DOCTORAL DISSERTATION

Understanding Impact of Social Contact Patterns in Epidemiological Models

Author: Evans Kiptoo Korir
Supervisor: Zsolt Vizi

Faculty of Science and Informatics
Department of Mathematics
Doctoral School of Mathematics
Bolyai Institute, University of Szeged

2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to God for His countless blessings and guidance throughout this journey. The completion of my thesis would not have been possible without the unwavering support of my supervisor, family, colleagues, and friends. I am truly fortunate to have such a supportive and loving circle, and I extend my heartfelt thanks to all those who have contributed to this work, including those whose names may not be mentioned here.

I am profoundly grateful to my supervisor, Dr. Zsolt Vizi, for his invaluable guidance, inspiration, and continuous support throughout my Ph.D. studies. His encouragement and mentorship have played a pivotal role in shaping my research journey. I am especially thankful for his patience, for introducing me to this fascinating field of science, and for pushing me to strive for excellence. His kindness and unwavering support have been instrumental in my academic and personal growth. I also sincerely appreciate his insights into machine learning and programming, which I believe will be of great benefit to me in the future.

I would like to extend my sincere appreciation to my coauthor, Dr. Péter Boldog, with whom I had the privilege of collaborating. His contributions have been immensely valuable to the development of this thesis, and I am truly grateful for his expertise and guidance.

Furthermore, I would like to express my profound gratitude to Professor Gergely Röst and the National Laboratory for Health Security for their generous support, particularly in facilitating my participation in conferences where I had the opportunity to present my research findings. Additionally, I extend my sincere thanks to the dedicated staff members of the department, Csilla, Andrea, and Szilvia, whose efforts and coordination ensured a smooth academic experience.

No words can adequately express my gratitude to my parents, siblings, and extended family for their unconditional love, encouragement, and prayers. Their constant support has been my strength through both joyous and challenging moments, and I will forever be indebted to them.

I would also like to extend my heartfelt thanks to my friends and colleagues, Ibrahim, Thuy, Ngoc, Endre, Dr. Nirmali Prabha Das, and Kata, for their invaluable friendship, moral support, and encouragement. Their presence has made this journey more fulfilling and memorable.

Lastly, I wish to acknowledge all others who have contributed to my academic and personal growth, directly or indirectly. Your support and kindness are deeply appreciated.

Contents

Chapter 1

Introduction

Respiratory infections such as measles, influenza, and COVID-19 are primarily transmitted through direct social interactions between individuals. Accurately modeling these interactions is essential for understanding epidemic dynamics and evaluating interventions such as school closures, social distancing, and vaccination strategies [28, 32, 40, 46, 50]. Age-structured deterministic models are commonly used for this purpose, relying on social contact matrices that quantify the frequency of interactions between different age groups.

Social contact patterns are shaped by a range of demographic and behavioral factors, including age, sex, and individual activity levels, as well as the settings where contacts occur, such as households, schools, workplaces, and public spaces [30]. Among these, age plays a particularly central role in disease transmission. Contact matrices often exhibit assortative mixing, where individuals interact more frequently within their age group and intergenerational interactions, especially in multigenerational households [27, 40, 46].

Extensive studies have documented social contact patterns across diverse geographic

and cultural contexts [2, 3, 14, 18, 21, 26, 31, 38, 47], revealing both common structural features, such as assortative mixing, and significant regional variation. These differences are shaped by demographic composition, cultural norms, and socioeconomic conditions, and they directly influence key epidemiological quantities such as the basic reproduction number ($\mathcal{R}_0$), the timing of epidemic peaks, and the overall disease burden.

Assessing the sensitivity of epidemic models to variation in input parameters is a critical step in understanding and managing infectious disease dynamics. Given the central role of contact structures in determining transmission pathways, it is important to quantify how changes in these inputs influence model outcomes. Sensitivity analysis provides a rigorous framework for evaluating the effects of uncertainty in parameters such as age-specific contact rates on key epidemiological indicators [15, 17, 34, 42, 54]. Commonly used techniques include Monte Carlo simulation, Latin Hypercube Sampling (LHS), and Partial Rank Correlation Coefficients (PRCC), which help identify the most influential factors affecting transmission dynamics and healthcare system burden [56].

Despite these advances, regional differences in contact behavior and their implications for disease control remain insufficiently explored. During the COVID-19 pandemic, countries that experienced later outbreaks often observed and adapted the interventions used in early-affected regions. Understanding which countries share similar social mixing patterns could facilitate more coordinated and context-appropriate responses [35].

Non-pharmaceutical interventions (NPIs) such as social distancing, school closures, and restrictions on movement have proven effective in limiting transmission. However, these measures impose significant socioeconomic burdens, particularly in low- and middle-income countries. Economic disruptions caused by NPIs disproportionately

affect vulnerable populations and sectors such as agriculture and manufacturing, which are more sensitive to mobility restrictions [8, 33, 43, 53].

To address these challenges, this thesis proposes a unified analytical framework that integrates sensitivity analysis and clustering to investigate how age-structured contact patterns and socioeconomic context influence epidemic outcomes. The approach aims to identify key transmission drivers, evaluate contact-based intervention strategies, and classify countries into risk-relevant groups. To achieve these aims, the thesis is structured around the following three methodological objectives:

1. **Develop an eigenvector-based sensitivity analysis framework to identify high-impact age-specific contact patterns.** This objective involves formulating a mathematical method to evaluate how perturbations in age-structured contact matrices influence the basic reproduction number ($\mathcal{R}_0$). Using the spectral properties of the Next Generation Matrix (NGM), the framework derives analytical expressions for the gradient of $\mathcal{R}_0$ with respect to contact rates, using the left and right eigenvectors of the NGM. These pairwise sensitivity values are then used to derive cumulative age-group sensitivity scores, quantifying the overall influence of each age group on transmission dynamics. The framework is further extended to evaluate mortality-related sensitivities by incorporating age-specific probabilities of clinical progression and death.

2. **Develop a statistical sensitivity analysis framework to assess the impact of age-specific contact variations on epidemic outcomes.** This objective introduces the Age Group Sensitivity Analysis (AGSA) method, which integrates Latin Hypercube Sampling (LHS) and Partial Rank Correlation Coefficients (PRCC) within an age-structured epidemic modeling framework. The

method systematically samples plausible reductions in non-household contacts and quantifies their influence on multiple epidemiological outcomes, including $\mathcal{R}_0$, infection peaks, ICU demand, and cumulative deaths. Sensitivity values are computed for individual contact pairs and then summarized into age-group-level scores through a statistically weighted aggregation that accounts for the significance of each pair, using corresponding p-values.

3. **Cluster countries based on standardized social contact patterns and socioeconomic characteristics.** This objective establishes a comparative framework for grouping countries according to similarities in social mixing behavior and structural conditions that influence disease transmission. Contact matrices are first standardized using age-structured epidemic models by calibrating baseline transmission rates to a fixed basic reproduction number ($\mathcal{R}_0$). For European countries, clustering is based solely on these standardized contact matrices. For African countries, contact data are combined with reduced socioeconomic feature sets derived using 1D PCA. The contact matrices themselves are processed using $(2D)^2$ PCA to preserve structural information while reducing dimensionality. The resulting feature vectors are concatenated and used to cluster countries into groups with shared transmission risk profiles using agglomerative hierarchical clustering.

Structure of the Dissertation

This dissertation is structured into 7 main chapters, each addressing a core component of the research aimed at improving the effectiveness of non-pharmaceutical interventions (NPIs) in managing infectious disease outbreaks. Chapter 2.3 focuses

on social contact patterns and the construction of contact matrices, which form the basis for modeling disease transmission. It explains the survey methodologies used to collect contact data, the structure and symmetrization of contact matrices, and their aggregation into broader age groups for computational efficiency. These matrices are central to the sensitivity analyses and clustering studies presented in later chapters.

Chapter 3 presents a series of age-structured epidemic models derived from compartmental ordinary differential equations (ODEs). These models, drawn from literature and adapted for this research, are used to simulate disease dynamics under varying social contact conditions. The chapter also introduces the Next Generation Matrix methodology for computing the basic reproduction number ($\mathcal{R}_0$), which is used as a primary metric throughout the sensitivity analyses.

In Chapter 4, statistical and data-driven methods are employed to analyze uncertainty in contact matrices and their influence on model outputs. Techniques such as Latin Hypercube Sampling (LHS), Partial Rank Correlation Coefficients (PRCC), and principal component analysis (PCA) are introduced. These tools enable both global sensitivity analysis and dimensionality reduction, preparing the data for simulation-based assessments and clustering.

Chapters 5 and 6 present two complementary approaches to sensitivity analysis, both aimed at evaluating how variations in age-specific contact rates influence the dynamics of infectious disease spread. Chapter 5 introduces an analytical method grounded in the spectral properties of the Next Generation Matrix, where the gradient of the basic reproduction number ($\mathcal{R}_0$) is computed with respect to contact elements. In contrast, Chapter 6 adopts a simulation-based approach using Latin Hypercube Sampling (LHS) and Partial Rank Correlation Coefficients (PRCC) to assess the global sensitivity of epidemic outcomes such as infection peaks, hospitalization rates,

and mortality, to changes in contact patterns. The results from both methods are applied to epidemic models introduced in Chapter 3 and are compared to highlight the strengths and limitations of each sensitivity analysis framework.

Finally, Chapter 7 presents a clustering analysis of countries using both social contact patterns and socioeconomic indicators. European countries are clustered based solely on contact matrices, while African countries are grouped using a combined feature set that includes demographic and economic data. This clustering aims to identify regions that may benefit from similar NPI strategies, thereby enhancing regional coordination in epidemic response.

The dissertation concludes with a summary of the key findings and contributions of the research, emphasizing the methodological innovations in sensitivity analysis and the practical implications for designing effective non-pharmaceutical interventions. It also includes a list of peer-reviewed publications on which this thesis is based. Additionally, information regarding data sources and availability is provided to facilitate the reproducibility of results. For completeness, the Appendix provides the ordinary differential equations underlying the age-structured models, along with tables summarizing the parameter values used in the epidemic models discussed in Chapter 3.

Chapter 2

Social Contact Patterns

2.1 Problem Setting

Social contact matrices play a fundamental role in understanding the transmission dynamics of infectious diseases. These matrices quantify interactions between different age groups and are essential for constructing realistic epidemiological models. Contact matrices are derived primarily from diary-based surveys in which participants record their daily interactions across various settings, such as home, work, school, and public places. Researchers estimate the frequency and intensity of contacts among different demographic groups by systematically collecting this information, providing crucial insights into potential transmission pathways.

The reliability of social contact matrices depends on accurate data collection and proper adjustments to ensure consistency. Several studies, including those by [2, 3, 14, 18, 21, 26, 31, 32, 38, 47], have constructed contact matrices using demographic data and survey responses, refining the estimates through statistical techniques such as Markov Chain Monte Carlo simulations. Once compiled, these matrices are adjusted

for reciprocity, ensuring that the total number of reported contacts between two groups remains symmetric. Additionally, matrices are often aggregated to align with broader population demographics, enhancing their applicability in large-scale modeling efforts. The resulting matrices serve as a foundation for evaluating the impact of NPIs, such as social distancing and school closures.

2.2 Survey Methodology for Collecting Social Contact Data

Researchers primarily derive social contact data from cross-sectional surveys conducted by commercial entities or public health organizations. Sometimes, they generate this information by constructing virtual populations that use highly detailed census and demographic data as parameters. Survey organizers ensure that participants represent the broader population in terms of age, sex, and geography. Recruitment methods include random-digit dialing via landlines, face-to-face interviews, or utilizing population registers linked to national seroepidemiology surveys. Researchers often oversample these groups, as children and young adults tend to have naturally high contact rates. Typically, only one household member is required to participate to gather household-level data.

Participants received paper diaries either through the mail or in person, often accompanied by detailed instructions on completing them. These diaries collected the participants' sociodemographic details, such as age, sex, household composition, and employment status. On a randomly selected day, participants recorded every person they interacted with from 7 a.m. to 7 a.m. the following morning. They were instructed to log each contact only once. Contacts fall into two categories: physical

contact (e.g., skin-to-skin interactions such as handshakes or kisses) and non-physical contact (e.g., a two-way conversation involving at least three words, conducted in the physical presence of another individual without any physical touch).

Participants provided additional details about each contact, including their age and sex. If participants could not determine a contact's exact age, they estimated an age range, with the midpoint used for analysis. They also recorded the location of each interaction (e.g., home, work, school, leisure, transport, or other), the duration of the contact (e.g., less than 5 minutes, 5–15 minutes, 15 minutes to 1 hour, 1–4 hours, or more than 4 hours), and the frequency of these interactions (e.g., daily, weekly, monthly, or for the first time).

Researchers translated the diaries into local languages to improve accessibility. For young children, a parent or guardian completes the diary on their behalf. Older children with parental consent used simplified diaries they could fill out independently.

2.3 Social Contact Matrices

The social contact hypothesis posits that the number of secondary infections an infected individual generates correlates with their social interactions [57]. Using survey methodologies outlined in Section 2.2, [40] estimated social contact matrices for European countries, including Belgium, Germany, Finland, Great Britain, Italy, Luxembourg, the Netherlands, and Poland. Expanding on this work, [45] used updated demographic and survey data from sources like the United Nations Population Division, International Labour Organization, and World Bank databases to estimate social contact matrices for 152 countries. These estimates were further projected to cover 177 countries in a subsequent study [46].

The contact matrices from [46] reflect interactions in four primary settings: home,

work, school, and other locations such as public transport and marketplaces. These matrices are structured to cover 16 age groups, with each group spanning five years (e.g., $0-4, 5-9, \ldots, 75+$). Markov Chain Monte Carlo simulations were employed to estimate the contact patterns and validate them against previously established matrices.

We define $\widetilde{c}_{i,j}$ as the daily per capita contact rate, which is the average number of contacts that an individual in age group i has with individuals in age group j. These values are derived from survey data and demographic estimates, and they form the entries of an initial (asymmetric) contact matrix $\widetilde{C} \in \mathbb{R}^{n_a \times n_a}$, where n_a is the number of age groups.

Social contact matrices are inherently asymmetric. However, under the principle of reciprocity, the total number of contacts between age groups should be balanced:

$$\widetilde{c}_{i,j} P_i = \widetilde{c}_{j,i} P_j,$$

where P_i and P_j denote the population sizes of groups i and j, respectively. To address asymmetries caused by sampling variability, we apply the symmetrization method described in [28]. For a contact matrix $\widetilde{C} \in \mathbb{R}^{n_a \times n_a}$, the symmetric total matrix is calculated as:

$$c_{i,j}^{\text{total}} = \frac{\widetilde{c}_{i,j} P_i + \widetilde{c}_{j,i} P_j}{2}. \tag{2.1}$$

Next, the adjusted contact matrix is computed as:

$$c_{i,j} = \frac{c_{i,j}^{\text{total}}}{P_i},$$

where $c_{i,j}$ represents the element in the adjusted social contact matrix corresponding to row i and column j. The complete adjusted contact matrix C is defined as:

$$C = [c_{i,j}]_{i,j=1}^{n_a} \in \mathbb{R}^{n_a \times n_a}.$$

If C^{H}, C^{S}, C^{W}, and C^{O} denote the contact matrices for home, school, work, and other settings respectively, the full contact matrix C combines all settings as:

$$C = C^{\mathrm{H}} + C^{\mathrm{S}} + C^{\mathrm{W}} + C^{\mathrm{O}}. \tag{2.2}$$

As an example, the adjusted social contact matrices for Hungary are illustrated in Figure 2.1.

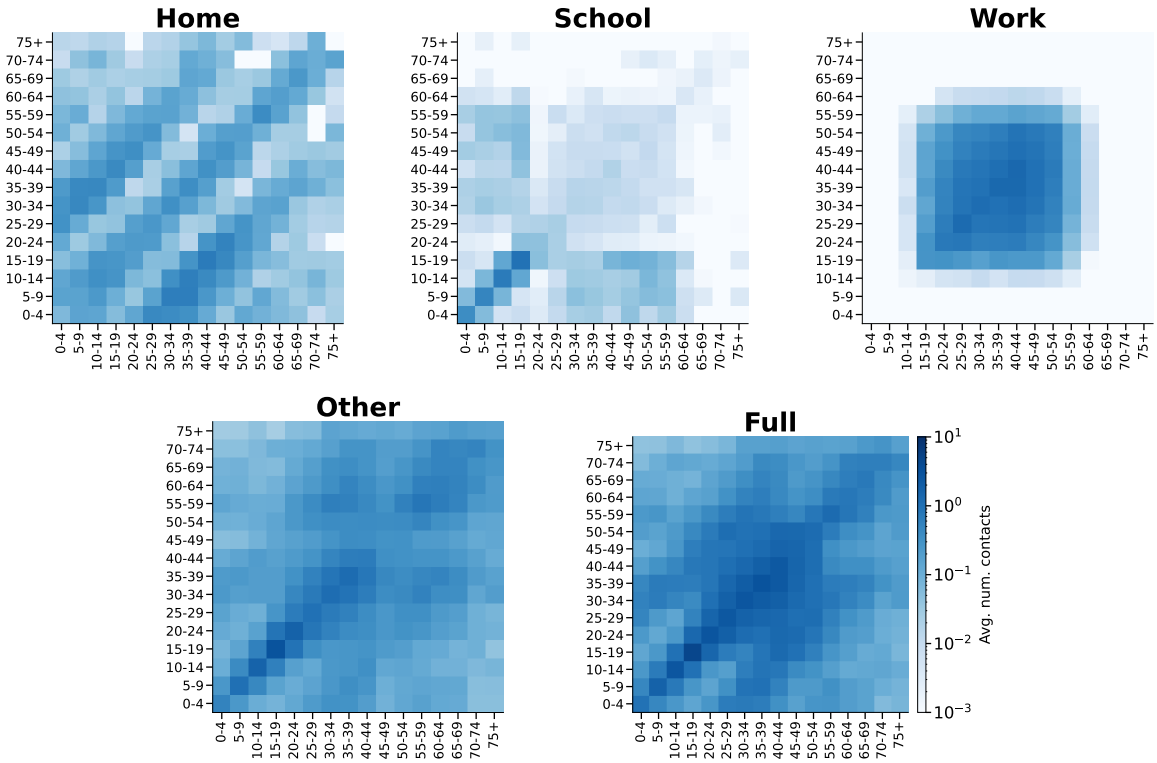

Figure 2.1: Age-specific contact matrices for Hungary across different settings: Home, School, Work, Other, and Full Contact, adapted from Prem et al. (2021) [46]. The full contact matrix is the cumulative sum of the four setting-specific matrices. Each matrix element represents the average number of contacts between individuals from different age groups. The horizontal axis represents survey participants (respondents), while the vertical axis corresponds to the individuals they reported having contact with (contactees). Color intensity reflects contact frequency, with darker blue shades indicating higher interaction levels, as visualized using a blue color map.

It is important to recognize that interventions such as school closures, social distancing, and remote work primarily impact school, work, and other contact settings.

Home interactions remain unaffected and must be treated separately. To account for this, we define C^{Δ} as the sum of the school, work, and other contact matrices:

$$C^{\Delta} = C^{\mathrm{S}} + C^{\mathrm{W}} + C^{\mathrm{O}}. \tag{2.3}$$

Thus, the full contact matrix can be expressed as $C = C^{\mathrm{H}} + C^{\Delta}$. The elements of these contact matrices are considered parameters in the sensitivity analysis presented in chapters 5 and 6. For the eigenvector-based sensitivity analysis in chapter 5, we reversed the scaling in Eq. (2.1) using the population to derive the adjusted full contact matrix defined in Eq. (2.2) and illustrated in Fig. 2.2.

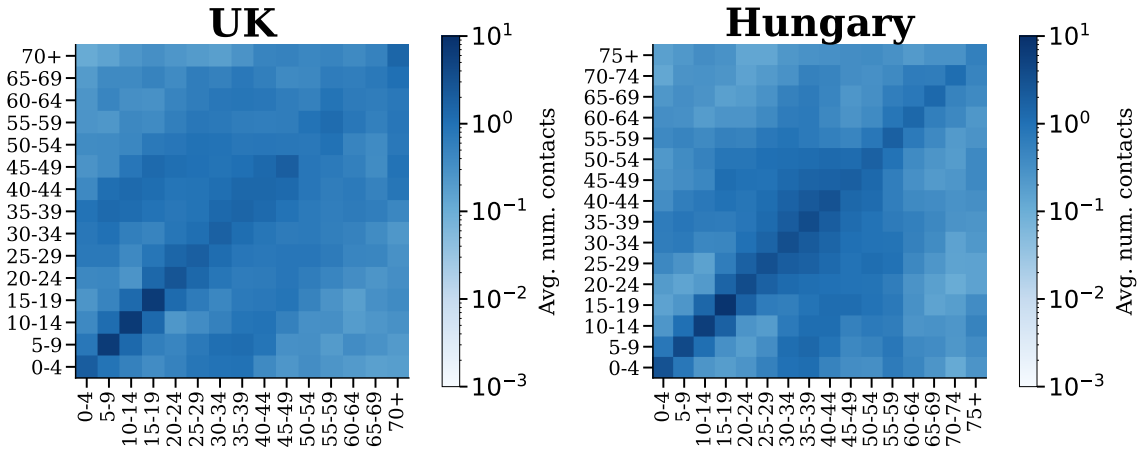

Figure 2.2: Adjusted age-specific full contact matrices for the UK [40] and Hungary [46]. The matrix elements represent the average number of contacts between individuals of different age groups. The horizontal axis represents respondents, while the vertical axis corresponds to the contactees. The color intensity indicates contact frequency, with light blue representing lower contact rates and dark blue denoting higher contact rates, as shown in the colormap. These matrices are used as input for the sensitivity analysis in Section 5 and 6.

For regional analysis, we consider age-structured contact patterns from 39 European and 32 African countries, as provided in [46]. For each country c, the full contact matrix is formed by aggregating setting-specific matrices:

$$C(\boldsymbol{c} = c) = C^{\mathrm{H}}(\boldsymbol{c} = c) + C^{\mathrm{S}}(\boldsymbol{c} = c) + C^{\mathrm{W}}(\boldsymbol{c} = c) + C^{\mathrm{O}}(\boldsymbol{c} = c),$$

where $C^t(\boldsymbol{c} = c)$ corresponds to the contact matrix for setting $t \in \{\mathrm{H}, \mathrm{S}, \mathrm{W}, \mathrm{O}\}$. We apply the symmetrization method from Eq. (2.1) and use the country-specific population vector $P(\boldsymbol{c} = c)$ to obtain the adjusted matrix $C(\boldsymbol{c} = c)$.

2.4 Aggregation of Social Contact Matrices

To adapt the social contact matrix $C(\boldsymbol{c} = c) \in \mathbb{R}^{n_a \times n_a}$ of the country c, where n_a represents the number of original age bins, to a new set of age bins, we define the aggregated contact matrix $\mathbf{C}(\boldsymbol{c} = c) \in \mathbb{R}^{n_g \times n_g}$, where n_g represents the number of new age bins. Additionally, we define the aggregated population vector $\mathbf{P}(\boldsymbol{c} = c) \in \mathbb{R}^{n_g}$, corresponding to these new age bins.

Let G_i denote the set of indices corresponding to the original age bins that are aggregated into the i-th new age bin. The population in the i-th new age bin is computed as:

$$\mathbf{P}_i(\boldsymbol{c} = c) = \sum_{m \in G_i} P_m(\boldsymbol{c} = c).$$

The elements of the aggregated contact matrix $\mathbf{c}_{i,j}(\boldsymbol{c} = c)$ are calculated by averaging the interactions between all combinations of original age bins $m \in G_i$ and $m' \in G_j$. The formula for the aggregated contact matrix element $\mathbf{c}_{i,j}(\boldsymbol{c} = c)$ is given by:

$$\mathbf{c}_{i,j}(\boldsymbol{c} = c) = \frac{1}{\mathbf{P}_i(\boldsymbol{c} = c)} \sum_{(m,m') \in G_i \times G_j} C_{m,m'}(\boldsymbol{c} = c) \cdot P_m(\boldsymbol{c} = c).$$

Here, $\mathbf{c}_{i,j}(\boldsymbol{c} = c)$ represents the country-specific aggregated adjusted full contact element, and the aggregated contact matrix is denoted by $\mathbf{C}(\boldsymbol{c} = c)$. This aggregation approach preserves the proportionality of the aggregated contact matrix to the original contact rates and population sizes, ensuring that $\mathbf{C}(\boldsymbol{c} = c)$ accurately represents the original demographic and contact data within the new age structure. In practice, the

new age structure is typically designed to balance the resolution of interaction patterns and computational feasibility. For instance, one might group the population into six new age bins: 0–4, 5–14, 15–24, 25–44, 45–64, and 65+ as shown in Fig. 2.3.

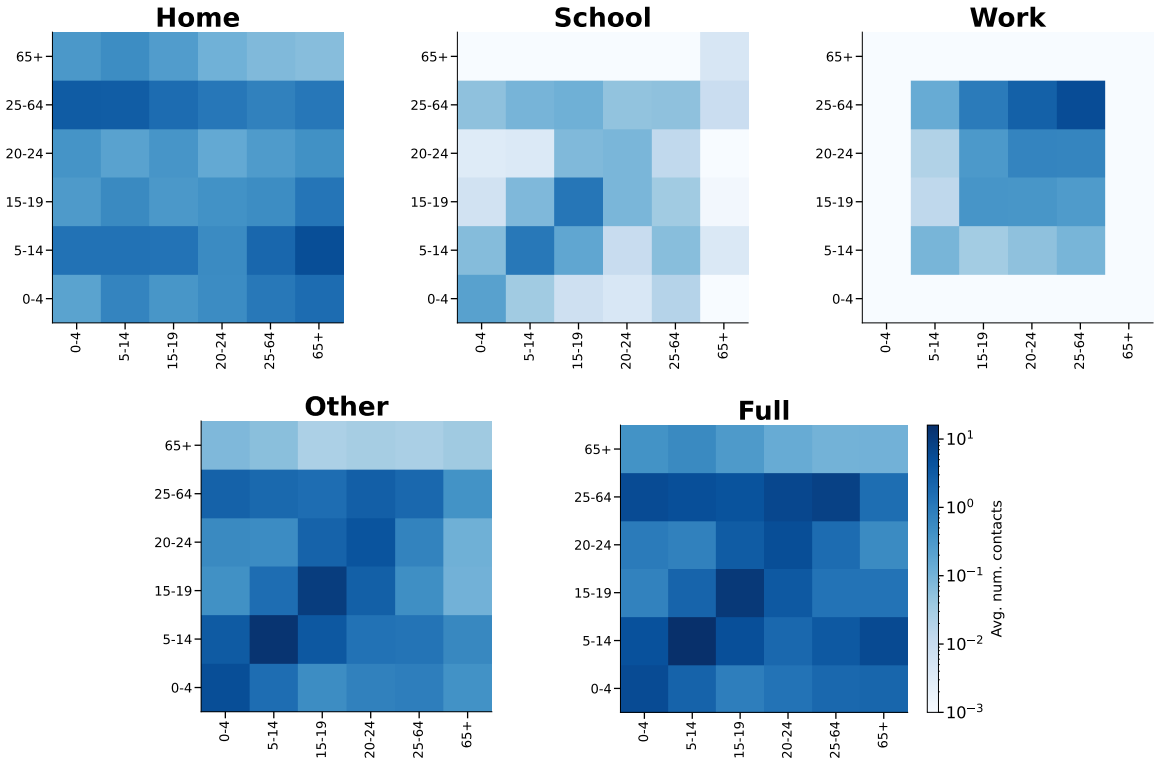

Figure 2.3: Aggregated age-specific contact matrices for Kenya across various settings: Home, School, Work, Other, and Full Contact matrix. The Full Contact matrix represents the combined sum of the matrices from the individual settings. Each matrix element indicates the average number of contacts between individuals from specific age groups, where the horizontal axis represents the survey participants (respondents) and the vertical axis corresponds to the individuals they reported having contact with (contactees). The color intensity, depicted using the blue color map, ranges from light blue for lower contact frequencies to dark blue for higher frequencies.

Chapter 3

Age-structured Epidemic Models

3.1 Problem Setting

Understanding the spread of infectious diseases requires robust mathematical models that capture key epidemiological dynamics. Among these, age-structured epidemic models play a critical role in analyzing disease transmission across different age groups and geographical regions. These models offer insights that are essential for designing effective public health interventions. Traditional epidemic models, such as the Susceptible-Infected-Recovered (SIR), Susceptible-Exposed-Infected-Recovered (SEIR), and Susceptible-Infected-Susceptible (SIS) frameworks, provide a foundation for studying disease propagation. These models are typically expressed as systems of ordinary differential equations (ODEs) and incorporate variables such as transmission rates, contact patterns, and demographic factors. However, standard models often assume a homogeneous population, overlooking the variability in social behavior and disease susceptibility across different age groups. Age-structured models address this limitation by integrating demographic stratification, allowing for more precise

epidemiological predictions.

To further enhance biological realism, we incorporate the linear chain trick, a widely used technique that approximates gamma-distributed waiting times through multiple exponential stages. Standard models often assume that transitions between compartments follow an exponential distribution, which can result in unrealistic scenarios, such as individuals immediately leaving a compartment or remaining there indefinitely. The linear chain trick mitigates these artifacts by subdividing each stage (e.g., latent or infectious periods) into two or more sequential sub-compartments [59]. This results in a more accurate representation of time spent in each disease state and better alignment with empirical data. The model presented in Section 3.4 uses this approach to capture more realistic disease progression dynamics.

