Rubem P. Mondaini *Editor*

Trends in Biomathematics: Modeling Epidemiological, Neuronal, and Social Dynamics

Selected Works from the BIOMAT Consortium Lectures, Rio de Janeiro, Brazil, 2022

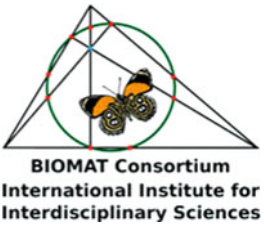
Springer

Trends in Biomathematics: Modeling Epidemiological, Neuronal, and Social Dynamics

Rubem P. Mondaini
Editor

Trends in Biomathematics: Modeling Epidemiological, Neuronal, and Social Dynamics

Selected Works from the BIOMAT Consortium Lectures, Rio de Janeiro, Brazil, 2022



Springer

Editor
Rubem P. Mondaini
BIOMAT Consortium
International Institute for Interdisciplinary
Mathematical and Biological Sciences
Rio de Janeiro, Brazil

Federal University of Rio de Janeiro
Rio de Janeiro, Brazil

ISBN 978-3-031-33049-0 ISBN 978-3-031-33050-6 (eBook)
https://doi.org/10.1007/978-3-031-33050-6

Mathematics Subject Classification: 92Bxx, 92-08, 92-10

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Preface

A BIOMAT International Series book is a collection of papers that have been selected for presentation in technical sessions at a BIOMAT Symposium, following a peer review analysis by members of the Editorial Board and other international senior reviewers. Another peer review is then performed to choose among the papers presented, which ones will be published as chapters of a BIOMAT book.

We have held 22 international conferences of the BIOMAT series since the first one in 2001, held in Rio de Janeiro, Brazil. The last three, in 2020, 2021, 2022, had to be held online, either for security reasons during the pandemic or due to economic constraints generated by it. The traditional pattern of BIOMAT conferences was however fully maintained: no parallel sessions; constant recommendation to participants to attend to the lectures of all their colleagues, in the best performance of a truly interdisciplinary conference; about 40 plenary lectures delivered in 5 conference days, between 9:00am and 5:00pm: Annual Plenary Meeting of BIOMAT Consortium and Meeting of the Consortium Board of Directors, representing the different regions of the Consortium: Western Europe, Eastern Europe, North America, South America, Africa, Asia and Oceania.

For this 3rd online meeting of BIOMAT Consortium, The International Symposium BIOMAT 2022, held with headquarters in Rio de Janeiro, from 07 to 11 November 2022, we had again the competent support of RNP – National Research Network, with the assistance of Dr. Beatriz Zoss, to whom once more we express our gratitude on behalf of BIOMAT Consortium.

The president of the BIOMAT Consortium also express his thanks for the collaboration in the organization to professors of the Institute for High-Performance Computing and Networking (ICAR) from Naples, Italy, Professors Lucia Maddalena and Ilaria Granata. We thank Dr. Simão C. de Albuquerque Neto and Carmem Lúcia S.C. Mondaini for their always competent and helpful assistance in all stages of the conference.

Rio de Janeiro, Brazil
November 11, 2022

Rubem P. Mondaini

Editorial Board of the BIOMAT Consortium

Contents

Clustering of Countries Based on the Associated Social Contact Patterns in Epidemiological Modelling

Evans Kiptoo Korir and Zsolt Vizi

1 Introduction

The spread of respiratory infections such as measles, influenza, and the recent COVID-19 occur predominantly through social contacts between susceptible and infected persons. Quantification of these contacts relevant is significant to predict disease dynamics and effect of introduced strategies that aim to control and curb further spread [15]. Epidemic models (especially deterministic ones) are widely used to understand and predict the disease dynamics and the impact of new strategies and interventions such as lockdown, school closure, or vaccination to control infectious disease transmission. Using these models, individuals are assumed to follow certain age-specific contact patterns [13, 19, 22, 25, 28]. These patterns are often inferred from diary-based surveys of self-recorded contacts which captures demographic information (age, household size, gender, and occupation) [12]. Such surveys are rare and limited in nature due to economic cost involved.

The rate of contacts between people is not same as it depends on factors such as individual behaviors, age groups, gender, and location of the contact (home, school, workplace, community, or other setting) [12, 13, 16, 19, 22, 25]. Age group is a critical determinant of disease transmission, and it is usually represented by contact matrices. The elements of these matrices are the average number of contacts an individual in some age group i makes with other individuals belonging to each age group j within a specified period of time, and they are commonly assortative [12, 13, 16, 19, 22, 25]. Age-structured models provide an understanding of age groups that are key epidemic drivers and the portion of people who are vulnerable to the disease. For an outbreak of infectious diseases such as COVID-

E. K. Korir · Z. Vizi (✉)
University of Szeged, Bolyai Institute, Szeged, Hungary
e-mail: zsvizi@math.u-szeged.hu

R. P. Mondaini (ed.), *Trends in Biomathematics: Modeling Epidemiological, Neuronal, and Social Dynamics*, https://doi.org/10.1007/978-3-031-33050-6_15

19 and influenza, public health introduces non-pharmaceutical interventions (school closure, contact tracing, social distancing, lockdowns, travel restrictions), or public health interventions such as vaccination and antiviral use to alter the pattern of contacts [19].

Multiple studies targeting to estimate social contacts have been published. The most referred study, POLYMOD, estimated contact matrices in 8 European countries based on surveys that involved 7290 participants with reported 97,904 contacts [19]. Authors in [21] estimated synthetic contact matrices for 152 countries in 2017 based on [19], demographic data, surveys, and other sources. The contact matrices were then updated in 2021 with recent data to include 177 countries with introduction of countries such as Hong Kong and the People's Republic of China [22]. Fumanelli et al. [6] estimated the contact matrices by constructing virtual populations in 26 European countries using detailed census and demographic data, an approach that can be adopted in the absence of specific experimental data. Iozzi et al. in their paper [10] entitled "little Italy" used agent-based model to compute social contact matrices by simulating individual-based model using demographic and Italian use data. Country-specific contact matrices have also been estimated, for instance, in Uganda [15], Kenya [11], Zimbabwe [17], India [14], China [23], Vietnam [9], Peru [7], and Russia [2].

As social contact patterns depend on the region of investigation, there is no evidence on spatial comparison of social contacts relevant for transmission between countries. Countries with unfavorable social structures and large population with risky behaviors are likely to have high transmission rates. Understanding this behavior is critical for timely introducing non-pharmaceutical interventions, monitoring the impacts of such decisions, and possible reintroduction of new measures in the event of cases resurgence in the countries [16]. During the COVID-19 world pandemic, decision-makers in countries, where an outbreak occurred later in time, monitored the actively intervening countries and analyzed the applied strategies to make efficient decisions with proper timing. In these situations, it is crucial to know which regions have to be monitored and which practices are recommended to be adapted. For this end, social contact patterns made in different countries are grouped into clusters and then examined to understand which countries have similar mixing patterns.

