A database on the socioeconomic and behavioural impacts in Sri Lanka through multiple waves of the COVID-19

Gayanthi A. Ilangarathna[1,2], Lakshitha Ramanayake[1,*], Neranjan Senarath[1], Yasiru Ranasinghe[1,3], Harshana Weligampola[1,4], Wathsala Dedunupitiya[1], Thanuja Wijesiri[1], Pabodha Rathnaweera[1], Roshan Godaliyadda[1], Parakrama Ekanayake[1], Vijitha Herath[1], Janaka Ekanayake[1], Sakunthala Yatigammana[5], Anuruddhika Rathnayake[6], Mallika Pinnawala[5], Muthucumaru Maheswaran[7], Ganga Thilakaratne[8], and Samath Dharmarathne[6]

[1]Faculty of Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka
[2]Department of Geography, Virginia Polytechnic Institute and State University, Blacksburg,24060, United States
[3]Whiting School of Engineering, Johns Hopkins University, Baltimore, 21218, United States
[4]Department of Electrical and Computer Engineering, Purdue University,West Lafayette, 47907, United States
[5]Faculty of Arts, University of Peradeniya, Peradeniya 20400, Sri Lanka
[6]Faculty of Medicine, University of Peradeniya, Peradeniya 20400, Sri Lanka
[7]Department of Electrical and Computer Engineering, McGill University, 845 Sherbrooke Street West, Montréal (Québec), Canada
[8]Institute of Policy Studies, Colombo 00700, Sri Lanka
[*]corresponding author: Lakshitha Ramanayake (lakshitharamanayake@eng.pdn.ac.lk)

ABSTRACT

The impact of the COVID-19 pandemic was diverse and disproportionate among nations, and population segments. The impacts of the disease and the containment strategies adopted are broad and cut across multiple facets of life, society, and the economy, which are intimately interlinked. Therefore, a large household survey was conducted to ascertain the socioeconomic impact and human behavior changes due to the pandemic and the containment strategies covering all provinces of Sri Lanka. The ramifications on mobility and human behavior, income, economic status, food consumption, education, access to health services and information, and cultural and psychological changes were explored, and the uncovered data are reported in this paper. The survey was conducted on 3020 households, selected using a multistage clustering technique, to access the impacts of the pandemic through three distinctly identified waves/phases of the pandemic in Sri Lanka. This dataset will enable researchers and policymakers to analyze the impact of the pandemic through a multifaceted perspective enabling a more holistic approach to decision-making.

Background & Summary

As of 23rd January 2023, more than 663 million cases and 6.7 million deaths were reported globally as a consequence of the COVID-19 pandemic (https://covid19.who.int/). In addition to affecting people's physical health, the pandemic also had an impact on each person's social, economic, and behavioral circumstances. Even though the impact of the pandemic on the populace's health is tangible, it is challenging to quantify the short-term and long-term effects of the social, economic, and behavioral impacts on people due to the disease itself as well as the containment strategies adopted. Furthermore, while the health impact of the pandemic is frequently monitored, the socioeconomic and behavioral impacts have not been monitored as exhaustively and frequently.

With this insight, a nationwide household survey was conducted to capture the socioeconomic and behavioral impact of the COVID- 19 in Sri Lanka. The survey was conducted as a face-to-face one-time survey, which started on 6th November 2021 and lasted till 10th December 2021. It spanned through 20 districts covering a total number of 3020 households. The impact on the aforementioned main areas was assessed by breaking down them into the categories mentioned below.

- Socioeconomic Impact
  - Education

  - Income and economic level
  - Access to health services and information

- Behavioral Impact
  - mobility patterns
  - human interactions
  - food consumption
  - religious and cultural
  - psychological

The importance of this kind of a data-set which expedites the impact of COVID-19 on such a wide array of aspects as mentioned above will enable researchers to explore and analyze the inter-connectivity, how the pandemic and the mitigation measures adopted by the governing bodies affected the socioeconomic state and behavior of the population. Furthermore, this opens up a gateway to explore the correlations and dependencies between the socioeconomic state and the behavior of the citizens in the country. Most importantly, this serves as a source to identify the groups of people who are disproportionately affected by the pandemic in underdeveloped nations.