This chapter presents a set of age-structured epidemic models aimed at examining disease transmission dynamics in various regions. A key focus is on understanding how social contact patterns influence the spread of infection. This analysis is grounded in the Next Generation Matrix (NGM) methodology, a powerful mathematical tool used to quantify disease transmissibility and evaluate intervention strategies. The NGM approach allows for the formulation of inter-age group transmission dynamics and facilitates the calculation of the basic reproduction number ($\mathcal{R}_0$), a critical measure of disease spread potential.

We apply these modeling approaches to a range of geographic settings, the United Kingdom (Section 3.3) and Hungary (Section 3.4). Additionally, we extend the methodology to assess disease transmission in broader international contexts, examining variations in contact patterns and transmission dynamics across European and African countries using the model outlined in Section 3.4. Leveraging the age-stratified contact matrices presented in Section 2.3, we investigate how demographic structures and

social behavior influence epidemic trajectories.

3.2 Next Generation Matrix Methodology

In this chapter, we introduce model-specific notations such as $\mathcal{R}_0(m)$, $K(m)$, and $F(m)$ to distinguish between different epidemiological models. For clarity, we use m consistently to denote the model under consideration. The basic reproduction number for model m, denoted as $\mathcal{R}_0(m)$, is a key threshold parameter in infectious disease modeling. It represents the expected number of secondary infections caused by a single infected individual in a fully susceptible population at the beginning of an epidemic (typically when there is a small number of infected individuals). If $\mathcal{R}_0(m) > 1$, the infection is expected to spread, whereas if $\mathcal{R}_0(m) < 1$, the disease is likely to die out.

To determine $\mathcal{R}_0(m)$ for different epidemic models m, the Next Generation Matrix methodology, as established by Diekmann et al. [11], is employed. This approach provides a structured framework to analyze disease transmission by examining the early-phase dynamics of an outbreak in a compartmental model. It is especially useful for age-structured and other heterogeneous population models. The NGM methodology begins by linearizing the infectious subsystem of model m around its disease-free equilibrium (DFE). The resulting linearized system is represented by a matrix that can be decomposed into two fundamental matrices:

- Transmission Matrix $F(m)$: quantifies the rate at which new infections are produced by currently infected individuals. This depends on key transmission parameters, including the baseline transmission rate $\beta(m)$ and the contact data (defined in Section 2.3).

- Transition Matrix $V(m)$: describes the shifting of individuals through different

infectious states, including latency, infectiousness, recovery, and death.

At the DFE, the linearized infectious subsystem is represented as:

$$X'(m)(t) = (F(m) + V(m)) \cdot X(m)(t),$$

where $X(m)(t)$ is the vector representing infectious compartments at time t.

In many compartmental models m, the infected population is distributed across multiple states, some of which may not immediately contribute to new infections. To account for all infectious states, the NGM with Large Domain, denoted as $K_L(m)$, is constructed and defined by:

$$K_L(m) = -F(m) \cdot V^{-1}(m). \tag{3.1}$$

This matrix captures all infectious states, including those that do not directly contribute to disease transmission. Importantly, the entries of $K_L(m)$ can be interpreted as pairwise reproduction numbers, that is, each entry (i, j) represents the expected number of new infections in compartment i, caused by a single individual who was initially infected in compartment j.

To focus only on those states directly responsible for generating new infections, an Auxiliary Matrix $E(m)$ is introduced to select these states. The reduced NGM is then given by:

$$K(m) = E(m) \cdot K_L(m) \cdot E^{\top}(m). \tag{3.2}$$

This reduction ensures that only the compartments occupied immediately after infection are retained in $K(m)$. Since $K_L(m)$ and $K(m)$ share the same eigenvalues, but the former may have additional dimensions that complicate computation, the reduced NGM provides a more practical formulation. Therefore, in this thesis, we will work with $K(m)$, which retains only the compartments occupied immediately after infection.

Each entry $k_{ij}(m)$ of $K(m)$ represents the expected number of secondary infections in newly infected compartment i, caused by a single infected individual who initially entered compartment j. These entries can be interpreted as pairwise reproduction numbers, typically across age groups or infectious classes, depending on the specific model structure.

The selection matrix $E(m)$ is structured to extract only those compartments that correspond to newly infected individuals immediately after infection. Specifically, it selects n_a compartments from a larger space of infectious states. It is defined as:

$$E(m) = \begin{bmatrix} —e_1— \\ —e_2— \\ \vdots \\ —e_{n_a}— \end{bmatrix}, \tag{3.3}$$

where each row vector e_i is a standard basis row vector with a 1 in the position corresponding to the newly infected compartment for age group i, and zeros elsewhere.

Once $K(m)$ is determined, the basic reproduction number is given by the spectral radius (i.e., the largest eigenvalue in magnitude) of $K(m)$:

$$\mathcal{R}_0(m) = \rho(K(m)). \tag{3.4}$$

The spectral radius quantifies the dominant growth rate of infection per generation, serving as a threshold parameter to assess whether an epidemic will persist or die out.

This methodology is applied to different epidemic models m in the sections that follow. Each model defines its matrices $F(m)$, $V(m)$, and $E(m)$, which are then used to calculate the corresponding $K(m)$. Sections 3.3 and 3.4 demonstrate this process with specific infectious disease examples, including detailed calculations of the basic reproduction number $\mathcal{R}_0(m)$ for each model.

3.3 Pitman et al. Model

The SIR model, first developed by Kermack and McKendrick [7], forms the foundation for understanding the spread of infectious diseases. Initially applied to study the transmission of diseases such as measles and influenza in populations, this model was later extended to include an exposed (latent) state, giving rise to the SEIR model. Pitman et al. further expanded this model to evaluate the impact of vaccination programs, specifically in the context of the UK and Wales [44].

The SEIR model describes the movement of individuals between four distinct states: susceptible (S), exposed (E), infectious (I), and recovered (R). These states are linked by transitions governed by key parameters such as the transmission rate, the rate at which individuals progress from exposed to infectious (α), and the recovery rate (γ). The model structure is visually represented in Fig. 3.1, where individuals flow between these compartments over time.

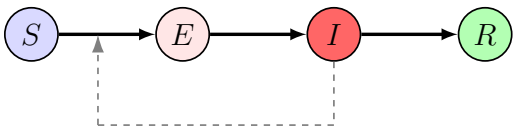

Figure 3.1: Schematic representation of the classical SEIR model used by Pitman et al. [44], showing the flow of individuals through the disease states. This version of the model does not include vital dynamics (births or deaths). The dashed lines indicate the feedback loop from infectious individuals contributing to new infections in the susceptible population, capturing the core transmission mechanism.

The model employs a *Who Acquires Infection From Whom* (WAIFW) matrix based on data from the POLYMOD study [40], which captures age-specific contact patterns across 15 groups, distinguishing between physical and non-physical interactions. This UK-specific contact structure plays a central role in shaping the model's transmission

dynamics. In our simulations, we adopt the original parameters of the model, calibrated for the population of the UK and Wales, with two modifications: we assume no waning immunity, thus reducing the framework to a standard SEIR-type model, and we exclude vaccination. The model assumes uniform susceptibility across all age groups and sets the baseline reproduction number $\overline{\mathcal{R}}_0(m = \mathrm{P})$ to 1.8. The governing equations are provided in Appendix A.1, with corresponding parameter values, including the latent period α and recovery rate γ, summarized in Table A.1.

Following the Next Generation Matrix methodology introduced in Section 3.2, the infectious state vector $X(m = \mathrm{P})(t)$ from Eq. (A.1) includes the age-structured compartments E_i and I_i, where $i = 1, \ldots, 15$. The transmission matrix $F(m = \mathrm{P})$ is defined based on the baseline transmission rate $\beta(m = \mathrm{P})$ and age-specific contact patterns, reflecting the rate of new infections entering the exposed compartment E_i. The structure of $F(m = \mathrm{P})$ is given by:

$$F_{j,i}(m = \mathrm{P}) = \beta(m = \mathrm{P}) \cdot c_{j,i}(m = \mathrm{P}) \cdot \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Here, only infectious I_j individuals are assumed to contribute to new infections entering the E_i compartment, consistent with the transmission structure illustrated in Fig. 3.1. The full matrix $F(m = \mathrm{P}) \in \mathbb{R}^{30\times30}$ consists of 15×15 blocks of size 2×2, encoding all pairwise age-specific transmission interactions.

The transition matrix $V(m = \mathrm{P})$ describes the progression dynamics between successive infectious states in the model. It is structured as a block-diagonal matrix with 15×15 diagonal blocks $V_{i,i}(m = \mathrm{P}) \in \mathbb{R}^{2\times2}$, each corresponding to age group

$i = 1, \ldots, 15$. The block $V_{i,i}(m = \mathrm{P})$ is defined as:

$$V_{i,i}(m = \mathrm{P}) = \begin{bmatrix} -\alpha & 0 \\ \alpha & -\gamma \end{bmatrix}$$

The full matrix $V(m = \mathrm{P}) \in \mathbb{R}^{30\times 30}$ consists of these blocks along its diagonal. The NGM with large domain is then given by $K_L(m = \mathrm{P}) \in \mathbb{R}^{30\times 30}$.

To isolate the compartments representing newly infected individuals (i.e., the exposed states E_i), we define the selection matrix $E(m = \mathrm{P}) \in \mathbb{R}^{15\times 30}$ as a block-diagonal matrix with 15 blocks of size 1×2. Each block corresponds to an age group and selects the E_i compartment from the local $[E_i, I_i]$ pair using the row vector [1 0]. The structure of E is:

$$E(m = \mathrm{P}) = \begin{bmatrix} \boxed{1\ 0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boxed{1\ 0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boxed{1\ 0} \end{bmatrix} \in \mathbb{R}^{15\times 30}.$$

This formulation ensures that Eq. (3.2) yields the reduced NGM, $K(m = \mathrm{P}) \in \mathbb{R}^{15\times 15}$, whose spectral radius defines $\mathcal{R}_0$.

3.4 Röst et al. Model

This model consists of 15 compartments, each representing a distinct stage in disease progression. The compartment S represents individuals who are susceptible to the disease, meaning they are at risk of infection. Once exposed, individuals move to the L compartment, which represents the latent period. During this phase, they are infected but remain asymptomatic and cannot transmit the disease.

After the latent period, individuals transition to the $I^{(p)}$ compartment, which

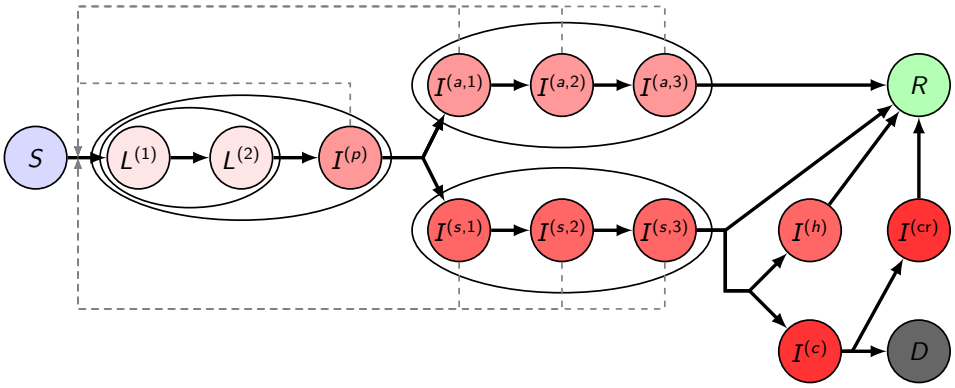

Figure 3.2: Flowchart illustrating the age-structured compartmental model adapted from [50], representing COVID-19 transmission dynamics in Hungary. Compartments are color-coded to reflect disease severity. Dashed grey lines denote the set of compartments contributing to onward transmission, including pre-symptomatic, asymptomatic, and symptomatic stages.

accounts for pre-symptomatic cases. At this stage, individuals are infected and capable of transmitting the disease, although they have not yet developed symptoms. The model distinguishes between those who remain asymptomatic or exhibit mild symptoms $\left(I^{(a)}\right)$ and those who develop symptomatic infections $\left(I^{(s)}\right)$. To more accurately capture variability in recovery times and avoid unrealistic assumptions about exponentially distributed durations, both asymptomatic and symptomatic cases are further divided into three sequential sub-stages.

Individuals from the $I^{(a,3)}$ compartment will all recover and transition to the recovered class R. Symptomatic individuals, depending on disease severity, either recover and enter R or require hospitalization. Hospitalized cases are categorized into two groups: those receiving standard hospital care $\left(I^{(h)}\right)$ and those requiring intensive care $\left(I^{(c)}\right)$. Individuals from $I^{(h)}$ will proceed to R after recovery. Patients in intensive care may recover and move to a rehabilitation unit $\left(I^{(\mathrm{cr})}\right)$ or face a fatal outcome, represented by the D compartment.

To improve the accuracy of disease progression dynamics, this model applies the

linear chain trick, as introduced in the problem setting (Section 3.1). The latent period is divided into two sub-compartments $\left(L^{(1)}\right)$ and $\left(L^{(2)}\right)$, and both asymptomatic and symptomatic infectious periods are modeled using three sequential stages $\left(I^{(a,1)} \text{ to } I^{(a,3)}\right)$ and $\left(I^{(s,1)} \text{ to } I^{(s,3)}\right)$, respectively. This effectively reduces variance in transition times and improves the model's alignment with empirical epidemiological data. To account for differences in disease dynamics across age groups, the Hungarian population is stratified into 16 age groups, as defined by available Hungarian contact data [46]. The system of ordinary differential equations (ODEs) governing this model is provided in Eq. (A.2) in Appendix A. The compartments in Fig. 3.2 and Eq. (A.2) are indexed by age group, with $i \in 1, \ldots, 16$. The parameter $\beta(m = \mathrm{R})$ denotes the baseline transmission probability. The model incorporates age-specific contact matrices $c_{i,j}(m = \mathrm{R})$, which capture variations in social interactions across age groups. Additionally, age-dependent parameters are used to model disease progression transitions, providing a more detailed understanding of how the disease spreads among different age categories. For a complete list of age-dependent parameters used in this model, refer to Table A.2 in Appendix A.

In the clustering analysis in Section 7, we will assume the model parameters are consistent across all countries, except for the baseline transmission rate, which is closely tied to the contact matrix of each country. We denote the baseline transmission rate for a given country c as $\beta(\boldsymbol{c} = c)$.

To compute the Next Generation Matrix (NGM) for the Röst et al. model, we follow the approach introduced in Section 3.2. The infectious state vector $X(m = \mathrm{R})(t)$ consists of the compartments $L_i^{(1)}, L_i^{(2)}, I_i^{(p)}, I_i^{(a,1)}, I_i^{(a,2)}, I_i^{(a,3)}, I_i^{(s,1)}, I_i^{(s,2)}, I_i^{(s,3)}$. The transmission matrix $F(m = \mathrm{R})$ captures the structure of the force of infection defined in Eq. (A.2). It is organized as a block matrix with 16×16 blocks $F_{j,i}(m = \mathrm{R}) \in \mathbb{R}^{9\times 9}$,

where each block represents transmission from infectious individuals in age group j to susceptibles in age group i. Each block is given by:

$$F_{j,i}(\mathfrak{m} = \mathrm{R}) = \beta(\mathfrak{m} = \mathrm{R}) \cdot \sigma_i \cdot c_{j,i}(\mathfrak{m} = \mathrm{R}) \cdot \begin{bmatrix} 0 & 0 & 1 & \inf^{(a)} & \inf^{(a)} & \inf^{(a)} & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Only the first row is nonzero, reflecting that new infections enter the first latent compartment $L_i^{(1)}$. The nonzero terms correspond to infectious compartments $I_j^{(p)}$, $I_j^{(a,m)}$, and $I_j^{(s,m)}$, with $\inf^{(a)}$ capturing reduced transmissibility from asymptomatic individuals. Thus, $F(\mathfrak{m} = \mathrm{R}) \in \mathbb{R}^{144\times144}$.

The transition matrix $V(\mathfrak{m} = \mathrm{R})$ captures the progression rates between successive infectious states. It is block-diagonal, with each diagonal block $V_{i,i}(\mathfrak{m} = \mathrm{R}) \in \mathbb{R}^{9\times9}$ representing within-group transitions for age group i. Each $V_{i,i}(\mathfrak{m} = \mathrm{R})$ is given by:

$$\begin{bmatrix} -\alpha^{(L)} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha^{(L)} & -\alpha^{(L)} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha^{(L)} & -\alpha^{(p)} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_i\alpha^{(p)} & -\gamma^{(a)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma^{(a)} & -\gamma^{(a)} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \gamma^{(a)} & -\gamma^{(a)} & 0 & 0 & 0 \\ 0 & 0 & (1-p_i)\alpha^{(p)} & 0 & 0 & \gamma^{(a)} & -\gamma^{(s)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma^{(s)} & -\gamma^{(s)} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \gamma^{(s)} & -\gamma^{(s)} \end{bmatrix}$$

The entries represent sequential progression within age group i, including transitions from latent to pre-symptomatic, and then to either asymptomatic or symptomatic

pathways, followed by recovery. The full matrix $V(m = \mathrm{R}) \in \mathbb{R}^{144\times144}$ consists of these blocks along its diagonal. Following the approach in Eq. (3.1), we construct NGM with large domain $K_L(m = \mathrm{R}) \in \mathbb{R}^{144\times144}$. However, not all infectious states directly contribute to new infections. The selection matrix $E(m = \mathrm{R}) \in \mathbb{R}^{16\times144}$ is constructed to isolate the compartments $L_i^{(1)}$ for each age group i, which represent individuals immediately after infection. It is a block-diagonal matrix, where each block is a row vector of the form $[1, 0, \ldots, 0] \in \mathbb{R}^{1\times9}$. These blocks are aligned along the row corresponding to each age group, so that the full matrix has the form:

$$E(m = \mathrm{R}) = \begin{bmatrix} \boxed{1\ 0\ \cdots\ 0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boxed{1\ 0\ \cdots\ 0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boxed{1\ 0\ \cdots\ 0} \end{bmatrix} \in \mathbb{R}^{16\times144}.$$

Then we can determine the NGM with a small domain for the model as $K(m = \mathrm{R}) \in \mathbb{R}^{16\times16}$.

Chapter 4

Statistical and Data-Driven Approaches

In computational modeling and the study of uncertainty quantification, statistical and data-driven techniques are indispensable for enhancing the accuracy and reliability of analytical findings. These methods enable efficient sampling, sensitivity analysis, and dimensionality reduction, which are crucial for interpreting complex relationships within high-dimensional parameter spaces.

This chapter examines fundamental statistical approaches used to evaluate uncertainties in contact patterns within epidemic modeling, with a particular emphasis on Latin Hypercube Sampling (LHS) and Partial Rank Correlation Coefficients (PRCC). LHS is a widely employed sampling method that optimizes uncertainty quantification by ensuring a well-distributed representation of parameter variations while reducing computational demands. Meanwhile, PRCC serves as a robust tool for sensitivity analysis based on LHS, offering insights into the extent to which variations in contact patterns influence epidemic outcomes. Beyond these primary techniques, the chapter

also explores dimensionality reduction methods, such as Principal Component Analysis (PCA), which are essential for addressing challenges posed by high-dimensional data. Furthermore, clustering techniques, especially hierarchical clustering, are utilized to group data based on inherent similarities, improving the interpretability of contact structures across different regions. Finally, the chapter introduces socioeconomic indicators, which are used to assess their impact on adherence to non-pharmaceutical interventions (NPIs) and to understand their role in shaping clustering patterns within epidemiological models.

4.1 Latin Hypercube Sampling (LHS)

LHS Parameter Samples				**Target Variable**
p_1	p_2	$\cdots$	p_{n_p}	t
$p_1^{(1)}$	$p_2^{(1)}$	$\cdots$	$p_{n_p}^{(1)}$	$t^{(1)}$
$p_1^{(2)}$	$p_2^{(2)}$	$\cdots$	$p_{n_p}^{(2)}$	$t^{(2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$p_1^{(n_s)}$	$p_2^{(n_s)}$	$\cdots$	$p_{n_p}^{(n_s)}$	$t^{(n_s)}$

Table 4.1: Structure of the LHS matrix and its associated outputs. The left section illustrates the parameter samples $p_q^{(i)}$, the i-th sample of the q-th parameter, generated for each of the n_s sample vectors. The right section presents the corresponding value $t^{(i)}$ of the target variable, obtained from the simulation performed using the i-th sample vector. Each row represents a unique sample-output pair.

Latin Hypercube Sampling (LHS) is an efficient statistical method used to generate representative samples from multidimensional parameter distributions. Introduced by McKay et al. in 1979 [36], LHS offers a more systematic and reliable alternative to random sampling, enabling similar accuracy with fewer samples. This technique is widely applied in Monte Carlo simulations and sensitivity analyses.

In LHS, each model parameter is assumed to follow a specific probability distribution

with an associated density function. To ensure comprehensive coverage of the parameter space, the range of each parameter is divided into n_s non-overlapping intervals of equal probability. According to McKay, the sample size n_s should satisfy $n_s > \frac{4}{3}n_p$, where n_p represents the number of parameters under consideration [6, 37]. A value is then randomly selected from each interval without repetition. This process is repeated across all parameters. The resulting LHS matrix has rows that represent individual sample vectors $\left(p_1^{(i)}, \ldots, p_{n_p}^{(i)}\right)$, with each vector used as input for a separate simulation. The outputs of these simulations are then generated, and the values $t^{(i)}$ of the investigated target variable are obtained for every parameter configuration. This is illustrated in Table 4.1.

4.2 Partial Rank Correlation Coefficients (PRCC)

The Partial Rank Correlation Coefficient (PRCC), commonly used in conjunction with Latin Hypercube Sampling (LHS), is a widely applied method for sensitivity analysis. Pearson's correlation coefficient, partial correlation coefficient, and standardized regression coefficients assume linear relationships between variables. In contrast, the rank correlation variants test for non-linear but monotonic relationships. This method works by replacing actual values with rank numbers, thereby mitigating the effect of non-linearity in the relationship. Examples of rank correlation methods include SRCC (Spearman Rank Correlation Coefficient), PRCC, and SRRC (Standardized Rank Regression Coefficient) [6, 12, 15].

The first step in calculating PRCC values is to replace each entry in the LHS matrix (Table 4.1) with ascending integers $1, 2, 3, \ldots, n_s$, corresponding to their ranks. If two or more values are equal, they are assigned the average of the ranks they would have occupied. This ranking process is applied to every input parameter and the simulation

output values, as shown in Table 4.2. We denote the ranked input parameters by r_i for $(i = 1, 2, \ldots, n_p)$ and the ranked output variable by τ. Subsequently, the procedure involves fitting $2n_p$ regression models in two rounds: using τ as a target in the first round and then fitting models with each r_i as the output variable. In the second round of linear regression, the rank parameter r_q is adjusted using the other rank parameters [34].

LHS Parameter Ranks				**Output Ranks**
r_1	r_2	$\cdots$	r_{n_p}	τ
$r_1^{(1)}$	$r_2^{(1)}$	$\cdots$	$r_{n_p}^{(1)}$	$\tau^{(1)}$
$r_1^{(2)}$	$r_2^{(2)}$	$\cdots$	$r_{n_p}^{(2)}$	$\tau^{(2)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$r_1^{(n_s)}$	$r_2^{(n_s)}$	$\cdots$	$r_{n_p}^{(n_s)}$	$\tau^{(n_s)}$

Table 4.2: Replacing the sampled values in Table 4.1 with their respective ranks. Specifically, the ith sample of the qth variable $p_q^{(i)}$ is replaced with its rank $r_q^{(i)}$, obtained by sorting each column. The ranked output variable is denoted by τ.

The pairwise PRCC value is computed as the Pearson correlation coefficient between the residuals of two linear regression models. Specifically, for the q-th pairwise PRCC parameter, it is given by:

$$\mathcal{P}_q = \rho_{\mathrm{Res}_{1,q},\mathrm{Res}_{2,q}} = \frac{\mathrm{Cov}(\mathrm{Res}_{1,q}, \mathrm{Res}_{2,q})}{\sqrt{\mathrm{Var}(\mathrm{Res}_{1,q}) \cdot \mathrm{Var}(\mathrm{Res}_{2,q})}}.$$

Here, $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot, \cdot)$ denote the variance and covariance, respectively. Note that $\mathcal{P}_q \in [-1, 1]$. To streamline the presentation, we assume a single output variable. The residuals are computed as:

$$\mathrm{Res}_{1,q} = \tau^{(i)} - \tau^{(i)}_{\mathrm{pred},q}, \quad \mathrm{Res}_{2,q} = r_q^{(i)} - r^{(i)}_{\mathrm{pred},q}, \quad i = 1, \ldots, n_s,$$

where $\tau^{(i)}_{\mathrm{pred},q}$ and $r^{(i)}_{\mathrm{pred},q}$ are the predicted values of $\tau^{(i)}$ and $r_q^{(i)}$, respectively, from linear

regression models that exclude the predictor r_q:

$$\tau^{(i)}_{\text{pred},q} = a^{(0)}_q + \sum_{\substack{j=1 \\ j\neq q}}^{n_p} a^{(j)}_q r^{(i)}_j, \qquad r^{(i)}_{\text{pred},q} = b^{(0)}_q + \sum_{\substack{j=1 \\ j\neq q}}^{n_p} b^{(j)}_q r^{(i)}_j, \quad i = 1, \ldots, n_s.$$

In both regression steps, the predictor r_q is excluded to isolate its partial contribution. The absolute value of the resulting PRCC quantifies the strength of the monotonic association between r_q and the output, controlling for the effects of all other parameters. Larger absolute values indicate stronger relationships.

4.3 Inference on PRCCs

To assess whether a particular Partial Rank Correlation Coefficient (PRCC) $\mathcal{P}_q$ reflects a statistically significant relationship, we formulate a hypothesis test. The null hypothesis $\mathcal{H}_0$ assumes that $\mathcal{P}_q = 0$, indicating no monotonic association between the ranked input parameter r_q and the ranked output τ, after accounting for the influence of other parameters.

Following Marino et al. [34], the test statistic $\mathcal{T}_q$ used to evaluate $\mathcal{H}_0$ is computed as:

$$\mathcal{T}_q = \mathcal{P}_q \sqrt{\frac{n_s - 2 - n_p}{1 - \mathcal{P}_q^2}} \sim t_{n_s-2-n_p},$$

where n_s is the sample size, n_p is the number of parameters, and $\mathcal{T}_q$ approximately follows a Student's t-distribution with $n_s - 2 - n_p$ degrees of freedom (dof). To quantify the strength of evidence against $\mathcal{H}_0$, we compute the two-sided p-value π_q using the cumulative distribution function (CDF) of the t-distribution:

$$\pi_q = 2 \cdot \Big(1 - \text{CDF}\left(|\mathcal{T}_q|, \text{dof}\right)\Big),$$

where $\pi_q \in [0, 1]$. PRCC values are interpreted as statistically significant if $\pi_q < 0.05$, indicating sufficient evidence to reject the null hypothesis. Larger p-values suggest

weaker evidence, supporting the conclusion that $\mathcal{P}_q$ may be zero.

4.4 Standardization of Contact Matrices

Since the basic reproduction number $\mathcal{R}_0$ corresponds to the dominant eigenvalue of the model-specific NGM, and this matrix depends on the structure of the contact matrix, a known value of $\mathcal{R}_0$ can be used to determine the corresponding baseline transmission rate $\beta(\boldsymbol{c} = c)$ for each country $c \in \{1, 2, \ldots, n_c\}$, assuming the contact matrix and model parameters are given.

To enable meaningful cross-country comparisons, we assume a fixed $\mathcal{R}_0$ value and calculate the country-specific transmission rate $\beta(\boldsymbol{c} = c)$ accordingly. These transmission rates are then used to scale the contact matrices for each country. To ensure comparability across countries, we standardize all contact matrices to lie on a common epidemiological scale using the region-specific form of the contact matrix. Specifically, for each country $\boldsymbol{c} = c$ in region $\boldsymbol{r} = r$, the standardized contact matrix is defined as:

$$S(\boldsymbol{r} = r, \boldsymbol{c} = c) = \beta(\boldsymbol{c} = c) \cdot C(\boldsymbol{r} = r, \boldsymbol{c} = c), \tag{4.1}$$

where the contact matrix $C(\boldsymbol{r}, \boldsymbol{c})$ is defined based on the region:

$$C(\boldsymbol{r}, \boldsymbol{c}) = \begin{cases} \mathbf{C}(\boldsymbol{c} = c) & \text{if } \boldsymbol{r} = \text{A (Africa)}, \\ C(\boldsymbol{c} = c) & \text{if } \boldsymbol{r} = \text{E (Europe)}. \end{cases}$$

This normalization step transforms the contact matrices onto a comparable scale, addressing disparities in data availability and quality, such as those noted in [46]. It also serves as an essential preprocessing step for the data-driven components of the framework discussed in Section 4.5.