Here, we employ hierarchical clustering technique to determine the clusters of countries based on only the social contact patterns. The approach can be extended to include country characteristics such as gross domestic product (GDP), population density, life expectancy, total fertility rate, mortality rates, and other metrics. Since optimization of learning algorithms are highly affected by dimensionality of the data, we handle this problem with usage of dimension reduction techniques. Additionally, for ensuring comparability of the underlying contact matrices, we use epidemic models to normalize these matrices as a preprocessing step before the abovementioned steps.

The structure of the paper is the following: in the second chapter, we introduce the underlying data and the different pieces of the proposed framework, namely, the concept of an age-structured epidemic model, the dimension reduction to avoid

curse of dimensionality, and possible approaches for clustering the data points represented by a reduced feature vector. The third section gives a demonstration about how the pipeline works with specific setup of the framework. In the last section, we discuss the considerable options for modifying and improving the presented methods.

2 Methods and Materials

2.1 Data

Prem et al. [22] updated synthetic contact matrices based on [19] with the most recent data from sources such as the United Nations Population Division, International Labor Organization, World Bank databases, demographic data, and surveys. The authors estimated contact patterns in four different locations (home, work, school, and other locations) for non-POLYMOD countries and made comparison to prior estimated contact matrices. Contact patterns were then projected for a total of 177 countries.

We used prior results of this constructed social contact matrices in [22] but targeting only all the European countries. The social contact matrices are of size 16×16 representing four settings (home, school, work, and other location) among 16 age groups: 0–4, 5–9, 10–14, ..., 75+. A total of 39 countries in Europe were available for analysis (countries are listed in Appendix). We considered the complete social contact mixing in the form of a *full social contact matrix*:

$$M_C = M_{C,H} + M_{C,S} + M_{C,W} + M_{C,O}$$

where $M_{C,X}$ is the social contact matrix with type $X \in \{H, W, S, O\}$ from country $C \in \{1, ..., 39\}$.

When social contacts are measured empirically, reciprocity is usually reported [12, 13, 16, 19, 22, 25, 28] since contacts are mutual. We corrected each matrices M_C to ensure that the total number of contacts from age group j to i is equal to the total number of contacts from age group i to age group j:

$$M_C^{(i,j)} P_i = M_C^{(j,i)} P_j \tag{1}$$

where $M_C^{(i,j)}$ is the element of M_C in the ith row and the jth column N_i and N_j represent the number of individuals in age class i and j, respectively.

Specifically, we want to work with contact matrices satisfying Eq. (1); thus we applied symmetrization for the underlying data to ensure the consistency of total number of contacts between two age groups:

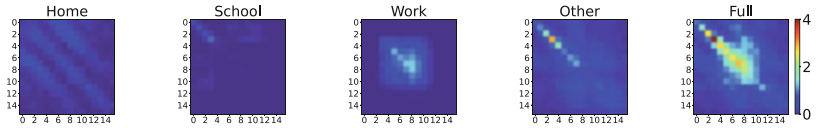


Fig. 1 Age-specific contact matrices in Hungary. Mixing patterns by age at home, school, work, and other places. The full matrix is obtained as the sum of the matrices represented in the four settings (home, school, work, and other places). In the horizontal axis, we have the age of the participant that was involved in the survey, and vertical axis shows the age of the reported contact. The elements of the matrices are the average number of contacts in the age group of the participant, which increases from dark blue to red on the images

$$M_C = \frac{1}{2P_i}\left(M_C^{(i,j)}P_i + M_C^{(j,i)}P_j\right)$$

where M_C is the full contact matrix now being corrected for reciprocity. Its elements are average number of contacts between age group i and j with very strong contacts along the diagonals characterizing strong age-assortative mixing (interactions between people of same age group). The off-diagonals capture inter-generational mixing (interactions between people of different age groups, for instance, children and their parents).

Figure 1 shows the contact matrices obtained for Hungary as an example of the social mixing patterns in European countries. The matrices are highly structured and varied considerably between settings. Reported contacts at home in Fig. 1 show a dominant diagonal (assortative pattern) representing contacts with people of the same age group such as young individuals. The off-diagonals reports inter-generational mixing contacts, for instance, between children and their parents. Notably, these contacts are least pronounced for people aged over 60. The contacts at school in Fig. 1 is evident as young people (people with age 0–19 years) interact mostly with other students of the same age and school. In the workplaces displayed in Fig. 1, we can see that most contacts occur between people in working population age (individuals with age group 20–64 years). These contacts are less assortative than at home. Overall contacts in other places (excludes home, work, and school) shows age assortative for younger groups and less assortative for older groups (Fig. 1). The full contacts in Fig. 1 which combines all the contacts in the four settings shows strong contacts between people of age 15–59 years in the country.

2.2 Epidemic Model

Epidemic models particularly age-structured models provide a better understanding of global spread and enable designing public health interventions. Deterministic compartmental models such as variants of SIR, $SEIR$, and SIS are formulated as a system of ordinary differential equations. The proposed framework can consider any

age-structured model for analyzing the similarities between countries with different contact matrices.

As a demonstration, we consider the model in Fig. 2 from [25]. The model has 15 compartments. S denotes the susceptibles, i.e., those who are at the risk of contracting the disease. Latents (L) are infected individuals but do not show symptoms and cannot infect. I_p compartment contains the pre-symptomatic individuals who do not show symptoms but are able to infect. Asymptomatic (or mildly symptomatic) and symptomatic infectious individuals, denoted by I_a and I_s, respectively, are divided into three groups (enabling the modelling of recovery time by an Erlang-distributed variable). Individuals from the last stage of asymptomatic compartment will all recover and hence proceed to the recovered class R, while those in the last stage of symptomatic compartment may recover without the need for hospital treatment (and thus move to R) or become hospitalized, either with normal hospital care (I_h) or intensive critical care (I_c). Those who need normal hospital care will recover, and those at the intensive care units may overwhelm the disease proceeding first to a rehabilitation unit (I_{cr}) before complete recovery, or for them fatal outcome might occur (D). The example model has an age-structured setup considering heterogeneity of contacts in the population via contact matrices and introduces age-specific parameters for some transitions in the transmission diagram. For the governing equations, see Appendix 2.

In this demonstration, we assume that the model parameters are the same for all countries except the baseline transmission rate, which is strongly related to the corresponding contact matrix; thus we denote by ($\beta_{0,C}$) the baseline transmission rate for country C.