Sri Lanka is an island nation located in South Asia to the south of India, reputed for its multi-ethnic and multi-religious population in addition to its natural and geographical diversity. As per the world bank data 2021 (https://databank.worldbank.org/data/402 download/POP.pdf), with a population of more than 21 million, Sri Lanka is ranked 57th in world population amongst 216 countries and autonomous regions, while 8th in population amongst island nations. The country has densely populated urban population hot spots, while more than 75 percent of it is rural areas that are largely dependent on agriculture. These features of the country provide the pathway to conduct a unique case study based on the adverse impact on socioeconomic factors as a result of the pandemic which can be extrapolated and applied to other countries in the world.

Pandemic In Sri Lanka

The first confirmed case of COVID-19 in Sri Lanka was reported on 27th January 2020. After a period of silence, the second confirmed case (the first local case) was identified on 11th March 2020. Since then, many cases were reported across the country. According to the epidemiology unit of Sri Lanka, as of 23rd January 2023 after three consecutive waves of the pandemic, Sri Lanka reports 671,987 confirmed cases with a death toll of 16,828. During the pandemic, the Sri Lankan government imposed different measures like social distancing, travel restrictions, isolation, and partial and complete lockdowns similar to other countries, to mitigate the spread of the disease. A detailed description of the distinct pandemic waves and the actions taken by the regulatory bodies to mitigate the disease spread during those periods are provided in the following sections.

The timeline of the three waves based on the data from the Epidemiology Unit (https://www.epid.gov.lk/), Ministry of Health, Sri Lanka is illustrated in Figure 1.

First Wave

Though the first wave was recorded to have been initiated in January 2020 the escalation of the number of reported cases started after finding the second patient in March 2020. As the outbreak was reported, the government imposed strict containment measures such as lockdown along with large-scale testing and thorough contact tracing[1]. The lockdown in this phase took a shape similar to a curfew where violations of guidelines and regulations were considered criminal offenses. The high stringency index reported[2] during this time represents the strict containment measures imposed by the government. In addition, state media launched many awareness campaigns which highlighted the severity of COVID-19 using numerical and graphical sources from different countries across the world to create an atmosphere of fear towards the disease. These unique features such as the severe lockdown, the legal repercussions associated, and the high level of awareness as which are listed in Figure 1 were used by the enumerators to trigger the lived experience of the respondents. During this time the citizens had to get adapted to new concepts such as working from home, online learning, online shopping, etc. Furthermore, the government declared the concept of essential services motivating the vendors to approach the consumers and enabling door-to-door sales. The first wave and the containment strategies adopted by the government affected the socioeconomic, behavioral, and educational state of the people throughout the country. The identification of the first wave by the public is the wave where effective elimination was done through highly stringent containment and regulations. A total number of 3396 cases was reported with 13 deaths during this period (https://www.epid.gov.lk/web/images/pdf/corona_monthly_summery/esummery-december.pdf).

Second Wave

The second wave of COVID-19 in the country started at the beginning of October 2020 after months of COVID-quiet. The cluster started centering Minuwangoda, Paliyagoda, and Diwlapitya in the western province marking the initiation of the second wave. The total number of cases rose by more than three times, after stagnating at 3300+ cases within three days. Infected cases were reported from different regions around the country. The number of cases increased, as new incidents were reported concentrating on textile industry workers, university students, police officers, and prisoners. As the government imposed strict containment strategies including an island-wide lockdown during the first wave, the government was reluctant to impose another lockdown during the second wave as it could result in undesirable repercussions on the livelihood of the citizens. Therefore, inter-district travel was restricted during the second wave, while cities with high case counts were isolated to control the epidemic. Regardless, several clusters were detected frequently around Sri Lanka, and the highest infections were reported mainly in the financial capital of the country and the surrounding western province areas. These features are summarized in Figure 1 and were used in the interviews to reassure the respondents' memory about the wave. During this period, schools and higher educational institutes were partially opened, students attended schools and tuition classes on a roster, and most activities gradually returned to normal. One of the marked features of this wave was how it presumably originated in the apparel, fisheries, and manufacturing sectors. The social and mass media used these distinct labels as identifiers to distinguish this period. The total number of cases resulting in this wave is reported as 92,341 and 591 deaths (https://www.epid.gov.lk/web/images/pdf/corona_monthly_summery/esummery-december.pdf).