Following the estimates in [19], we assume a fixed $\mathcal{R}_0 = 3.68$ for African countries,

based on an SEIR model combined with Bayesian inference, where the mean and median values were reported as 3.68 and 3.67, respectively. For European countries, we adopt a fixed value of $\mathcal{R}_0 = 2.2$. Using these region-specific assumptions, we calculate the baseline transmission rates $\beta(\boldsymbol{c} = c)$ and apply them to standardize all contact matrices.

4.5 Dimensionality Reduction

High-dimensional data pose significant challenges to learning methods due to the curse of dimensionality, which can negatively impact performance [5, 29]. To address this, dimensionality reduction techniques or feature selection methods are applied [16, 48]. Principal Component Analysis (PCA) is a widely used method for projecting high-dimensional data onto a lower-dimensional subspace while preserving the most variance in the data. Depending on the structure of the data, different forms of PCA can be applied. In this thesis, we consider both one-dimensional PCA (1D PCA) and two-directional two-dimensional PCA ((2D)^2 PCA) for effective dimensionality reduction.

4.5.1 1D Principal Component Analysis (1D PCA)

Classical Principal Component Analysis (PCA), also referred to as 1D PCA, is a widely used technique for dimensionality reduction and feature extraction on vectorized datasets. It identifies directions (principal components) along which the data exhibits the most variance, facilitating compact representations of complex data while retaining the most informative structures [13, 23, 29, 52, 55]. In the context of this study, 1D PCA is applied to a set of socioeconomic indicators compiled across African countries [24]. These indicators, discussed in detail in Section 4.7, encompass a wide range of

dimensions including health, education, infrastructure, and economic performance.

Let $\mathcal{F} \in \mathbb{R}^{n_c \times n_f}$ denote the matrix of centered socioeconomic data, where n_c is the number of countries and n_f the number of features (indicators). We assume that $\mathcal{F}$ has been centered so that the mean of each column is zero. We seek a projection matrix, $\text{Proj}_{\mathcal{F}}(r) \in \mathbb{R}^{n_f \times r}$, with orthonormal columns, that projects the data onto a lower-dimensional subspace of dimension $r < n_f$. The projected (reduced) representation of the data is then given by

$$\widetilde{\mathcal{F}}_r = \mathcal{F} \cdot \text{Proj}_{\mathcal{F}}(r) \in \mathbb{R}^{n_c \times r}. \tag{4.2}$$

To compute $\text{Proj}_{\mathcal{F}}$, we perform the Singular Value Decomposition (SVD) of the matrix $\mathcal{F}$:

$$\mathcal{F} = U_{\mathcal{F}} \Sigma_{\mathcal{F}} V_{\mathcal{F}}^{\top},$$

where $U_{\mathcal{F}} \in \mathbb{R}^{n_c \times n_f}$, $V_{\mathcal{F}} \in \mathbb{R}^{n_f \times n_f}$ are orthonormal matrices, and $\Sigma_{\mathcal{F}} \in \mathbb{R}^{n_f \times n_f}$ is a diagonal matrix of singular values in descending order. The projection matrix onto the principal subspace of dimension r is defined as:

$$\text{Proj}_{\mathcal{F}}(r) = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix} \in \mathbb{R}^{n_f \times r},$$

where $v_1, \ldots, v_r$ are the first r columns of $V_{\mathcal{F}}$, corresponding to the top r right singular vectors of $\mathcal{F}$.

4.5.2 Two-Directional Two-Dimensional PCA ($(2D)^2$ PCA)

Two-directional two-dimensional principal component analysis, $(2D)^2$ PCA extends the classical principal component analysis to operate directly on matrix-structured data, mitigating the limitations of traditional PCA when dealing with high-dimensional datasets. Unlike conventional PCA, which requires vectorization of data matrices, leading to loss of spatial correlations and an increase in computational complexity,

$(2D)^2$ PCA applies transformations along both column and row directions, preserving the inherent structure of the data. This method is particularly advantageous when working with structured data representations such as social contact matrices, where the relationships between rows and columns capture meaningful interaction patterns across populations [58, 61].

Let $\{S(\boldsymbol{c} = c) \in \mathbb{R}^{n_a \times n_a}\}_{c=1}^{n_c}$ represent a collection of centered, scaled social contact matrices for n_c countries, where n_a is the number of age groups. To reduce dimensionality while preserving structural structure, we introduce two projection matrices with orthonormal columns:

$$\text{Proj}_{\text{row}}(d_{\text{row}}) \in \mathbb{R}^{n_a \times d_{\text{row}}}, \quad \text{Proj}_{\text{col}}(d_{\text{col}}) \in \mathbb{R}^{n_a \times d_{\text{col}}},$$

where $d_{\text{row}}, d_{\text{col}} < n_a$. To organize the data for two-directional projection, we construct two vertically stacked data matrices. In the column-wise direction, the original matrices $S(\boldsymbol{c} = c)$ are stacked directly (without transposition), yielding $\mathcal{F}_{\text{col}}(\boldsymbol{r}) \in \mathbb{R}^{(n_c \cdot n_a) \times n_a}$. In the row-wise direction, each matrix $S(\boldsymbol{c} = c)$ is first transposed before stacking, giving $\mathcal{F}_{\text{row}}(\boldsymbol{r}) \in \mathbb{R}^{(n_c \cdot n_a) \times n_a}$. For example, in the European region: $\mathcal{F}_{\text{col}}(\boldsymbol{r} = E)$, $\mathcal{F}_{\text{row}}(\boldsymbol{r} = E) \in \mathbb{R}^{(39 \cdot 16) \times 16}$, and in the African region: $\mathcal{F}_{\text{col}}(\boldsymbol{r} = A)$, $\mathcal{F}_{\text{row}}(\boldsymbol{r} = A) \in \mathbb{R}^{(32 \cdot 6) \times 6}$.

Dimensionality reduction is then carried out in two stages. First, we apply a column-space projection:

$$\mathcal{S}_{\text{col}}(\boldsymbol{c} = c) = S(\boldsymbol{c} = c) \cdot \text{Proj}_{\text{col}}(d_{\text{col}}),$$

which results in $\mathcal{S}_{\text{col}}(\boldsymbol{c} = c) \in \mathbb{R}^{n_a \times d_{\text{col}}}$. This is followed by a row-space projection:

$$\mathcal{S}_{\text{row}}(\boldsymbol{c} = c) = \text{Proj}_{\text{row}}(d_{\text{row}})^\top \cdot S(\boldsymbol{c} = c),$$

where $\mathcal{S}_{\text{row}}(\boldsymbol{c} = c) \in \mathbb{R}^{d_{\text{row}} \times n_a}$.

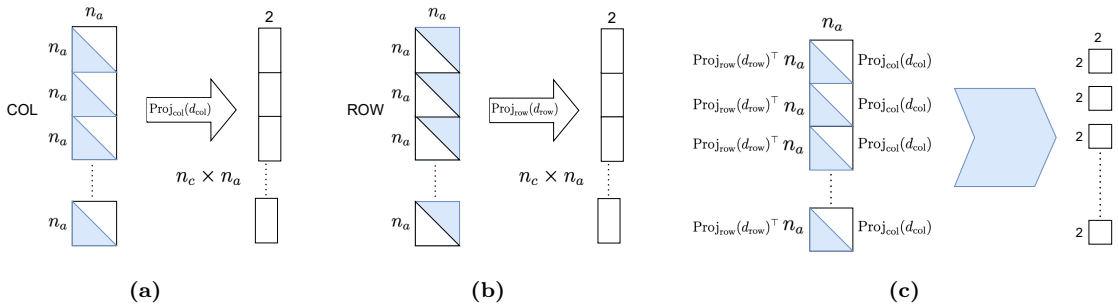

Figure 4.1: Illustration of the dimensionality reduction process using $(2D)^2$ PCA. In (a), the vertically stacked contact matrices $\mathcal{F}_{\text{col}}(r)$ are projected along the column direction using the projection matrix $\text{Proj}_{\text{col}}(d_{\text{col}})$, retaining two principal components. In (b), the transposed contact matrices are stacked to form $\mathcal{F}_{\text{row}}(r)$, which are then projected along the row direction using $\text{Proj}_{\text{row}}(d_{\text{row}})$, also retaining two components. In (c), each scaled contact matrix $S(\boldsymbol{c} = c)$ is jointly projected using both $\text{Proj}_{\text{row}}(d_{\text{row}})$ and $\text{Proj}_{\text{col}}(d_{\text{col}})$, resulting in a reduced matrix $\mathcal{S}(\boldsymbol{c} = c) \in \mathbb{R}^{2\times 2}$.

Combining these gives a two-directional reduction:

$$\mathcal{S}(\boldsymbol{c} = c) = \text{Proj}_{\text{row}}(d_{\text{row}})^\top \cdot S(\boldsymbol{c} = c) \cdot \text{Proj}_{\text{col}}(d_{\text{col}}),$$

resulting in a compact representation $\mathcal{S}(\boldsymbol{c} = c) \in \mathbb{R}^{d_{\text{row}} \times d_{\text{col}}}$. For European countries, this projection uses $\text{Proj}_{\text{col}}(\boldsymbol{r} = E), \text{Proj}_{\text{row}}(\boldsymbol{r} = E) \in \mathbb{R}^{16\times 2}$, producing reduced matrices

$$\mathcal{S}(\boldsymbol{r} = E, \boldsymbol{c} = c) \in \mathbb{R}^{2\times 2}. \tag{4.3}$$

Likewise, for African countries, using $\text{Proj}_{\text{col}}(\boldsymbol{r} = A), \text{Proj}_{\text{row}}(\boldsymbol{r} = A) \in \mathbb{R}^{6\times 2}$, we obtain

$$\mathcal{S}(\boldsymbol{r} = A, \boldsymbol{c} = c) \in \mathbb{R}^{2\times 2}. \tag{4.4}$$

The row projection, column projection, and final reduced matrices are illustrated in Figures 4.1b, 4.1a, and 4.1c, respectively.

4.6 Clustering

Clustering plays a pivotal role in data analysis across diverse fields such as health, medicine, social sciences, and spatial analysis. Its primary objective is to group similar data points together into clusters, such that the intra-cluster similarity is maximized,

while inter-cluster similarity is minimized [49]. The underlying principle is to uncover hidden structures in data without prior knowledge of group labels. Numerous clustering methods exist, each tailored to different data characteristics and application goals. These include distance-based techniques, connectivity-based strategies, density-oriented methods, and probabilistic models [5, 16, 48].

The choice of clustering algorithm depends largely on the data structure, distribution, and size. In this thesis, due to the relatively small sample size (i.e., a limited number of observations), algorithms that perform well with large or dense data sets may not be appropriate. Many density-based or probabilistic clustering methods rely on sufficient data density to define cluster boundaries, which is not feasible in this context. As such, agglomerative hierarchical clustering was selected for its robustness, interpretability, and computational efficiency in low-data regimes. Furthermore, this method is known for its flexibility in accommodating various dissimilarity measures and linkage criteria, making it suitable for a wide range of practical applications [9, 41, 51].

Agglomerative hierarchical clustering operates in a bottom-up fashion. Initially, each data point is treated as an individual cluster. At each iteration, the algorithm identifies the pair of clusters with the smallest dissimilarity and merges them. The dissimilarity matrix is updated to reflect the new configuration, and this process is repeated iteratively. The procedure continues until all data points are merged into a single, all-encompassing cluster.

One of the strengths of hierarchical clustering lies in its customizable linkage criteria, which determine how the distance between clusters is computed. Common options include single linkage (minimum distance), complete linkage (maximum distance), average linkage (mean distance), and Ward's criterion (minimization of variance within clusters). Each method introduces different clustering behaviors and assumptions

about the structure of the data. In this thesis, the complete linkage method was chosen because it emphasizes compact and well-separated clusters, avoiding the chaining effects often seen in single linkage clustering [39]. Formally, for two clusters Γ_i and Γ_j, complete linkage defines their dissimilarity as the greatest distance between any two observations, one from each cluster:

$$D_{\text{complete}}(\Gamma_i, \Gamma_j) = \max\{\mathrm{d}(c_1, c_2) : c_1 \in \Gamma_i, c_2 \in \Gamma_j\},$$

where d is the distance metric applied to the data. We used Euclidean distance due to its simplicity and wide applicability. Nonetheless, other metrics such as L_p norms or domain-specific dissimilarity measures can be integrated based on the context and the nature of the data. The choice of metric is critical, as it can significantly influence the resulting cluster structure and interpretability.

The outcome of hierarchical clustering is typically visualized using a dendrogram. To determine the optimal number of clusters, the dendrogram can be "cut" horizontally at a point that maximizes separation between clusters. A common heuristic is to identify the largest vertical gap in the dendrogram that does not intersect with any merging lines, signaling a natural division in the data [48]. This technique offers a straightforward and intuitive way to select the number of clusters without the need to re-run the algorithm or specify the number of clusters a priori.

4.7 Socioeconomic Indicators

The study [24] analyzed the selected countries' socioeconomic indicators using data from the World Bank database [60]. These indicators encompass various critical dimensions, including social, economic, environmental, institutional, governance, health, education, well-being, and gender inequality. A total of 28 distinct indicators were

considered and categorized across key sectors to ensure consistency and comparability for each country. The socioeconomic indicators analyzed in this study include:

- **Demographic Indicators:** Population Density, Urban Population, Population Growth Rate
- **Labour Market Indicators:** Labour Force Participation, Unemployment Rate, GDP Per Person Employed
- **Education Indicators:** School Enrollment, Literacy Rate
- **Economic Indicators:** GDP Growth Rate, Inflation Rate, Tax Revenue, Exports, Foreign Direct Investment (FDI), Value Added Percent
- **Social Protection Indicators:** Social Protection Coverage, Poverty Headcount Ratio, Food Insecurity Prevalence, Electricity Access
- **Health and Well-being Indicators:** Life Expectancy, Maternal Mortality Ratio, Under-five Mortality Rate, HIV Incidence Rate, Fertility Rate
- **Governance and Institutional Indicators:** Bribery Incidence, Debt Service
- **Technological and Infrastructure Indicators:** Internet Penetration Rate, Personal Remittances, Slums Proportion

These indicators were carefully chosen to provide a comprehensive view of each country's socio-economic landscape, capturing key aspects of social and economic well-being. They are crucial for contextualizing variations in public health outcomes and social contact patterns across nations. In clustering the selected African countries (Section 7.3.2, Chapter 7), socioeconomic data played a central role, grouping nations

with similar socio-economic profiles and contact patterns. In contrast, the clustering of European countries was primarily based on similarities in contact patterns, without incorporating these socioeconomic factors.

Chapter 5

Eigenvector-Based Sensitivity Analysis

5.1 Problem Setting

Understanding how variations in contact patterns affect epidemic outcomes is crucial for identifying the age groups that drive uncertainty in disease transmission models, thereby improving model accuracy and informing targeted interventions. Sensitivity analysis provides a structured approach to quantify how changes in parameters, such as age-specific contact rates, affect key metrics like the basic reproduction number, $\mathcal{R}_0$, and clinical outcomes such as hospitalizations, ICU admissions, and mortality. Traditional sensitivity analysis methods such as Monte Carlo simulations, Latin Hypercube Sampling, and Partial Rank Correlation Coefficients are useful for exploring high-dimensional parameter spaces [15].

In age-structured epidemic models, NGM provides a mathematical representation of transmission dynamics. $\mathcal{R}_0$ is defined as the dominant eigenvalue (i.e., the spectral radius) of the NGM. This spectral formulation enables analytical sensitivity analysis by linking perturbations in the contact matrix to changes in $\mathcal{R}_0$. This chapter

introduces an eigenvector-based sensitivity analysis method that uses the left and right eigenvectors of the Next Generation Matrix to quantify changes in age-specific contact rates affecting $\mathcal{R}_0$. This approach identifies which interactions between age groups contribute most to disease transmission. Compared to simulation-based techniques, it is both computationally efficient and interpretable, especially when applied to structured compartmental models. Following the framework developed by Diekmann et al. [11] and extended by others [4, 20, 22, 25], we formulate sensitivity measures directly from the spectral properties of the NGM.

5.2 Methods

5.2.1 Sensitivity Measures

To streamline the presentation in this section, we omit the model-specific indexing m from key quantities. That is, we refer to the NGM simply as K, its associated left and right eigenvectors as $\mathbf{v}$ and $\mathbf{w}$, and the basic reproduction number as $\mathcal{R}_0$. All references to these quantities are implicitly tied to the specific model under analysis.

To analyze how variations in contact rates among age groups affect $\mathcal{R}_0$, we compute its gradient with respect to the elements of the contact matrix. The population is divided into n_a age groups, and their contact rates are represented in an $n_a \times n_a$ matrix, as detailed in Chapter 2.3. Since the matrix captures symmetric interactions between age groups, the number of independent contact pairs is given by

$$n_p = \frac{n_a(n_a + 1)}{2}.$$

Using the age-stratified contact data from Prem et al. [46], which consists of 16 age groups, results in 136 unique parameters as shown in Table 5.1.

Age Group	**1**	**2**	**3**	**4**	**5**	$\cdots$	**15**	**16**
1	p_1	p_2	p_3	p_4	p_5	$\cdots$	p_{15}	p_{16}
2	0	p_{17}	p_{18}	p_{19}	p_{20}	$\cdots$	p_{30}	p_{31}
3	0	0	p_{32}	p_{33}	p_{34}	$\cdots$	p_{44}	p_{45}
4	0	0	0	p_{46}	p_{47}	$\cdots$	p_{57}	p_{58}
5	0	0	0	0	p_{59}	$\cdots$	p_{69}	p_{70}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
15	0	0	0	0	0	$\cdots$	p_{134}	p_{135}
16	0	0	0	0	0	$\cdots$	0	p_{136}

Table 5.1: Each value p_q denotes a particular element in the upper triangular part of the matrix. The rows and columns are numbered from 1 to 16, corresponding to the age groups, with all entries below the diagonal being zero.

For a given matrix such as K, $\mathcal{R}_0$ is defined as its dominant eigenvalue. It satisfies the following generalized eigenvalue problem:

$$K\mathbf{w} = \mathcal{R}_0\mathbf{w}, \quad \mathbf{v}^{\mathrm{T}}K = \mathcal{R}_0\mathbf{v}^{\mathrm{T}},$$

where $\mathbf{w}$ and $\mathbf{v}$ are the right and left eigenvectors corresponding to $\mathcal{R}_0$, respectively.

To compute the sensitivity of $\mathcal{R}_0$ to small perturbations in K, we start by introducing a perturbation to the matrix:

$$(K + \Delta K)(\mathbf{w} + \Delta\mathbf{w}) = (\mathcal{R}_0 + \Delta\mathcal{R}_0)(\mathbf{w} + \Delta\mathbf{w}).$$

Expanding this expression and neglecting higher-order terms gives:

$$K\Delta\mathbf{w} + \Delta K\mathbf{w} = \mathcal{R}_0\Delta\mathbf{w} + \Delta\mathcal{R}_0\mathbf{w}.$$

By rearranging terms to isolate $\Delta\mathcal{R}_0$, we have

$$\Delta K\mathbf{w} = \Delta\mathcal{R}_0\mathbf{w}.$$

Multiply both sides from the left by $\mathbf{v}^{\mathrm{T}}$, to project the perturbation onto $\mathcal{R}_0$, we

gather

$$\mathbf{v}^{\mathrm{T}}\Delta K\mathbf{w} = \Delta\mathcal{R}_0 \cdot \mathbf{v}^{\mathrm{T}}\mathbf{w}.$$

Solving this equation for $\Delta\mathcal{R}_0$ gives:

$$\Delta\mathcal{R}_0 = \frac{\mathbf{v}^{\mathrm{T}}\Delta K\mathbf{w}}{\mathbf{v}^{\mathrm{T}}\mathbf{w}}.$$

If the perturbation affects only one element of K denoted by k_{ij}, then:

$$\Delta\mathcal{R}_0 = \frac{v_i w_j}{\mathbf{v}^{\mathrm{T}}\mathbf{w}}\Delta k_{ij}.$$

Taking the limit as $\Delta k_{ij} \to 0$, the sensitivity of $\mathcal{R}_0$ with respect to k_{ij} is:

$$\frac{\partial\mathcal{R}_0}{\partial k_{ij}} = \frac{v_i w_j}{\mathbf{v}^{\mathrm{T}}\mathbf{w}}. \tag{5.1}$$

Here:

- k_{ij} is the element of the K,
- $\frac{\partial\mathcal{R}_0}{\partial k_{ij}}$ represents the sensitivity of $\mathcal{R}_0$ to the k_{ij},
- v_i and w_j are the i-th and j-th components of the left and right eigenvectors, respectively.

The entries of K are often expressed as functions of the contact matrix C introduced in Eq. (2.2), whose elements are denoted as $c_{i,j}$. In this study, we use the upper triangular representation of the contact matrix as defined in Table 5.1. Specifically, we define $\mathbf{z} \in \mathbb{R}^{n_p}$ to contain the n_p independent upper triangular elements of C. To determine the sensitivity of $\mathcal{R}_0$ with respect to these parameters, we employ the chain rule:

$$\frac{\partial\mathcal{R}_0}{\partial\mathbf{z}_q} = \sum_{i=1}^{n_a}\sum_{j=1}^{n_a}\frac{\partial\mathcal{R}_0}{\partial k_{ij}} \cdot \frac{\partial k_{ij}}{\partial\mathbf{z}_q}, \tag{5.2}$$

where,

- $\mathbf{z}_q$ is an individual element of the vector $\mathbf{z}$, corresponding to and not equal to an independent contact rate from C.

- $\frac{\partial k_{ij}}{\partial \mathbf{z}_q} \in \mathbb{R}$ is a scalar that describes how the entry k_{ij} of K changes with respect to the contact matrix elements.

The resulting gradient $\frac{\partial \mathcal{R}_0}{\partial \mathbf{z}} \in \mathbb{R}^{n_p}$ quantifies the sensitivity of the basic reproduction number $\mathcal{R}_0$ to perturbations in the n_p independent elements of $\mathbf{z}$, which correspond to contact rates in C. Each entry in this gradient reflects how a small change in the contact rate between a specific pair of age groups affects disease transmission, yielding pairwise sensitivity values that quantify the influence of contact patterns on $\mathcal{R}_0$.

The Röst et al. model, introduced in Eq. (A.2) and illustrated in Fig. 3.2, includes detailed clinical compartments that allow for outcome-specific analyses such as *hospitalization* ($I^{(h)}$), *ICU* admission ($I^{(c)}$), and *death* (D). However, it is important to emphasize that there is no distinct NGM or $\mathcal{R}_0$ associated with these clinical outcomes.

To explore how age-specific contact patterns influence mortality, we propose a simple modification to the original K, which incorporates the probability that an individual progresses to death after infection. Specifically, for each age group i, we define the age-specific probability of death following a presymptomatic infection as:

$$\delta_i^{(\mathrm{d})} = \Pr\left(I_i^{(p)} \to D_i\right) = (1 - p_i)\, h_i\, \xi_i\, \mu_i, \tag{5.3}$$

where each parameter corresponds to the probability of progression along the symptomatic and clinical pathway leading to death. This conditional probability is derived from the model's compartmental transitions. Using these death probabilities, we construct a new matrix $M = [m_{ij}]$ by scaling K:

$$m_{ij} = k_{ij} \cdot \delta_j^{(\mathrm{d})}. \tag{5.4}$$

The matrix M does not define a new NGM, nor does it yield a mortality-specific basic reproduction number. Rather, it is introduced as a heuristic to study how contact patterns influence mortality outcomes through transmission. This approach is not grounded in a formal derivation, but instead provides an intuitive method to adapt the sensitivity framework in Eq. (5.1) for outcome-related quantities. Although the spectral radius of M could be computed, our primary interest lies in the corresponding left and right eigenvectors of M, which we substitute into Eq. (5.1) to obtain a sensitivity metric related to mortality rather than to transmission.

5.2.2 Age-Group-Level Sensitivity Measures

To comprehensively evaluate how changes in contact rates across different age groups affect the basic reproduction number $\mathcal{R}_0$, we extend the sensitivity analysis by introducing age-group sensitivity measures, following the approach in [4]. These provide cumulative metrics that summarize the proportional change in $\mathcal{R}_0$ resulting from proportional changes in contact rates associated with each age group. Building on the gradient formulation in Eq. (5.2), we define a sensitivity score $\mathbf{S}_j$ for each age group j, as shown in Eq. (5.5). This score quantifies the total influence of contact perturbations involving age group j and reflects its overall contribution to transmission dynamics:

$$\mathbf{S}_j = \left\| \frac{\partial \mathcal{R}_0}{\partial \mathbf{z}} \right\|_1 = \sum_{i=1}^{n_a} \left| \frac{\partial \mathcal{R}_0}{\partial \mathbf{z}} \right| = \sum_{i=1}^{n_a} \left| \frac{\partial \mathcal{R}_0}{\partial k_{ij}} \cdot \frac{\partial k_{ij}}{\partial \mathbf{z}} \right|. \tag{5.5}$$

5.2.3 Framework for Sensitivity Analysis

We implemented the eigenvector-based sensitivity analysis using a modular Python framework tailored to age-structured epidemic models. This framework integrates contact data, model parameters, and epidemiological assumptions to compute the basic

reproduction number $\mathcal{R}_0$ and its sensitivities to changes in social mixing patterns.

The primary input consists of scaled contact matrices that capture age-specific social interactions across various environments (e.g., home, school, workplace, and other settings). These matrices are symmetrized and reduced to their upper triangular form, producing a vector of n_p independent contact parameters, denoted $\mathbf{z} \in \mathbb{R}^{n_p}$ as illustrated in Table 5.1.

As illustrated in Figure 5.1, the framework processes these inputs through a sequence of computational modules. The *Next Generation Matrix Calculator* constructs the transmission matrix F, the transition matrix V, and the auxiliary matrix E, which are then used to derive the next generation matrix K. The basic reproduction number $\mathcal{R}_0$ is computed as the dominant eigenvalue of K. The *Eigenvector Calculator* determines the left and right eigenvectors corresponding to this dominant eigenvalue. Next, the *Gradient Calculator* applies the formulation in Eq. (5.2) to evaluate the sensitivity of $\mathcal{R}_0$ to each element of $\mathbf{z}$. These values are subsequently summed using Eq. (5.5) to produce group-level sensitivity scores $\mathbf{S}_j$, which summarize the influence of each age group on transmission dynamics.

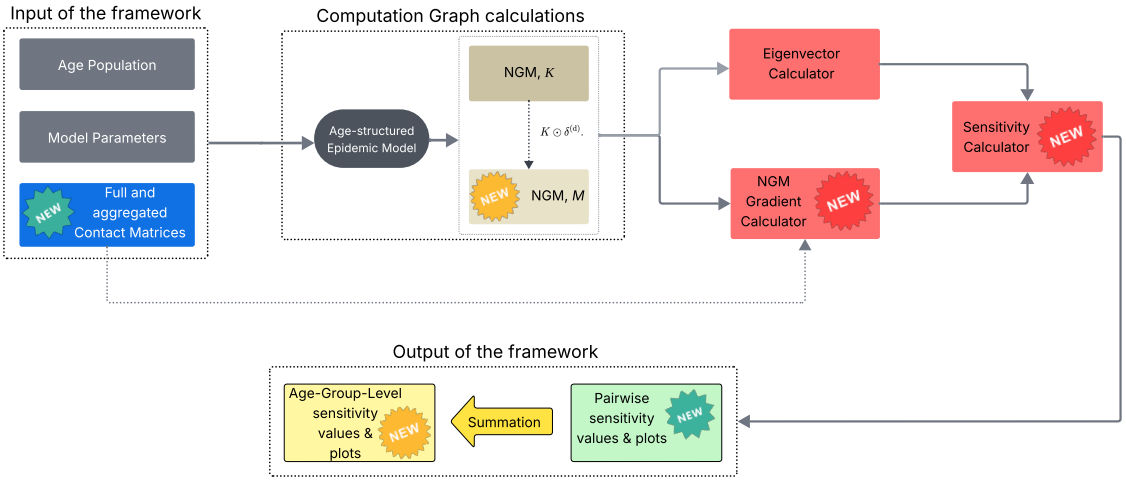

Figure 5.1: Schematic of the sensitivity analysis framework. The pipeline begins with scaled contact matrices and proceeds through eigenvalue and gradient computation to cumulative sensitivity evaluation and visualization. All components marked *NEW* were developed as part of this thesis. The structure is modular and adaptable to any age-structured compartmental model.

All gradient computations are performed through a computation graph constructed with `PyTorch's autograd` engine, enabling efficient evaluation via backpropagation. The framework is fully automated, with `PyTorch` handling the numerical backend and `matplotlib` used for visualization.

5.3 Demonstrations

This section presents the results of applying the methodology from Section 5 to analyze transmission dynamics, with a focus on age-specific contact patterns. The contact matrices serve as key inputs to the models and are the primary focus of the sensitivity analysis. As a case study, we apply the method to the Pitman et al. model from Section 3.3 using age-structured contact data from the UK. To validate the framework in a different epidemiological context, we repeat the analysis using Röst et al. model described in Section 3.4, using the Hungarian contact data.

To incorporate age-dependent susceptibility, specific values (σ_i) were assigned to different age groups. In the UK model, susceptibility was assumed to be uniform across all age groups. In contrast, the Hungarian model incorporated age-dependent susceptibility by assigning specific values (σ_i) to different age groups: individuals aged 20 and older were given a susceptibility of 1.0, while those under 20 had a reduced value of 0.5. Model parameters and assumptions were adopted directly from the original studies [44, 50].

