that Random Forest has the lowest RMSE by employing the original 17 variables (0.079) and further the ESR index based on the DFM approach presents the minimum RMSE (0.447).

Complete results for all the fitted regressions are reported in Appendix D.1.

Table 3
RMSE in predicting Unemployment rate using continuous index as regressor. RMSE for regression with original variables is reported in parenthesis.

	RMSE index (RMSE original)	
Algorithm	DFM	Robust PCA
Elastic-Net	0.999(0.859)	0.995(0.859)
MARS	1(0.583)	0.924(0.583)
Random Forest	0.447(0.079)	0.7(0.079)
Single Layer NN	0.994(0.31)	0.932(0.31)
SVM-RBF	1.024(0.083)	0.936(0.083)

Further, we can provide useful visual insights by exploring the temporal evolution of the ESR index values for each country in a world map. The animated Figures 3 and 4 report the global distribution of the ESR index for both the PCA and DFM methods, respectively.

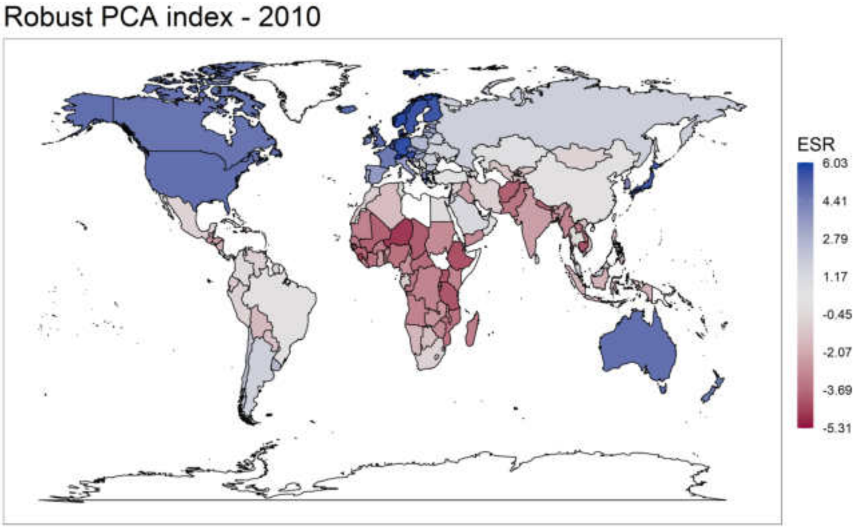


Fig. 3. Robust PCA index evolution over years. Shades of red color refer to riskier countries, while shades of blue to safer ones.

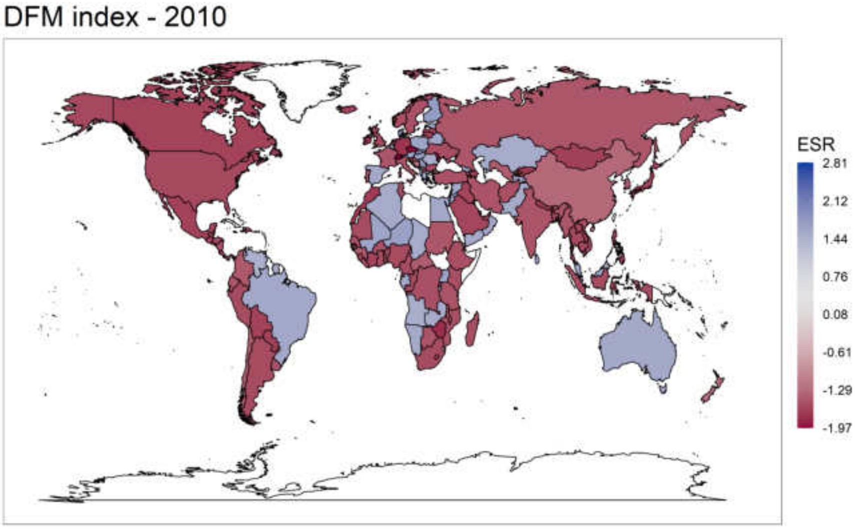
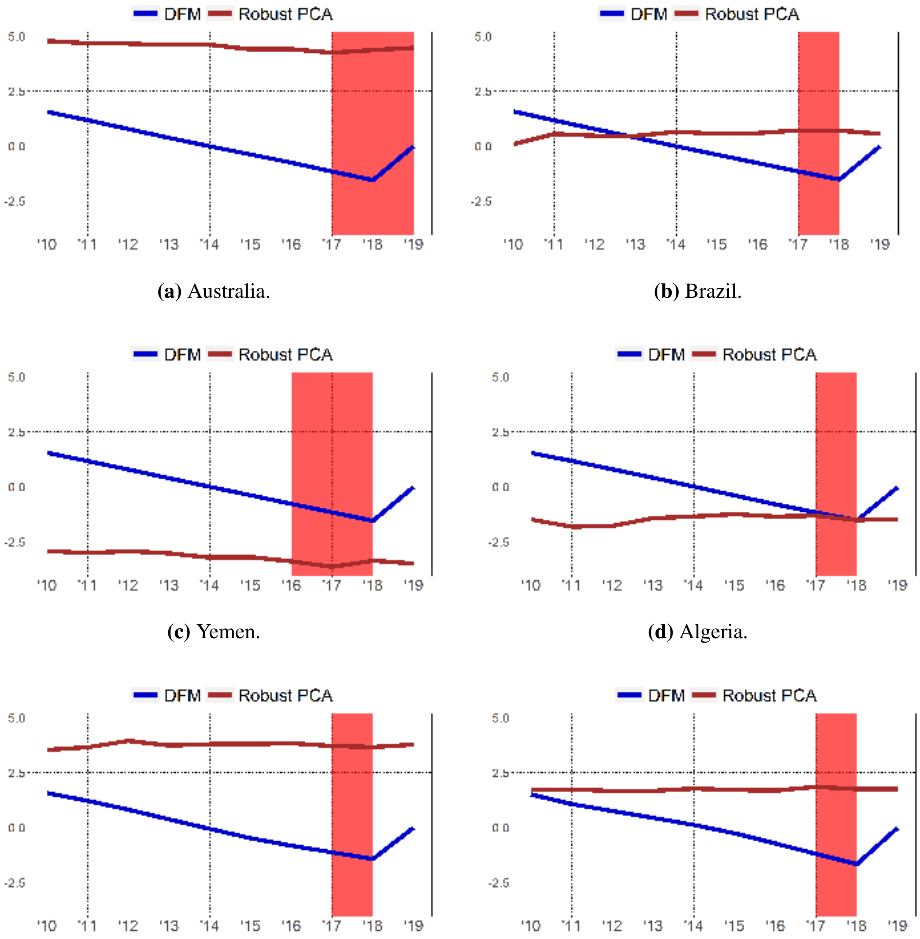


Fig. 4. DFM index evolution over years.Shades of red color refer to riskier countries, while shades of blue to safer ones.

Indeed, the native characteristic of DFM of properly modeling the temporal dynamics is reflected in the world map which presents more variability in the colour change compared to PCA.

Finally, Figure 5 shows the evolution over time of the ESR index for some individual countries, comparing the PCA and DFM methods. The PCA index is quite stable over time, whilst the DFM index captures the time dynamics of underlying latent factors. For example, Figure 5a shows that our index can capture the abnormal increase of Influenza cases in 2018-19 in Australia. In Figure 5b ESR index highlights the Zika virus outbreak of 2018 in Brazil. In Figure 5c the index underlines the Cholera spread between 2016 and 2018 in Yemen. Cholera outbreak in 2018 is captured for Algeria as well as shown in Figure 5d. Similarly Figure 5e and Figure 5f show how the index is able to capture the abnormal Influenza spread of 2018 and the increase of Measles case in 2018 in Spain and Romania respectively. Figure D.17 to D.20 in the Appendix provide the detailed evolution of the ESR index per country during the 2010-2019 period using both PCA and DFM methods.



(a) Australia.

(b) Brazil.

(c) Yemen.

(d) Algeria.

(e) Spain.

(f) Romania.

Fig. 5. Index evolution over years for some countries. Disease outbreaks are shaded in red.

Discussion

Epidemic outbreaks are extreme events that become less rare and more severe. The COVID-19 pandemic is an extreme risk event that has unfolded with tremendous speed and breadth. Epidemics cause huge economic costs for firms and countries. It is therefore important to evaluate the extent to which countries can identify and manage epidemiological risks adequately. Despite significant improvements in infrastructure and governance worldwide, many countries remain unprepared to adequately identify and manage epidemiological risks. In this study, we have proposed a country preparedness evaluation framework that countries and firms could use to manage the contagion and consequences of epidemic risks. The framework is based on the development of a composite indicator, which we call epidemiological susceptibility risk index (ESR), for 168 countries during 2010-2019.

In constructing our ESR measure, we use objective and regularly reproduced information that accounts for the role of infrastructure, economic activity, demographics and governance institutions. This integrated view of measuring epidemiological risk is in line with the general directions proposed by the WHO. We complement previous efforts at assessing country preparedness by proposing a methodological framework that makes the assessment of preparedness more policy-driven and expanded around the world. Importantly, our proposed framework uses a data-driven approach to constructing the index that utilizes both PCA and DFA methods and their variants for achieving dimensionality reduction. The results show that, after accounting for data characteristics and missing values, the robust PCA method shows very good performance whereby the first dimension explains about 90% of total variability. However, the nature of its construction prevents it from capturing properly the temporal latent dynamic of the data. We therefore use the alternative DFA method for this purpose. Albeit somewhat less efficient in comparative terms (the first component explains about 76% of the total variability), the DFA method must be considered as the benchmark model since it properly models the temporal dynamics, which are important in capturing epidemic outbreaks across a wide range of countries during the 10 available years. Our ESR index is fully data-driven that does not allow for arbitrary and subjective choice of weights that could impair its predictive efficiency.