An important quantity to calculate in epidemiology is the basic reproduction number $\mathcal{R}_0$, which represents the number of secondary infections caused by a single infective introduced into a wholly susceptible population [5, 12, 13, 16, 25]. For an age-stratified model, the basic reproduction number can be obtained via using the Next-Generation Matrix (shortly NGM) method. As a summary (see details

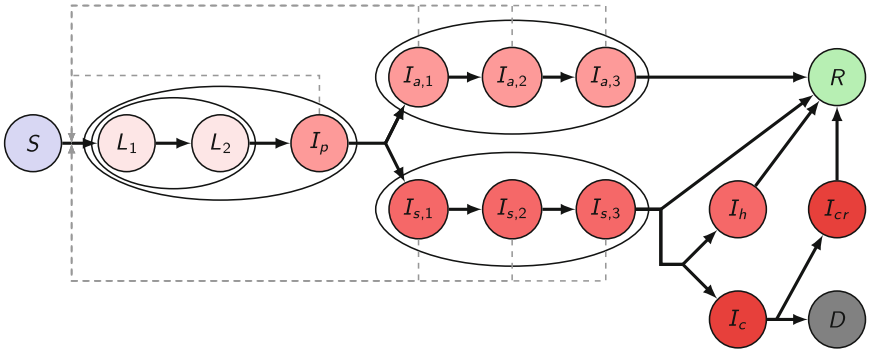


Fig. 2 Flowchart of the transmission model from [25] used for demonstrating the proposed framework

in Appendix 2), $\mathcal{R}_0$ can be calculated as the dominant eigenvalue of the model-specific NGM and particularly baseline transmission rate can be take out in the computations. Since calculation of the NGM depends on the underlying contact matrix, thus for a fix value of $\mathcal{R}_0$, we can determine the baseline transmission rate $\beta_{0,C}$ for each country $C \in \{1, 2, \ldots, 39\}$ (assuming that model parameters and contact matrix are known for the calculations). Using these transmission rates, we apply a standardization to all *full* contact matrices enabling comparability of the countries for a fixed value of $\mathcal{R}_0$, and we denote this new matrix by S_C:

$$S_C = \beta_{0,C} \cdot M_C.$$

This standardization is used as a normalization for transforming matrices to be distributed on the same interval. Significantly large differences can appear for the matrices depending on how much data was available for the estimation in [22]. This data preprocessing step is needed for the next data science-specific parts of the framework.

2.3 Dimensionality Reduction

Data with high-dimensional input vectors (i.e., order of sample size is at least the order of the dimensionality) pose problems to learning methods, since training performance of these methods is affected by the curse of dimensionality [3]. This severe problem can be solved by the introduction of dimension reduction technique or feature selection [1, 8].

In the classical theory of clustering, the features are arranged into vectors, and we often use projection-based algorithms for reducing the dimensionality of this vector. Principal component analysis (shortly PCA) is a popular technique to project the data vectors onto an affine hyperplane. The principal components are the vectors spanning an optimal projection plane and can be computed effectively using covariance matrix or SVD (former solves the optimization problem to maximize explained variance; latter gives the optimal low-rank approximation for the data matrix). We will refer to the classical PCA as 1D PCA.

Matrix-valued data are becoming increasingly common with advancement in the technology, e.g., images in computer vision problems. Driven by 1D PCA, several techniques that aim to preserve the $2D$ structure of the matrix have been developed [27]. These includes 2D variants of PCA, which consider projection along rows/columns of the matrices and $(2D)^2$ PCA, which applies projection along both dimensions. Recent approaches were developed such as the population value decomposition (PVD) that is precisely equivalent to a two-step SVD [29].

Using $(2D)^2$ PCA approach, we concatenate each country-specific full contact matrices to obtain the matrix $A_{\text{col}} = [M_1^\top, M_2^\top, \ldots, M_{39}^\top]^\top \in \mathbb{R}^{(39\cdot 16)\times 16}$. For the demonstration, in the column direction, two principal components were retained by projecting A onto the plane spanned by the columns of $Q \in \mathbb{R}^{16\times 2}$, thus having

$$\hat{A}_{\text{col}} = A_{\text{col}} Q \in \mathbb{R}^{(39\cdot 16)\times 2}.$$

In the next step, along the row direction, we retained two principal components as well and projected the concatenation of the transposed matrices $A_{\text{row}} = [M_1, M_2, \ldots, M_{39}]^\top \in \mathbb{R}^{(39\cdot 16)\times 16}$ using matrix $P \in \mathbb{R}^{16\times 2}$ obtaining

$$\hat{A}_{\text{row}} = A_{\text{row}} P \in \mathbb{R}^{(39\cdot 16)\times 2}.$$

Using the two projection matrices P and Q, we project all the matrices M_C with size 16×16 applying both P and Q to get the reduced matrix

$$\hat{M}_C = Q^\top M_C P \in \mathbb{R}^{2\times 2},$$

where $C \in \{1, 2, \ldots, 39\}$. In the clustering, for country C we use the vector $m_C \in \mathbb{R}^4$ calculated by flattening the $\hat{M}_C$ matrix as a feature vector.

Since the contact matrices satisfy Eq. (1), the lower and upper triangular parts of a contact matrix are not independent from each other. For applying 1D PCA technique, it is beneficial to consider only the independent elements of the contact matrix; thus we will take only the upper triangular part consists of 136 elements. In summary, the data matrix for 1D PCA will be $A \in \mathbb{R}^{39\times 136}$, where Cth row of A is the flattened upper triangular part of matrix M_C. In the demonstration, we can compare the result from 1D PCA projecting this data matrix to $\tilde{A} \in \mathbb{R}^{39\times 4}$ and use the row vector $\left(\tilde{A}^{(C)}\right)^\top \in \mathbb{R}^4$ associated to country C in the clustering.

2.4 Clustering Algorithm

Grouping of data points has become significant in various fields such as health, medicine, social, and spatial [24]. Solving this problem requires, for instance, a clustering technique to identify groups within a data set such that the similarity within the cluster is high and is low in-between [1]. Notably, there are several methods for clustering such as distance-based, connectivity-based methods, density-based, and probabilistic techniques [1, 3, 8].

Since we have a very small data set (i.e., number of data points is very low), the approaches considering dense set of data points do not work well now. In this study, agglomerative hierarchical clustering algorithm has been used due to its wide range of applications, simplicity, and ease of implementation relative to other clustering methods. This technique is expected to give better results in comparison to other methods, where large amount of data is considered. In hierarchical clustering, the closest sets of clusters are merged at each level using a linkage method, and then the dissimilarity matrix is updated correspondingly [1, 4, 18, 20, 24, 26]. This iterative process continues until only one maximal cluster remains.