5.3.1 SEIR model for influenza epidemic by Pitman et al.

We apply the methodology from Section 5 to analyze the sensitivity of contact patterns in the UK. The SEIR model introduced in Section 3.3, is used to assess the impact of pediatric vaccination programs in the UK and Wales. This model incorporates detailed age-specific contact matrices (Fig. 2.2, left panel), offering crucial insights into how transmission dynamics vary across age groups. The population is assumed to have uniform susceptibility, and the baseline reproduction number, $\overline{\mathcal{R}}_0(m = \mathrm{P})$, is set to 1.8, consistent with the original authors' assumptions.

The social contact matrix reveals frequent interactions among individuals aged 5–19, reflecting the high level of social mixing within school-aged groups (see Fig. 2.2, left panel). The left panel of Fig. 5.2 presents the sensitivity values of $\mathcal{R}_0$ with respect to the contacts, as defined in Eq. (5.2). The heatmap shows that the highest sensitivity values are concentrated in the age groups 5–19 and 35–49, suggesting that even small changes in contact rates within this demographic have a significant impact on transmission dynamics. This aligns with the frequent and intense social interactions observed among school-aged children and adolescents [1].

Additionally, the sensitivity analysis (Fig. 5.2, right panel) highlights distinct

transmission dynamics across age groups. The results show that younger individuals, especially those aged 5–19, as well as middle-aged adults in the 35–49 age range, contribute disproportionately to secondary infections.

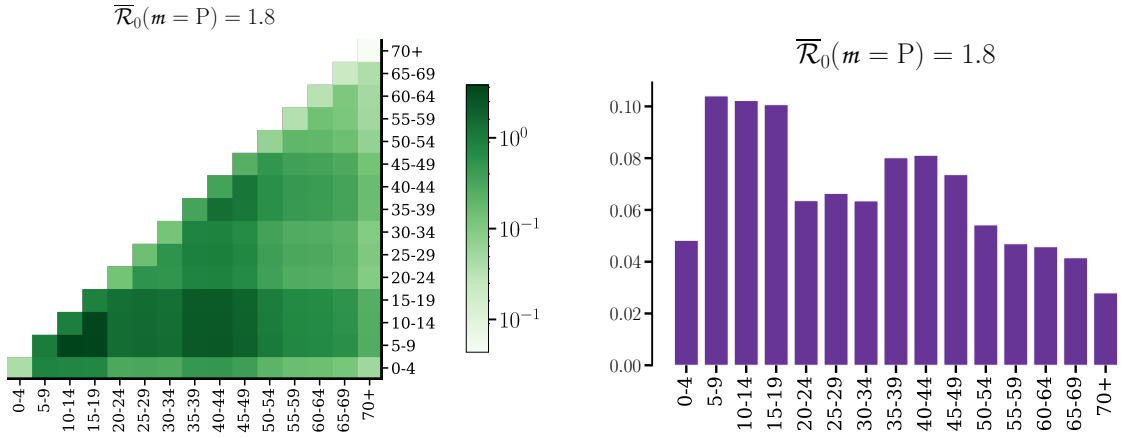

Figure 5.2: Sensitivity analysis of $\mathcal{R}_0(m = \mathbf{P})$ under uniform susceptibility assumptions. (Left) Pairwise sensitivity values based on Eq. (5.2), showing how small changes in contact rates between specific age group pairs influence $\mathcal{R}_0$. (Right) Age-group-level sensitivity values, $\mathbf{S}_j(m = \mathrm{P})$, from Eq. (5.5), summarizing the overall impact of each age group on transmission dynamics.

5.3.2 Covid-19 model by Röst et al.

We apply the eigenvector-based sensitivity analysis to the age-structured COVID-19 model introduced by Röst et al. (Eq. A.2) using age-specific contact data from Hungary. The analysis is performed using the adjusted full contact matrix shown on the right panel of Fig. 2.2. We examine how small perturbations in contact rates influence two critical epidemiological outcomes: the basic reproduction number $\mathcal{R}_0$ and cumulative mortality. In addition to the full-resolution analysis, we also investigate the effect of contact matrix aggregation on the sensitivity results. Specifically, we compare the original 16-age-group contact structure with an aggregated, 7-group representation to assess how age group simplification influences the interpretation of transmission and mortality dynamics.

Figure 5.3 summarizes the sensitivity results under the assumption that individuals under 20 years of age have reduced susceptibility ($\sigma_i = 0.5$). The top left panel shows the pairwise sensitivity matrix of $\mathcal{R}_0$, revealing that contact interactions among individuals aged 30–49 have the largest impact on transmission. In contrast, the top right panel displays the pairwise sensitivity matrix for cumulative mortality. Here, the greatest sensitivities are associated with contact patterns involving older adults, particularly those aged 60 and above. This shift reflects the increased severity and fatality rates in older populations, despite their generally lower contact rates.

The bottom panels of Fig. 5.3 present the corresponding group-level sensitivity values, as defined in Eq. (5.5). For transmission (bottom left), individuals aged 20–54 dominate the sensitivities, consistent with their central role in driving $\mathcal{R}_0$. Younger individuals ($<$20) and older adults ($>$54) contribute less to transmission dynamics, although the elderly show slightly higher sensitivities than the young due to greater clinical vulnerability. In terms of cumulative mortality (bottom right), the sensitivity landscape shifts markedly toward the elderly. Individuals aged 70 and above exhibit the highest mortality-related sensitivities, underscoring the critical importance of accurately characterizing clinical parameters for these groups. Even with relatively fewer contacts, their elevated risk magnifies their influence on mortality outcomes.

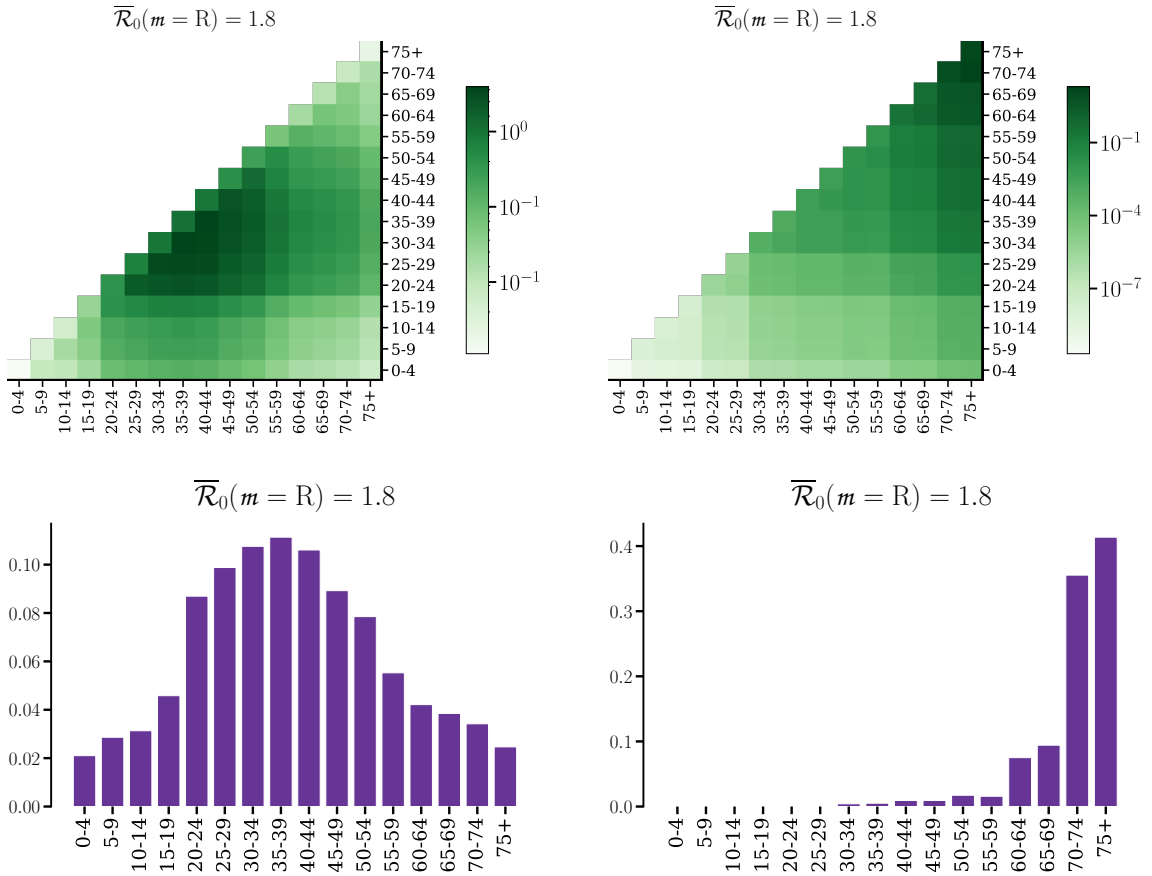

Figure 5.3: Eigenvector-based sensitivity analysis for Hungary under reduced susceptibility scenario ($\sigma_i = 0.5$). Top left: Pairwise sensitivity of the basic reproduction number $\mathcal{R}_0(m = \mathrm{R})$ with respect to contact rates between age groups, based on Eq. (5.2). Top right: Pairwise sensitivity of cumulative mortality with respect to age-specific contact rates. Bottom left: Age-group-level sensitivity values $\mathbf{S}_j(m = \mathrm{R})$, summarizing the overall influence of each age group on transmission. Bottom right: Age-group-level sensitivity values for mortality, identifying which age groups have the strongest influence on death outcomes.

Under the uniform susceptibility assumption ($\sigma_i = 1.0$), the distribution of influence on transmission shifts toward younger individuals. Adolescents, especially those aged 15–19, become the primary drivers of transmission, as indicated by their elevated pairwise and group-level sensitivity values (Fig. 5.4, top left and bottom left panels). This reflects both their high contact frequencies and the absence of susceptibility scaling across age groups. For mortality, the sensitivity patterns remain concentrated among

older adults. Contacts involving individuals aged 70 and above continue to dominate both the pairwise and cumulative sensitivity metrics (Fig. 5.4, top and bottom right panels), driven by their disproportionately high probabilities of death. These results emphasize the divergent roles of age groups in transmission versus mortality outcomes and remain consistent across scenarios with different $\overline{\mathcal{R}}_0(m = \mathrm{R})$.

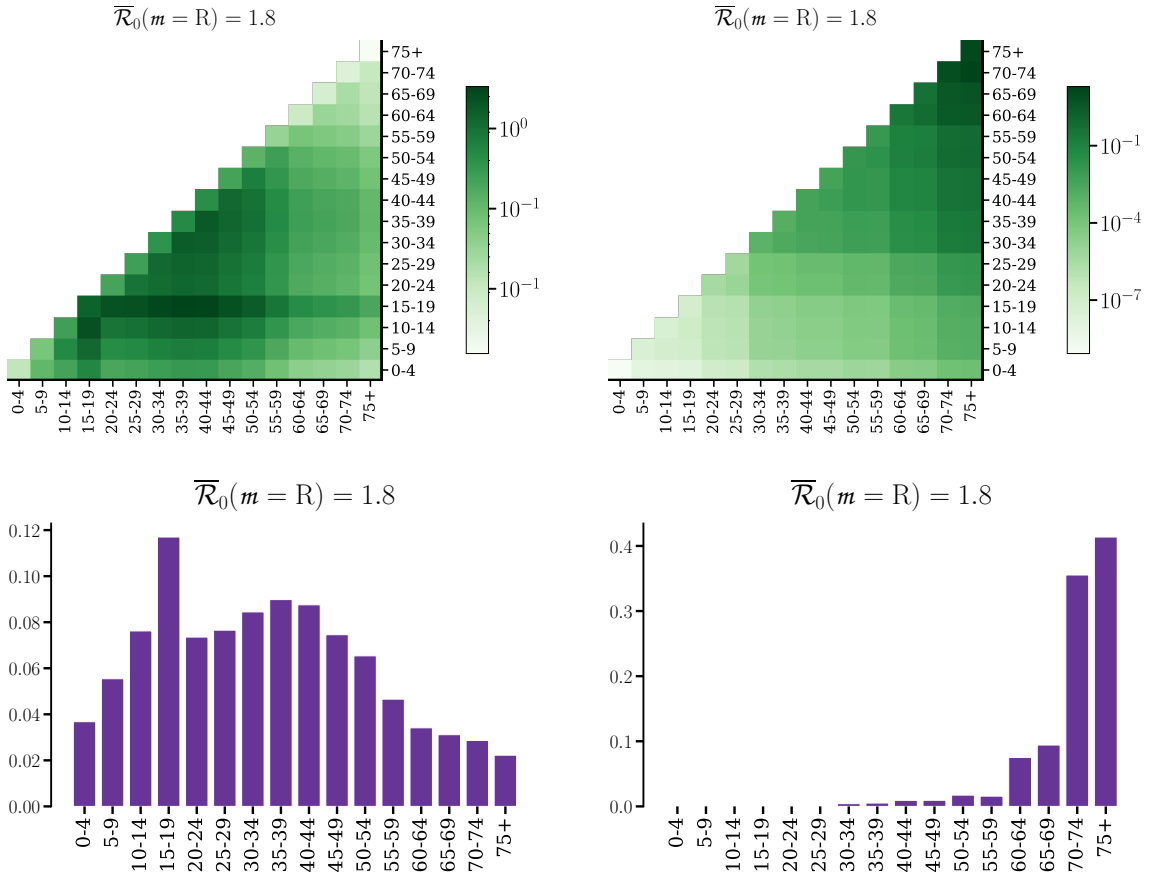

Figure 5.4: Eigenvector-based sensitivity analysis for Hungary under uniform susceptibility ($\sigma_i = 1.0$). Top row: Pairwise sensitivities of $\mathcal{R}_0(m = \mathrm{R})$ (left) and cumulative mortality (right) with respect to age-specific contact rates. Bottom row: Corresponding group-level sensitivity values $\mathbf{S}_j(m = \mathrm{R})$, summarizing the influence of each age group on transmission (left) and mortality (right).

To examine how age-group aggregation affects model insights, we reapply our framework to the model using an aggregated contact matrix with seven broader age

groups from the adjusted full contact matrix shown in the right panel of Fig. 2.2. The resulting pairwise and group-level sensitivities are shown in Fig. 5.5.

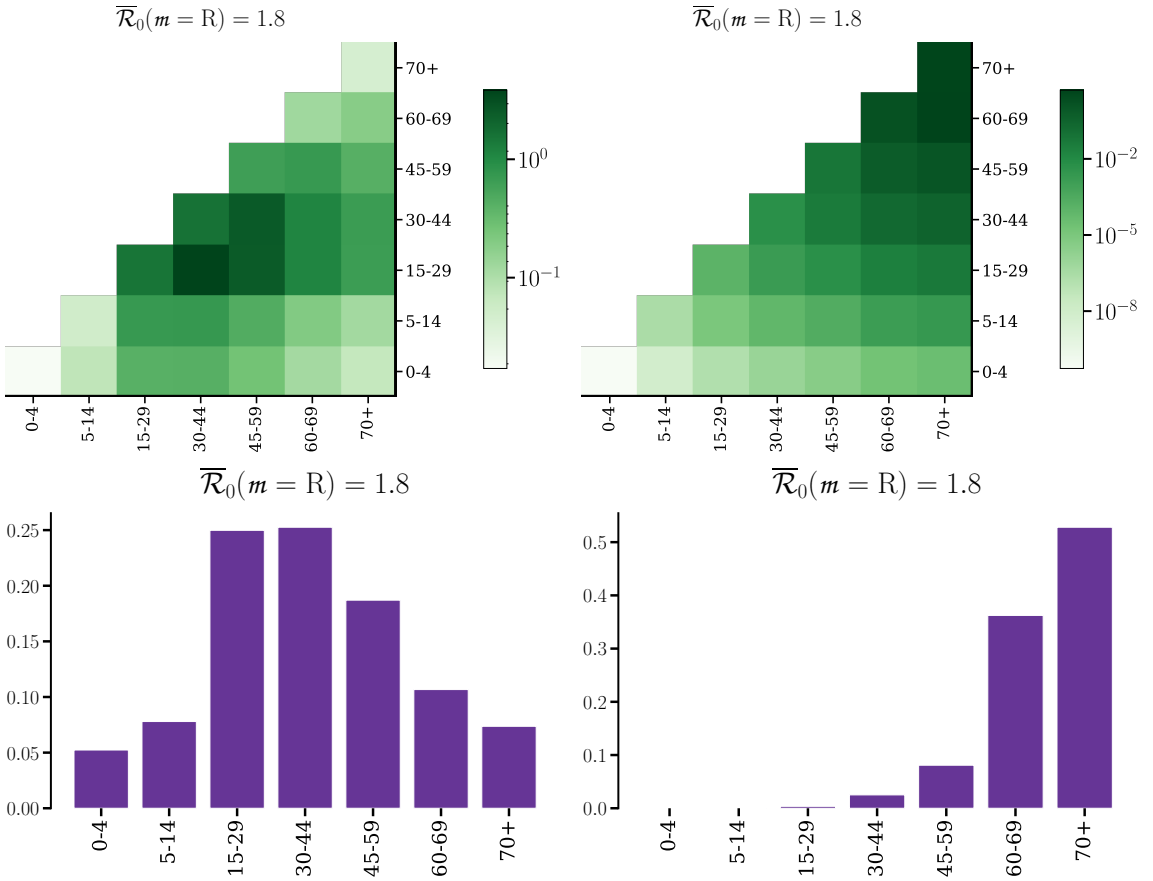

Figure 5.5: Sensitivity analysis with aggregated contact matrix. Top row: Pairwise sensitivities of the basic reproduction number $\mathcal{R}_0(m = \mathrm{R})$ (left) and *mortality* outcomes (right), assuming a 50% reduction in susceptibility for the 0–14 age group. Bottom row: Corresponding group-level sensitivities for transmission (left) and mortality (right). While aggregation reduces age resolution, it preserves core sensitivity patterns. Transmission remains driven by contacts involving ages 15–44, whereas mortality sensitivity concentrates in the 60+ age group.

Despite the reduced resolution, the qualitative patterns remain consistent. Transmission sensitivity remains highest for contacts within the 15–44 age range. However, aggregation blends finer age distinctions, shifting the dominant cumulative sensitivity toward a broader 15–44 group. For mortality, the highest sensitivities still concentrate

in older adults. However, compared to the original matrix (Fig. 5.3), the peak shifts downward from 70+ to the 60+ group, highlighting how aggregation can subtly affect the interpretation of age-targeted outcomes.

5.4 Conclusion

This study introduces a flexible, analytically grounded framework for sensitivity analysis in age-structured epidemic models. By leveraging the spectral properties of the NGM, the method quantifies how changes in age-specific contact rates influence both transmission and clinical outcomes, such as mortality. Unlike simulation-heavy approaches, our method provides direct, interpretable gradients that offer insights into which age group interactions drive epidemic dynamics.

A central strength of the framework lies in its ability to integrate detailed contact data with outcome-specific parameters. The inclusion of group-level sensitivity metrics can help prioritize where empirical data collection and parameter refinement are most needed. Rather than proposing interventions directly, the method highlights the contact patterns and age groups that most affect model outputs, indirectly informing where improved surveillance, calibration, or contextual information could enhance predictive accuracy.

The proposed sensitivity analysis framework offers several advantages for modeling disease transmission. Its deterministic nature ensures consistent and reproducible results, improving the reliability of predictions. Beyond transmission ($\mathcal{R}_0$), the method accommodates downstream health outcomes such as hospitalization and mortality. Despite its advantages, the framework has limitations. It provides point-based sensitivity estimates without confidence intervals, which restricts direct quantification of uncertainty. Moreover, the analysis assumes local linearity around the baseline

scenario and does not capture nonlinear threshold effects that may arise in more complex models. The current implementation also relies on static contact matrices, which may not reflect real-world variations in social behavior due to time-varying interventions or behavioral adaptation. Nevertheless, the framework is extensible. If contact matrices are available over multiple periods, the analysis can be repeated for each period independently, allowing for dynamic sensitivity profiling for an epidemic.

Chapter 6

Statistical Age Group Sensitivity Analysis of Epidemic Models

6.1 Problem Setting

To address the previously mentioned limitations, this chapter introduces the statistical Age Group Sensitivity Analysis (AGSA) method [56], a novel framework for quantifying the sensitivity of epidemic outcomes to variations in age-specific contact structures. The approach integrates age-structured epidemic models with Latin Hypercube Sampling and Partial Rank Correlation Coefficient (PRCC). These techniques, introduced in Sections 4.1 and 4.2, are well-suited for exploring high-dimensional parameter spaces and capturing nonlinear and monotonic relationships between input variables and model outputs. Unlike the eigenvector method, this approach allows us to incorporate parameter variability and uncertainty, enabling a more robust and comprehensive sensitivity analysis. The methodology involves generating a large number of plausible parameter combinations via Latin Hypercube Sampling and running

corresponding simulations of the epidemic model. The output metrics such as $\mathcal{R}_0$, peak incidence, or total infections, are then statistically analyzed to assess which input parameters (e.g., contact rates between age groups) exert the most influence on model outcomes.

In line with the symmetry of social contact matrices, we restrict variations to the upper triangular elements of the matrix, as structured in Table 5.1. To ensure the sampling process introduced in Chapter 4 is epidemiologically relevant, we make the following assumptions:

A1 We assume that no intervention can alter home contacts, as a complete lockdown restricts individuals to their households. Thus, we only sample changes for $C^{\Delta} = C^{S} + C^{W} + C^{O}$ (see Eq. (2.3)).

A2 The target parameters for our PRCC analysis include the reproduction number $\mathcal{R}_0$, the infected peak, the hospitalized peak, the peak of the intensive care unit, and the final death size. Therefore, to assess the effect of variations on outbreaks, we must ensure that our sampling keeps $\mathcal{R}_0 \geq 1$. Thus, we only consider sampled contact matrices for which the basic reproduction number is greater than 1.

A3 We aim to examine the overall sensitivity of all contacts characteristic of a given age group, rather than limiting the analysis to specific contacts stored in workplace, school, or other matrices. Thus, it is more practical to sample the reduction ratios of the elements of C^{Δ} rather than the elements of the contact matrix directly.

To assess **A2**, we first obtain the basic reproduction number $\mathcal{R}_0$ using the Next Generation Matrix. Then, we introduce a parameter κ and sample proportions for reduction from $[0, \kappa]$. Our objective is to find the value of κ so that $\mathcal{R}_0 = 1$ with the

contact matrix:

$$C^{\mathrm{H}} + (1 - \kappa) \cdot C^{\Delta}.$$

The appropriate value of κ can be identified using basic interval logic or more advanced techniques.

To assess **A3**, we randomly sample a vector of n_p unique elements independently within the interval $[0, \kappa]$ using a uniform distribution as illustrated in Table 4.1. We use a uniform distribution because we assume no prior knowledge about the distribution of individual contact pairs. If such prior information is available, the appropriate distribution can be selected accordingly. These elements represent proportional changes and are arranged in a square symmetric matrix, $M_{\text{ratio}} \in [0, \kappa]^{n_a \times n_a}$, and it is true that $M_{\text{ratio}}^{(i,j)} = M_{\text{ratio}}^{(j,i)}$.

The mapping between the sampled vector n_p and M_{ratio} can be seen in Table 6.1. We adjust the matrices by multiplying elementwise the total matrix C^{Δ} by $1 - M_{\text{ratio}}$, resulting in:

$$\widehat{C}^{\Delta} = (\mathbb{1} - M_{\text{ratio}}) \odot C^{\Delta},$$

where $\mathbb{1} \in \mathbb{R}^{n_a \times n_a}$ represents the matrix $[1]_{i,j=1}^{n_a}$. The operator $\odot$ indicates element-wise multiplication. The modified total matrix $\widehat{C}$ then has the form:

$$\widehat{C} = C^{\mathrm{H}} + \widehat{C}^{\Delta}. \tag{6.1}$$

In the extreme case, where the maximum change is applied, $M_{\text{ratio}} = 1$, resulting in $\widehat{C} = C^{H}$. Using Eq. (6.1), reducing the contacts below the home contacts is impossible.

To evaluate the sensitivity of the contact parameter, we employ the PRCC methodology, as introduced in Section 4.2. We define the target variables, $\mathcal{R}_0$, *infected peak*, *intensive care unit peak*, and the *final death size* as outputs and compute their

corresponding pairwise PRCC values ($\mathcal{P}_q$).

To assess whether $\mathcal{P}_q$ significantly differs from zero and to compare the differences between two PRCC values, we employ the method outlined in Section 4.3. This enables us to compute the p-values (π_q) associated with ($\mathcal{P}_q$). The n_p sensitivity measures for $\mathcal{P}_q$ and their corresponding π_q are then arranged in a square symmetric matrix to be used in Section 6.2. A upper triangular of a matrix like this can be seen in Fig. 6.2. For each $\mathcal{P}_q$ element, we determine the corresponding (i, j) age group pair to which the sensitivity value belongs as shown in Table 6.1. This value is then assigned to the $\mathcal{P}_{i,j}$ and $\mathcal{P}_{j,i}$ elements of the $\boldsymbol{\mathcal{P}}$ matrix. We repeat this process for the π_q values as well, assigning each to the corresponding $\pi_{i,j}$ and $\pi_{j,i}$ elements of the $\boldsymbol{\pi}$ matrix. Formally, we define the matrices $\mathcal{P}$ and $\boldsymbol{\pi}$ as follows:

$$\boldsymbol{\mathcal{P}} = [\mathcal{P}_{i,j}]_{i,j=1}^{n_a} \quad \text{with} \quad \mathcal{P}_{i,j} = \mathcal{P}_{j,i}, \quad (1 \leq i \leq j \leq n_a) \tag{6.2}$$

as the PRCC matrix and

$$\boldsymbol{\pi} = [\pi_{i,j}]_{i,j=1}^{n_a} \quad \text{with} \quad \pi_{i,j} = \pi_{j,i}, \quad (1 \leq i \leq j \leq n_a) \tag{6.3}$$

as the p-value matrix.

6.2 Age Group–Level PRCC Approach

In real-world scenarios, it is often impractical to adjust only a single pair of contacts between age groups. Instead, it is more effective to focus on one or more age groups and all their associated interactions with other age groups, or to target specific types of interactions (e.g., the effects of school closures on social contacts within schools). For example, independently modifying the contact rates between individuals aged 35-40 and 50-55 is rarely feasible. This limitation necessitates methods that analyze changes affecting larger groups of interactions simultaneously.

Age Group	**1**	**2**	**3**	**4**	**5**	$\cdots$	**15**	**16**
1	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\cdots$	$\mathcal{P}_{15}$	$\mathcal{P}_{16}$
2	$\mathcal{P}_2$	$\mathcal{P}_{17}$	$\mathcal{P}_{18}$	$\mathcal{P}_{19}$	$\mathcal{P}_{20}$	$\cdots$	$\mathcal{P}_{30}$	$\mathcal{P}_{31}$
3	$\mathcal{P}_3$	$\mathcal{P}_{18}$	$\mathcal{P}_{32}$	$\mathcal{P}_{33}$	$\mathcal{P}_{34}$	$\cdots$	$\mathcal{P}_{44}$	$\mathcal{P}_{45}$
4	$\mathcal{P}_4$	$\mathcal{P}_{19}$	$\mathcal{P}_{33}$	$\mathcal{P}_{46}$	$\mathcal{P}_{47}$	$\cdots$	$\mathcal{P}_{56}$	$\mathcal{P}_{57}$
5	$\mathcal{P}_5$	$\mathcal{P}_{20}$	$\mathcal{P}_{34}$	$\mathcal{P}_{47}$	$\mathcal{P}_{59}$	$\cdots$	$\mathcal{P}_{68}$	$\mathcal{P}_{69}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
15	$\mathcal{P}_{15}$	$\mathcal{P}_{30}$	$\mathcal{P}_{44}$	$\mathcal{P}_{56}$	$\mathcal{P}_{68}$	$\cdots$	$\mathcal{P}_{134}$	$\mathcal{P}_{135}$
16	$\mathcal{P}_{16}$	$\mathcal{P}_{31}$	$\mathcal{P}_{45}$	$\mathcal{P}_{57}$	$\mathcal{P}_{69}$	$\cdots$	$\mathcal{P}_{135}$	$\mathcal{P}_{136}$

Table 6.1: Symmetric Matrix Construction. This table presents the arrangement of elements $\mathcal{P}_q$ into a symmetric matrix by combining the entries from the age-stratified contact data from Prem et al. [46], which consists of 16 age groups. The resulting 16×16 matrix, known as the PRCC matrix $\boldsymbol{\mathcal{P}}$, contains pairwise PRCC values $\mathcal{P}_{i,j}$. Similarly, a corresponding p-value matrix $\boldsymbol{\pi}$ can be created to display the associated p-values $\pi_{i,j}$.