This framework and index could provide the basis for developing risk assessments of epidemiological risk contagion after the outbreak of an epidemic but also for ongoing monitoring of its spread and social and economic effects. It would also allow for useful comparisons in country preparedness and performance. This framework and index could be used by firms to assess likely economic consequences of epidemics and could therefore have managerial implications. For example, in addition to help managing epidemiological risk, the framework could be useful in aligning country and corporate policy to environmental sustainability considerations and responsible behavior. Further, it takes into consideration ongoing regulatory initiatives that stress the importance of non-financial risks due to climate change.

Finally, our framework could be revised and extended towards various directions to support decision making. One way to improve it is to increase the data series availability mindful of the missing data problem using more advanced techniques. Another way to extend it includes the addition of new relevant dimensions that may capture other aspects of epidemiological risk. As research on the sources and spread of Covid-19 continues, new information is being revealed, which might inform the re-construction of our ESR index. Another way would be to apply alternative data dimensionality reduction techniques and compare the predictive results. The extensive check on the index's predictive power remains to be accomplished by applying it to diverse real-world situations.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request.

Code availability

The custom code that supports the findings of this study is available from the corresponding author upon request and available alongside the data in the permanent repository indicated above.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program "PERISCOPE: Pan European Response to the ImpactS of COvid-19 and future Pandemics and Epidemics”, under the Grant Agreement No. 101016233, H2020-SC1-PHE-CORONAVIRUS-2020-2-RTD.

References

Adda, J., 2015. Economic activity and the spread of viral diseases. evidence from high frequency data. *IZA Discussion Paper*, 9326.

Boyce, T. and Brown, C., 2019. Economic and social impacts and benefits of health systems. *World Health Organization, Regional Office*.

Candes, E. J., Li, X., Ma, Y., and Wright, J., 2009. Robust principal component analysis?

CDCP, 2014. Board of scientific counselors bsc meeting: Summary report/record of the proceedings. *Centers for Disease Control and Prevention.* URL http://www.cdc.gov/phpr/science/documents/bsc_ophpr_meeting_minutes_04_07_14.pdf.

Debellut, F., Hendrix, N., Ortiz, J. R., Lambach, P., Neuzil, K. M., and Bhat, N., 2018. Forecasting demand for maternal influenza immunization in low- and lower-middle income countries. *PLoS One*, 13.

Doherty, B., Marin-Ferrer, M., and Vernaccini, L., 2018. The inform epidemic risk index.

D.Phelps, M., Azman, A. S., Lewnard, J. A., Antillon, M., Simonsen, L., Andreasen, V., and andV. E. Pitze, P. K. M. J., 2017. The importance of thinking beyond the water supply in cholera epidemics. a historical urban case study. *PLoS Neglected Tropical Diseases*, 11:1–15.

E.Marcozzi, D., Lawler, R. P. J. V., French, M. T., Mecher, C., Baehr, J. P. N. E., and Browne, B. J., 2020. Development of a hospital medical surge preparedness index using a national hospital survey. *Health Services and Outcomes Research Methodology*, 20:60–83.

Erichson, N. B., Zheng, P., Brunton, K. M. S. L., Kutz, J. N., and Aravkin, A. Y., 2018. Sparse principal component analysis via variable projection.

Garfield, R., Bartee, M., and Mayigane, L. N., 2019. Validating joint external evaluation reports with the quality of outbreak response in ethiopia, nigeria and madagascar. *The BMJ Global Health*, 4:1–8.

Gayer, M., Legros, D., Formenty, P., and Connolly, M. A., 2007. Conflict and emerging infectious diseases. *Emerging Infectious Diseases*, 13:1625–31.

Geard, N., Glass, K., McCaw, J. M., McBryde, E. S., Korb, K. B., Keeling, M. J., and McVernon, J., 2015. The effects of demographic change on disease transmission and vaccine impact in a household structured population. *Epidemics*, 13:56–64.

Gupta, V., Kraemer, J. D., Katz, R., Jha, A. K., Kerry, V. B., Sane, J., Ollgren, J., and Salminen, M. O., 2018. Analysis of results from the joint external evaluation, examining its strength and assessing for trends among participating countries. *Journal of Global Health*, 8:1–9.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R., 2015. Matrix completion and low-rank svd via fast alternating least squares.

Holmes, E. E., Ward, E. J., and Scheuerell, M. D., 2018. Analysis of multivariate time-series using the marss package. URL https://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf.

Hufnagel, L., Brockmann, D., and Geisel, T., 2004. Forecast and control of epidemics in a globalized world. *PNAS*, 101:15124–15129.

Im, K. S., Pesaran, M. H., and Shin, Y., 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115:53–74.

Johns Hopkins University Centre for Health Security, 2019. Forecasting demand for maternal influenza immunization in low- and lower-middle income countries. *Global Health Security Index.* URL https://www.ghsindex.org/.

Johnson, D. R. and Young, R., 2011. Toward best practices in analyzing datasets with missing data, comparisons and recommendations. *Journal of Marriage and Family*, 73:926–45.

Kafertein, F. K., 1997. Foodborne disease control: A transnational challenge. *Emerging Infectious Diseases*, 3:503–510.

Kaiser, H. F., 1970. A second generation little jiffy. *Psychometrika*, 35:401–415.

Kaji, A. H., Langford, V., and Lewis, R. J., 2008. Assessing hospital disaster preparedness: A comparison of an on-site survey, directly observed drill performance, and video analysis of teamwork. *Annals of Emergency Medicine*, 52: 195–201.

Khan, S. A. and Ammad-ud-din, M., 2016. tensorbf: an r package for bayesian tensor. URL https://www.biorxiv.org/content/biorxiv/early/2016/12/29/097048.full.pdf.

King, G., Honaker, J., Joseph, A., and Scheve, K., 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95:49–69.

Koplan, J. P., Milstein, R., and Wetterhall, S., 1999. Framework for program evaluation in public health. *Morbidity and Mortality Weekly Report*, 48:1–40.

Kruk, M. E. and Freedman, L. P., 2008. Assessing health system performance in developing countries. a review of the literature. *Health Policy*, 85:263–76.

KWR, 2020. What we learn about the corona virus through waste water research. URL https://www.kwrwater.nl/en/actueel/what-can-we-learn-about-the-corona-virus-throughwaste-water-research.

Lang, T., 2001. Trade, public health and food. *International Cooperation in Health*, pages 81–108.

Li, R., Richmond, P., and Roehner, B. M., 2018. Effect of population density on epidemics. *Physica A: Statistical Mechanics and its Applications*, 510:713–724.

Lin, W. Y., Hu, Y. H., and Tsai, C. F., 2012. Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 42(4):421–436.

Mertzanis, C. and Papastathopoulos, A., 2021. Epidemiological susceptibility risk and tourist flows around the world. *Annals for Tourism Research*, 86.

Meyer, M. D. and Elrahman, O. A., 2019. transportation and public health. an integrated approach to policy, planning, and implementation. *The National Academy of Sciences, Engineering and medicine.* URL https://www.elsevier.com/books/transportation-and-public-health/elrahman/978-0-12-816774-8.

Morse, S. S., 2007. Global infectious disease surveillance and health intelligence.

Health Affairs, 26:1069–77.

Najmul, H., 2020. The global health security index and joint external evaluation score for health preparedness are not correlated with countries' covid-19 detection response time and mortality outcome. *Epidemiology and Infection*, 148:1–8.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., and Giovannini, E., 2005. Handbook on constructing composite indicators: methodology and user guide. *OECD Statistics Working Paper STD/DOC.*

Neiderud, C. J., 2014. How urbanization affects the epidemiology of emerging infectious diseases. *Infection Ecology and Epidemiology*, 5:1–9.

NHSPI, 2015. The national health security preparedness index. *The BMJ Global Health.* URL http://www.nhspi.org/about/.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., and Vasilakis, C., 2021. Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, 290:99–115.

Pew Research Center, 2020. The role of the internet during the covid-19 outbreak. *Pew Research.* URL https://www.pewresearch.org/internet/2020/04/30/53-of-americans-say-the-internet-has-been-essential-during-the-covid-19-outbreak.

Polonsky, J. A., Baidjoe, A., Kamvar, Z. N., Cori, A., Durski, K., and Edmunds, W. J., 2019. Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 374.

Pritchett, L., Woolcock, M., and Andrews, M., 2013. Looking like a state, techniques of persistent failure in state capability for implementation. *Journal of Development Studies*, 49:1–18.

Quah, S. R., 2007. Crisis preparedness: Asia and the global governance of epidemics. *Washington, D.C.: The Brookings Institution.*

Rainwater-Lovett, K., Rodriguez-Barraquer, I., and Moss, W. J., 2016. Tracking viruses with smartphones and social media. in katze, m. g., korth, m. j., lyn, g. nathanson, n. (eds) viral pathogenesis. *Basics to Systems Biology.*

Rajabi, M., Mansourian, A., Pilesjo, P., Hedefalk, F., Groth, R., and Bazmani, A.,

2014. Comparing knowledge-driven and data-driven modeling methods for susceptibility mapping in spatial epidemiology: a case study in visceral leishmaniasis. in huerta, j., schade, s., granell, c. (eds) connecting a digital europe through location and place. *Proceedings of the AGILE'2014. International Conference on Geographic Information Science*, pages 3–6.

Razavi, A., Erondu, N. A., and Okereke, E., 2020. The global health security index: what value does it add? *The BMJ Global Health*, 5:1–3.

Reich, N. G., Brooks, L. C., Fox, S. J., Kandula, S., McGowan, C. J., and Moore, E., 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proc. Natl. Acad. Sci. U. S. A.*, 116:3146–3154.

Relman, D. A., Choffnes, E. R., and Mack, R., 2020. Infectious disease movement in a borderless world: Workshop summary. *Forum on Microbial Threats; Institute of Medicine of the National Academies*.

Rivers, C., Chretien, J. P., Riley, S., Pavlin, J. A., Woodward, A., and Brett-Major, D., 2019. Using outbreak science to strengthen the use of models during epidemics. *Nat. Commun.*, 10.