Several proximity measures are available for this clustering technique, e.g., single, complete, average, and ward's criterion. We have used complete linkage method as it takes the cluster structure into consideration (nonlocal in behavior) and generally obtains compact-shaped clusters [1, 18]. For any clusters C_i and C_j, this linkage measures the similarity of clusters as the similarity of their most dissimilar members, that is, it considers the longest distance among all their observations:

$$D_{\text{complete}}(C_i, C_j) = \max\{d(x, y) : x \in C_i, y \in C_j\}$$

where d is a chosen distance measure. In this work, we have considered the Euclidean distance, but other L_p norms or specialized distance metrics can be defined (in many applications, the distance metric is the key to a good clustering).

Using the dissimilarity matrix and the linkage method, we can construct a dendrogram showing the agglomeration process during the running of the algorithm. The optimal number of clusters is obtained by manually cutting the tree at any given level and obtaining the clusters correspondingly without rerunning the algorithm. The rule of thumb we applied considers the longest vertical distance without any horizontal line passing through it, i.e., cutting the dendrogram where we have the largest gap between two successive merges [1].

2.5 The Proposed Framework

The complete pipeline of the framework can be seen in Fig. 3. The main input of the pipeline is the list of country-specific contact matrices (or any list of contact matrices for which the grouping needs to be executed). The symmetrization of the matrices is done at loading the data to ensure consistent total number of contacts between the age groups. The resulted matrices inputted to an epidemic model, which is specified by the user/domain expert, and then the base transmission rate β_0 is calculated along a concept. The concept determines the indicator for the calculation of β_0: in this paper, we use $\mathcal{R}_0$ to calculate this parameter, but a viable concept would be to obtain that β_0 for which the final epidemic size or final number of deaths is a pre-defined fraction of the population (this can be done for all countries). Using the symmetrized matrices and the previously calculated β_0s, we apply the standardization for the matrices via scaling the elements of them.

In the next phase of the procedure, we apply the data science-related steps on the standardized matrices. First a dimension reduction is applied to avoid curse of dimensionality in the next steps. The outputted low-dimensional vectors (i.e., the feature vectors) are used in the clustering algorithm, which is preferred to be a connectivity-based approach for very small datasets. In Fig. 3, yellow-colored shapes show the generic parts of the pipeline (i.e., these parts are determined by the expert specifying the choices based on the needs and goals). The rectangular shapes give the output data from the modules depicted by ellipsoids.

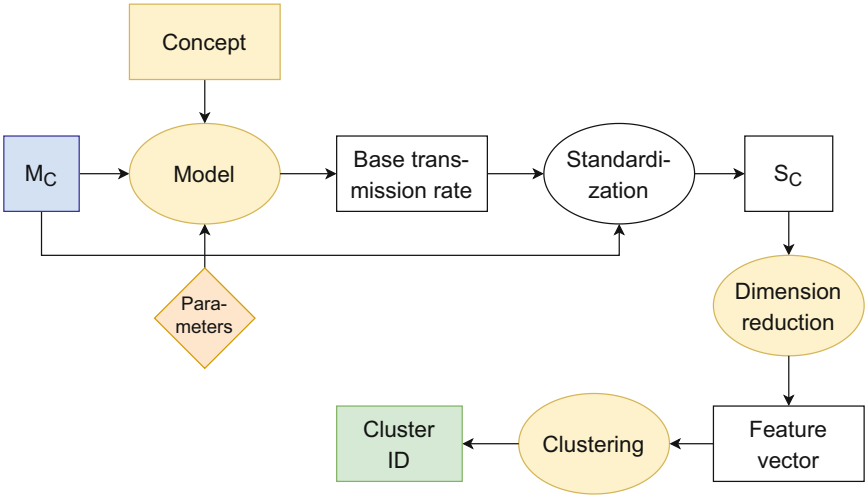


Fig. 3 Flowchart showing a summary of the proposed framework. Rectangles show the output of a module depicted by ellipsoids, which contain function calls executing a specific part of the pipeline. The yellow shapes are generic and can be replaced by considering improved/application-specific alternatives. As it can be seen, the input of the procedure is the contact matrices and cluster correspondence computed at the output. The concept associated with the model describes along which indicator the standardization has to be done

The proposed framework is implemented in Python using object-oriented programming to keep the code generic and easily extendable. We used open source packages to solve differential equations and eigenvalue problems and execute dimensionality reduction and clustering. The code is publicly available in Github [30].

3 Results

In this section, we demonstrate the previously introduced framework on the age-structured model of [25] with conceptually simple and correct algorithms for dimensionality reduction and clustering. Since [22] offers contact matrices for 16 age groups, we aligned the age-dependent parameters of the model. Here we investigate a total of 39 countries in Europe (see Appendix 1) to avoid significant differences between the countries from viewpoint of economics, location, demographics, and culture. Following the methodology, both 1D PCA and $(2D)^2$ PCA approaches were applied to project into $\mathbb{R}^4$ to highlight importance of the proper reduction for matrix-valued data. The clustering algorithm was chosen to be a connectivity-based one with agglomerative execution and using Euclidean distance and complete linkage for merging the clusters during the procedure. For

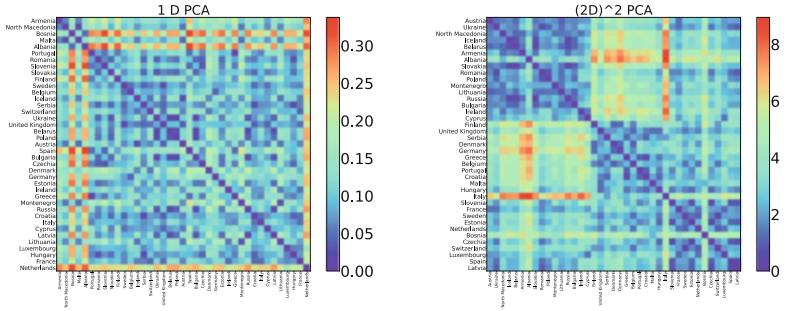


Fig. 4 The pairwise distance matrices computed for all countries based on the feature vector from the dimensionality reduction step. The left figure corresponds to 1D PCA, while the right is calculated from the result of the $(2D)^2$ PCA approach. Here we used Euclidean distance during the computations. We applied the dendrogram from the hierarchical clustering to arrange the rows and columns of the matrices providing a nicer presentation of the elements

the simulation we chose $\mathcal{R}_0 = 2.2$ for calculating country-specific transmission rates.

Applying standardization for the matrices and executing dimensionality reduction, the computed feature vectors can be used to calculate pairwise Euclidean distance for the countries. For a nicer representation, the rows and columns of the distance matrix can be permuted using the dendrogram from the hierarchical clustering. The left figure in Fig. 4 shows the distance matrix computed from the result of 1D PCA: as it can be seen, countries with dark-blue color entries such as Bulgaria, Latvia, and Belarus have very close social mixing patterns, and such countries are expected to belong to the same cluster. Notably, countries such as Netherlands, Albania, and Bosnia have orange and red cell colors, respectively, with most countries indicating somehow different social contact patterns. Such countries can have different cluster and seemingly different intervention strategies. The same intuition applies to $(2D)^2$ PCA approach in Fig. 4.