In this chapter, we propose a group-level method that incorporates a probability distribution derived from the p-values $\pi_{i,j}$. The method assigns higher weight to stronger correlations by leveraging the relationship between smaller p-values and higher PRCC values. Specifically, the formula uses $1 - \pi_{i,j}$, where $\pi_{i,j}$ is the p-value corresponding to the PRCC value $\mathcal{P}_{i,j}$. The approach is performed by constructing a discrete probability distribution over the set of PRCC values for age group i, denoted as $\mathcal{X}_i = \{\mathcal{P}_{i,1}, \mathcal{P}_{i,2}, \ldots, \mathcal{P}_{i,n_a}\}$, with the probability of selecting each $\mathcal{P}_{i,j}$ defined as:

$$\text{Probability}(\mathcal{X}_i = \mathcal{P}_{i,j}) = \frac{1 - \pi_{i,j}}{\sum_{m=0}^{n_a}(1 - \pi_{i,m})}, \quad j \in \{0, \ldots, n_a\}. \tag{6.4}$$

This formulation ensures that sensitivity values with higher reliability $(1 - \pi_{i,j})$ contribute more to the overall distribution of $\mathcal{X}_i$. The age-group-level PRCC value for the i-th age group, denoted as $\overline{\mathcal{P}}_i$, is then given by the median of the distribution $\mathcal{X}_i$:

$$\overline{\mathcal{P}}_i = \text{median}(\mathcal{X}_i).$$

We use the median rather than the mean since the distribution of $\mathcal{X}_i$ is typically skewed, and the mean may provide a misleading estimate in such cases. To quantify

the uncertainty in the values, we define the confidence interval ($\boldsymbol{\iota}_i$) as the interquartile range:

$$\boldsymbol{\iota}_i = [Q_1(\mathcal{X}_i), Q_3(\mathcal{X}_i)],$$

where $Q_1(\mathcal{X}_i)$ and $Q_3(\mathcal{X}_i)$ represent the lower and upper quartiles of the distribution $\mathcal{X}_i$, respectively.

6.2.1 Summary of the Framework

The statistical Age Group Sensitivity Analysis (AGSA) framework provides a structured approach to quantifying the impact of age-specific contact patterns on key epidemic outcomes. It is designed to integrate social mixing data with an age-structured epidemic model, offering a statistically grounded method for exploring sensitivity across age groups.

The framework requires three essential inputs. First, it takes as input a *contact matrix* that describes the structure of interactions between different age groups in the population. This matrix captures the frequency and context of contacts such as those occurring at home, school, work, or in other settings. Second, the framework employs an *age-structured epidemic model*, introduced in Section 3, which simulates the transmission dynamics of the disease based on both biological parameters and social contact patterns. Third, the model relies on a set of *additional parameters*, including epidemiological rates and initial conditions, which are defined in Appendix A.

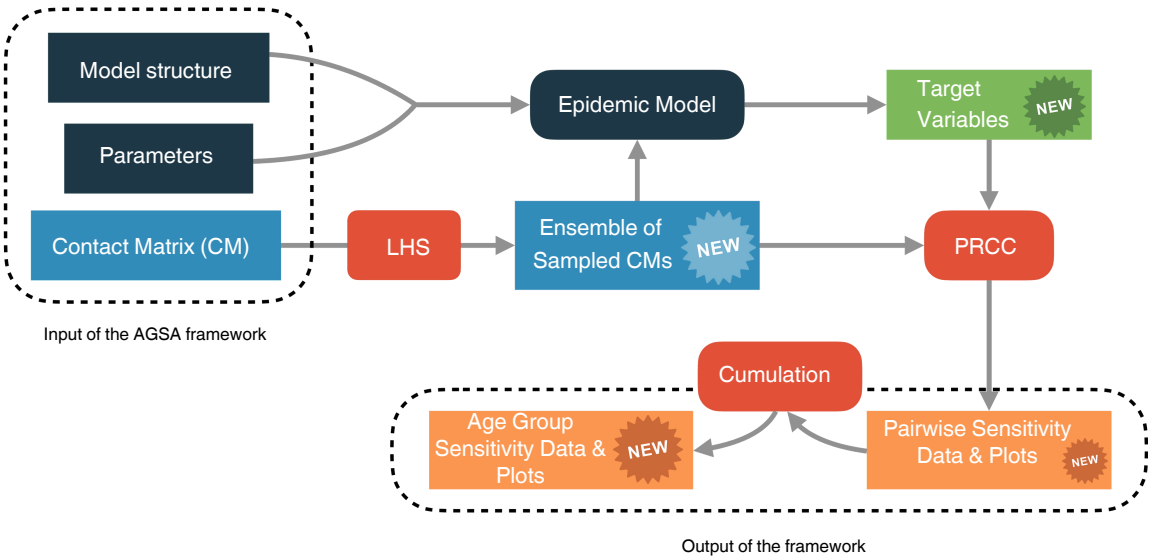

Figure 6.1: Overview of the AGSA framework. The framework begins with input parameters, including the contact matrix (CM) that captures age-specific interaction patterns and a set of model parameters. Latin Hypercube Sampling (LHS) is used to generate an ensemble of modified contact matrices, which are then passed through an age-structured epidemic model to compute key outcome metrics. Partial Rank Correlation Coefficient (PRCC) analysis is applied to the simulation results to estimate the sensitivity of outcomes to contact changes. These sensitivity values are subsequently summarized by age group to identify the most influential groups. The framework outputs both numerical sensitivity measures and visual representations. Components labeled *NEW* represent original contributions of this thesis, including ensemble generation, pairwise and cumulative sensitivity computation, and summarization modules.

The analysis begins by generating multiple variants of the contact matrix using *Latin Hypercube Sampling (LHS)*, as described in Section 6. Each sampled matrix represents a plausible scenario of reduced or altered contact patterns. For each of these scenarios, the epidemic model is simulated, producing corresponding values for one or more *target variables*, such as the basic reproduction number ($\mathcal{R}_0$), peak infections, intensive care unit (ICU) burden, or cumulative deaths.

Once all simulations are completed, the framework computes *Partial Rank Correlation Coefficients (PRCC)* to assess the statistical relationship between each sampled input and the resulting outputs. This step, detailed in Section 4.2, yields a sensitivity measure for each parameter, reflecting its relative influence on the model's behavior.

To make these results interpretable at the age group level, the PRCC values are organized into a symmetric matrix and then derived using the method described in Section 6.2. This approach accounts for statistical reliability by weighting sensitivity values according to their associated p-values. The result is a single, representative sensitivity value for each age group, along with a confidence interval to reflect uncertainty.

The entire workflow is illustrated in Figure 6.1. The AGSA framework is implemented in Python using an object-oriented approach and leverages scientific computing libraries such as `numpy`, `scipy`, and `smt`. This ensures modularity, flexibility, and efficiency, allowing the framework to be adapted to a wide range of epidemiological applications.

6.3 Demonstrations

To evaluate the applicability and robustness of our proposed sensitivity analysis framework, we applied it to two age-structured epidemic models representing different levels of epidemiological complexity. These include an SEIR model for influenza developed by Pitman et al. [44], and a more detailed COVID-19 model from Röst et al. [50]. Full descriptions of the models and their mathematical formulations are provided in Section 3.

Our analysis uses Latin Hypercube Sampling (LHS) with $n_s = 10{,}000$ samples, stratifying the uncertainty space of the contact matrix entries. Each simulation runs until the number of infected individuals falls below one. We explore two epidemic intensity scenarios characterized by basic reproduction numbers $\mathcal{R}_0 = 1.2$ (mild) and $\mathcal{R}_0 = 2.5$ (severe).

Model-specific susceptibility settings were implemented: in the influenza model, we assumed uniform susceptibility across all age groups; for the COVID-19 model,

we introduced non-uniform susceptibility by assigning susceptibility $\sigma_i = 1.0$ for individuals aged 20 and older, and $\sigma_i = 0.5$ for those under 20.

The outcomes of interest in our sensitivity analysis include several key epidemiological indicators: the basic reproduction number $\mathcal{R}_0$, the peak number of infected individuals, the intensive care unit (ICU) peak, and total deaths. Among these, $\mathcal{R}_0$ is computed via the NGM approach [11]. In this framework, we first calibrate the baseline transmission rate β to achieve a desired reproduction number $\overline{\mathcal{R}}_0$ using the original contact matrix C. We then perturb the matrix to obtain $\widehat{C}$ (see Eq. (6.1)) and evaluate the resulting modified reproduction number $\mathcal{R}_0'$. As a result, $\mathcal{R}_0$ for any given model m can be considered a target variable within this method.

6.3.1 SEIR model for influenza epidemic by Pitman et al.

This model, $m = \mathrm{P}$, lacks the complexity required to define multiple target functions. Therefore, we focus on analyzing the sensitivity of contact patterns with respect to the basic reproduction number $\mathcal{R}_0(m = \mathrm{P})$ and the peak number of infected individuals.

Figure 6.2 presents the results of the contact matrix sensitivity analysis under different susceptibility assumptions. The top panels display the pairwise sensitivity values and their corresponding p-values, using $\mathcal{R}_0(m = \mathrm{P})$ as the target variable in our simulations. In contrast, the bottom panels illustrate the results when considering the peak number of infected individuals as the target variable. The analysis is conducted for mild and severe influenza outbreaks to assess variations in sensitivity across different epidemic scenarios.

When using $\mathcal{R}_0(m = \mathrm{P})$ as the target outcome, contact pairs within the same younger age groups, specifically 5–9 with 5–9, 10–14 with 10–14, and 15–19 with 15–19, exhibit the strongest influence (indicated by dark green), and correspondingly have

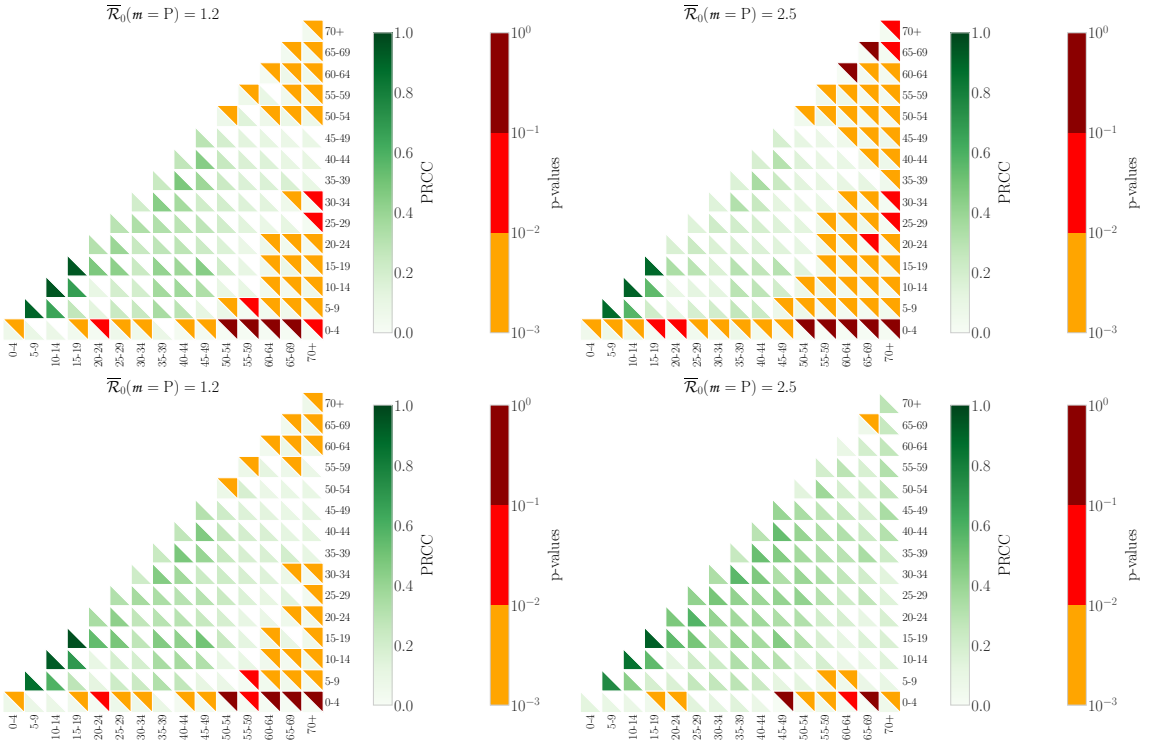

Figure 6.2: Illustration of the framework applied to the Influenza model by Pitman et al. [44] for mild and severe outbreaks. **Top row:** Sensitivity analysis using $\mathcal{R}_0(m = \mathrm{P})$ as the target variable. **Bottom row:** Sensitivity analysis considering the peak number of infected individuals as the target parameter. Pairwise sensitivity values (green bars) are shown on the right axis, and associated p-values are represented by different colors: white (not shown, $< 0.1\%$), orange ($0.1\% - 1\%$), red ($1\% - 10\%$), and dark red ($> 10\%$).

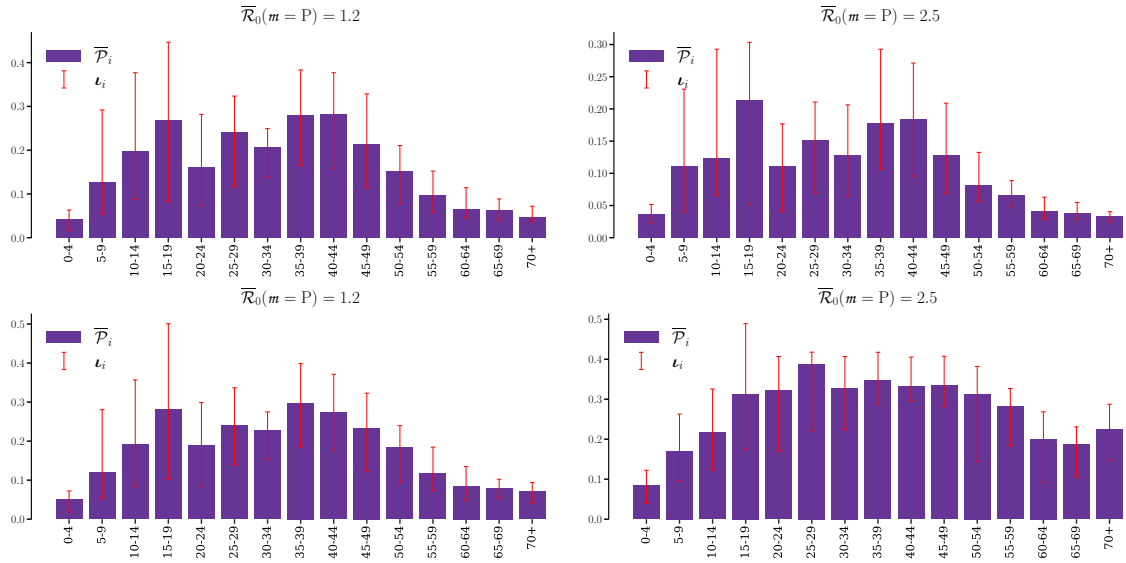

Figure 6.3: Illustration of the PRCC-based sensitivity framework applied to the influenza model by Pitman et al. under both mild and severe outbreak scenarios. **Top row:** Age-group-level sensitivity values with respect to the basic reproduction number $\mathcal{R}_0(m = \mathrm{P})$. **Bottom row:** Age-group-level sensitivity values using the *peak number of infected individuals* as the outcome of interest. Red error bars indicate the confidence intervals associated with the summarized PRCC values.

the lowest p-values, indicating high statistical significance. In contrast, contact pairs involving older age groups tend to be less significant (shown in red and dark red), with higher p-values. This pattern remains consistent across different levels of disease severity. However, when shifting the target variable to the *peak number of infected individuals*, contact interactions across nearly all age group pairs become statistically significant, especially under severe outbreak scenarios (see Figure 6.2, bottom panel).

After deriving age-group-level sensitivity values, Figure 6.3 highlights the groups with the strongest influence on the model's target variables. When considering $\mathcal{R}_0(m = \mathrm{P})$ as the target variable (top row), the sensitivity analysis reveals that in a mild outbreak, the system is most sensitive to the contacts of individuals aged 15–19 and 35–44. In a severe outbreak, individuals aged 5–49 play a dominant role, with the 15–19 age group continuing to exert a strong influence.

In contrast, the sensitivity pattern shifts when using the *peak number of infected individuals* as the target variable (bottom row). For a mild outbreak, sensitivity is again highest for the 35–39 age group. As outbreak severity increases, more age groups become influential. In a severe outbreak, a broader range of age groups from 15–59 years becomes significant. This insight underscores that, depending on the target variable, the sensitivity of different age groups varies significantly with epidemic severity. In particular, the consistent influence of the 15–19 age group across all scenarios can be attributed to the high volume and strong assortative nature of contacts within this cohort.

6.3.2 COVID-19 Model by Röst et al.

This model's complexity allows for defining multiple target functions. We examine the sensitivity of contacts with respect to the basic reproduction number $\mathcal{R}_0(m = \mathrm{R})$, ICU peak, and cumulative fatalities under both mild and severe epidemic scenarios, along with their aggregated sensitivity values.

For $\mathcal{R}_0(m = \mathrm{R})$, the system is most sensitive to contacts among individuals aged 15–59, as indicated by high sensitivity values and low p-values (Fig. 6.4). Older age groups contribute minimally to transmission due to limited interactions. However, under uniform susceptibility, the 15–19 age group becomes the most significant (see bottom panel of Fig. 6.4), aligning with high contact rates in the full contact matrix (Fig. 2.1).

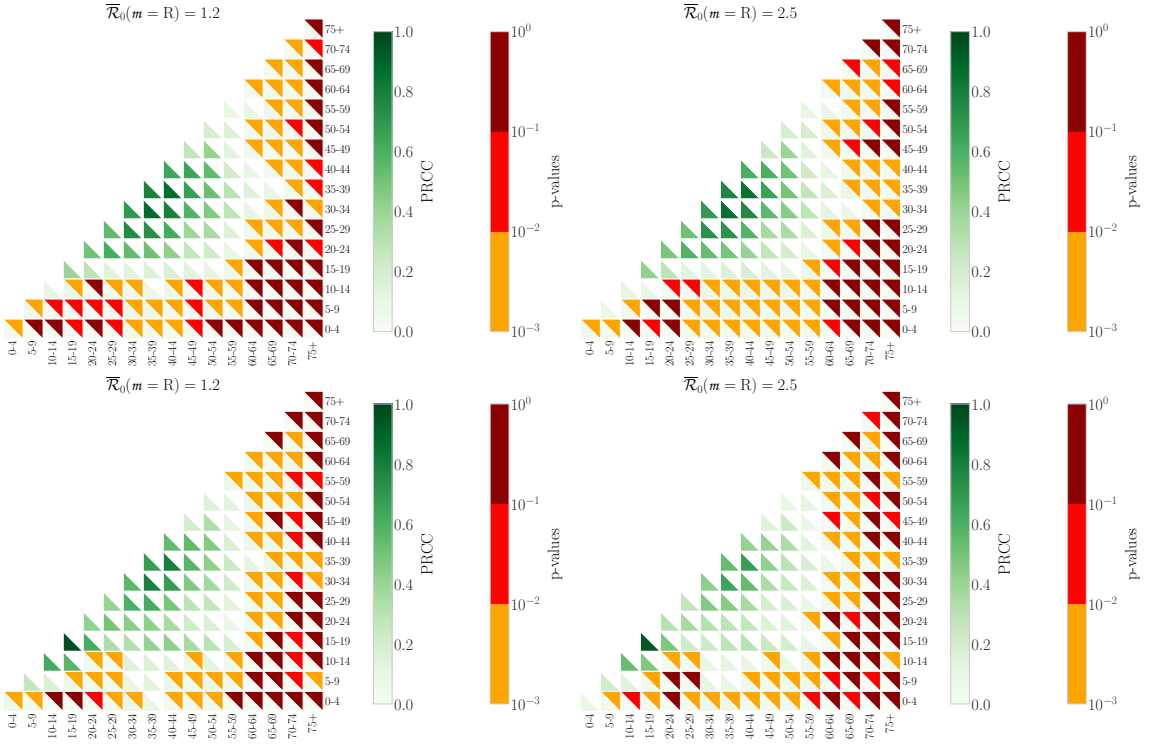

Figure 6.4: Sensitivity of age-group-specific contacts with respect to the $\mathcal{R}_0(m = \mathrm{R})$ under different assumptions of $\overline{\mathcal{R}}_0(m = \mathrm{R})$. The top and bottom panels correspond to scenarios with $\sigma_i = 0.5$ and 1.0, respectively. Green bars indicate pairwise sensitivity values (right axis), while p-values are color-coded for statistical significance: white (not shown, $< 0.1\%$), orange ($0.1\% - 1\%$), red ($1\% - 10\%$), and dark red ($> 10\%$). The figures reveal how transmission dynamics vary across age groups depending on $\overline{\mathcal{R}}_0(m = \mathrm{R})$.

When analyzing the ICU peak, sensitivity shifts towards older age groups under severe scenarios (Fig. 6.5). Under uniform susceptibility, younger groups, particularly 15–19, also become significant.

For cumulative fatalities, contacts among older (60+) and younger (<15) age groups have little impact during mild epidemics but become crucial in severe cases, especially when uniform susceptibility is assumed. High pairwise sensitivity values in older age groups persist due to their elevated mortality risk (Fig. 6.6). Reducing interactions between high-risk groups and the general population is essential for mitigating fatalities.

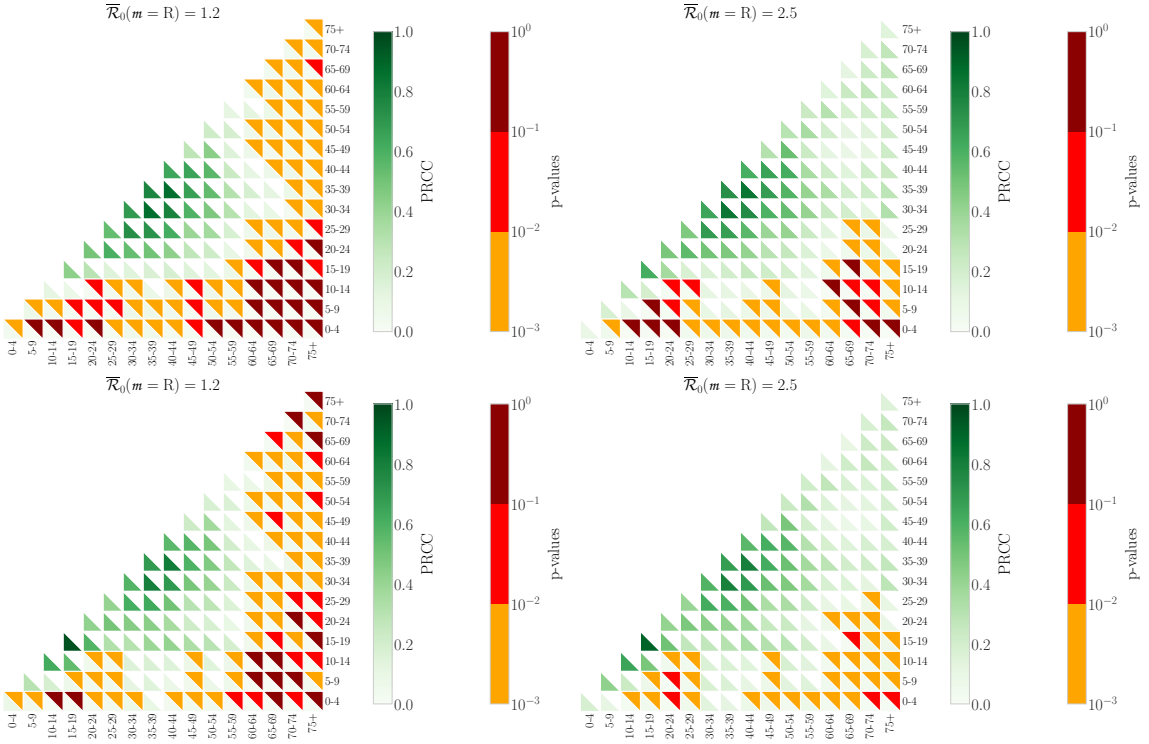

Figure 6.5: Sensitivity of age-group-specific contacts with respect to ICU peak occupancy under different values of $\overline{\mathcal{R}}_0(m = \mathrm{R})$. The top and bottom panels represent cases with $\sigma_i = 0.5$ and $\sigma_i = 1.0$, respectively. Pairwise sensitivity values (green bars, right axis) indicate the strength of association, while p-values are color-coded: white (not shown, $< 0.1\%$), orange ($0.1\% - 1\%$), red ($1\% - 10\%$), and dark red ($> 10\%$). These results highlight the role of different age groups in driving severe cases requiring *intensive care*.

Overall, Figures 6.4, 6.5, and 6.6 demonstrate that contacts within the 15–59 age range are key across all severity levels. However, under uniform susceptibility, the 15–19 age group becomes particularly influential due to their high interaction rates. This highlights how sensitivity varies across age groups depending on epidemic severity and target metrics.

To quantify each age group's contribution to disease transmission, we summarized the pairwise sensitivity values using Eq. (6.4), assigning higher probabilities to larger values due to their statistical significance. The results are presented in Figures 6.7, 6.8, and 6.9.

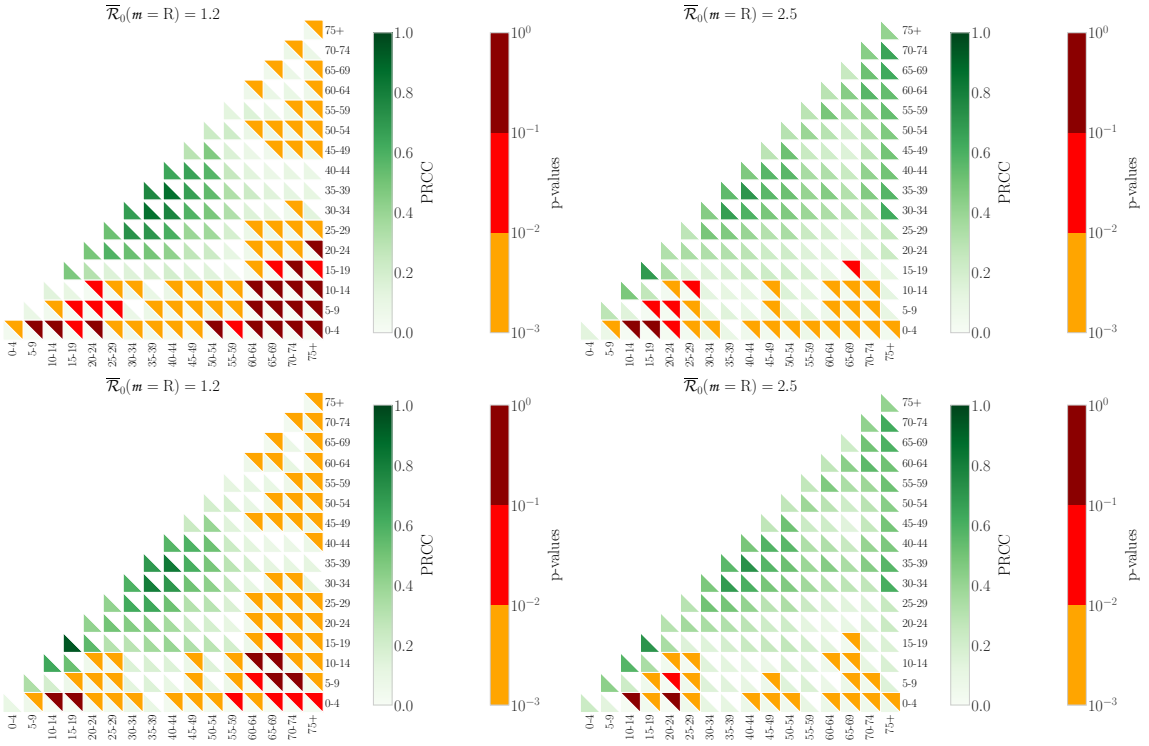

Figure 6.6: Sensitivity of age-group-specific contacts with respect to *cumulative fatalities* across different $\overline{\mathcal{R}}_0(m = \mathrm{R})$. The top and bottom panels show cases with $\sigma_i = 0.5$ and $\sigma_i = 1.0$, respectively. Pairwise sensitivity values (green bars, right axis) represent the strength of association, with p-values color-coded: white (not shown, $< 0.1\%$), orange ($0.1\% - 1\%$), red ($1\% - 10\%$), and dark red ($> 10\%$). These figures emphasize the impact of different age groups on overall mortality.

Figure 6.7 shows that, despite higher contact rates among younger individuals, the 30–44 age group has the most significant impact on transmission. However, under uniform susceptibility, the 15–19 age group also becomes influential. Notably, the age-group-level sensitivity values exhibit higher uncertainty, particularly in the most impactful age groups.