Third Wave

Since mid-April 2021, when the country's COVID-19 death count stood at 609, the spread re-emerged as the third wave. This was caused by the lifting of travel restrictions alongside the Sinhala and Tamil New Year (one of the largest cultural celebrations in the country) holiday seasons. The population of the country considered this to be a relief from the intense containment restriction and quickly adapted to the regular lifestyle. As this was a festive season for Sri Lanka, the majority of the population was busy traveling during the holidays and the markets and shops were filled with the locals getting ready for the celebration. As a result, the most destructive third wave of COVID-19 started in the mid part of April. This can be clearly seen from Figure 2 where the take-off number of the newly infected cases and deaths soon after the low stringency period appeared from the 1st to the 21st of April. At this point in time, six strains of the COVID-19 virus were detected throughout the island. The two peaks in the newly reported cases during this wave as seen in Figure 2, were reported as a result of the successive spread of delta and omicron variants. In this stage, the virus has been spread across the entire country in contrast to the geographical or occupation level spread in the previous waves. At this stage entire "*Grama Niladari*" (village level administrative) divisions were put under lockdown once infected patients were reported from the locality. However, as some lockdowns lasted only a few days or weeks this strategy was not sufficient to stop the pandemic spread. As there was no coordinated effort by the stakeholders, and the prominent variants were highly transmissible, the total number of reported cases were over nearly five hundred thousand while 14, 375 were deceased as a result (https://www.epid.gov.lk/web/images/pdf/corona_monthly_summery/esummery-december.pdf).

The reluctance of the government to impose strict regulations, because the country has already experienced huge losses during the first two waves, the public and private sectors continued to work while adhering to roster-based strategies. Many organizations followed health protocols and Standard Operating Procedures (SOPs). The manufacturing industry of the country adopted bio-bubble-based strategies for uninterrupted operation. Furthermore, the education system of the country slowly started the transition from online to physical education towards the end of this wave. More importantly, the national COVID-19 vaccination program was initiated during the third phase. According to the epidemiology unit, the conclusion of the wave is reported to be in late December 2021.

In Sri Lanka, the dynamics of the pandemic spread is concisely described through the above mentioned three waves. These waves are often recalled through the use of triggering events which initiated the spread, and also by the unique features of each period as listed in Figure 1. Generally, the population of the country bears a good understanding of the wave-wise description as the regulatory bodies and the mass media often used this approach when reporting and sharing information regarding the pandemic. Each wave of the pandemic enforced a distinct way of life for the population based on the interventions adopted and the features of COVID-19 itself. Based on these facts, it was decided that the best way to capture the social, economic, and behavioral impacts on the population during the three waves of the pandemic is to use the already well-known wave-wise differentiation rather than using absolute time stamps which can eventually frustrate the respondents. Enumerators used certain keywords, aspects, features, and labels of these waves (see Figure 1) to facilitate the recollection of the respondent. Furthermore, the questionnaire was designed to capture a general answer rather than being time specific and exact, so that the answer would reflect the respondents' relative experiences during the waves of interest.

The policies and containment strategies adopted by the government of the country have proven to be effective in risk mitigation. However, these containment strategies disrupted the social and economic aspects of the people to a great extent. Due to travel restrictions, people who engaged in wage-earning employment lost their primary income and economic status. Further, self-employed people were affected excessively due to interruptions in their supply and value chains. As a result, a

significantly disproportionate impact on the income of people was observed. Further, the education system was drastically changed, creating inequalities among students due to economic disparities and access to online resources[3]. Consequently, these impacts led to sudden changes when accessing even basic needs, attending work, and participating in leisure activities leading to possible changes in behavior, attitudes, access to services, and quality of life.