The advantage of using hierarchical clustering method is that it allows for manually cutting the hierarchy at any given level and obtaining the clusters correspondingly just by defining a threshold. The choice of threshold depends completely on the user. Dividing the dendrogram from 1D PCA at a cluster distance of 0.25, three separate clusters were formed (Fig. 5), Netherlands became a single cluster. In the same way, cutting the dendrogram at a distance of 5.5 between the clusters calculated on reduced data from $(2D)^2$ PCA technique, the dendrogram formed three meaningful clusters (Fig. 6). The final list of clusters is provided in Tables 1 and 2 (for the former one, the one-element cluster with Netherlands was omitted from the table). Notice, that cluster sizes are highly imbalanced for 1D PCA, which seems to show the flaw of this reduction approach compared to the result using $(2D)^2$ PCA, where cluster sizes are more comparable, which indicates better separation after the projection.

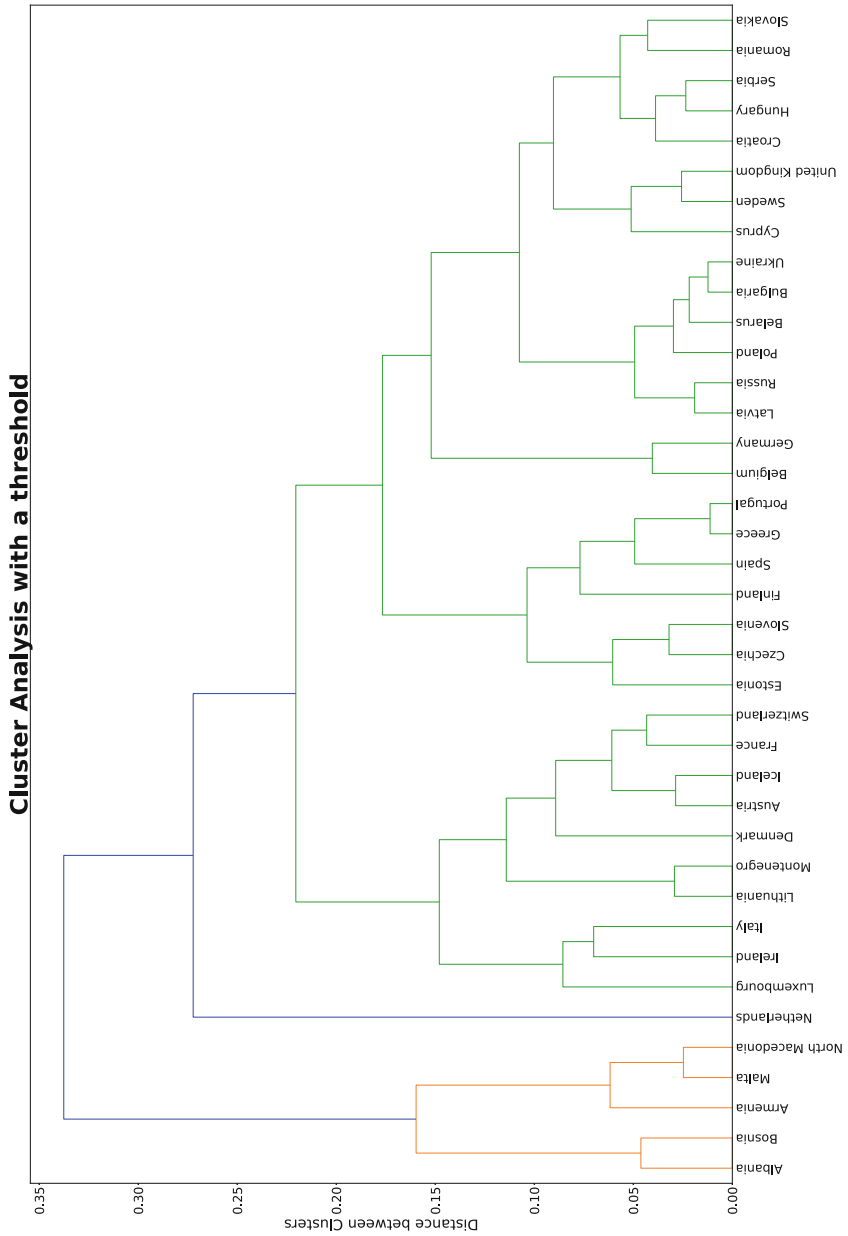


Fig. 5 Dendrogram associated with the clustering of countries using feature vector generated by 1D PCA. Vertical axis indicates the distance between the two clusters being connected. The agglomerative manner of the algorithm can be read from following the tree from bottom to top