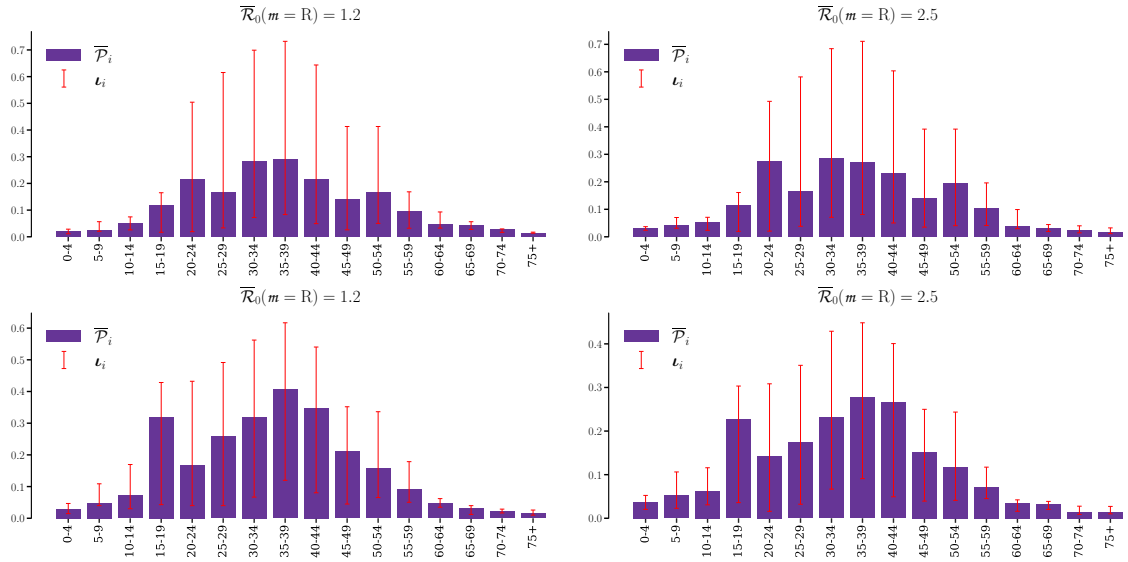

Figure 6.7: Age-group-level sensitivity values for different $\overline{\mathcal{R}}_0(m = \mathrm{R})$ and σ_i conditions, targeting the basic reproduction number $\mathcal{R}_0(m = \mathrm{R})$. The top and bottom rows represent $\sigma_i = 0.5$ and $\sigma_i = 1.0$, respectively. Error bars (red lines) represent confidence intervals, while all bars are uniformly shaded in purple to indicate the sensitivity values.

For *ICU peak*, sensitivity is highest in the 30–44 age group when younger individuals have lower susceptibility. However, under uniform susceptibility, the 15–19 age group also plays a significant role. In severe outbreaks, the 75+ age group becomes critical regardless of susceptibility assumptions (Fig. 6.8). Interestingly, cumulative ICU peaks are more sensitive to the 40–44 age group than the elderly (75+).

For *cumulative fatalities*, a similar pattern to *ICU peak* emerges. The 30–44 age group remains the most sensitive in mild scenarios when younger individuals have

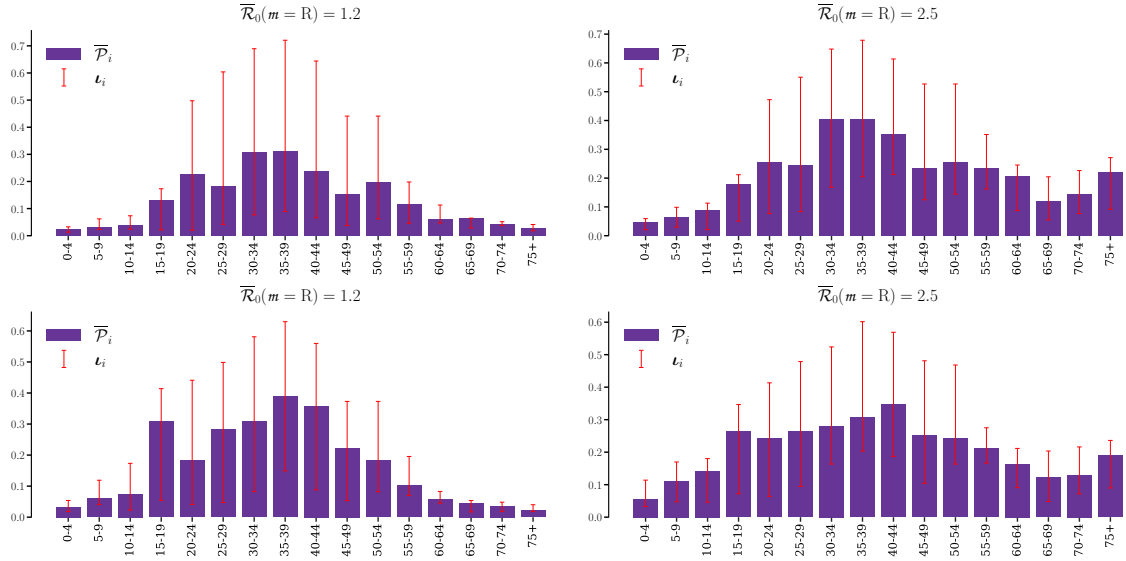

Figure 6.8: Age group–level sensitivity values for age-specific contact rates, using *ICU peak occupancy* as the target outcome. Each column corresponds to a different basic reproduction number $\overline{\mathcal{R}}_0(m = \mathrm{R})$ (1.2 and 2.5), while the top and bottom rows represent susceptibility scenarios of $\sigma_i = 0.5$ and $\sigma_i = 1.0$, respectively. Red error bars indicate confidence intervals for the summarized sensitivity values, which are uniformly shaded in purple. This analysis highlights the age groups most influential in determining ICU demand.

lower susceptibility. However, under uniform susceptibility, sensitivity spreads more broadly across younger and middle age groups, with noticeable contributions from the 15–19 and 25–44 age groups. As severity increases, older populations, particularly the 75+ group, become dominant in determining fatality outcomes. Confidence intervals reflect higher uncertainty in mild scenarios, while the elderly's influence appears more consistent and robust in severe settings (Fig. 6.9). These findings highlight a dynamic age-risk landscape, where younger and middle-aged individuals shape early outbreak outcomes, while older adults dominate fatality sensitivity in severe epidemics.

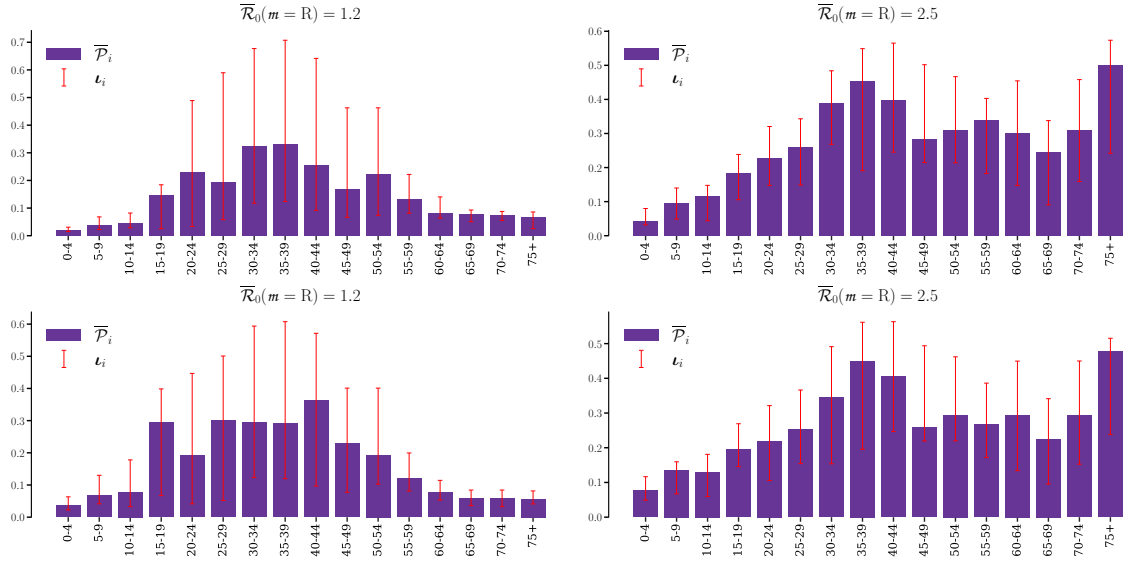

Figure 6.9: Age-group-level sensitivity values for different age groups, targeting *cumulative fatalities*. The layout follows the same structure as Figs. 6.7 and 6.8, with $\overline{\mathcal{R}}_0(m = \mathrm{R})$ values increasing from left to right and $\sigma_i = 0.5$ (top) vs. $\sigma_i = 1.0$ (bottom). Confidence intervals (red error bars) while the bars indicate the sensitivity values. This figure highlights the age groups most associated with overall mortality risk.

A comparison of outbreak scenarios shows that in mild epidemics, reducing contacts among adolescents (15–19) and middle-aged adults (30–44) is effective in limiting transmission and easing hospital burden, owing to their high contact rates. In more severe outbreaks, however, the impact of reducing adolescent contacts diminishes, and the 30–44 age group becomes the dominant contributor to transmission, likely due to their central role in workplace and community interactions. These findings point to the importance of adjusting intervention strategies to outbreak intensity, with targeted contact reductions tailored to the most influential age groups for controlling ICU demand and reducing fatalities.

6.4 Conclusion

The AGSA framework offers a broadly applicable and effective approach for analyzing age-structured epidemic models. Its application to both influenza and COVID-19, two diseases differing significantly in transmission dynamics and complexity, demonstrated its capacity to identify age groups that most influence epidemic outcomes. This underscores the framework's flexibility and potential utility in a wide range of infectious disease contexts, including those with non-respiratory transmission modes such as vector-borne or sexually transmitted infections. Future research could explore the consistency of age-related sensitivities across such varied settings.

A key implication of this study is its contribution to improving data collection strategies. Social contact matrices are foundational to age-structured epidemic modeling, yet they often suffer from biases and insufficient representation of certain age groups. AGSA addresses this gap by providing a data-driven method to prioritize empirical efforts, identifying the age groups that introduce the greatest uncertainty into model predictions. Targeting these groups for improved contact data collection can significantly enhance model accuracy and the reliability of intervention planning.

AGSA also proves valuable in assessing a diverse set of epidemic outcomes, including basic reproduction number ($\mathcal{R}_0$), ICU demand, and cumulative mortality. This versatility allows for alignment of interventions with outbreak severity. For example, in milder epidemics, reducing contacts among adolescents (15–19) and middle-aged adults (30–44) effectively lowers transmission and healthcare burden. In contrast, in more severe scenarios, middle-aged adults emerge as the primary drivers of transmission, while the relative role of younger individuals diminishes. These findings emphasize the importance of tailoring public health strategies to epidemic severity.

AGSA could be enhanced by incorporating additional demographic and behavioral factors such as socioeconomic status, occupational exposure, and comorbidities. Integrating these variables would enable a more comprehensive understanding of population vulnerability. Moreover, accounting for dynamic behavioral responses such as compliance fatigue and evolving policy measures would further refine the model's ability to capture the complexities of real-world epidemics.

Chapter 7

Clustering Analysis

7.1 Problem setting

Social contact patterns differ significantly across regions, yet there is limited comparative evidence on how these patterns influence disease transmission between countries. Countries with larger populations, higher-risk behaviors, and less favorable social structures are more likely to experience elevated transmission rates. Understanding these variations is essential for the timely implementation of non-pharmaceutical interventions (NPIs), assessing their effectiveness, and reintroducing measures in response to case resurgences [35]. During the COVID-19 pandemic, countries with later outbreaks often observed the strategies of others to make informed, timely decisions. Identifying regions with similar social mixing patterns can guide the adoption of effective interventions.

Although non-pharmaceutical interventions (NPIs) have proven effective in curbing disease transmission, they have also disrupted economies and disproportionately impacted vulnerable populations [8]. Developed countries have mitigated some of these

effects through digitization and alternative services, whereas developing countries have faced challenges such as disrupted supply chains, unemployment, and income loss [10]. Balancing public health and economic stability during pandemics is therefore critical. Social distancing measures, including remote work, school closures, and restrictions on gatherings, influence both epidemic outcomes and economic activity. Younger and older populations, in particular, are heavily impacted [43, 53]. In developing countries, reduced mobility and social interactions can intensify economic instability. At the same time, limited healthcare access, crowded living conditions, and reliance on informal labor markets may contribute to continued disease transmission despite restrictions [33]. Socioeconomic indicators, together with public health measures, must be integrated to promote resilience and ensure adherence to NPIs.

In this chapter, countries are grouped according to their social contact patterns and socioeconomic characteristics to identify regions with similar dynamics and to support the design of tailored non-pharmaceutical intervention (NPI) strategies. European countries are analyzed based solely on their contact patterns, as described in [30], whereas African countries are clustered using a method that incorporates both social contact patterns and the socioeconomic indicators presented in Section 4.7, following the approach outlined in [24]. Due to the high dimensionality of the data, dimensionality reduction techniques, as discussed in Section 4.5, are applied.

7.2 Methods

Socioeconomic data were integrated with social contact information to explore their implications for non-pharmaceutical interventions (NPIs). Given the high dimensionality of the socioeconomic data, dimensionality reduction techniques were applied to mitigate the curse of dimensionality. Specifically, we employed the 1D Principal

Component Analysis (PCA) method introduced in Eq. (4.2) in Section 4.5.1. The objective was to reduce the socioeconomic data from $n_f = 28$ features per country to a lower-dimensional representation. Using 1D PCA, the dataset comprising 28 features for each of the $n_c = 32$ African countries was reduced to $r = 4$ principal components, forming the transformed feature matrix:

$$\widetilde{\mathcal{F}}_r(\boldsymbol{r} = A) = \mathcal{F}(\boldsymbol{r} = A) \cdot \mathrm{Proj}_{\mathcal{F}}(r) \in \mathbb{R}^{32\times 4}. \tag{7.1}$$

The social contact matrices from [46], introduced in Chapter 2.3, serve as the input for analyzing contact patterns across different countries. These matrices facilitate comparative studies of social interactions. For the European region, contact data were available for $n_c = 39$ countries, with each country's population stratified into $n_a = 16$ age groups. As introduced in Section 2.3, the corresponding contact matrices are denoted by $C(\boldsymbol{c} = c) \in \mathbb{R}^{16\times 16}$. These matrices are then standardized using the baseline transmission rates, following the procedure described in Equation (4.1) in Section 4.4. To perform dimensionality reduction, we applied the $(2D)^2$ PCA method introduced in Section 4.5 and derived the reduced matrix for each country, $\mathcal{S}(\boldsymbol{r} = E, \boldsymbol{c} = c) \in \mathbb{R}^{2\times 2}$ in Eq. (4.3). Each reduced matrix was then flattened into a feature vector:

$$f(\boldsymbol{r} = E, \boldsymbol{c} = c) = \mathrm{vec}\left(\mathcal{S}(\boldsymbol{r} = E, \boldsymbol{c} = c)\right) \in \mathbb{R}^4. \tag{7.2}$$

This feature vector was used for clustering analysis in Section 7.3.1.

For African countries, contact data from [46] were available for $n_c = 32$ countries, initially structured into $n_a = 16$ age groups. As described in Section 2.4, this data was aggregated into $n_g = 6$ broader age groups, as illustrated in Figure 2.3. The resulting aggregated contact matrices are denoted by $\mathbf{C}(\boldsymbol{c} = c) \in \mathbb{R}^{6\times 6}$. These matrices were then scaled using the baseline transmission rates, following the standardization procedure outlined in Equation (4.1) in Section 4.4. Subsequently, the scaled matrices were

concatenated row-wise, and $(2D)^2$ PCA was applied to extract a reduced representation $\mathcal{S}(r = A) \in \mathbb{R}^{2\times 2}$, as shown in Equation (4.4). Flattening this reduced contact matrix results in a feature vector:

$$f(r = A) = \text{vec}\left(\mathcal{S}(r = A)\right) \in \mathbb{R}^4. \tag{7.3}$$

Since socioeconomic indicators were also incorporated into the analysis of African countries, the reduced feature vectors from Eq. (7.1) were concatenated with the contact matrix-derived feature vectors in Eq. (7.3):

$$\boldsymbol{X}(\boldsymbol{c} = c) = \left[f(r = A, \boldsymbol{c} = c) \;\middle|\; \widetilde{\mathcal{F}}_r(r = A)\right] \in \mathbb{R}^8. \tag{7.4}$$

This combined representation was subsequently used for the clustering analysis.

To cluster countries from both regions, we apply agglomerative hierarchical clustering using the complete linkage method, as described in Section 4.6. The feature vector of each country, derived from Equations (7.2) and (7.4), is considered an individual observation in the clustering process. Pairwise dissimilarities between countries are calculated using the Euclidean distance between their feature vectors, resulting in dissimilarity matrices that are iteratively updated as clusters are merged. This process continues until all countries are grouped into a single cluster. The result is a dendrogram that visualizes the hierarchy of merges. The optimal number of clusters is determined by identifying the largest vertical gap, or linkage distance, in the dendrogram, which reflects the most significant separation between clusters.

7.2.1 Framework Implementation

The complete pipeline of the proposed framework is illustrated in Figure 7.1. The primary inputs are country-specific contact matrices for European countries and aggregated contact matrices for African countries. As a preprocessing step, all matrices

are symmetrized upon loading to ensure consistent total contact volumes across age groups.

These contact matrices serve as inputs to the age-structured epidemic model described in Section 3. For each country $\boldsymbol{c} = c$, we calculate the baseline transmission rate $\beta(\boldsymbol{c} = c)$ using the NGM approach. The basic reproduction number $\mathcal{R}_0$ is fixed based on the region. Alternatively, one could derive $\beta(\boldsymbol{c} = c)$ by calibrating to a predefined epidemiological outcome, such as a fixed final epidemic size or mortality burden. Once $\beta(\boldsymbol{c} = c)$ is computed, we standardize the contact matrices by scaling them as defined in Equation (4.1).

In the next stage, we apply data-driven methods to the standardized matrices, as discussed in Section 4.5. First, we perform dimensionality reduction to address the curse of dimensionality and simplify the data structure. For European countries, the resulting low-dimensional feature vectors are defined in Equation (7.2) and subsequently used for clustering. For African countries, additional socioeconomic indicators relevant to NPI adherence are integrated with the reduced features from scaled contact matrices. These combined feature vectors, presented in Equation (7.4), are then used for clustering as outlined in Section 4.6.

The framework is implemented in Python using an object-oriented design, ensuring modularity and ease of extension. We leverage open-source libraries for solving differential equations, computing eigenvalues, and performing dimensionality reduction and clustering.

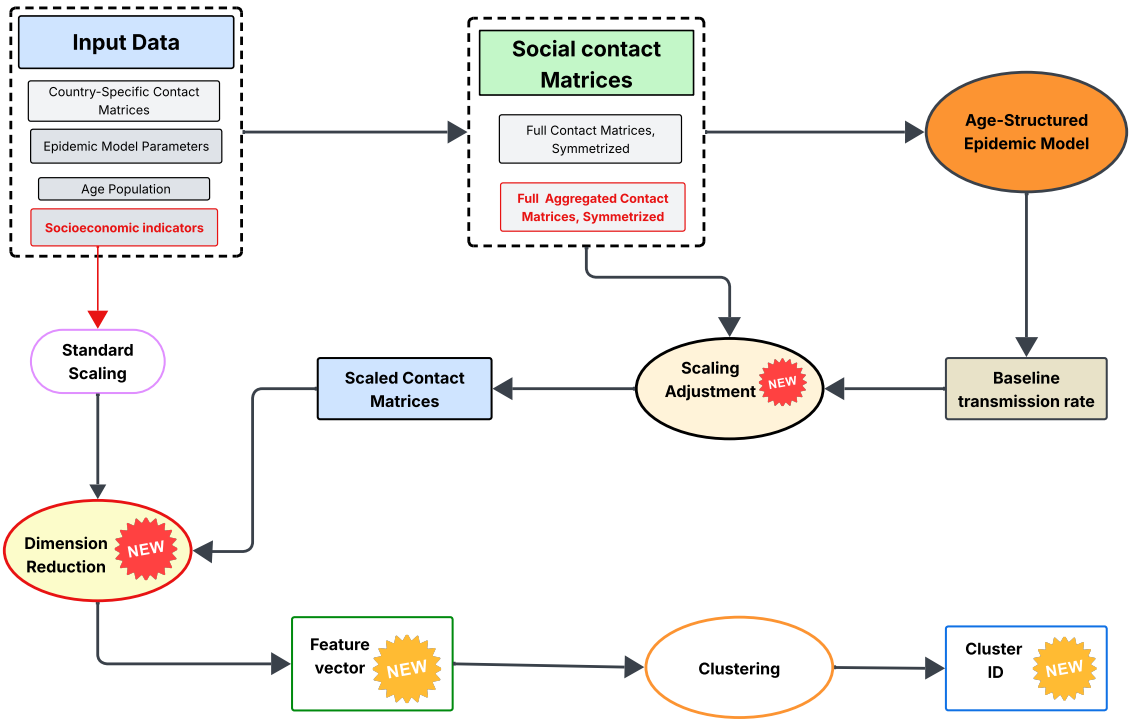

Figure 7.1: Pipeline of the proposed framework. The procedure begins with loading and symmetrizing region-specific contact matrices, followed by the computation of baseline transmission rates using the NGM method under a fixed $\mathcal{R}_0$. These rates are then used to standardize the contact matrices for each country. In the next stage, dimensionality reduction is applied to the standardized matrices to obtain low-dimensional feature representations. For African countries, socioeconomic indicators are incorporated into these feature vectors. The resulting representations are then used for clustering countries based on their contact structure and contextual factors. Newly developed components introduced in this section are marked with the *NEW* label.

7.3 Demonstrations

To ensure comparability across countries within each region, contact matrices were standardized using region-specific baseline reproduction numbers: $\mathcal{R}_0 = 3.68$ for African countries and $\mathcal{R}_0 = 2.2$ for European countries, as described in Section 4.4. The resulting scaled contact matrices provided a consistent basis for analyzing cross-country similarities in social mixing patterns before clustering.

7.3.1 Clustering of European Countries Based on Social Contact Patterns

Pairwise Euclidean distances between countries were computed from their scaled contact matrices. These distances were then used to generate a hierarchical clustering dendrogram, which was used to reorder the distance matrix for clearer interpretation, as illustrated in Figure 7.2.

In this reordered distance matrix, darker shades of blue and purple represent greater similarity in contact patterns between European countries, clustering similar social behaviors together. For instance, countries like Austria, Ukraine, and North Macedonia form a cohesive group with low inter-country distance, suggesting aligned interaction structures. In contrast, bright yellow areas indicate higher dissimilarity, such as between Germany and Italy, and countries like Albania or Armenia. This stratification underscores the existence of distinct social mixing patterns across the continent and supports the case for region-specific public health strategies that account for localized contact behaviors.

One advantage of hierarchical clustering is its flexibility in defining clusters at different levels of granularity. By setting a linkage distance threshold of 5.5, we identified three distinct clusters of countries. The dendrogram illustrating the hierarchical structure is shown in Figure 7.3, and the resulting clusters are listed in Table 7.1.

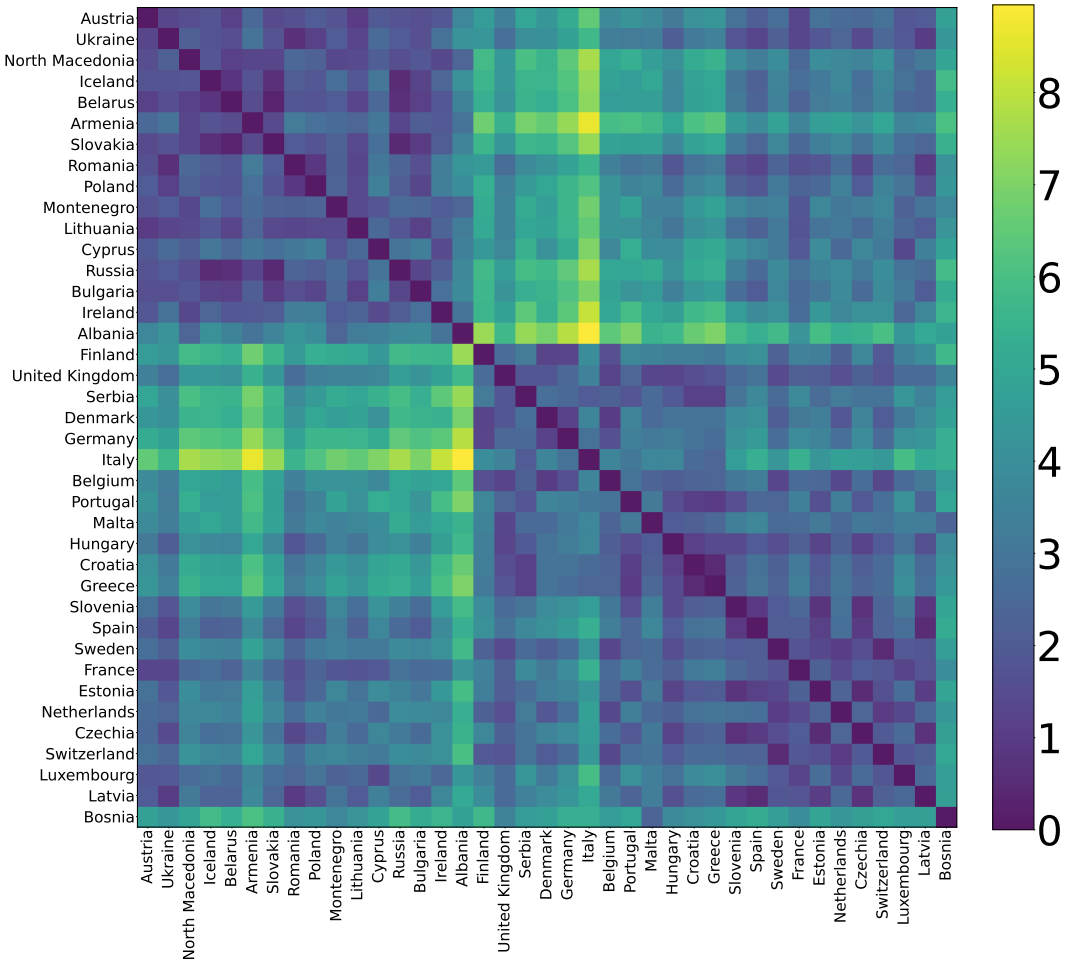

Figure 7.2: Pairwise distance matrix computed for all countries based on feature vectors from the dimensionality reduction step. Euclidean distance was used for the calculations. The dendrogram from hierarchical clustering was applied to reorder the rows and columns, improving the interpretability of the matrix.

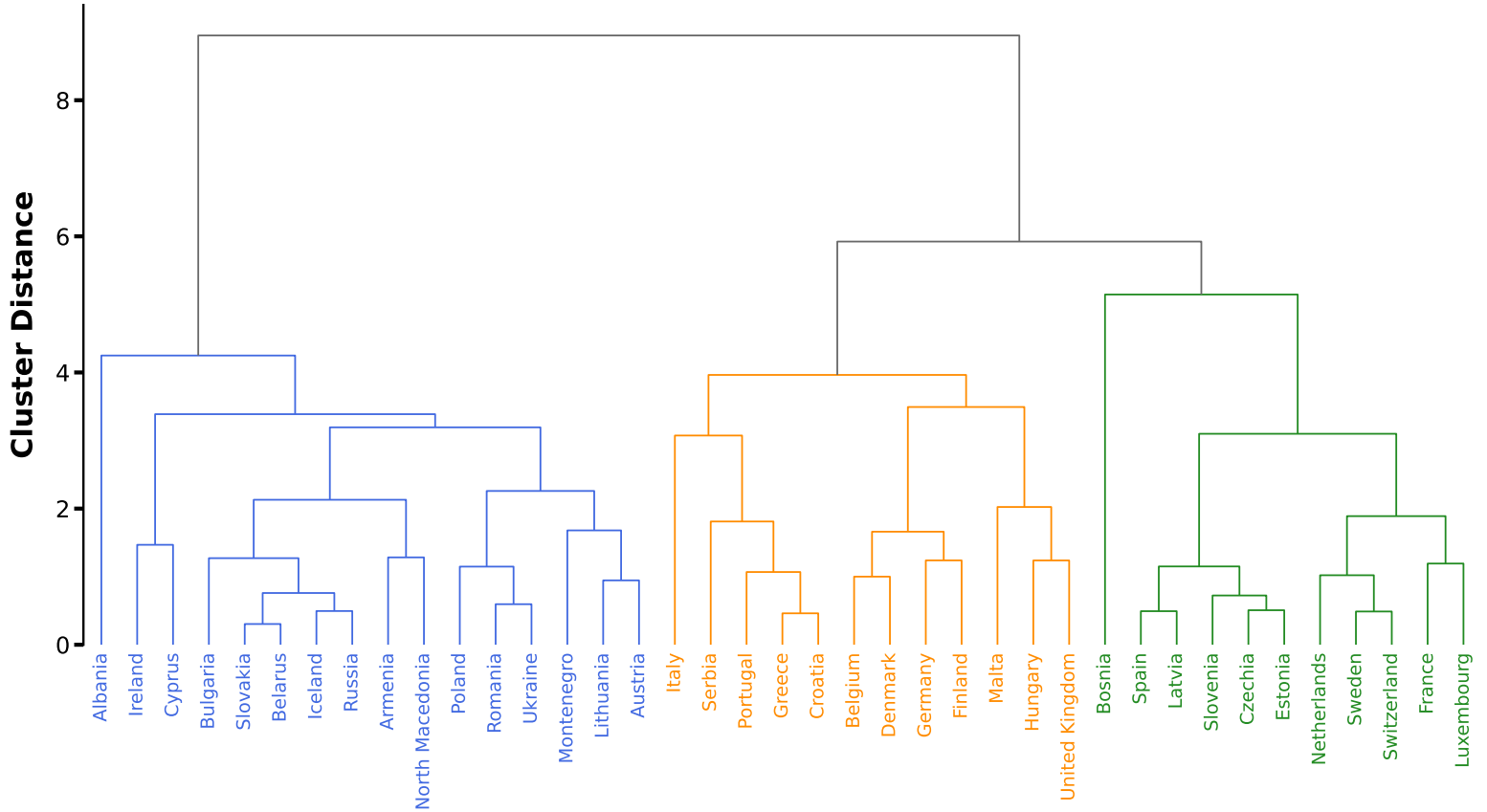

Figure 7.3: Hierarchical dendrogram illustrating the clustering of European countries based on feature vectors obtained from the $(2D)^2$ PCA method. The vertical axis represents the linkage distance, indicating how dissimilar two clusters are before merging. The hierarchical clustering was performed using an agglomerative approach, where individual countries initially form separate clusters and progressively merge as one moves up the tree.