Shahpar, C., 2019. Protecting the world from infectious disease threats: now or never. *BMJ Global Health*, 4.

Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, I., and Hassanien, A. E., 2020. Forecasting models for coronavirus disease covid-19: A survey. *SN Computer Science 1*, 197.

Simiyu, C. N., Odhiambo-Otieno, G., and Okero, D., 2014. Capacity indicators for disaster preparedness in hospitals within nairobi county, kenya. *Pan Afr Med Journal*, 18:349.

Spreco, A., Eriksson, O., Dahlstrom, O., Cowling, B. J., and Timpka, T., 2018. Evaluation of nowcasting for detecting and predicting local influenza epidemics, sweden, 2009-2014. *Emerging Infect. Dis.*, 24:1868–1873.

Suhrcke, M., Stuckler, D., Suk, J. E., Desai, M., Senek, M., McKee, M., Tsolova, S., Basu, S., Abubakar, I., and Hunter, P., 2011. The impact of economic crises on communicable disease transmission and control, a systematic review of the evidence. *PLoS ONE*.

Tarwater, P. M. and Martin, C. F., 2001. Effects of population density on the spread of disease. *Complexity*, 6:29–36.

T.Craig, A., Heywood, A. E., and Hall, J., 2020. Risk of covid-19 importation to the pacific islands through global air travel. *Epidemiology and Infection*, 148.

The INFORM Annual Meeting Report, 2017. URL https://docs.google.com/document/d/13B8L7_XQLWNxp_vZa8Dbow2cWLUOo2UGoSN083yKSZs/edit?usp=sharing.

The International Monetary Fund, 2020. Exceptional times, exceptional action. *Opening Remarks for Spring Meetings Press Conference*. URL https,//www.imf.org/en/News/Articles/2020/04/15/sp041520-exceptional-times-exceptional-action.

The OECD, 2010. Social capital, human capital and health: What is the evidence?

The OECD, 2020. Environmental health and strengthening resilience to pandemics.

The World Health Organization, 2017. Joint external evaluation jee. *Zoonotic Diseases Action Package Conference.*

The World Health Organization, 2020d. Emergencies preparedness, response. sars-cov-2 variants. URL https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/.

Appendix A. List of variables and countries

Table A1
List of used variable. Sources are World Health Organization (WHO), World Bank's Development Indicators (WDI), Penn Tables (PT) and World Bank's Worldwide Governance Indicators (WGI).

Variable	Description	Source	Total Obs.	Missing Values	Min	Max	Mean	Median	Standard Deviation
var1	health care exenditure per capita	WHO	1,680	523 (31%)	12.64	10,014.71	1,077.66	317.86	1,821.28
var2	health care access and quality	WHO	1,680	20 (1.2%)	28.60	93.60	62.97	62.55	16.49
var3	response level (%) to public health hazards	WHO	1,680	670 (40%)	0.00	100.00	66.15	73.00	30.61
var4	num of physicians per 1000 people	WDI	1,680	941 (56%)	0.00	6.11	2.01	2.05	1.42
var5	num of hospital beds per 1000 people	WDI	1,680	1175 (70%)	0.10	13.40	3.25	2.70	2.31
var6	num of air passengers to population ratio	WDI	1,680	397 (24%)	0.00	34.53	1.18	0.29	2.87
var7	num of urban pop (% of total)	WDI	1,680	168 (10%)	10.64	100.00	58.91	59.48	22.19
var8	num of people per Km2 (pop density)	WDI	1,680	177 (11%)	1.75	7,953.00	231.38	81.13	808.73
var9	num of people age 65% (% of total)	WDI	1,680	177 (11%)	0.69	27.58	8.38	6.20	5.93
var10	num of people using drinking water services (% of pop)	WDI	1,680	340 (20%)	33.05	100.00	86.51	94.72	16.70
var11	num of people using safely-managed drinking water services (% of pop)	WDI	1,680	952 (57%)	6.19	100.00	76.98	91.52	26.98
var12	num of people using safely-managed sanitation services (% of pop)	WDI	1,680	1024 (61%)	7.45	100.00	66.76	76.01	28.97
var13	human capital index	WDI	1,680	218 (13%)	0.00	4.01	2.51	2.64	0.84
var14	num of people using the internet (% of pop)	WDI	1,680	257 (15%)	0.25	100.00	45.73	45.96	29.14
var15	value of trade (% GDP)	PT	1,680	103 (6.1%)	0.20	442.62	91.30	79.51	58.12
var16	government effectiveness index	WGI	1,680	20 (1.2%)	-2.28	2.24	0.02	-0.08	0.97
var17	rule of law index	WGI	1,680	20 (1.2%)	-2.32	2.10	-0.03	-0.24	0.98

Table A2
Correlation matrix of input variables.
var1 is health care exenditure per capita, var2 is health care access and quality, var3 is response level (%) to public health hazards, var4 is num of physicians per 1000 people, var5 is num of hospital beds per 1000 people, var6 is num of air passengers to population ratio, var7 is num of urban pop (% of total), var8 is num of people per Km2 (pop density), var9 is num of people age 65% (% of total), var10 is num of people using drinking water services (% of pop), var11 is num of people using safely-managed drinking water services (% of pop), var12 is num of people using safely-managed sanitation services (% of pop), var13 is human capital index, var14 is num of people using the internet (% of pop), var15 is value of trade (% GDP), var16 is government effectiveness index, var17 is rule of law index.

	var1	var2	var3	var4	var5	var6	var7	var8	var9	var10	var11	var12	var13	var14	var15	var16
var2	0.66*															
var3	0.35*	0.5*														
var4	0.57*	0.75*	0.44*													
var5	0.32*	0.52*	0.27*	0.64*												
var6	0.33*	0.32*	0.17*	0.21*	0.01											
var7	0.48*	0.7*	0.39*	0.58*	0.31*	0.22*										
var8	-0.04	0.15*	0.13*	-0.04	-0.02	0.2*	0.19*									
var9	0.61*	0.79*	0.38*	0.76*	0.67*	0.13*	0.46*	0.07*								
var10	0.41*	0.79*	0.42*	0.65*	0.37*	0.22*	0.64*	0.12*	0.61*							
var11	0.43*	0.53*	0.22*	0.5*	0.31*	0.26*	0.5*	0.14*	0.49*	0.36*						
var12	0.38*	0.29*	0.22*	0.27*	0.28*	0.16*	0.29*	0.14*	0.27*	-0.09*	0.72*					
var13	0.53*	0.67*	0.43*	0.54*	0.46*	0.16*	0.5*	0.12*	0.62*	0.57*	0.35*	0.34*				
var14	0.63*	0.86*	0.51*	0.66*	0.48*	0.36*	0.69*	0.15*	0.69*	0.73*	0.53*	0.31*	0.6*			
var15	0.12*	0.31*	0.06*	0.16*	0.19*	0.32*	0.29*	0.55*	0.23*	0.26*	0.2*	0.14*	0.19*	0.32*		
var16	0.71*	0.81*	0.47*	0.6*	0.39*	0.36*	0.58*	0.24*	0.72*	0.66*	0.4*	0.28*	0.63*	0.79*	0.37*	
var17	0.73*	0.76*	0.42*	0.57*	0.37*	0.38*	0.53*	0.22*	0.69*	0.58*	0.41*	0.32*	0.56*	0.75*	0.37*	0.95*

$*$ p-val $<$ 0.05

Table A3
Complete list of selected countries and relative missing values count and percentage over total number of observations.