The presented data set covers Social, Economical and Behavioral impact on the population of the country across the three distinctly identified pandemic waves which can be used in multi-disciplinary avenues. This article provides a detailed description of the data set presented and the methodology followed in gathering the data.

Literature Review

According to the recent literature, multiple efforts of researchers have produced different types of data repositories and open-access data sets related to the COVID-19 pandemic across countries. The data sets focus on capturing the impact on people through multiple directions. In the recent literature, there are data sets that focus on the impact of interventions and other measures taken by the government to mitigate the further spread of the disease which comprises public health[4] and non-pharmaceutical interventions[5–8], and economic measures[9]. Furthermore, Sugawara, D. et al.[8] attempt to identify and examines resilience factors and how they protect individuals from COVID-19-related fear and sustain their mental health. Cunningham T.J. et. al[10] attempt to identify the effects of the pandemic on human well-being through accumulating data related to sleep and mental health. Mondino et. al[11] tried to capture the public risk perception in Italy and Sweden for proper management and promotion of public health and safety. In addition to that, multiple data repositories can be found which are concentrated on psychological and behavioral responses[12]; the factors associated with psychological distress[13] and instant anxiety[14]; mood changes happen as a result of emotional responses to specific measures and policies[15]. Besides that, it is apparent that both the measures implemented by the authorities and the pandemic itself have significantly altered the mobility patterns and behavior of individuals. Attempts to observe the changes in the motion pattern of individuals have been taken by capturing the trajectories through GPS navigation systems[16,17]. This data captures dynamic changes in mobility and patterns of interaction at various geographical levels[17], enabling an evaluation of the effects of different interventions implemented by regulatory bodies. Furthermore, there has been an exploration of the impact on overall behavioral patterns of populations, including grocery shopping[18], travel-related behavior and attitudes[19], and access to education[20].

From Table 1 it can be seen that a large number of pandemic impact data sets were collected following an online survey[10–13] or through a mobile application to capture the mobility patterns[16,17]. Apart from that there are some studies based on the data collected from organizations such as WHO[4] and sources such as official government sources, peer-reviewed and non-peer-reviewed scientific papers, web pages of public health institutions (World Health Organization, Centers for Disease Control and Prevention, and European Centre for Disease Prevention and Control), press releases, newspaper articles, and government communication through social media[21]. Further, most of the studies gather data targeting a specific period of the pandemic (e.g.: Initial Wave[10,15]) with the exception of Desvars-Larrive, A., Dervic, E., Haug, N. et al.[14] where it covers four waves of the pandemic to investigate intolerance to distress and instant anxiety. What is primarily apparent from the available data is that the focus of the studies has been on a particular area such as psychological impact[10,12–14], behavioral impact[11,12,15], impact on mobility[8,16,17], impact on public health and economy[9] and quantification of interventions put forth[21].

Sri Lanka is a country where access to the internet is not uniform among its regions thus making online surveys biased. As the authors try to capture the impact on the people affected by COVID-19 throughout the country, it is vital that the data collected to be unbiased and reliable. The door-to-door, one-time surveying method adopted in the study has advantages over an online survey as it helps the enumerators to take a personalized approach to build rapport with the respondent while capturing nonverbal/verbal feedback, and emotional and behavioral clues[22]. Most importantly, the reliability of this kind of data set is higher as it is not subjected to survey fatigue, fraud, and sampling issues. One of the major contributions of this data is that it captures the impact on people throughout the pandemic waves in Sri Lanka rather than targeting a particular time period. Furthermore, it assesses the impact over multiple domains namely, education, income and economy, food consumption, mobility and human interactions, health, culture, and psychology. As the data consists of the impact on the population under the aforementioned areas for all three waves of COVID-19 in Sri Lanka it enables the users to capture the dynamics of the population during the pandemic and how it varied across different waves. To summarize, the data that has been proposed is multi-faceted and covers all stages of the pandemic in Sri Lanka through a comprehensive Door-to-Door survey.

Our data set is unique compared to other publicly available data sets as it contains,

1. Different impact domains in a single study: Mobility and human behavior, income and economic status, food consumption, education, access to health services and related information, cultural and psychological changes.