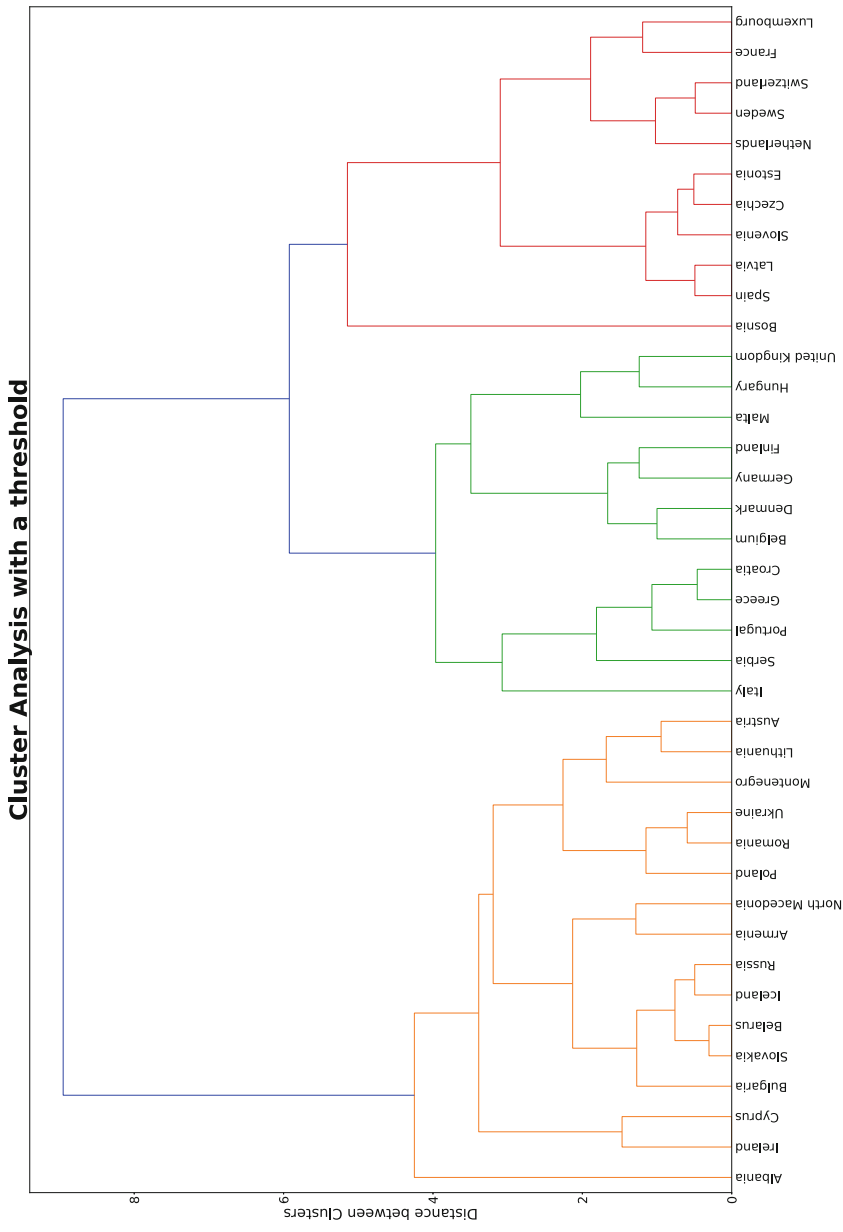


Fig. 6 Dendrogram associated with the clustering of countries using feature vector generated by $(2D)^2$ PCA. Vertical axis indicates the distance between the two clusters being connected. The agglomerative manner of the algorithm can be read from following the tree from bottom to top

Table 1 Clusters produced using 1D PCA. Netherlands is one element cluster, and thus it is omitted in this listing. Clusters produced by the agglomerative hierarchical clustering algorithm based on social contact patterns for the 39 European countries applying projection 1D PCA. Clusters 1 and 2 have 5 and 33 elements, respectively. Netherlands is one element cluster, and thus it is omitted in this listing

Cluster 1	Cluster 2
Albania, Bosnia, Armenia, Malta, North Macedonia	Luxembourg, Ireland, Italy, Lithuania, Montenegro, Denmark, Austria, Iceland, France, Switzerland, Estonia, Czechia, Slovenia, Finland, Spain, Greece, Portugal, Belgium, Germany, Latvia, Russia , Poland, Belarus, Bulgaria, Ukraine, Cyprus, Sweden, United Kingdom, Croatia, Hungary, Serbia, Romania, Slovakia

Table 2 Clusters produced by the agglomerative hierarchical clustering algorithm based on social contact patterns for the 39 European countries applying projection $(2D)^2 PCA$). Clusters 1, 2, and 3 have 16, 12, and 11 elements, respectively

Cluster 1	Cluster 2	Cluster 3
Albania, Ireland, Cyprus, Bulgaria, Slovakia, Belarus, Iceland, Russia, Armenia, North Macedonia, Poland, Romania, Ukraine, Montenegro, Lithuania, Austria.	Italy, Serbia, Portugal, Greece, Croatia, Belgium, Denmark, Germany, Finland, Malta, Hungary, United Kingdom.	Bosnia, Spain, Latvia, Slovenia, Czechia, Estonia, Netherlands, Sweden, Switzerland, France, Luxembourg.

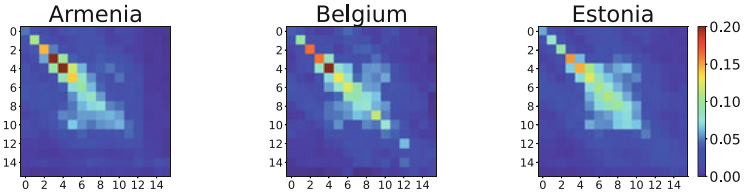


Fig. 7 Standardized contact matrices correspond to Armenia (left), Belgium (center), and Estonia (right), selected from different clusters resulted by clustering based on the feature vectors generated by the $(2D)^2 PCA$ technique. These countries are located "in the middle" of the ordered list of the cluster elements, based on the dendrograms. Differences between these matrices appear along the elements in the main diagonal (larger number for younger and old age groups for Belgium), in the elements along diagonals parallel to the main one (expressing higher contact rates between "adult children" and older/retired parents) and middle square of the matrices (which come mainly from work contact)

Figure 7 shows one element from each of the clusters calculated from reduced vectors resulted by $(2D)^2$ PCA. These countries (Armenia, Belgium, and Estonia) were chosen, since they are located in the middle of the blocks along the horizontal axis. Considering these matrices, we see that differences occur between the matrices along the main diagonal (for Armenia, the elements for the older age groups are not as strong as the rest of the elements are not that large), along the diagonal parallel to the main diagonal (these express the contacts between the adult parents and retired

grandparents; here Belgium and Estonia show more concentrated rates) and in the middle square of the matrix (this part comes mainly from workplace contacts, the distribution of the elements differ for all three of the samples). It is important to mention that distribution of the elements is directly related to the statistical method described in [22], since stronger smoothing effect can be recognized for estimation with less amount of data; thus distance measuring will be highly affected by the quality of the estimations. Nevertheless, the dimensionality reduction techniques developed for matrices aim to preserve the patterns in the structure of the matrices, for which the previously mentioned set of elements (diagonals, middle submatrix) give examples.

4 Discussion

Since emerging infectious diseases do not occur at the same time in all countries, it is important to understand which countries are affected in the same way with a non-pharmaceutical intervention. The government of a country where the outbreak happens later in time can benefit from this information to make better decision in prevention and damage control. This paper proposes a framework which can support decision-makers in determining, which other countries have to be primarily monitored and analyzed based on social mixing patterns. The underlying pipeline assumes only the contact matrices in the population and transforms it via an epidemic model and the concept of the analysis and then applies dimension reduction on this data and executes clustering on the projected data points associated with the considered countries. A realization of the framework was demonstrated with simple and correct approaches for the generic parts such as the epidemic model, the dimensionality reduction and the clustering.

In this procedure we can replace the concept of the analysis, since $\mathcal{R}_0$ describes only the transmission spread of a disease, but if a country primarily wants to minimize the number of fatal outcomes or the number of patients admitted to the hospital, these aspects can be also used for determining the country-specific scaling parameter in the abovementioned standardization. As an example, for all countries we can determine that β_0, for which the ratio of fatal outcomes is a fixed proportion. Clearly, the epidemic model can be replaced by any pre-defined model written for any infectious disease. For dimensionality reduction we applied a simple technique developed for matrices, but other approaches might be suitable for this part; see [27] for similar but more sophisticated methods, but convolution-based techniques can be developed, or modified optimization problem can be written down for the matrix reconstruction technique. For very small data, connectivity-based algorithms perform well, and many different approaches are available (agglomerative and divisive types, different similarity measures, and linkage methods). Globally, we can extend the feature vectors after the dimension reduction with other descriptors about the countries, which can be relevant in the analysis, e.g., some economical indicators, which can be relevant for the NPI strategy. Finally, the full contact matrix

used as a main input to the procedure can also be computed with other weighting in the summation to give more importance to specified type of contacts, e.g., home contacts are really important, since typical NPIs reducing contacts (school closure, home office, closing malls) leave this contact unchanged (or even slightly increase them).