Country	Missing Values	Country	Missing Values	Country	Missing Values
Nigeria	91 (6.7%)	Philippines	276 (20.3%)	Slovak Republic	336 (24.7%)
Sri Lanka	92 (6.8%)	Costa Rica	277 (20.4%)	Latvia	337 (24.8%)
Armenia	103 (7.6%)	St. Vincent and the Grenadines	278 (20.4%)	Serbia	337 (24.8%)
Lao PDR	105 (7.7%)	Mali	280 (20.6%)	Spain	337 (24.8%)
Mongolia	106 (7.8%)	Yemen, Rep.	280 (20.6%)	Austria	341 (25.1%)
Bolivia	113 (8.3%)	Guinea-Bissau	283 (20.8%)	Trinidad and Tobago	342 (25.1%)
Honduras	117 (8.6%)	China	285 (21%)	Belgium	347 (25.5%)
Moldova	122 (9%)	Indonesia	285 (21%)	Tunisia	347 (25.5%)
Nicaragua	123 (9%)	Liberia	286 (21%)	Eswatini	348 (25.6%)
Sierra Leone	129 (9.5%)	Croatia	288 (21.2%)	Romania	350 (25.7%)
Tanzania	130 (9.6%)	Ecuador	289 (21.2%)	Qatar	354 (26%)
Mauritania	134 (9.9%)	Malaysia	288 (21.2%)	Mauritius	355 (26.1%)
Benin	138 (10.1%)	Chile	292 (21.5%)	Kazakhstan	357 (26.2%)
India	139 (10.2%)	Hungary	292 (21.5%)	Bulgaria	359 (26.4%)
Kenya	142 (10.4%)	Singapore	293 (21.5%)	Malta	359 (26.4%)
Togo	141 (10.4%)	Djibouti	296 (21.8%)	Fiji	362 (26.6%)
Cote d'Ivoire	146 (10.7%)	Malawi	298 (21.9%)	Turkey	362 (26.6%)
Cameroon	150 (11%)	Sweden	300 (22.1%)	Luxembourg	363 (26.7%)
Burundi	151 (11.1%)	Peru	302 (22.2%)	Uruguay	365 (26.8%)
Mozambique	151 (11.1%)	Egypt, Arab Rep.	303 (22.3%)	Ukraine	367 (27%)
Tajikistan	152 (11.2%)	Brazil	305 (22.4%)	Finland	369 (27.1%)
Georgia	159 (11.7%)	South Africa	305 (22.4%)	Botswana	373 (27.4%)
Burkina Faso	163 (12%)	Thailand	304 (22.4%)	Denmark	372 (27.4%)
Niger	163 (12%)	Iran, Islamic Rep.	310 (22.8%)	Lebanon	372 (27.4%)
Bangladesh	165 (12.1%)	Switzerland	310 (22.8%)	Israel	375 (27.6%)
Angola	169 (12.4%)	Dominica	311 (22.9%)	Oman	376 (27.6%)
Rwanda	169 (12.4%)	Canada	313 (23%)	Portugal	376 (27.6%)
Zimbabwe	168 (12.4%)	Lithuania	313 (23%)	Norway	377 (27.7%)
Sudan	170 (12.5%)	Argentina	316 (23.2%)	Saudi Arabia	379 (27.9%)
Vietnam	170 (12.5%)	Jordan	315 (23.2%)	Germany	381 (28%)
Senegal	175 (12.9%)	Uzbekistan	315 (23.2%)	Iceland	382 (28.1%)
Bosnia and Herzegovina	179 (13.2%)	Australia	317 (23.3%)	Algeria	386 (28.4%)
Central African Republic	185 (13.6%)	Czech Republic	317 (23.3%)	Ireland	386 (28.4%)
Lesotho	185 (13.6%)	Guatemala	317 (23.3%)	Namibia	391 (28.7%)
Cambodia	194 (14.3%)	Jamaica	318 (23.4%)	Suriname	390 (28.7%)
Ethiopia	197 (14.5%)	Japan	318 (23.4%)	United Kingdom	392 (28.8%)
Sao Tome and Principe	198 (14.6%)	Dominican Republic	320 (23.5%)	North Macedonia	396 (29.1%)
Kyrgyz Republic	200 (14.7%)	Iraq	320 (23.5%)	Cyprus	399 (29.3%)
Pakistan	201 (14.8%)	Morocco	319 (23.5%)	Italy	403 (29.6%)
Bhutan	210 (15.4%)	France	322 (23.7%)	Gabon	410 (30.1%)
Nepal	209 (15.4%)	New Zealand	322 (23.7%)	Azerbaijan	411 (30.2%)
Ghana	214 (15.7%)	Panama	322 (23.7%)	Belarus	425 (31.2%)
Guinea	223 (16.4%)	Estonia	323 (23.8%)	Greece	430 (31.6%)
Uganda	223 (16.4%)	Grenada	325 (23.9%)	El Salvador	437 (32.1%)
Zambia	230 (16.9%)	United States	325 (23.9%)	United Arab Emirates	439 (32.3%)
Cabo Verde	232 (17.1%)	Kuwait	326 (24%)	Bahamas, The	444 (32.6%)
Chad	236 (17.4%)	Netherlands	327 (24%)	Belize	455 (33.5%)
Myanmar	237 (17.4%)	Russian Federation	326 (24%)	Brunei Darussalam	464 (34.1%)
Congo, Rep.	244 (17.9%)	Venezuela, RB	327 (24%)	Seychelles	477 (35.1%)
Gambia, The	243 (17.9%)	Paraguay	329 (24.2%)	Hong Kong SAR, China	482 (35.4%)
Madagascar	250 (18.4%)	Poland	330 (24.3%)	Montenegro	483 (35.5%)
Haiti	254 (18.7%)	Korea, Rep.	333 (24.5%)	Antigua and Barbuda	492 (36.2%)
St. Lucia	265 (19.5%)	Comoros	335 (24.6%)	Albania	507 (37.3%)
Congo, Dem. Rep.	268 (19.7%)	Slovenia	335 (24.6%)	Equatorial Guinea	523 (38.5%)
Mexico	274 (20.1%)	Bahrain	336 (24.7%)	Syrian Arab Republic	529 (38.9%)
Colombia	276 (20.3%)	Barbados	336 (24.7%)	Afghanistan	535 (39.3%)

In table A4 we report the distribution over time of the missing values quota, as to evaluate the impact of missing data imputation. It clearly emerges the highest quota for the last two available years.

Table A4
Missing values over years.

Year	Total Observations	Missing Values
2010	22,848	4,367 (19.1%)
2011	22,848	3,959 (17.3%)
2012	22,848	4,273 (18.7%)
2013	22,848	4,072 (17.8%)
2014	22,848	4,019 (17.6%)
2015	22,848	4,494 (19.7%)
2016	22,848	4,404 (19.3%)
2017	22,848	4,245 (18.6%)
2018	22,848	7,478 (32.7%)
2019	22,848	8,218 (36.0%)

Appendix B. Missing values imputation methodology

To assess imputation performances and to choose the best method, we test the algorithm in three settings. In the first (named *Original*) we consider the whole dataset made of 168 countries by 17 variables for 10 years for a total of 28,560 entries. It contains 25% of missing values, thus we randomly remove some additional values representing 10%, 20% and 30% of the initial dataset. In the second (named *No missing*) we drop all entries with missing values and apply the same incremental sampling procedure on the remaining subset. In the last (named *Some missing*) we drop all countries with at least 3 missing values for any year and apply again the incremental sampling procedure on the remaining subset. Furthermore, we fit the two methods, MC-SVD and BTF, on the previous 3 cases with different sampling percentages and we evaluate the Mean Absolute Reconstruction Error (MARE) on the excluded observations:

$$MARE = \frac{1}{M}\sum_{i}^{M}|x_{ablated} - x_{reconstructed}|$$

where M is the total number of excluded values. Moreover, we check the sensitivity to the original percentage of missing values by comparing the MARE on *No missing* and *Some missing* with the one on *Original*. Figure B.6 shows bar plot of MARE values for all settings for each increasing percentage of added missing values. Bar whiskers are scaled value of $max(MARE)$, defined as:

$$RM = \frac{max(MARE)}{\text{Average value of Original matrix}}$$

In order to grasp the magnitude of the impact of MARE we also report its ratio R with the average value of the non-missing entries of original matrix:

$$R = \frac{MARE}{\text{Average value of Original matrix}}$$

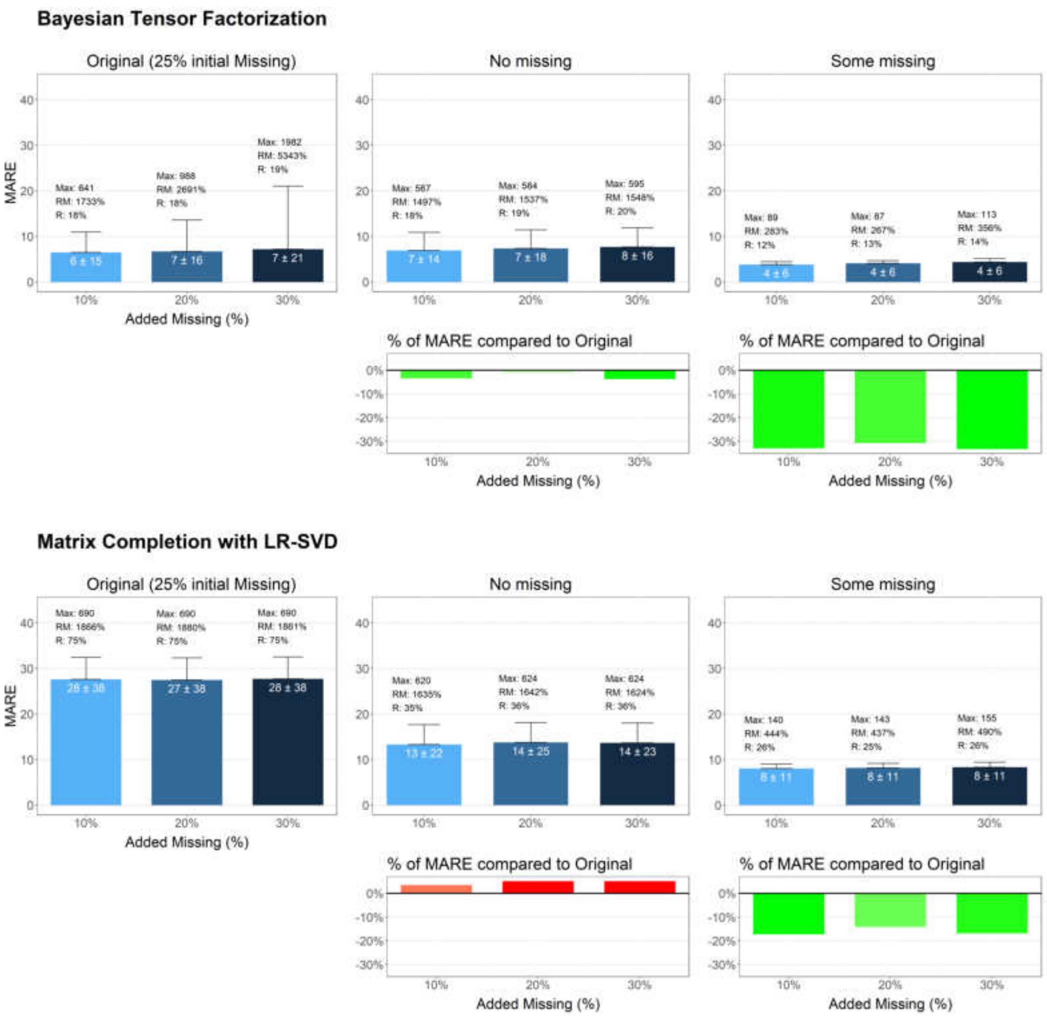


Fig. B.6. Testing missing values imputation methodologies. Blue bars report the Mean Absolute Reconstruction Error (MARE), green/red bars report the percent decrease/increase of MARE compared to the one evaluated on the *Original* setting.

Finally, for the *No missing* and *Some missing* setting we show the green/red bar plot reporting the percent decrease/increase, respectively, of MARE compared to the one evaluated in the *Original* setting so to evaluate the impact of missing data in the matching entries subset. BTF has lower MARE and higher percent decrease compared to MC-SVD implying a better data reconstruction ability and reliability.