2. Data for three identified pandemic waves and pre-pandemic period which enables the study of the evolution of human behavior and impacts on multiple socioeconomic segments in day-to-day life as the pandemic progressed.

3. 283 impact-related variables/features for the 3020 households scattered around the country covering diverse groups of people with different socioeconomic backgrounds, occupations, income levels, education levels, and ethnic identities.

4. Gauge the effectiveness or response of the population to the enforced containment strategies and its evolution across different pandemic waves.

5. Beneficial information for policymakers and researchers to analyze the impact of the pandemic through a multidomain perspective enabling a more holistic approach to decision-making.

Methods

Survey Initiation

The survey was initiated on 6th November 2021, two years after the first COVID-19 patient was reported from Sri Lanka[23], and lasted till 10th December 2021. The motivation of capturing the socioeconomic and behavioral impact across the three waves and the pre-pandemic period demands the proper placement of the household survey to capture reliable information when the respondent still can contrast the pandemic waves and the pre-pandemic period. The survey was conducted in the late stage of the third wave when the country was returning to the norm. This placement allowed the authors to capture the detail of the entire pandemic, and the progressive evolution of impacts and attitudes with a thorough wave-wise description from the respondent. As mentioned in the section: Pandemic in Sri Lanka, the understanding of the population on the three waves aided the process of gathering information on distinct pandemic waves. This approach of narrowing down the entire pandemic to wave-wise portions while triggering the respondents' memory using the aforementioned re-collective tools highlighting unique features of each wave (see Figure 1), allows the enumerators to capture the respondents' experiences evolved throughout the pandemic across distinctly identified waves.

Study Instrument

The data collection was conducted through a field household survey by providing an interviewer-administered semi-structured questionnaire that consisted of close-ended and open-ended questions HARVARD Dataverse[24]. The questionnaire is modeled to capture different impact areas, namely, 1) education, 2) income and economic status, 3) food consumption, 4) Impact on mobility and human interactions, 5) health, 6) Cultural, and 7). Psychological. The questionnaire was model to capture the impact on these sections throughout the three waves, in contrast to the pre-pandemic period. The initial section covers the basic demographic details of the family. The person who responded to this section is considered the main respondent of the survey. The questions related to different impact areas were answered by a relevant person who is associated with the particular area. (e.g: the questions on educational impact were answered by a member of the family who is engaged in studies.). A detailed description of the impact areas covered and their subsections are tabulated in Table 2.

A Likert scale was used to collect responses for the provided questionnaire. Throughout the questions, three different Likert scales were used to evaluate the level of agreement, level of change, and frequency of attending/visiting. For the level of agreement and level of change, the conventional scale was adapted while a slightly different scale was adapted to capture the frequencies in a probabilistic sense. Here, **Never** is taken as the lower bound and, from there on-wards the scale is built. On a scale of zero to ten, never is taken as zero and the scale is divided into three equal segments to represent the remaining scale. (e.g. 1-3.3 : **Less Often**, 3.4-6.6 : **often**, 6.7-10: **very often**). The respondents were informed by the enumerators thoroughly about the scales prior to conducting the interviews. This attempts to accommodate the variance of uncertainty imposed through recollection, using a Likert scale rather than specific numbers. As this was a door-to-door survey the enumerators were given the tools and time to explain these scales and labels, in order to avoid any confusion.

Selecting the Sample

Sri Lanka is a South Asian island of $65610 km^2$ area. The country is divided into 9 provinces, which are subdivided into 25 administrative districts. (http://apps.moha.gov.lk:8090/moha/web/index.php/en). The sample in this study was selected using a multistage clustering technique which was used across the different administrative divisions in Sri Lanka. Figure 3 illustrates the top-down administrative structure of Sri Lanka. Further, information regarding the total number of administrative units at each level, as well as the specific number of selected areas, was also provided. Accordingly, this survey was conducted in 202 village-level administrative divisions in the country which were distributed over all 9 provinces in the country. Table 3 provides the criteria used to select each administrative division for the multistage clustering.