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.

Table 7.1: Clusters obtained from the agglomerative hierarchical clustering algorithm, based on social contact patterns of 39 European countries using the $(2D)^2$ PCA projection. The analysis resulted in three distinct clusters: Cluster 1 contains 16 countries, Cluster 2 includes 12 countries, and Cluster 3 consists of 11 countries. These groupings reflect similarities in social interaction patterns among the countries, which may have implications for public health interventions and policy planning.

To visualize structural differences in social contact patterns across clusters, we selected one representative country from each group: Armenia (Cluster 1), Belgium (Cluster 2), and Estonia (Cluster 3). These countries were chosen because they are positioned near the center of their respective clusters along the horizontal axis of the dendrogram, making them representative of the typical contact patterns within each group. The standardized contact matrices for these countries are presented in Figure 7.4.

Several distinguishing patterns emerge across the matrices. Along the main diagonal, which captures intra-age group contact, Armenia displays weaker interactions among older age groups compared to Belgium and Estonia. This suggests that in Armenia, older individuals have fewer contacts within their own age group. Intergenerational interactions, reflected in the off-diagonal bands, are more prominent in Belgium and Estonia, indicating stronger ties between adult children and older adults in these countries. The central region of the matrices, typically representing working-age contacts, also shows clear variation. Belgium demonstrates a higher concentration of interactions in this area, suggesting a denser pattern of workplace-related contacts. Estonia exhibits a similar structure, albeit slightly less concentrated, while Armenia's contact intensity in this region appears more diffused. These variations underscore

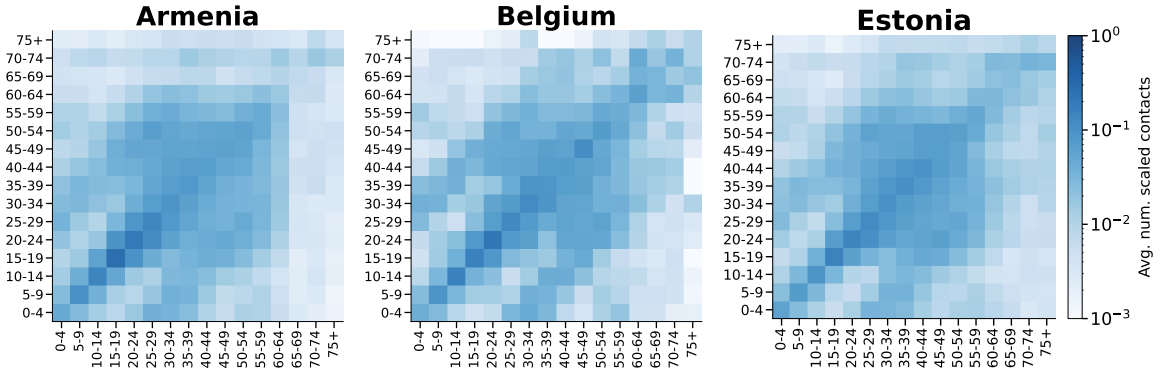

Figure 7.4: Standardized contact matrices for Armenia (left), Belgium (center), and Estonia (right), selected from different clusters formed using feature vectors generated by the $(2D)^2$ PCA technique. These countries were chosen as representative examples from the central positions of their respective clusters based on the hierarchical dendrograms. The age groups are structured in five-year intervals: age group 0 represents 0–4 years, age group 1 corresponds to 5–9 years, ..., and the last age group 16 represents individuals aged 75+. Notable differences among the matrices appear in several key regions: (i) along the main diagonal, where Belgium shows higher contact intensities for both younger and older age groups, (ii) along the diagonals parallel to the main diagonal, reflecting stronger interactions between adult children and their older counterparts, and (iii) in the central region, which primarily represents workplace contacts and exhibits distinct distribution patterns across the three matrices.

the diversity of social mixing behaviors, which are shaped by cultural, demographic, and economic factors across different regions. Although the estimation of contact matrices can be affected by smoothing techniques, especially in contexts with limited data availability [46], the use of $(2D)^2$ PCA helps retain essential structural features. Consequently, the resulting clustering analysis captures meaningful distinctions in contact dynamics among countries.

7.3.2 Clustering of African Countries Based on Social Contacts and Socioeconomic Indicators

For African countries, the clustering analysis incorporated both social contact patterns and socioeconomic indicators to better capture the multifaceted drivers of disease transmission. To reduce the complexity of the data, we applied Principal Component Analysis (PCA) to the socioeconomic dataset, reducing the dimensionality

from 28 to 4 components. As illustrated in Figure 7.5, the first principal component is associated with variables like internet access, electricity coverage, and GDP per worker. The second is driven by unemployment and HIV incidence, while the third reflects macroeconomic metrics such as exports and debt service. The fourth component captures variation in GDP growth.

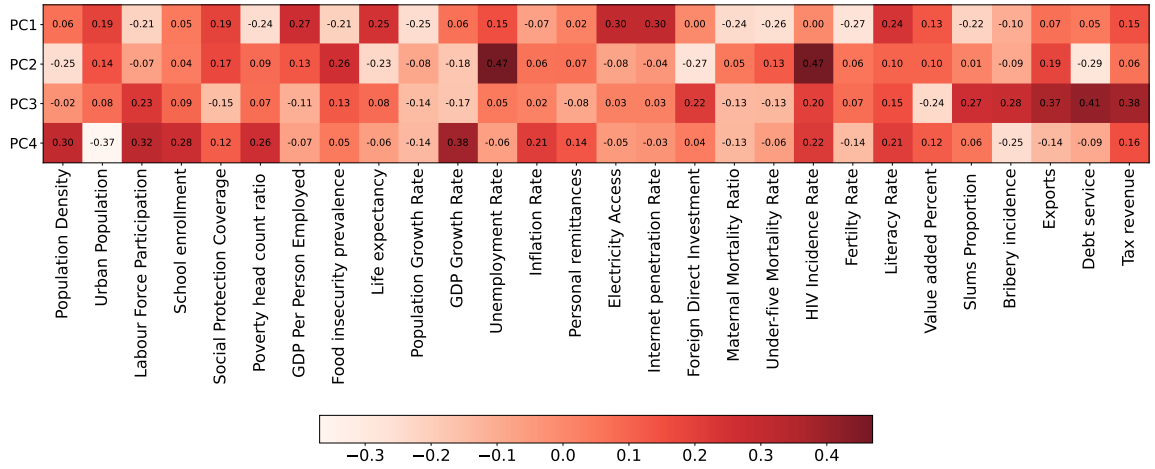

Figure 7.5: Heatmap showing the loadings of socioeconomic indicators on the first four principal components (PCs), obtained via 1D PCA applied to standardized indicator data. Each cell represents the contribution (loading) of a given indicator to a principal component. Darker red shades indicate higher absolute loadings, while lighter shades indicate weaker contributions. The axes are labeled with indicator names and PC numbers, and each cell is annotated with the exact loading value. The 'Reds' colormap provides a visual reference for the relative importance of each variable in each component.

In parallel, we normalized and reduced the age-structured contact matrices using the $(2D)^2$ PCA method. The resulting features were then concatenated with the reduced socioeconomic indicators to form a unified representation for each country. Based on these combined features, we computed pairwise Euclidean distances and reordered the resulting dissimilarity matrix using hierarchical clustering. The heatmap in Figure 7.6 highlights how countries group according to similarities in both behavioral and

structural characteristics.

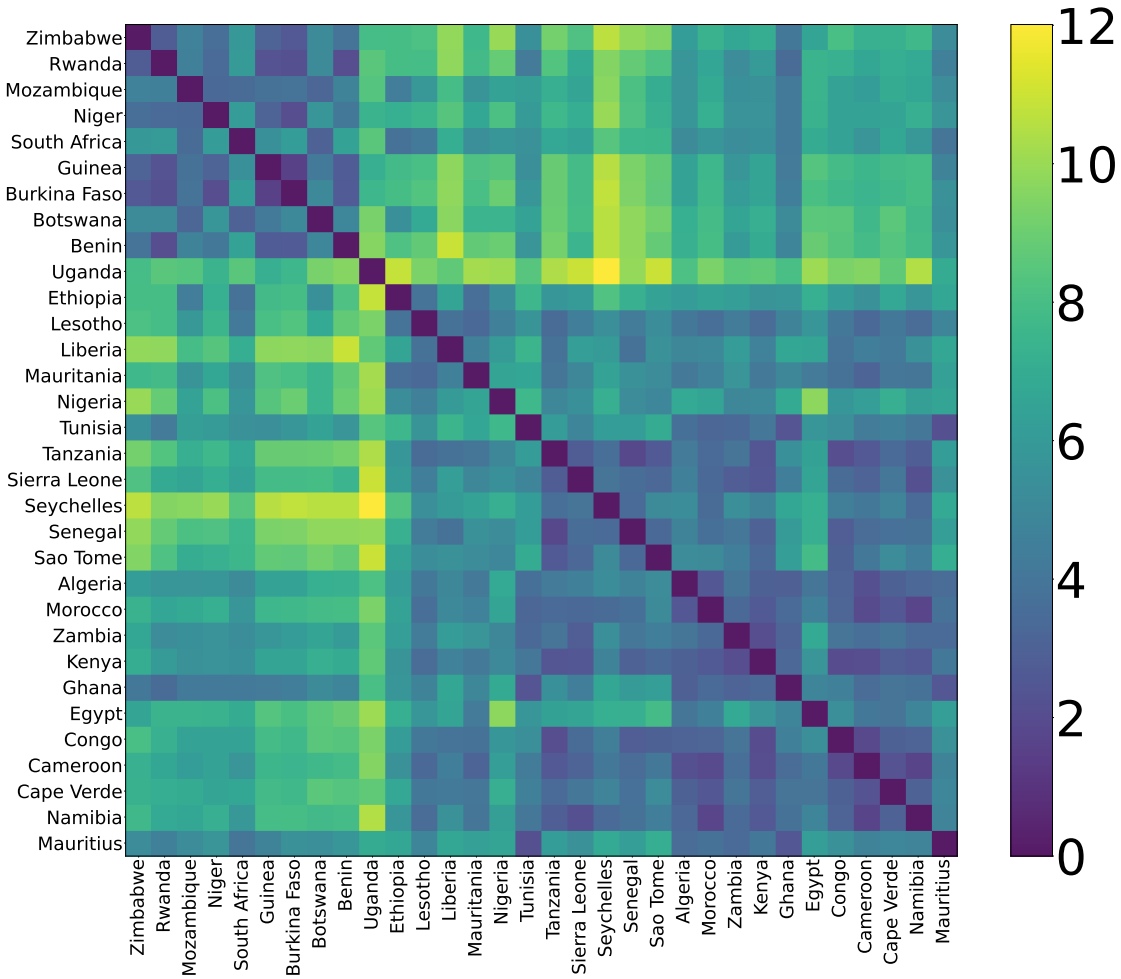

Figure 7.6: Heatmap of pairwise distances between African countries, computed from their social contact matrices and socioeconomic indicators using the $(2D)^2$ PCA technique. Each cell represents the distance between two countries, with darker blue and purple shades indicating greater similarity (shorter distances), and brighter yellow tones representing greater dissimilarity (longer distances). The axes list the analyzed countries, and the color bar on the right provides a quantitative reference. This visualization reveals clusters of countries with similar social contact patterns and socioeconomic characteristics.

The matrix reveals strong similarities among several countries, especially in central

and southern Africa, while countries like Uganda appear as outliers due to their unique profiles. To determine the optimal number of clusters, we identified the largest vertical gap in the dendrogram (Figure 7.7) and cut the hierarchy at a height of 8. This yielded four distinct clusters.

The resulting clusters reflect meaningful regional and socioeconomic groupings:

Cluster 1 includes countries like Botswana, Mozambique, and South Africa, characterized by high labor force participation but limited social security. Some, such as Zimbabwe and Mozambique, also face high HIV prevalence and informal housing. South Africa stands out within this group due to its advanced infrastructure and industrial base.

Cluster 2 contains Nigeria, Ethiopia, Lesotho, Liberia, and Mauritania. These countries share moderate access to electricity (around 50%), low internet penetration, and relatively high maternal health indicators. Economic and digital infrastructure challenges are key differentiators here.

Cluster 3 consists of 17 countries, including Algeria, Egypt, Morocco, Ghana, and Mauritius. This group has the highest average life expectancy (around 77 years), widespread access to health and education services, and strong social development metrics. It includes both North African nations and more developed island states.

Uganda due to its unique combination of low social protection, rapid population growth, and high foreign investment, forms a separate cluster of one. Despite sharing some contact similarities with Sierra Leone, its socioeconomic profile is distinct enough to warrant individual classification. Table 7.2 summarizes the country composition of

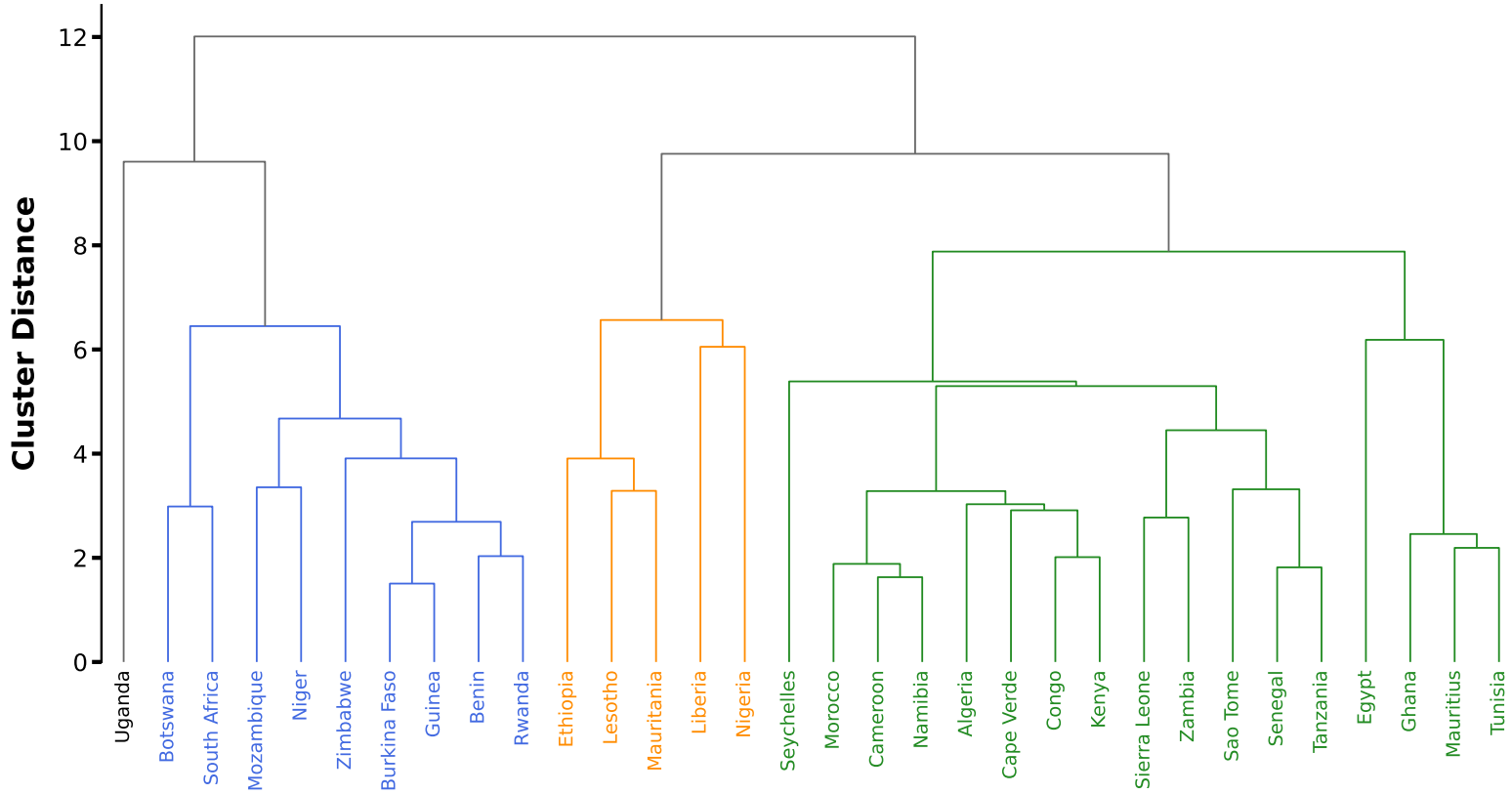

Figure 7.7: Dendrogram illustrating the hierarchical clustering of countries based on feature vectors derived from social contact matrices and socioeconomic indicators. The clustering was performed using the $(2D)^2$ PCA and PCA techniques. The vertical axis represents the linkage distance, indicating the dissimilarity between merged clusters. The clustering process follows an agglomerative approach, where countries and clusters merge progressively from the bottom to the top. Cutting the hierarchy at a height of 8 results in four distinct clusters, visually represented by black, blue, orange, and green. Countries sharing the same color belong to the same cluster, highlighting similarities in social contact patterns and socioeconomic characteristics.

each cluster (excluding Uganda, which stands alone).

Cluster 1	Cluster 2	Cluster 3
Botswana, Mozambique, South Africa, Burkina Faso, Guinea, Benin, Rwanda, Niger, Zimbabwe.	Nigeria, Ethiopia, Lesotho, Liberia, Mauritania.	Sao Tome, Seychelles, Senegal, Tanzania, Egypt, Mauritius, Ghana, Tunisia, Sierra Leone, Kenya, Zambia, Algeria, Congo, Cape Verde, Namibia, Cameroon, Morocco.

Table 7.2: Clusters obtained through agglomerative hierarchical clustering based on social contact patterns and socioeconomic indicators for 32 African countries, using PCA and $(2D)^2$ PCA projections. The analysis resulted in three main clusters, containing 9, 5, and 17 countries, respectively. Uganda, forming a distinct single-element cluster, is excluded from the list.

To further illustrate the diversity of contact structures across clusters, we selected one representative country from each group, Zimbabwe (Cluster 1), Mauritania (Cluster 2), and Sierra Leone (Cluster 3), based on their central positioning in the dendrogram (Figure 7.7). Figure 7.8 presents their standardized contact matrices.

The matrices highlight variations in age-based interactions. In Zimbabwe, there is strong contact among school-age children and extensive workplace interactions, with limited elderly engagement, reflecting demographic and labor dynamics. Mauritania and Sierra Leone show higher contact rates among younger age groups but fewer interactions in secondary and university-age populations. Workplace interactions also vary, with Zimbabwe showing broader engagement across age groups.

Across all three countries, contact with individuals aged 65 and above is minimal. This pattern likely reflects lower life expectancy in the region, as well as limited institutional or familial support for regular intergenerational interactions. Although the overall structure of contact matrices in African countries exhibits broad similarities, particularly due to the predominance of younger populations, the inclusion of socioeconomic indicators significantly alters the clustering results. Countries that appear similar based on contact behavior alone diverge once factors such as income, infrastruc-

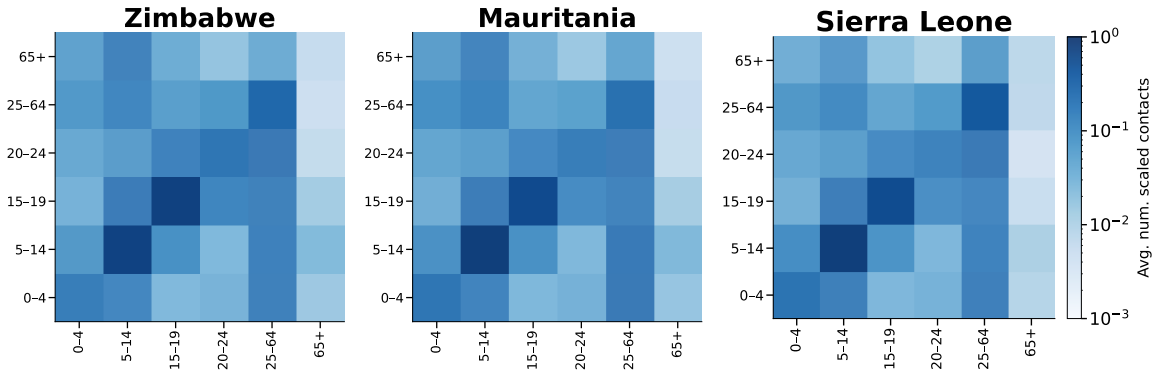

Figure 7.8: Standardized contact matrices for Zimbabwe (a), Mauritania (b), and Sierra Leone (c), selected as representative countries from the three clusters based on the contact matrices in Figure 7.7. These countries are centrally positioned within their respective clusters. The matrix patterns vary across clusters, with Cluster 1 reflecting the pattern observed in (a), Cluster 2 following (b), and Cluster 3 aligning with (c).

ture, education, and healthcare access are taken into account. These findings highlight the importance of a multidimensional approach to analyzing disease transmission risk and demonstrate how structural inequalities influence both social behavior and public health vulnerability.

7.4 Conclusion

The global spread of emerging infectious diseases varies significantly across countries, highlighting the importance of tools that can identify regions with similar transmission dynamics and comparable responses to non-pharmaceutical interventions (NPIs). This chapter introduced a clustering framework designed to group countries based on their social contact patterns and, in the case of African countries, their socioeconomic conditions. The framework is intended to support policymakers in anticipating transmission risks and adopting interventions informed by experiences from countries with similar profiles.

One of the main strengths of this framework is its adaptability to different modeling

goals. The contact matrices were scaled using the basic reproduction number ($\mathcal{R}_0$) to reflect differences in transmissibility. However, alternative public health priorities, such as minimizing hospital burden or reducing mortality, can be incorporated by adjusting the weighting of model parameters. The clustering process can also be adapted to include factors such as case fatality rates or healthcare system capacity. In addition, the underlying epidemic model is modular and can be replaced with disease-specific models to extend the framework to other pathogens.

Further methodological improvements could enhance the performance of the framework. Dimensionality reduction, although effective in this study, may benefit from the use of advanced techniques such as deep learning or convolutional models that capture nonlinear structures in the data. The clustering step can also be refined by exploring alternative distance metrics or linkage methods, depending on the nature of the dataset.

Another important consideration is the role of different contact settings, including home, school, workplace, and other environments. Contact patterns across these settings change during the implementation of NPIs. For example, household interactions tend to remain stable or increase during lockdowns, while contacts at schools and workplaces often decline. Assigning appropriate weights to these settings could improve the responsiveness of the framework to real-world behavioral shifts.

In conclusion, the proposed clustering framework provides a practical, scalable, and flexible approach for identifying countries with similar transmission-related characteristics. The integration of behavioral and structural data supports more informed public health decision-making. This approach can help improve epidemic preparedness and guide context-specific responses to emerging infectious diseases.

Chapter 8

Concluding remarks

Mathematical models have long played a central role in understanding the spread of infectious diseases, particularly in identifying mechanisms that drive transmission and in designing effective interventions. Although classical compartmental ODE models have provided valuable insights into disease dynamics, more recent challenges such as the COVID-19 pandemic, have underscored the importance of incorporating structured heterogeneity into these models. Motivated by this need, the second part of this thesis has focused on how age-specific contact patterns, susceptibility profiles, and socioeconomic structures influence epidemic outcomes. Through the development and application of sensitivity and clustering frameworks, we have demonstrated how these elements interact to shape disease dynamics in both deterministic and stochastic settings.

In Chapter 5, we introduced an eigenvector-based sensitivity analysis that quantifies the marginal effect of changes in contact rates on the basic reproduction number, $\mathcal{R}_0$. The approach builds directly on the spectral properties of the Next Generation Matrix and yields a computationally efficient and interpretable method for evaluating how

individual interactions between age groups influence transmission potential. Unlike simulation-heavy methods, this analytical framework is capable of identifying the most influential contact pairs and aggregating sensitivity over age groups. Applying the method to multiple epidemic models, including a COVID-19 model for Hungary and an SEIR model for influenza in the UK, we observed that middle-aged adults, particularly those aged 30–44, often dominate in their contribution to transmission due to their dense intra-group and inter-group contact networks.

To complement this local sensitivity analysis, Chapter 6 introduced a statistical, simulation-based framework combining Latin Hypercube Sampling (LHS) with Partial Rank Correlation Coefficients (PRCC). This method accounts for uncertainty in model inputs and allows for the global exploration of parameter space, capturing nonlinear and monotonic relationships between contact patterns and key epidemic outcomes. We proposed a novel method to quantify the overall influence of each age group, incorporating statistical significance via a weighting scheme derived from p-values. Applying this to both influenza and COVID-19 models, we found that contact patterns among working-age adults again played a central role, especially in moderate and severe outbreaks. Notably, while adolescents (15–19) were highly influential in mild scenarios due to their high contact frequency, their importance diminished as disease severity increased, reinforcing the dynamic nature of intervention effectiveness.

In Chapter 7, we shifted from within-population dynamics to cross-country comparisons, developing a clustering framework to group countries based on social contact matrices and, where available, socioeconomic indicators. To manage the high dimensionality of contact and socioeconomic data, we applied one and two-dimensional principal component analysis (PCA), including the $(2D)^2$ PCA method for contact matrix reduction. Hierarchical clustering revealed meaningful groupings of countries

that reflect not only similarities in age-specific mixing behaviors but also underlying structural characteristics. In the European context, countries were grouped based on similarities in how different age groups interact. Although some regional trends were visible, for instance, many Eastern and Central European countries appeared in the same group, the clusters were not strictly based on geography. Countries from different parts of Europe, including the North, South, East, and West, were sometimes found in the same group, suggesting that social behavior patterns can cut across regional lines. For African countries, the integration of socioeconomic data substantially altered the clustering outcomes, highlighting the importance of variables such as electricity access, health expenditure, and employment in shaping public health vulnerabilities. The resulting clusters offer an empirical basis for cross-national learning, where countries facing similar structural and behavioral risks can adopt and adapt effective non-pharmaceutical interventions from one another.

The sensitivity and clustering frameworks introduced in this thesis provide complementary approaches to epidemic modeling. Sensitivity analysis offers detailed insights within a specific population while clustering analysis provides a broader perspective that supports policy transferability between regions. Both methods are highly flexible and can be extended to address additional epidemic outcomes, such as hospitalizations, fatalities, or overall healthcare burden. The techniques developed are modular and broadly applicable: the eigenvector-based framework works with any structured compartmental model, the LHS-PRCC method is suitable for high-dimensional epidemiological models, and the clustering approach is adaptable to new data sources or disease-specific factors. Together, these tools offer a robust foundation for enhancing both the precision and applicability of epidemic modeling in diverse populations.

Looking ahead, several promising directions for future research are evident. One key

opportunity lies in incorporating behavioral and temporal dynamics, such as evolving mobility patterns and time-dependent contact matrices, into sensitivity analyses. Accounting for uncertainties in contact data or parameter estimates could also enhance the robustness of these analyses by generating confidence intervals for sensitivity measures, thereby strengthening their utility for policy decisions. On the clustering side, exploring alternative distance metrics or leveraging more advanced unsupervised learning techniques beyond hierarchical clustering could uncover hidden structures that are not apparent using standard Euclidean measures.

In conclusion, this thesis contributes both methodological innovations and practical perspectives to the field of structured epidemic modeling. The analysis emphasizes the importance of contact patterns and socioeconomic conditions in shaping disease transmission dynamics, providing a foundation for the development of more accurate, flexible, and data-informed public health strategies. Tools of this kind play a vital role in a globally connected world where emerging pathogens pose ongoing threats. They support not only the management of current outbreaks but also the long-term goal of building resilient health systems capable of anticipating and responding to future epidemics.

Publications The dissertation is based on the following five scientific papers:

- Korir, Evans Kiptoo, and Zsolt Vizi. "Clustering of countries based on the associated social contact patterns in epidemiological modeling." International Symposium on Mathematical and Computational Biology . Cham: Springer Nature Switzerland, 2022, [30].

- Korir, Evans Kiptoo, and Zsolt Vizi. "Clusters of African countries based on the

social contacts and associated socioeconomic indicators relevant to the spread of the epidemic." Journal of Mathematics in Industry 14.1 (2024): 24, [24].

- Korir, Evans Kiptoo. Comparative clustering and visualization of socioeconomic and health indicators: A case of kenya. Socio-Economic Planning Sciences, 95:101961, 2024, [29].

- Korir, E. K., and Vizi, Z. (2025). Eigenvector-Based Sensitivity Analysis of Contact Patterns in Epidemic Modeling. ArXiv. https://arxiv.org/abs/2502.20117[25].