Appendix C. Scree Plot and Loadings Plot for PCA method

In this appendix we report scree plots and loadings of all the competing PCA approaches: Original PCA, Robust PCA, Robust Sparse PCA. If we pay attention to loadings results available in C.9, we can notice that Original PCA and Robust PCA are very similar to each other, while Robust Sparse PCA appears different for several variables (namely var2, var3, var6, var8, var12, var13, var14) because by construction it aims to a sparse and parsimonious representation. In the Robust PCA almost all the variables have a meaningful positive contribution to the first Principal Component, that constitutes our ESR index (var8 (num of people per Km2) and var15 (value of trade as % of GDP) appear to be less significant).

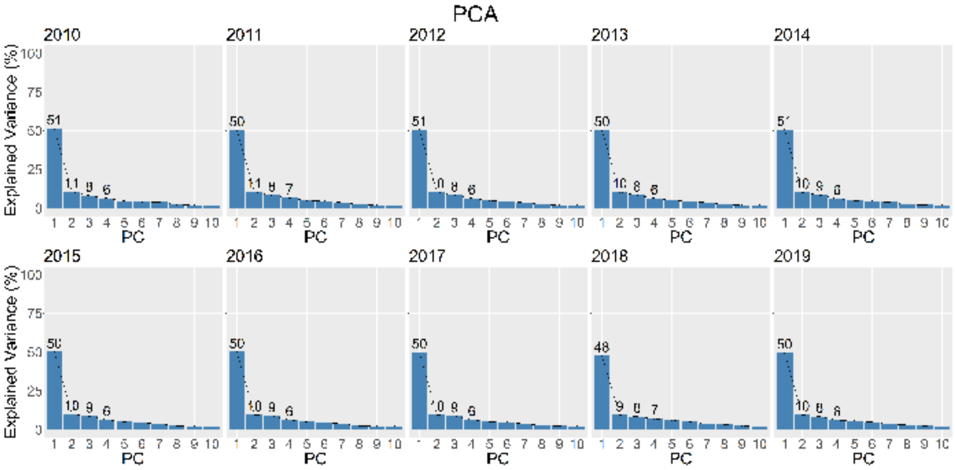


Fig. C.7. Scree plot for PCA method.

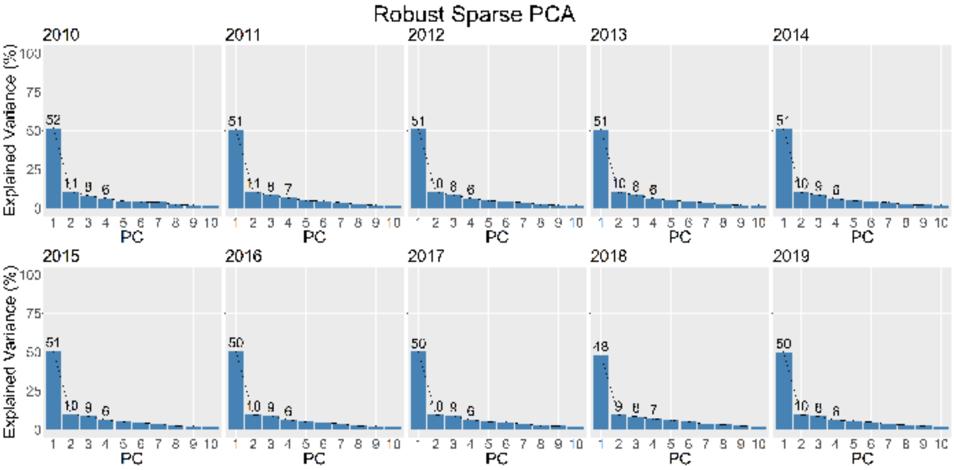


Fig. C.8. Scree plot for Robust Sparse PCA method.

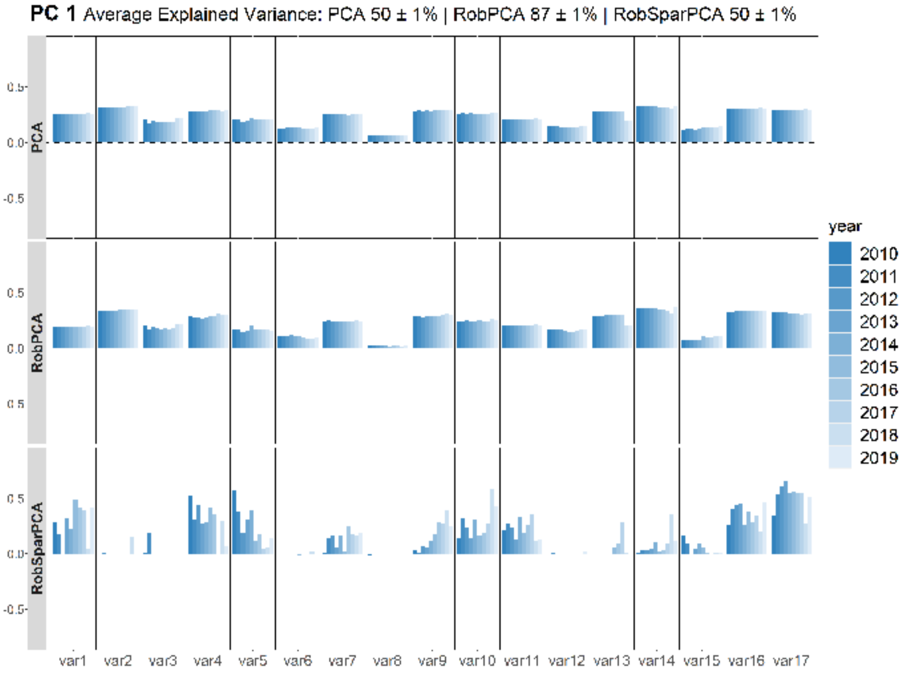


Fig. C.9. Loading plot for all PCA methods.

Appendix D. Loadings Plot for DFM method

In this appendix we report loadings of the DFM approach. As described in section the loadings $\mathbf{C}^i$ for the i-th country are stacked into the diagonal matrix

$\mathbf{C}$, whereas the cross-country interactions are introduced by the matrix $\hat{\boldsymbol{A}}$ estimated with VAR. Our setting force the $\mathbf{C}^i$ to be constant so we can estimate loadings for each country-variable pair. Therefore, for ease of visualization, figure D.10 reports the distribution of the loadings for each input variable over the 168 countries, representing the average trend over the years. The bimodal shape of all distributions implies a clear discriminative power of the index between less risky countries and riskier ones.

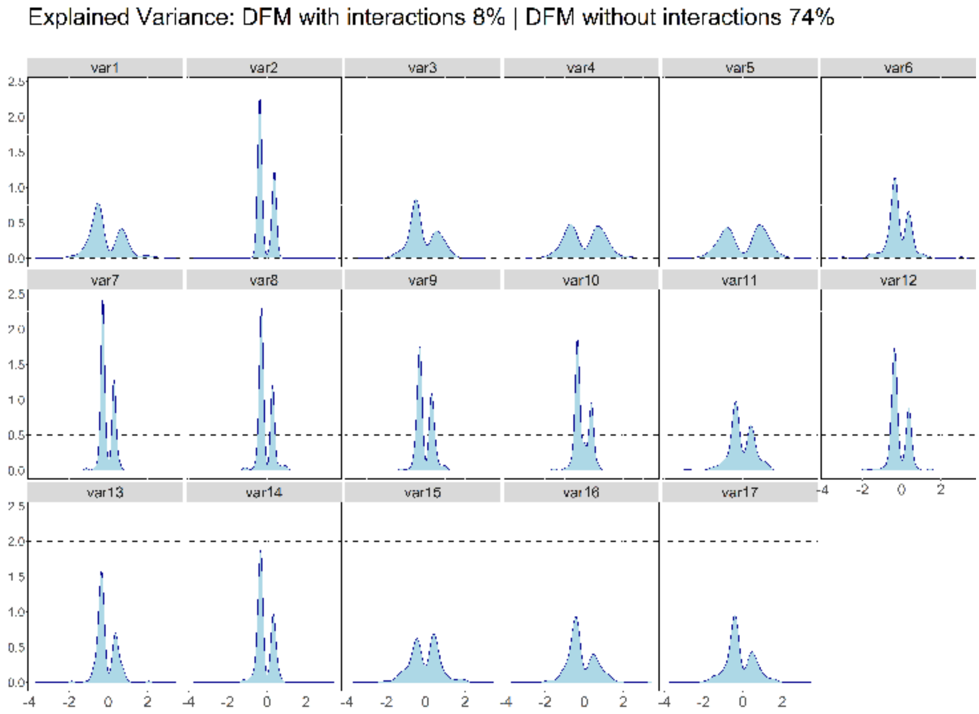


Fig. D.10. Loading plot for DFM method. On x-axis is reported the logarithm of loading values.

Appendix D.1. Index Robustness Check

Our robustness check is performed by using the ESR index as an input variable in supervised regressions. The aim is to evaluate the fitting power of the summary index compared to the original variables in modeling some relevant macro economic indicators. From Figure D.11 through Figure D.16 we report percent increase of RMSE in predicting macro economic indicators of interest (Unenployment, Real GDP per capita, Share of government consumption, Price level of capital information, Trade

Volume, Outstanding Loans of Commercial banks) due to the employment of the ESR index. The graphs report comparison between regressions with the single continuous ESR index as regressor and the one with original variables. In this way we assess how much the RMSE increases by substituting 17 variables with our summary index. In table A5 we report numerical results for the regressions above described.