Acknowledgments This research was supported by project TKP2021-NVA-09 and implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

Appendix 1: First Appendix

List of the European Countries

We considered the following European countries in the demonstration:
Albania, Armenia, Austria, Belarus, Belgium, Bosnia, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, North Macedonia, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom.

Appendix 2: Second Appendix

The Governing Equations of the Epidemic Model

The governing equations of the disease model described in Sect. 2.2 take the form

$$S^{i\,\prime}(t) = -\beta_0 \frac{S^i(t)}{N_i} \cdot \sum_{k=1}^{16} M^{(k,i)} \left[I_p^k(t) + \inf_a \sum_{j=1}^{3} I_{a,j}^k(t) + \sum_{j=1}^{3} I_{s,j}^k(t) \right]$$

$$L_1^{i\,\prime}(t) = \beta_0 \frac{S^i(t)}{N_i} \cdot \sum_{k=1}^{16} M^{(k,i)} \left[I_p^k(t) + \inf_a \sum_{j=1}^{3} I_{a,j}^k(t) + \sum_{j=1}^{3} I_{s,j}^k(t) \right] - 2\alpha_l L_1^i(t)$$

$$L_2^{i\,\prime}(t) = 2\alpha_l L_1^i(t) - 2\alpha_l L_2^i(t),$$

$$I_a^{i\,\prime}(t) = 2\alpha_l L_2^i(t) - \alpha_p I_p^i(t)$$

$$I_{a,1}^{i\,\prime}(t) = p^i \alpha_p I_p^i(t) - 3\gamma_a I_{a,1}^i(t)$$

$$I_{a,2}^{i\,\prime}(t) = 3\gamma_a I_{a,1}^i(t) - 3\gamma_a I_{a,2}^i(t)$$

$$
\begin{aligned}
{I^i_{a,3}}'(t) &= 3\gamma_a I^i_{a,2}(t) - 3\gamma_a I^i_{a,3}(t) \\
{I^i_{s,1}}'(t) &= (1-p^i)\alpha_p I^i_p - 3\gamma_s I^i_{s,1}(t) \\
{I^i_{s,2}}'(t) &= 3\gamma_s I^i_{s,1}(t) - 3\gamma_s I^i_{s,2}(t) \\
{I^i_{s,3}}'(t) &= 3\gamma_s I^i_{s,2}(t) - 3\gamma_s I^i_{s,3}(t) \\
{I^i_h}'(t) &= h^i(1-\xi^i)3\gamma_s I^i_{s,3}(t) - \gamma_h I^i_h(t) \\
{I^i_c}'(t) &= h^i\xi^i 3\gamma_s I^i_{s,3}(t) - \gamma_c I^i_c(t) \\
{I^i_{\mathrm{cr}}}'(t) &= (1-\mu^i)\gamma_c I^i_c(t) - \gamma_{\mathrm{cr}} I^i_{\mathrm{cr}}(t) \\
{R^i}'(t) &= 3\gamma_a I^i_{a,3}(t) + (1-h^i)3\gamma_s I^i_{s,3}(t) + \gamma_h I^i_h(t) + \gamma_{\mathrm{cr}} I^i_{\mathrm{cr}}(t) \\
{D^i}'(t) &= \mu^i\gamma_c I^i_c(t),
\end{aligned}
\tag{2}
$$

where the index $i \in 1, ..., 16$ represents the corresponding age group. As it can be seen, the model contains age-dependent parameters (probabilities p, h, ξ, μ, for which upper index shows the age group) and age-independent ones (fraction $\inf_a$ and transition parameters α and γ_X, where $X \in \{a, s, h, c, \mathrm{cr}\}$). Notation here is aligned with the parameter file located in the repository of the framework. Here $\inf_a$ denotes the relative infectiousness of I_a compared to I_s, for more details about the other parameters and the methodology for parametrization, see [25].

Next-Generation Matrix

To calculate $\mathcal{R}_0$ for the previously mentioned epidemic model, we consider the infectious subsystem for

$$L^i_1(t), L^i_2(t), I^i_p(t), I^i_j(t),$$

with $j \in \{a, s\} \times \{1, 2, 3\}$, $i \in \{1, ..., 16\}$, thus

$$X(t) = \left[L^i_1(t)\ L^i_2(t)\ I^i_p(t)\ I^i_{a,1}(t)\ I^i_{a,2}(t)\ I^i_{a,3}(t)\ I^i_{s,1}(t)\ I^i_{s,2}(t)\ I^i_{s,3}(t)\right]^\top$$

and linearization gives

$$X'(t) = (\beta_0 \cdot T + \Sigma) \cdot X(t),$$

where $T \in \mathbb{R}^{144\times144}$ is the transmission part and $\Sigma \in \mathbb{R}^{144\times144}$ represents the transition mechanisms in the model. The matrix Σ is a block-diagonal matrix, where blocks have size of 9×9 containing transition parameters related to the linear

terms of the system. On the other hand, the transmission matrix T is partitioned into blocks of size 9×9, and each of this blocks have nonzero elements only in their first rows, since transmission between individuals affects only the classes $L_1^i, i \in \{1, 2, \ldots, 16\}$, and these nonzero elements are related to the corresponding elements of the contact matrix and the transmission-related parameter $\inf_a$. The Next-Generation Matrix (shortly NGM) can be calculated as

$$\mathrm{NGM} = -\beta_0 \cdot T\,\Sigma^{-1},$$

and the basic reproduction number is the dominant eigenvalue of the NGM, i.e.,

$$\mathcal{R}_0 = \beta_0 \cdot \rho(-T\,\Sigma^{-1}).$$

On one hand, the model parameters are the same for all countries in the paper; on the other hand, this calculation has to be executed for each countries separately, since social contact matrices (thus T matrices) are different. For more details about NGM method, see [5].