- Vizi, Z., Korir, E. K., Bogya, N., Rosztóczy, C., Makay, G., and Boldog, P. (2025). Age Group Sensitivity Analysis of Epidemic Models: Investigating the Impact of Contact Matrix Structure. ArXiv. https://arxiv.org/abs/2502.19206 [56].

Data availability All the code and data required to reproduce the dissertation are based on the following four GitHub repositories:

- Code and data repository for reproducing Chapter 5, https://github.com/Evanskorir/adjoint-contact-sensitivity

- Code and data repository for reproducing Chapter 6, https://github.com/zsvizi/sensitivity-contact-epidemics

- Code and data repositories for reproducing Chapter 7;

  1. https://github.com/zsvizi/clustering-social-patterns-epidemic

  2. https://github.com/Evanskorir/African-social-contact-patterns

Chapter 9

Összefoglalás

A matematikai modellek régóta központi szerepet játszanak a fertőző betegségek terjedésének megértésében, különösen a terjedést előidéző mechanizmusok feltárásában és a hatékony beavatkozási stratégiák kidolgozásában. Bár a klasszikus, determinisztikus differenciálegyenleteken alapuló kompartmentális modellek fontos eszközök a járványdinamika működésébe, a közelmúlt kihívásai, különösen a COVID–19 világjárvány, rávilágítottak arra, hogy elengedhetetlen a társadalmi heterogeneitás figyelembevétele ezen modellekben. E felismerés nyomán dolgozatunk második része azt vizsgálta, miként befolyásolják a korcsoport-specifikus kacsolati mintázatok, a fertőzékenységi profilok és a társadalmi-gazdasági tényezők a járványok kimenetelét.

A 5. fejezetben egy saját fejlesztésű, sajátvektor-alapú érzékenységvizsgálati módszert vezettünk be, amely a kontaktusok számának változásainak marginális hatását méri az alap reprodukciós számra ($\mathcal{R}_0$) nézve. A megközelítés a következő generációs mátrix spektrális tulajdonságain alapul, és számítási szempontból hatékony, jól értelmezhető elemzést tesz lehetővé az egyes korcsoport-párok közötti kontaktusok járványterjedésre gyakorolt hatásának meghatározására. A módszer előnye, hogy nem igényli nagyszámú

szimuláció futtatását, mégis képes azonosítani a legkritikusabb kontaktuspárokat, illetve korcsoportonkénti, ún. aggregált érzékenységi mutatót is szolgáltat. A módszert különböző járványmodelleken – többek között egy magyarországi COVID–19 és egy brit influenzamodellen – alkalmazva azt találtuk, hogy a 30–44 éves középkorú felnőttek jellemzően meghatározó szerepet játszanak a terjedésben, kiterjedt saját és más korcsoportokkal való kapcsolatuk miatt.

A 6. fejezet ezt a lokális elemzést egy statisztikai, szimuláción alapuló keretrendszerrel egészíti ki, amely a Latin Hiperkocka Mintavételezés (LHS) és a Részleges Rangkorrelációs Együttható (PRCC) módszerét kombinálja. Ez a megközelítés figyelembe veszi a bemenetként megadott paraméterek bizonytalanságát, és lehetővé teszi a paramétertér globális feltérképezését, miközben képes kezelni nemlineáris és monoton kapcsolatok elemzését. Új eljárást javasoltunk, amely p-értékek alapján súlyozva számszerűsíti az egyes korcsoportok teljes hatását. Influenza- és COVID–19 modellekre alkalmazva ismét azt tapasztaltuk, hogy a munkaképes korú felnőttek kontaktusai meghatározó szerepet töltenek be, különösen a közepes és súlyos járványforgatókönyvek esetén. Érdekesség, hogy az enyhébb esetekben a serdülők (15–19 év) szignifikáns szerepet játszottak a kontaktussűrűségük révén, ám súlyosabb helyzetekben jelentőségük visszaszorult, rámutatva az intervenciók hatékonyságának kontextusfüggő természetére.

A 7. fejezetben a populáción belüli elemzésekről országok közötti összehasonlításokra tértünk át. Olyan klaszterezési eljárást dolgoztunk ki, amely társadalmi kontaktusmátrixok, valamint – ahol elérhetők voltak – társadalmi-gazdasági mutatók alapján csoportosította az országokat. A kontaktusmintázatok és szocioökonómiai adatok magas dimenzionalitásának kezelésére egy- és kétdimenziós főkomponens-analízist (PCA), valamint a $(2D)^2$ PCA módszert alkalmaztuk. A hierarchikus klaszterezés

interpretálható csoportosítást eredményezett az országok között, amelyek nemcsak a hasonló életkor-specifikus kontaktusviselkedéseket tükrözték, hanem strukturális jellemzőkre is utaltak. Az európai országok klaszterezésekor bár regionális mintázatok is megfigyelhetők voltak – például Kelet- és Közép-Európa országai gyakran egy csoportba kerültek –, a klaszterek nem követték szigorúan a földrajzi határokat. Különböző régiókból (Észak, Dél, Kelet, Nyugat) származó országok is azonos csoportba kerülhettek, ami arra utal, hogy a társadalmi viselkedésminták gyakran függetlenek a régióktól.

Afrikai országok esetében a szocioökonómiai adatok integrálása jelentősen megváltoztatta a klaszterstruktúrát, különösen az olyan változók szerepe révén, mint az elektromos áramhoz való hozzáférés, az egészségügyi kiadások és a foglalkoztatottság. Az így kapott csoportosítások empirikus alapot nyújtanak a nemzetközi szintű monitorozáshoz: azok az országok, amelyek hasonló strukturális és viselkedési kockázatokkal szembesülnek, egymástól adaptálhatnak hatékony, nem gyógyszeres beavatkozási stratégiákat.

Az ebben a dolgozatban bemutatott érzékenységi és klaszterezési keretrendszerek egyfajta kiegészítő eszközként szolgálnak a járványmodellezésben. Az érzékenységi elemzés részletes betekintést nyújt egy adott populáció sajátosságaiba, míg a klaszterezés tágabb perspektívát biztosít, támogatva a politikai intézkedések térbeli átvihetőségét. Mindkét megközelítés rugalmasan bővíthető további járványkimenetelek – például kórházi kezelések, halálozások vagy az egészségügyi rendszer terhelése – figyelembevételével. A kidolgozott módszerek modulárisak és széles körben alkalmazhatók: a sajátvektor-alapú keretrendszer bármely strukturált modellre alkalmazható, az LHS-PRCC módszer alkalmas magas dimenziójú modellekre, míg a klaszterezési technika könnyen adaptálható új adathalmazokhoz vagy betegség-specifikus jellemzőkhöz.

A jövőre nézve több ígéretes kutatási irány is megfogalmazható. Fontos lehet például a viselkedési és időbeli dinamikák – például változó mobilitási minták vagy időfüggő kontaktusmátrixok – integrálása az érzékenységi elemzésekbe. A kontaktusadatok és paraméterbecslések bizonytalanságainak kezelése lehetővé tenné érzékenységi mértékek konfidenciaintervallumokkal való ellátását, ami tovább növelné ezek szakpolitikai hasznosíthatóságát. A klaszterezési oldalon alternatív távolságmértékek alkalmazása vagy fejlettebb nem felügyelt tanulási technikák bevonása – túlmutatva a hierarchikus klaszterezésen – új összefüggések feltárását teheti lehetővé.

Összességében a dolgozat módszertani újításokat és gyakorlati szempontokat egyaránt kínál a strukturált járványmodellezés területén. Az elemzés hangsúlyozza a kontaktusmintázatok és társadalmi-gazdasági feltételek szerepét a fertőzések terjedésének alakításában, hozzájárulva a pontosabb, rugalmasabb és adatvezérelt közegészségügyi stratégiák kialakításához. Az ilyen típusú eszközök különösen fontosak a globálisan összekapcsolt világban, ahol az új kórokozók állandó fenyegetést jelentenek. Ezek nemcsak a jelenlegi járványok kezelését segítik, hanem a hosszú távú célhoz is hozzájárulnak: egy olyan ellenálló egészségügyi rendszer kialakításához, amely képes előrejelezni és kezelni a jövőbeli járványokat is.

A disszertáció az alábbi öt tudományos publikáción alapul:

- Korir, Evans Kiptoo, and Zsolt Vizi. "Clustering of countries based on the associated social contact patterns in epidemiological modeling." International Symposium on Mathematical and Computational Biology . Cham: Springer Nature Switzerland, 2022, [30].

- Korir, Evans Kiptoo, and Zsolt Vizi. "Clusters of African countries based on the

social contacts and associated socioeconomic indicators relevant to the spread of the epidemic." Journal of Mathematics in Industry 14.1 (2024): 24, [24].

- Korir, Evans Kiptoo. Comparative clustering and visualization of socioeconomic and health indicators: A case of kenya. Socio-Economic Planning Sciences, 95:101961, 2024, [29].

- Korir, E. K., and Vizi, Z. (2025). Eigenvector-Based Sensitivity Analysis of Contact Patterns in Epidemic Modeling. ArXiv. https://arxiv.org/abs/2502.20117[25].

- Vizi, Z., Korir, E. K., Bogya, N., Rosztóczy, C., Makay, G., and Boldog, P. (2025). Age Group Sensitivity Analysis of Epidemic Models: Investigating the Impact of Contact Matrix Structure. ArXiv. https://arxiv.org/abs/2502.19206 [56].

Chapter A

Appendix

A.1 Pitman et al. Model

The transmission dynamics in the model proposed by Pitman et al. [44], as depicted in Figure 3.1, follow a standard age-structured SEIR framework. The corresponding system of ordinary differential equations (ODEs) is given by:

$$\begin{aligned}
S_i'(t) &= -\beta(m = \mathrm{P}) \cdot \frac{S_i(t)}{P_i(m = \mathrm{P})} \sum_{j=1}^{n_a} c_{i,j}(m = \mathrm{P}) \cdot I_j(t), \\
E_i'(t) &= \beta(m = \mathrm{P}) \cdot \frac{S_i(t)}{P_i(m = \mathrm{P})} \sum_{j=1}^{n_a} c_{i,j}(m = \mathrm{P}) \cdot I_j(t) - \alpha E_i(t), \\
I_i'(t) &= \alpha E_i(t) - \gamma I_i(t), \\
R_i'(t) &= \gamma I_i(t).
\end{aligned} \tag{A.1}$$

The model incorporates age-specific contact patterns through the matrix $c_{i,j}(m = P)$, where contacts are based on pre-pandemic estimates. Table A.1 summarizes the parameters used in the model.

Parameter	Description	Value
n_a	Number of age groups	15
σ_i	Susceptibility parameter	1.0
α	Rate of latent individuals becoming infectious	0.5
γ	Recovery rate	0.5
$\overline{\mathcal{R}}_0(m = \mathrm{P})$	Baseline reproduction number	[1.2, 1.8, 2.5]

Table A.1: Summary of model parameters for the Pitman et al. model. The contact matrices are based on UK population data and social contact patterns reported in Mossong et al. [40].

A.2 Röst et al. Model

The dynamics of the system, corresponding to the transmission structure illustrated in Figure 3.2, are described by the following set of ordinary differential equations:

$$
\begin{aligned}
S_i'(t) &= -\beta(m=\mathrm{R}) \cdot \frac{S_i(t)}{P_i(m=\mathrm{R})} \cdot \\
&\qquad \sigma_i \sum_{j=1}^{n_a} c_{i,j}(m=\mathrm{R}) \left[I_j^{(p)}(t) + \inf{}^{(a)} \sum_{m=1}^{3} I_j^{(a,m)}(t) + \sum_{m=1}^{3} I_j^{(s,m)}(t) \right], \\
{L^{(1)}}_i'(t) &= \beta(m=\mathrm{R}) \cdot \frac{S_i(t)}{P_i(m=\mathrm{R})} \cdot \\
&\qquad \sigma_i \sum_{j=1}^{n_a} c_{i,j}(m=\mathrm{R}) \left[I_j^{(p)}(t) + \inf{}^{(a)} \sum_{m=1}^{3} I_j^{(a,m)}(t) + \sum_{m=1}^{3} I_j^{(s,m)}(t) \right] - 2\alpha^{(L)} L_i^{(1)}(t), \\
{L^{(2)}}_i'(t) &= 2\alpha^{(L)} L_i^{(1)}(t) - 2\alpha^{(L)} L_i^{(2)}(t), \\
{I^{(p)}}_i'(t) &= 2\alpha^{(L)} L_i^{(2)}(t) - \alpha^{(p)} I_i^{(p)}(t), \\
{I^{(a,1)}}_i'(t) &= p_i \alpha^{(p)} I_i^{(p)}(t) - 3\gamma^{(a)} I_i^{(a,1)}(t), \\
{I^{(a,2)}}_i'(t) &= 3\gamma^{(a)} I_i^{(a,1)}(t) - 3\gamma^{(a)} I_i^{(a,2)}(t), \\
{I^{(a,3)}}_i'(t) &= 3\gamma^{(a)} I_i^{(a,2)}(t) - 3\gamma^{(a)} I_i^{(a,3)}(t),
\end{aligned}
\tag{A.2}
$$

$$
\begin{aligned}
{I^{(s,1)}}'_i(t) &= (1-p_i)\alpha^{(p)} I_i^{(p)}(t) - 3\gamma^{(s)} I_i^{(s,1)}(t),\\
{I^{(s,2)}}'_i(t) &= 3\gamma^{(s)} I_i^{(s,1)}(t) - 3\gamma^{(s)} I_i^{(s,2)}(t),\\
{I^{(s,3)}}'_i(t) &= 3\gamma^{(s)} I_i^{(s,2)}(t) - 3\gamma^{(s)} I_i^{(s,3)}(t),\\
{I^{(h)}}'_i(t) &= h_i(1-\xi_i)\cdot 3\gamma^{(s)} I_i^{(s,3)}(t) - \gamma^{(h)} I_i^{(h)}(t),\\
{I^{(c)}}'_i(t) &= h_i\xi_i \cdot 3\gamma^{(s)} I_i^{(s,3)}(t) - \gamma^{(c)} I_i^{(c)}(t),\\
{I^{(\mathrm{cr})}}'_i(t) &= (1-\mu_i)\gamma^{(c)} I_i^{(c)}(t) - \gamma^{(\mathrm{cr})} I_i^{(\mathrm{cr})}(t),\\
R'_i(t) &= 3\gamma^{(a)} I_i^{(a,3)}(t) + (1-h_i)3\gamma^{(s)} I_i^{(s,3)}(t) + \gamma^{(h)} I_i^{(h)}(t) + \gamma^{(\mathrm{cr})} I_i^{(\mathrm{cr})}(t),\\
D'_i(t) &= \mu_i\gamma^{(c)} I_i^{(c)}(t).
\end{aligned}
$$

Age Group	Asymptomatic Probability (p_i)	ICU Given Hospitalization (ξ_i)	Fatal Outcome Probability (μ_i)	Hospitalization Rate (h_i)
0-4	0.95	0.333	0.2	0.0003
5-9	0.8	0.333	0.2	0.0003
10-14	0.8	0.333	0.2	0.0003
15-19	0.7	0.333	0.216	0.0003
20-24	0.7	0.282	0.216	0.0039
25-29	0.7	0.282	0.216	0.0039
30-34	0.5	0.297	0.3	0.0145
35-39	0.5	0.297	0.3	0.0145
40-44	0.5	0.294	0.3	0.0255
45-49	0.5	0.294	0.3	0.0255
50-54	0.5	0.293	0.3	0.0495
55-59	0.5	0.293	0.3	0.0495
60-64	0.4	0.293	0.582	0.0775
65-69	0.4	0.293	0.582	0.0775
70-74	0.3	0.294	0.678	0.1788
75+	0.2	0.294	0.687	0.3297
Susceptibility (σ_i)		[0.5, 1.0]		
Baseline Reproduction Number ($\overline{\mathcal{R}}_0(m = \mathrm{R})$)		[1.2, 1.8, 2.5]		

Table A.2: Age-dependent model parameters used for simulating the spread of COVID-19 in Hungary based on the model described in Eq. (A.2). The equations are indexed by $i \in \{1, \ldots, 16\}$, where each i corresponds to a specific age group.

Bibliography

[1] Steven Abrams, James Wambua, Eva Santermans, Lander Willem, Elise Kuylen, Pietro Coletti, Pieter Libin, Christel Faes, Oana Petrof, Sereina A Herzog, et al. Modelling the early phase of the belgian covid-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics*, 35:100449, 2021.

[2] P Adu, Mawuena Binka, Bushra Mahmood, Dahn Jeong, T Buller-Tylor, M Jean Damascene, Sarafa Iyaniwura, Notice Ringa, Hector Velasquez, Stanley Wong, et al. Quantifying contact patterns: development and characteristics of the british columbia covid-19 population mixing patterns survey. *International Journal of Infectious Diseases*, 116:S30–S31, 2022.

[3] Marco Ajelli and Maria Litvinova. Estimating contact patterns relevant to the spread of infectious diseases in russia. *Journal of theoretical biology*, 419:1–7, 2017.

[4] Leonardo Angeli, Constantino Pereira Caetano, Nicolas Franco, Steven Abrams, Pietro Coletti, Inneke Van Nieuwenhuyse, Sorin Pop, and Niel Hens. Who acquires infection from whom? a sensitivity analysis of transmission dynamics during the early phase of the covid-19 pandemic in belgium. *Journal of Theoretical Biology*, 581:111721, 2024.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[6] Sally M Blower and Hadi Dowlatabadi. Sensitivity and uncertainty analysis of complex models of disease transmission: an hiv model, as an example. *International Statistical Review/Revue Internationale de Statistique*, pages 229–243, 1994.

[7] Dimitri Breda, Odo Diekmann, Wilfred F de Graaf, Andrea Pugliese, and Rossana Vermiglio.

On the formulation of epidemic models (an appraisal of kermack and mckendrick). *Journal of biological dynamics*, 6(sup2):103–117, 2012.

[8] Jamaica Briones, Yi Wang, Juthamas Prawjaeng, Hwee Lin Wee, Angela Kairu, Stacey Orangi, Edwine Barasa, and Yot Teerawattananon. A data-driven analysis of the economic cost of non-pharmaceutical interventions: A cross-country comparison of kenya, singapore, and thailand. *International journal of public health*, 67:1604854, 2022.

[9] Rodrigo M Carrillo-Larco and Manuel Castillo-Cara. Using country-level variables to classify countries according to the number of confirmed covid-19 cases: An unsupervised machine learning approach. *Wellcome open research*, 5:56, 2020.

[10] Asli Demirgüç-Kunt, Michael Lokshin, and Iván Torre. The sooner, the better: The economic impact of non-pharmaceutical interventions during the early stage of the covid-19 pandemic. *Economics of Transition and Institutional Change*, 29(4):551–573, 2021.

[11] Odo Diekmann, Johan Andre Peter Heesterbeek, and Michael G Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the royal society interface*, 7(47):873–885, 2010.

[12] Boloye Gomero. Latin hypercube sampling and partial rank correlation coefficient analysis applied to an optimal control problem. Master of science thesis, University of Tennessee, August 2012. Available at: https://trace.tennessee.edu/utk_gradthes/1278/.

[13] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[14] Carlos G Grijalva, Nele Goeyvaerts, Hector Verastegui, Kathryn M Edwards, Ana I Gil, Claudio F Lanata, Niel Hens, and RESPIRA PERU project. A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of peru. *PloS one*, 10(3):e0118457, 2015.

[15] David M Hamby. A review of techniques for parameter sensitivity analysis of environmental models. *Environmental monitoring and assessment*, 32:135–154, 1994.

[16] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[17] Jon Craig Helton and Freddie J Davis. Sampling-based methods for uncertainty and sensitivity analysis. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia ..., 2000.

[18] Peter Horby, Pham Quang Thai, Niel Hens, Nguyen Thi Thu Yen, Le Quynh Mai, Dang Dinh Thoang, Nguyen Manh Linh, Nguyen Thu Huong, Neal Alexander, W John Edmunds, et al. Social contact patterns in vietnam and implications for the control of infectious diseases. *PloS one*, 6(2):e16965, 2011.

[19] Sarafa A Iyaniwura, Musa Rabiu, Jummy F David, and Jude D Kong. The basic reproduction number of covid-19 across africa. *Plos one*, 17(2):e0264455, 2022.

[20] Steven G Johnson. Notes on adjoint methods for 18.335. *Introduction to Numerical Methods*, 2012.

[21] Gergely Röst Júlia Koltai, Orsolya Vásárhelyi. Reconstructing social mixing patterns via weighted contact matrices from online and representative surveys. *Scientific reports 12.1*, 2022.

[22] Nathan Keyfitz, Hal Caswell, et al. *Applied mathematical demography*, volume 47. Springer, 2005.

[23] Ferath Kherif and Adeliya Latypova. Principal component analysis. In *Machine learning*, pages 209–225. Elsevier, 2020.

[24] Evans Kiptoo and Zsolt Vizi. Clusters of african countries based on the social contacts and associated socioeconomic indicators relevant to the spread of the epidemic. *Journal of Mathematics in Industry 14.1*, 2024.

[25] Evans Kiptoo Korir and Zsolt Vizi. Eigenvector-based sensitivity analysis of contact patterns in epidemic modeling. *arXiv e-prints*, pages arXiv–2502, 2025.

[26] Moses Chapa Kiti, Timothy Muiruri Kinyanjui, Dorothy Chelagat Koech, Patrick Kiio Munywoki, Graham Francis Medley, and David James Nokes. Quantifying age-related rates of social contact using diaries in a rural coastal population of kenya. *PloS one*, 9(8):e104786, 2014.

[27] Petra Klepac, Adam J Kucharski, Andrew JK Conlan, Stephen Kissler, Maria L Tang, Hannah Fry, and Julia R Gog. Contacts in context: large-scale setting-specific social mixing matrices from the bbc pandemic project. *MedRxiv*, pages 2020–02, 2020.

[28] Diána Knipl and Gergely Röst. Modelling the strategies for age specific vaccination scheduling during influenza pandemic outbreaks. *Mathematical Biosciences and Engineering*, 8(1):123–139, January 2011. Also available as arXiv:0912.4662 [q-bio.PE].

[29] Evans Kiptoo Korir. Comparative clustering and visualization of socioeconomic and health indicators: A case of kenya. *Socio-Economic Planning Sciences*, 95:101961, 2024.

[30] Evans Kiptoo Korir and Zsolt Vizi. Clustering of countries based on the associated social contact patterns in epidemiological modelling. In *International Symposium on Mathematical and Computational Biology*, pages 253–271. Springer, 2022.

[31] Supriya Kumar, Mudita Gosain, Hanspria Sharma, Eric Swetts, Ritvik Amarchand, Rakesh Kumar, Kathryn E Lafond, Fatimah S Dawood, Seema Jain, Marc-Alain Widdowson, et al. Who interacts with whom? social mixing insights from a rural population in india. *Plos one*, 13(12):e0209039, 2018.

[32] O Le Polain de Waroux, Sandra Cohuet, Donny Ndazima, AJ Kucharski, Aitana Juan-Giner, Stefan Flasche, Elioda Tumwesigye, Rinah Arinaitwe, Juliet Mwanga-Amumpaire, Yap Boum, et al. Characteristics of human encounters and social mixing patterns relevant to infectious diseases spread by close contact: a survey in southwest uganda. *BMC infectious diseases*, 18:1–12, 2018.

[33] William Maloney and Temel Taskin. Determinants of social distancing and economic activity during covid-19: A global view. *Covid Economics*, 13:157–177, 2020.

[34] Simeone Marino, Ian B Hogue, Christian J Ray, and Denise E Kirschner. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of theoretical biology*, 254(1):178–196, 2008.

[35] Zachary McCarthy, Yanyu Xiao, Francesca Scarabel, Biao Tang, Nicola Luigi Bragazzi, Kyeongah Nah, Jane M Heffernan, Ali Asgary, V Kumar Murty, Nicholas H Ogden, et al. Quantifying the shift in social contact patterns in response to non-pharmaceutical interventions. *Journal of Mathematics in Industry*, 10:1–25, 2020.

[36] Michael D McKay. Latin hypercube sampling as a tool in uncertainty analysis of computer models. In *Proceedings of the 24th conference on Winter simulation*, pages 557–564, 1992.

[37] Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

[38] Alessia Melegaro, Emanuele Del Fava, Piero Poletti, Stefano Merler, Constance Nyamukapa, John Williams, Simon Gregson, and Piero Manfredi. Social contact structures and time use patterns in the manicaland province of zimbabwe. *PloS one*, 12(1):e0170459, 2017.

[39] C. E. Mongi, Y. A. R. Langi, C. E. J. C. Montolalu, and N. Nainggolan. Comparison of hierarchical clustering methods (case study: Data on poverty influence in north sulawesi). In *IOP Conference Series: Materials Science and Engineering*, volume 567, page 012048. IOP Publishing, 2019.

[40] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74, 2008.

[41] Charles Nicholson, Lex Beattie, Matthew Beattie, Talayeh Razzaghi, and Sixia Chen. A

machine learning and clustering-based approach for county-level covid-19 analysis. *Plos one*, 17(4):e0267558, 2022.

[42] Elaine O Nsoesie, Richard J Beckman, and Madhav V Marathe. Sensitivity analysis of an individual-based model for simulation of influenza epidemics. *PloS one*, 7(10):e45414, 2012.

[43] Anne Osterrieder, Giulia Cuman, Wirichada Pan-Ngum, Phaik Kin Cheah, Phee-Kheng Cheah, Pimnara Peerawaranun, Margherita Silan, Miha Orazem, Ksenija Perkovic, Urh Groselj, et al. Economic and social impacts of covid-19 and public health measures: results from an anonymous online survey in thailand, malaysia, the uk, italy and slovenia. *BMJ open*, 11(7):e046863, 2021.

[44] RJ Pitman, LJ White, and M Sculpher. Estimating the clinical impact of introducing paediatric influenza vaccination in england and wales. *Vaccine*, 30(6):1208–1224, 2012.

[45] Kiesha Prem, Alex R Cook, and Mark Jit. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS computational biology*, 13(9):e1005697, 2017.

[46] Kiesha Prem, Kevin van Zandvoort, Petra Klepac, Rosalind M Eggo, Nicholas G Davies, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Alex R Cook, and Mark Jit. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the covid-19 era. *PLoS computational biology*, 17(7):e1009098, 2021.

[47] Jonathan M Read, Justin Lessler, Steven Riley, Shuying Wang, Li Jiu Tan, Kin On Kwok, Yi Guan, Chao Qiang Jiang, and Derek AT Cummings. Social mixing patterns in rural and urban areas of southern china. *Proceedings of the Royal Society B: Biological Sciences*, 281(1785):20140268, 2014.

[48] Chandan K Reddy. *Data clustering: algorithms and applications*. Chapman and Hall/CRC, 2018.

[49] Syeda Amna Rizvi, Muhammad Umair, and Muhammad Aamir Cheema. Clustering of countries for covid-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*, 151:111240, 2021.

[50] Gergely Röst, Ferenc A Bartha, Norbert Bogya, Péter Boldog, Attila Dénes, Tamás Ferenci, Krisztina J Horváth, Attila Juhász, Csilla Nagy, Tamás Tekeli, et al. Early phase of the covid-19 outbreak in hungary and post-lockdown scenarios. *Viruses*, 12(7):708, 2020.

[51] Banafsheh Sadeghi, Rex CY Cheung, and Meagan Hanbury. Using hierarchical clustering analysis to evaluate covid-19 pandemic preparedness and performance in 180 countries in 2020. *BMJ open*, 11(11):e049844, 2021.

[52] Nema Salem and Sahar Hussein. Data dimensional reduction and principal components analysis. *Procedia Computer Science*, 163:292–299, 2019.

[53] Agustí Segarra-Blasco, Mercedes Teruel, and Sebastiano Cattaruzzo. The economic reaction to non-pharmaceutical interventions during covid-19. *Economic Analysis and Policy*, 72:592–608, 2021.

[54] Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

[55] Sudeep Tanwar, Tilak Ramani, and Sudhanshu Tyagi. Dimensionality reduction using pca and svd in big data: A comparative case study. In *Future Internet Technologies and Trends: First International Conference, ICFITT 2017, Surat, India, August 31-September 2, 2017, Proceedings 1*, pages 116–125. Springer, 2018.

[56] Zsolt Vizi, Evans Kiptoo Korir, Norbert Bogya, Csaba Rosztóczy, Géza Makay, and Péter Boldog. Age group sensitivity analysis of epidemic models: Investigating the impact of contact matrix structure. *arXiv preprint arXiv:2502.19206*, 2025.

[57] Jacco Wallinga, Peter Teunis, and Mirjam Kretzschmar. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American journal of epidemiology*, 164(10):936–944, 2006.

[58] Dong Wang, Haipeng Shen, and Young Truong. Efficient dimension reduction for high-dimensional matrix-valued data. *Neurocomputing*, 190:25–34, 2016.

[59] Helen J Wearing, Pejman Rohani, and Matt J Keeling. Appropriate models for the management of infectious diseases. *PLoS medicine*, 2(7):e174, 2005.

[60] World Bank. World bank country data. https://data.worldbank.org/country. Accessed: 2020-02-29.

[61] Daoqiang Zhang and Zhi-Hua Zhou. (2d) 2pca: Two-directional two-dimensional pca for efficient face representation and recognition. *Neurocomputing*, 69(1-3):224–231, 2005.