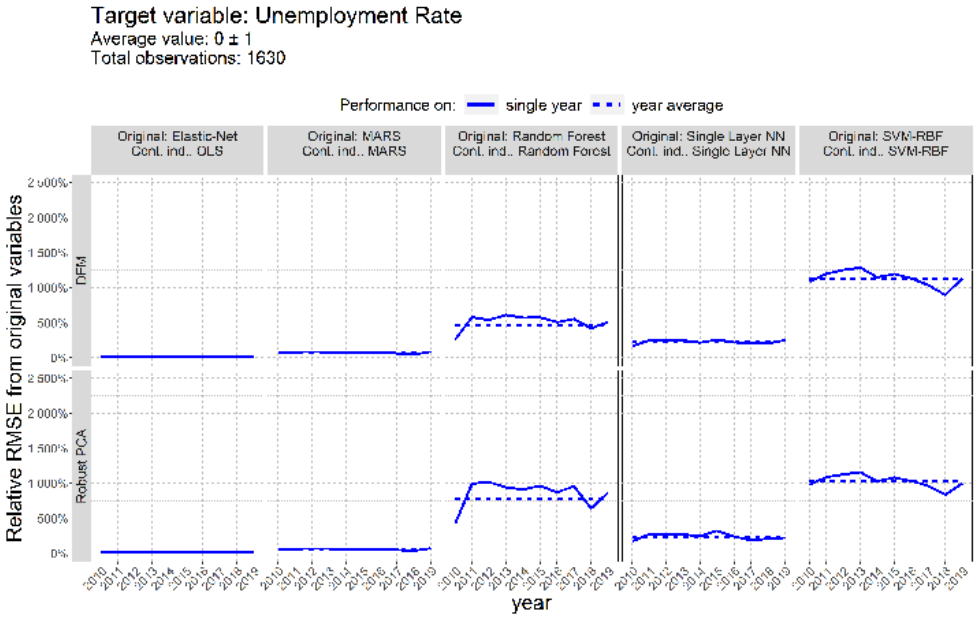


Fig. D.11. RMSE percent increase in predicting Unemployment rate. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

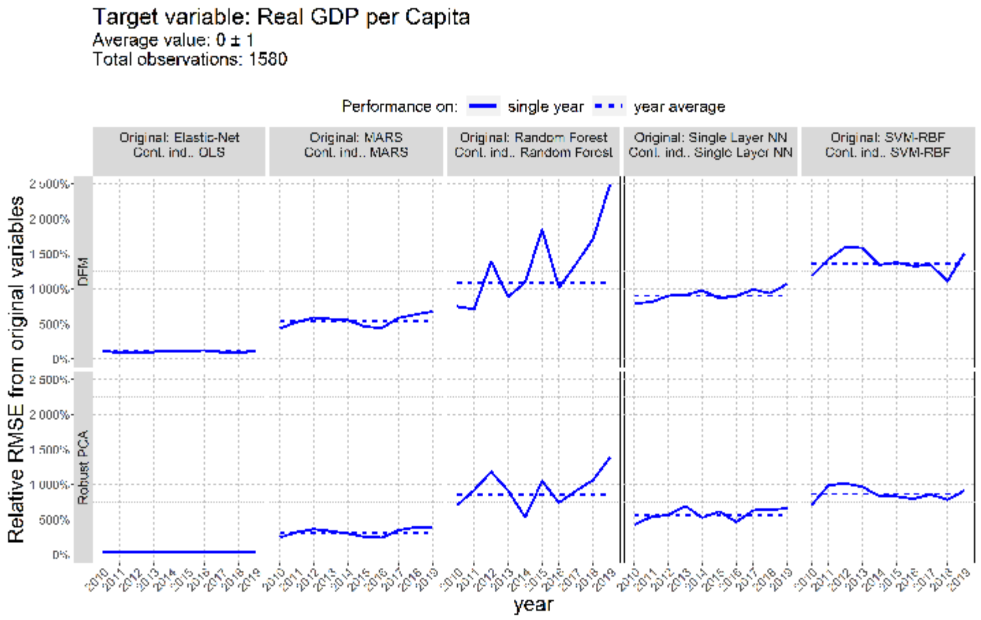


Fig. D.12. RMSE percent increase in predicting Real GDP per capita. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

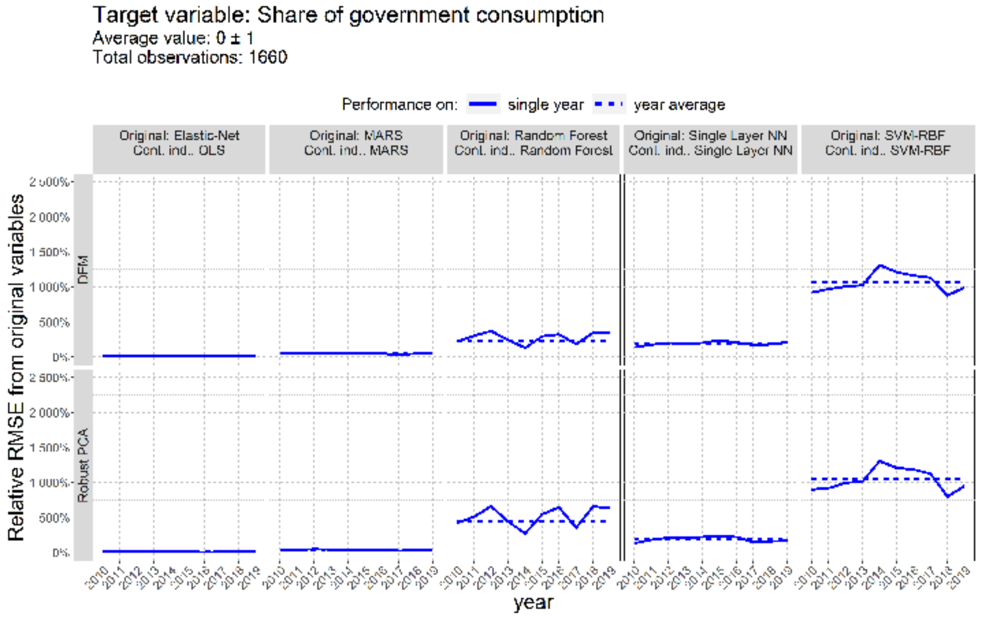


Fig. D.13. RMSE percent increase in predicting Share of government consumption. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

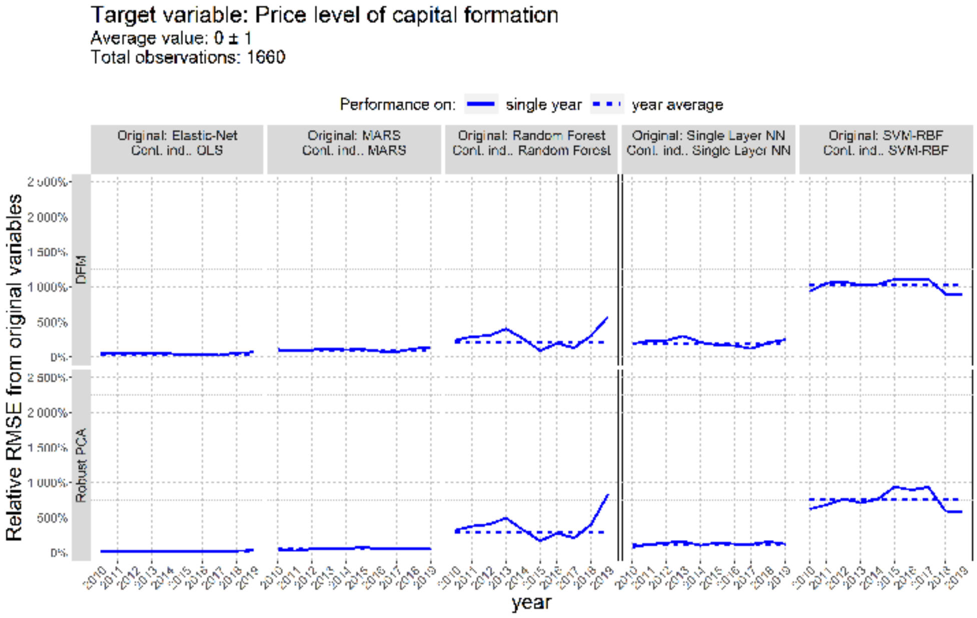


Fig. D.14. RMSE percent increase in predicting Price level of capital formation. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

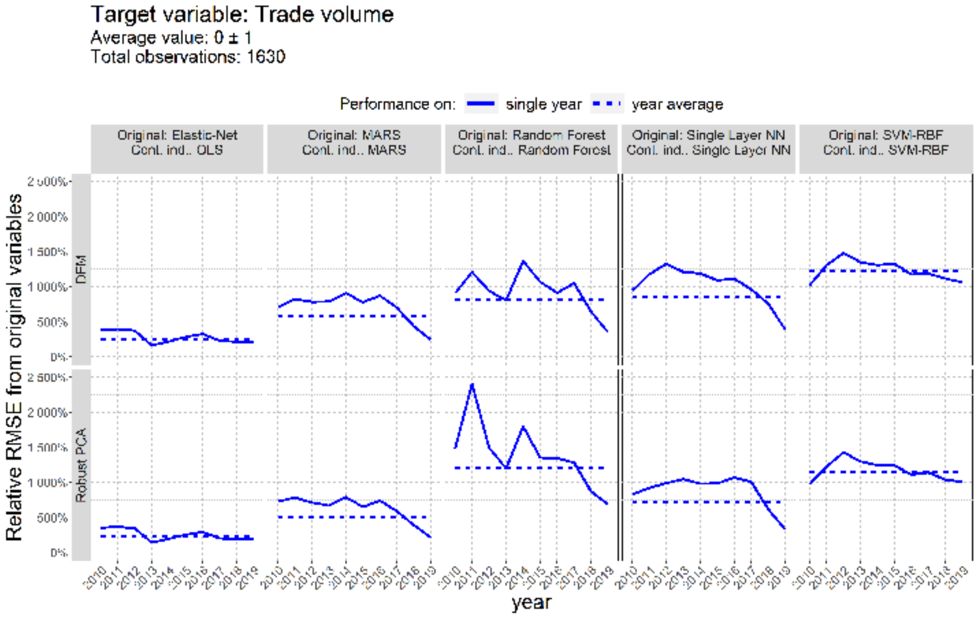


Fig. D.15. RMSE percent increase in predicting Trade volume. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