References

1. Aggarwal, C.C. Reddy, C.K., 2014. Data clustering. Algorithms and applications. Chapman and Hall/CRC Data mining and Knowledge Discovery series, Londra. https://www.taylorfrancis.com/books/edit/10.1201/9781315373515/data-clustering-chandan-reddy-charu-aggarwal
2. Ajelli, M. and Litvinova, M., 2017. Estimating contact patterns relevant to the spread of infectious diseases in Russia. Journal of theoretical biology, 419, pp.1–7. Available online: https://www.sciencedirect.com/science/article/pii/S0022519317300504
3. Bishop, C.M. and Nasrabadi, N.M., 2006. Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer. https://link.springer.com/book/9780387310732
4. Carrillo-Larco, R.M. and Castillo-Cara, M., 2020. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome open research, 5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7308996/
5. Diekmann, O., Heesterbeek, J.A.P. and Roberts, M.G., 2010. The construction of next-generation matrices for compartmental epidemic models. Journal of the royal society interface, 7(47), pp.873–885. https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2009.0386
6. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A. and Merler, S., 2012. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002673
7. Grijalva, C.G., Goeyvaerts, N., Verastegui, H., Edwards, K.M., Gil, A.I., Lanata, C.F., Hens, N. and RESPIRA PERU project, 2015. A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru. PloS one, 10(3), p.e0118457. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118457
8. Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1–758). New York: springer. https://link.springer.com/book/10.1007/978-0-387-21606-5

9. Horby, P., Thai, P.Q., Hens, N., Yen, N.T.T., Mai, L.Q., Thoang, D.D., Linh, N.M., Huong, N.T., Alexander, N., Edmunds, W.J. and Duong, T.N., 2011. Social contact patterns in Vietnam and implications for the control of infectious diseases. PloS one, 6(2), p.e16965. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016965
10. Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F.C., Zagheni, E., Merler, S., Ajelli, M., Del Fava, E. and Manfredi, P., 2010. Little Italy: an agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. PLoS computational biology, 6(12), p.e1001021. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001021
11. Kiti, M.C., Kinyanjui, T.M., Koech, D.C., Munywoki, P.K., Medley, G.F. and Nokes, D.J., 2014. Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya. PloS one, 9(8), p.e104786. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104786
12. Klepac, P., Kucharski, A.J., Conlan, A.J., Kissler, S., Tang, M.L., Fry, H. and Gog, J.R., 2020. Contacts in context: large-scale setting-specific social mixing matrices from the BBC Pandemic project. MedRxiv. https://covid-19.conacyt.mx/jspui/handle/1000/232
13. Knipl, D. and Röst, G., 2009. Modelling the strategies for age specific vaccination scheduling during influenza pandemic outbreaks. arXiv preprint arXiv:0912.4662. https://pubmed.ncbi.nlm.nih.gov/21361404/
14. Kumar, S., Gosain, M., Sharma, H., Swetts, E., Amarchand, R., Kumar, R., Lafond, K.E., Dawood, F.S., Jain, S., Widdowson, M.A. and Read, J.M., 2018. Who interacts with whom? Social mixing insights from a rural population in India. PLoS One, 13(12), p.e0209039. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209039
15. Le Polain de Waroux, O., Cohuet, S., Ndazima, D., Kucharski, A.J., Juan-Giner, A., Flasche, S., Tumwesigye, E., Arinaitwe, R., Mwanga-Amumpaire, J., Boum, Y. and Nackers, F., 2018. Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: a survey in Southwest Uganda. BMC infectious diseases, 18(1), pp.1–12. Available online: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-018-3073-1.
16. McCarthy, Z., Xiao, Y., Scarabel, F., Tang, B., Bragazzi, N.L., Nah, K., Heffernan, J.M., Asgary, A., Murty, V.K., Ogden, N.H. and Wu, J., 2020. Quantifying the shift in social contact patterns in response to non-pharmaceutical interventions. Journal of Mathematics in Industry, 10(1), pp.1–25. https://link.springer.com/article/10.1186/s13362-020-00096-y
17. Melegaro, A., Del Fava, E., Poletti, P., Merler, S., Nyamukapa, C., Williams, J., Gregson, S. and Manfredi, P., 2017. Social contact structures and time use patterns in the Manicaland Province of Zimbabwe. PloS one, 12(1), p.e0170459. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0170459
18. Mongi, C.E., Langi, Y.A.R., Montolalu, C.E.J.C. and Nainggolan, N., 2019, July. Comparison of hierarchical clustering methods (case study: Data on poverty influence in North Sulawesi). In IOP Conference Series: Materials Science and Engineering (Vol. 567, No. 1, p. 012048). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/567/1/012048/meta
19. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J. and Heijne, J., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine, 5(3), p.e74. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050074&s=09
20. Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T. and Chen, S., 2022. A machine learning and clustering-based approach for county-level COVID-19 analysis. Plos one, 17(4), p.e0267558. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267558
21. Prem, K., Cook, A.R. and Jit, M., 2017. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS computational biology, 13(9), p.e1005697. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005697

22. Prem, K., Zandvoort, K.V., Klepac, P., Eggo, R.M., Davies, N.G., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Cook, A.R. and Jit, M., 2021. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the COVID-19 era. PLoS computational biology, 17(7), p.e1009098. Available online: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009098
23. Read, J.M., Lessler, J., Riley, S., Wang, S., Tan, L.J., Kwok, K.O., Guan, Y., Jiang, C.Q. and Cummings, D.A., 2014. Social mixing patterns in rural and urban areas of southern China. Proceedings of the Royal Society B: Biological Sciences, 281(1785), p.20140268. https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2014.0268
24. Rizvi, S.A., Umair, M. and Cheema, M.A., 2021. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. Chaos, Solitons and Fractals, 151, p.111240. Available online: https://www.sciencedirect.com/science/article/pii/S0960077921005944
25. Röst, G., Bartha, F.A., Bogya, N., Boldog, P., Dénes, A., Ferenci, T., Horváth, K.J., Juhász, A., Nagy, C., Tekeli, T. and Vizi, Z., 2020. Early phase of the COVID-19 outbreak in Hungary and post-lockdown scenarios. Viruses, 12(7), p.708. https://www.mdpi.com/1999-4915/12/7/708
26. Sadeghi, B., Cheung, R.C. and Hanbury, M., 2021. Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020. BMJ open, 11(11), p.e049844. https://bmjopen.bmj.com/content/11/11/e049844.abstract
27. Wang, D., Shen, H. and Truong, Y., 2016. Efficient dimension reduction for high-dimensional matrix-valued data. Neurocomputing, 190, pp.25–34. Available online: https://www.sciencedirect.com/science/article/pii/S0925231216000084
28. Weerasuriya, C.K., Harris, R.C., McQuaid, C.F., Gomez, G.B. and White, R.G., 2022. Updating age-specific contact structures to match evolving demography in a dynamic mathematical model of tuberculosis vaccination. PLoS computational biology, 18(4), p.e1010002. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010002
29. Zhang, D. and Zhou, Z.H., 2005. (2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. Neurocomputing, 69(1–3), pp.224–231. Availabe online: https://www.sciencedirect.com/science/article/pii/S0925231205001785
30. GitHub repository containing code for the framework proposed in this study https://github.com/zsvizi/clustering-social-patterns-epidemic