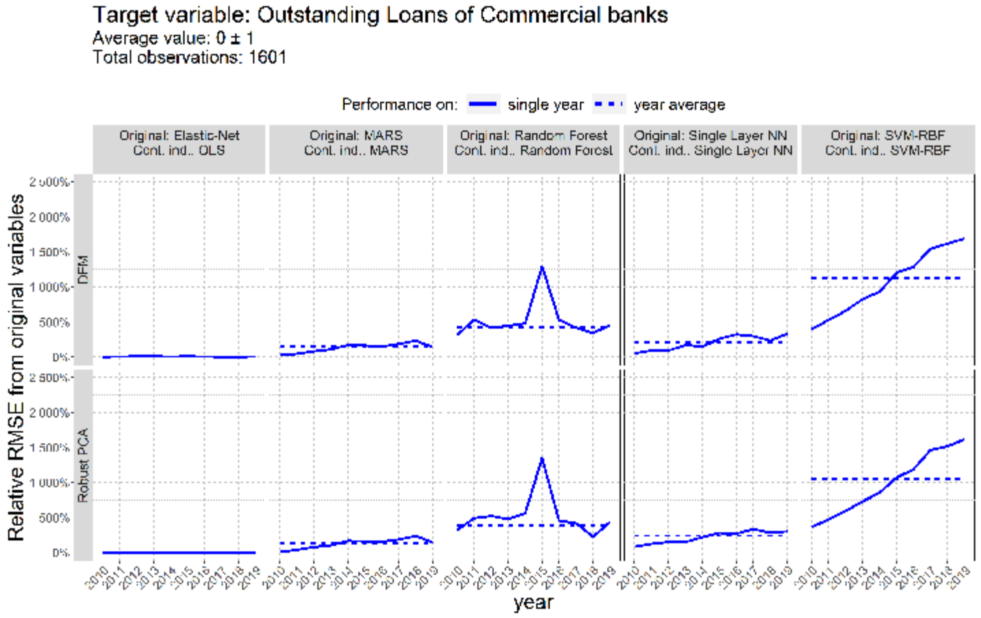


Fig. D.16. RMSE percent increase in predicting Outstanding Loans of Commercial banks. Comparison between regression with the single continuous index as regressor and the one with original variables. Solid lines show the single year metrics, dashed lines show the full dataset, i.e. average over years, metric.

Table A5

RMSE in predicting macro-economic variables with continuous index as regressor. RMSE for regression with original variables is reported in parenthesis.

	RMSE index (RMSE original)					
Target variable	Outstanding Loans of Commercial banks		Price level of capital formation		Real GDP per Capita	
Algorithm	DFM	Robust PCA	DFM	Robust PCA	DFM	Robust PCA
Elastic-Net	0.998(0.962)	1(0.962)	1(0.705)	0.839(0.705)	1(0.492)	0.663(0.492)
MARS	0.995(0.409)	0.986(0.409)	1(0.502)	0.764(0.502)	0.987(0.155)	0.634(0.155)
Random Forest	0.854(0.163)	0.914(0.163)	0.432(0.137)	0.549(0.137)	0.479(0.04)	0.395(0.04)
SVM-RBF	1.001(0.081)	1(0.081)	1.005(0.089)	0.764(0.089)	1.027(0.07)	0.664(0.07)
Single Layer NN	0.991(0.321)	0.976(0.321)	0.994(0.347)	0.768(0.347)	0.997(0.099)	0.663(0.099)

Target variable	Share of government consumption		Trade volume	
Algorithm	DFM	Robust PCA	DFM	Robust PCA
Elastic-Net	1(0.887)	0.999(0.887)	1(0.283)	0.935(0.283)
MARS	1(0.679)	0.948(0.679)	1(0.148)	0.909(0.148)
Random Forest	0.445(0.133)	0.719(0.133)	0.437(0.048)	0.637(0.048)
SVM-RBF	0.992(0.085)	0.977(0.085)	1.011(0.077)	0.954(0.077)
Single Layer NN	0.994(0.35)	0.973(0.35)	0.995(0.105)	0.926(0.105)

Appendix D.2. Index evolution over years

From D.17 through D.20 we report the evolution across time of the ESR index based on the two competing techniques for the different considered countries. It clearly emerges the higher sensitivity of the ESR index based on the DFM approach to the temporal dynamics which are explicitly modelled. PCA instead produces a rather flat pattern in line with the no direct modelling of the available years.

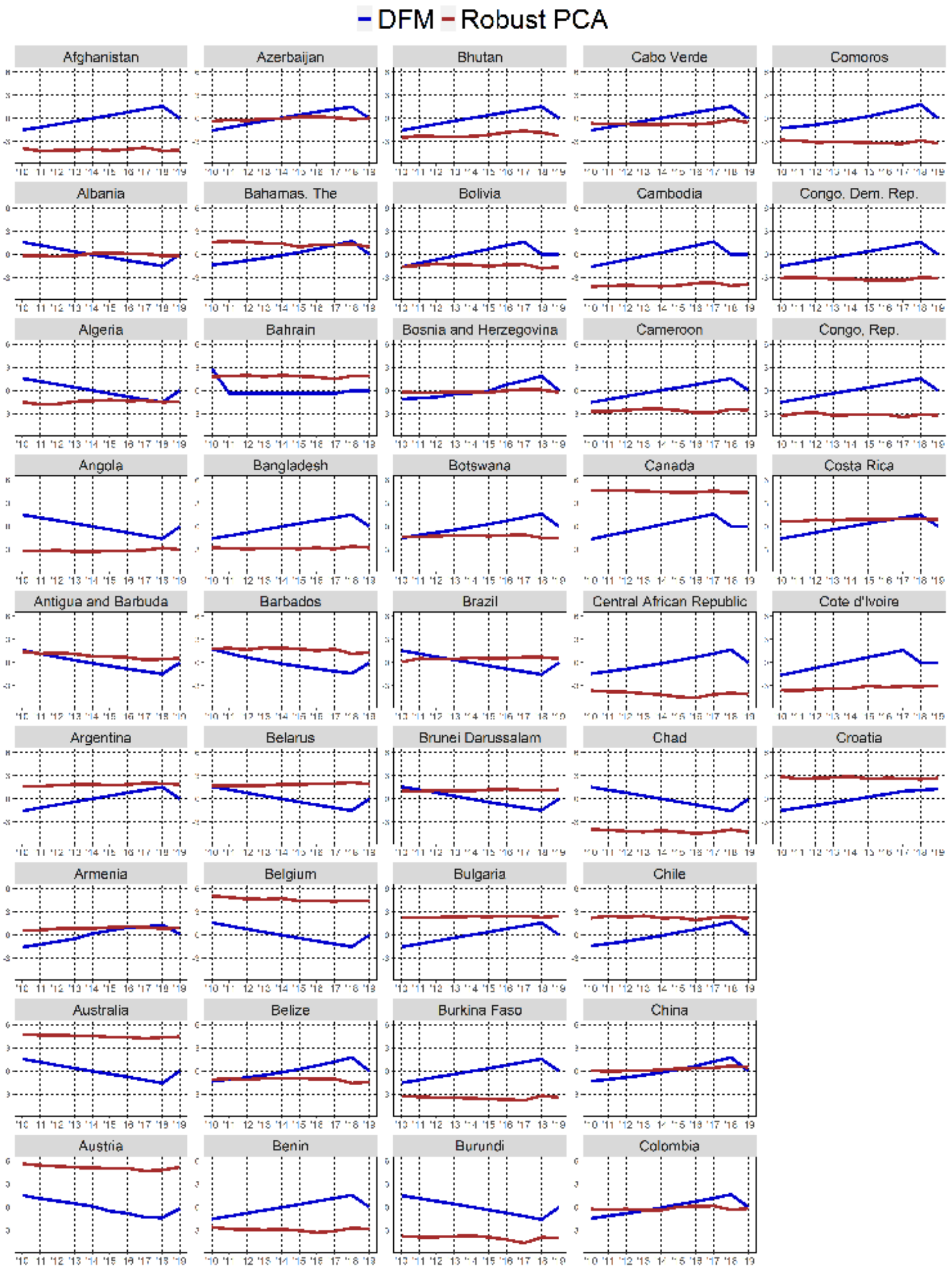


Fig. D.17. Index evolution over years

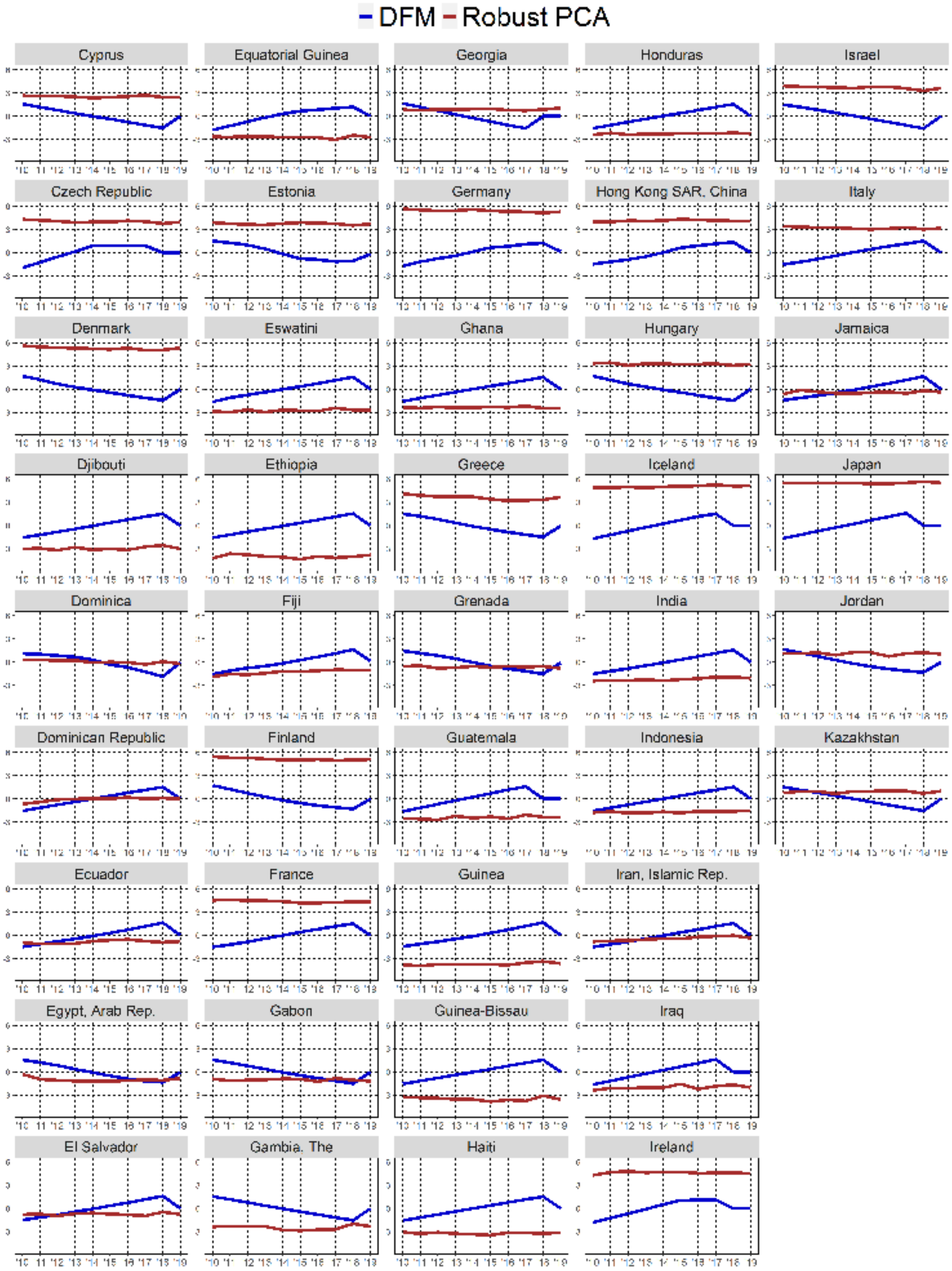


Fig. D.18. Index evolution over years

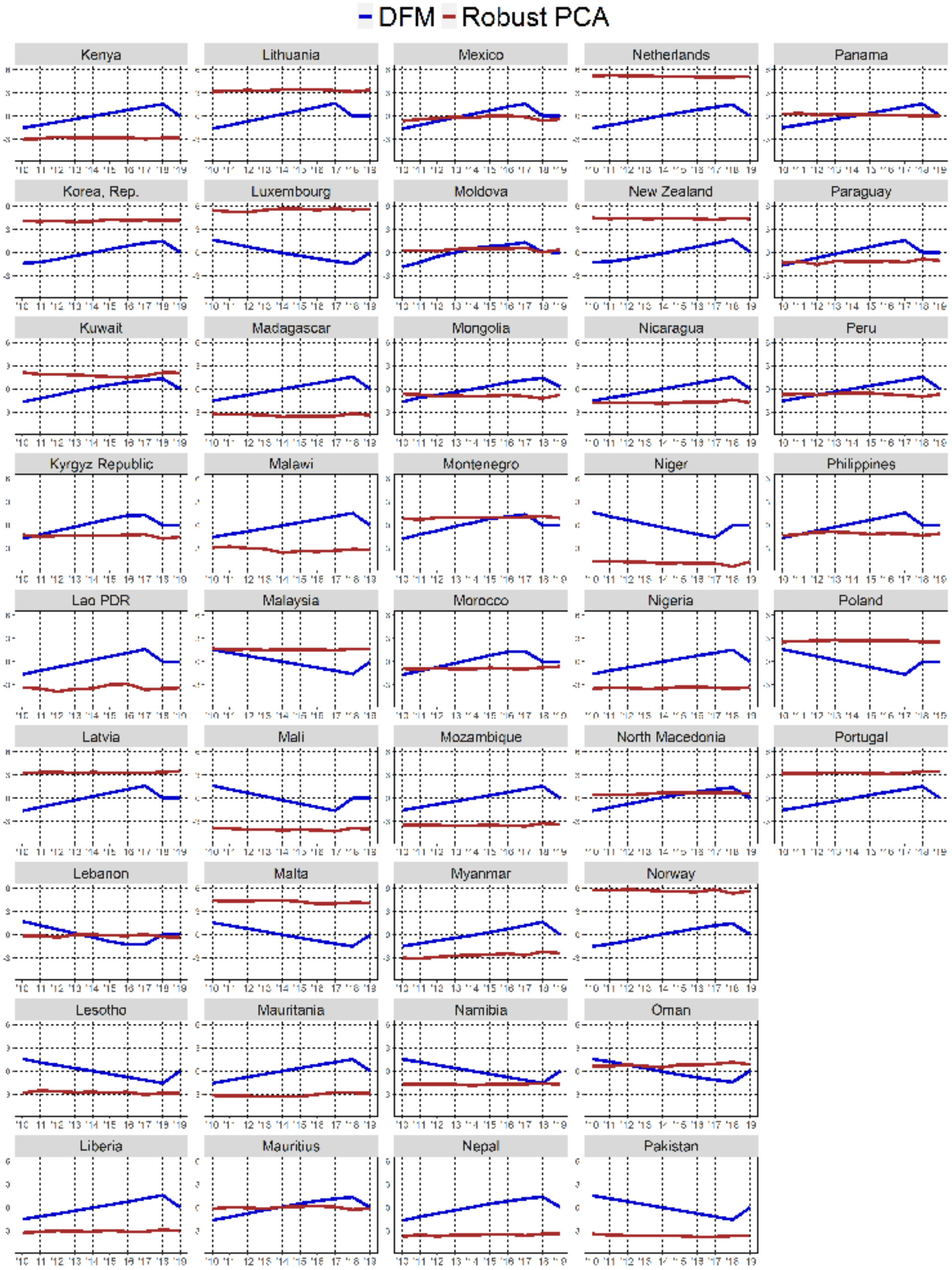


Fig. D.19. Index evolution over years

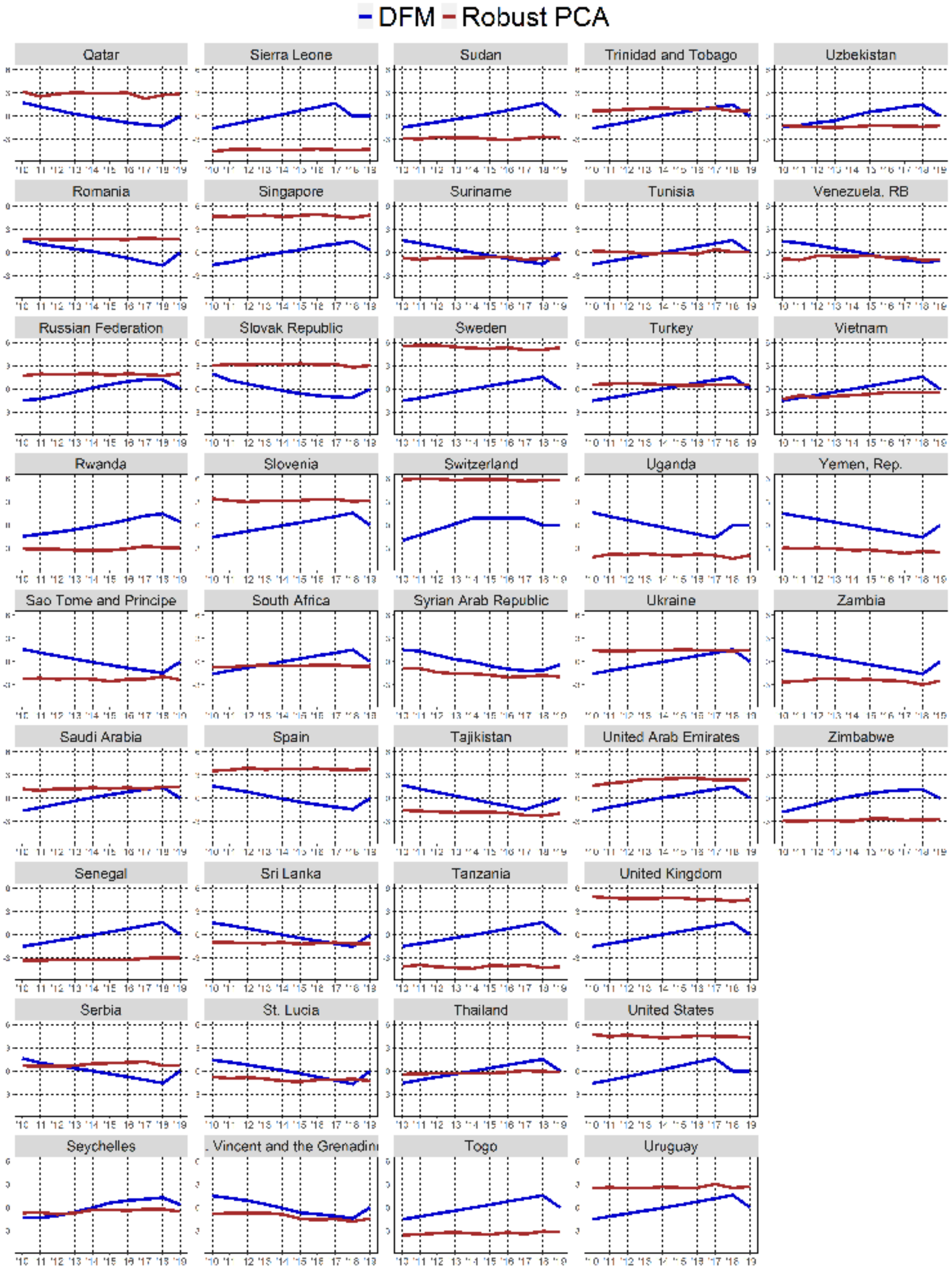


Fig. D.20. Index evolution over years