According to Table 3, initially,20 districts were selected based on the rate of COVID-19 infection. Divisional Secretariats were selected based on the data by the census and statistics department in such a way that it includes regions containing urban areas, rural areas, estate sectors, industrial zones, fisheries zones, agriculture zones, etc. to ensure that the data is unbiased. Furthermore, the poverty head count of the population was also considered to identify the economic condition of the

population. Next, the Grama Niladari (GN) divisions were selected based on the highest proportion of risk due to COVID-19 while accounting for the majority of ethnicity, the majority of religion, and minorities included in that GN. The main purpose of the presented data is to analyze the extent of the impact given that a household is affected by the pandemic. The selection of high-risk regions is based on this fact. Finally, households were selected randomly with the help of village-level administrative officers using the lists of registered voters. As a result, the selected households represent diverse groups of individuals across diverse ethnicities, age groups, gender, employment sectors, geographical locations, and other demographic characteristics. A total of 3020 households, with on average four members in a family, were selected from the village levels randomly to capture the variance in impacts. Figure 4 illustrates the geographical diversity of the survey.

The sample unit of the survey is a single household consisting of responses from several members. The questions related to different impact sections were answered by a relevant member of the household. Finally, all these responses were considered to represent the entire household (a data point). This is further elaborated through the execution of the field survey section. The data set eventually provides a total number of 3020 data points representing the households.

Execution of the field survey

Data collection commenced at the beginning of November 2021 and ended in the mid of December 2021. Computer Assisted Personal Interviews (CAPI) were used for data recording. Enumerators were trained before the survey, which included a role-play through a workshop at the University of Peradeniya, Sri Lanka. All enumerators are social sciences graduates from several state universities in Sri Lanka. Three enumerator teams including a Tamil (the second official language in Sri Lanka) speaking team were engaged in the data collection process. Every enumerator team was under the supervision of an experienced supervisor and a research team member. Enumerators were recruited to overcome religious, cultural, and gender-related barriers during the data collection. Frequent meetings with the enumerators and supervisors were conducted to overcome the obstacles by discussing the challenges and experiences at length throughout the data collection phase. The demographic balance, especially gender balance, among enumerators was maintained throughout the survey. Each household was interviewed for an average time span of 45 minutes providing a lot of elaboration on the questions to avoid any misunderstandings or confusion. The enumerators used phrases, and examples to trigger the memory of the respondent to get more precise answers.

The survey was conducted by taking the household as a unit. Therefore, in the data set a household is taken as a data point. A data point is fully formed from the information gathered through the relevant members of the family. Therefore, individual-based analysis using this data is not recommended. Simultaneously conducting the interview with all the members of a household encouraged the members relevant for each section to answer when needed. For example, a full-time or a part-time student (school student/university student/postgraduate student/diploma student/student engaged in vocational training) was selected to answer the education impact section. The respondent to the questions related to accessing financial localities (second subsection of the mobility impact), was selected based on mostly attending to financial activities. The person who is mainly responsible for buying groceries for the household was selected to answer the mobility section related to food access (the first subsection of the mobility impact). The member of the family who answered the questions on the demographic information on the household is considered the main respondent.

Ethical and administrative considerations

The ethical clearance for this study was obtained from the Ethical Review Committee, Faculty of Arts, University of Peradeniya, Sri Lanka, with the support of the Department of Sociology. Administrative clearance was obtained by the Ministry of Home Affairs of Sri Lanka, relevant DS offices, and relevant GN offices.

Informed verbal and written consent was obtained from all the participants. Voluntary participation was ensured. Privacy and confidentiality of data were maintained during each step of data collection, adhering to standard practices and protocols.

Data Records

The dataset from the survey and the survey instrument is available in HARVARD Dataverse[24]. The repository consists of three ".xlsx" files

- Socio-economic Impact of COVID-19 AI4COVID.xlsx
- Variable definition and coding system.xlsx
- Demographics.xlsx

The description of variables and the coding used can be found in the 'variable definition and coding system.xlsx' file. Further, Demographics.xlsx contain the demographic information associated a household. A household is considered a data point, which is represented in a single row of the dataset. Each column represents a feature or a variable.

The interviews to collect the data were conducted for 3020 households. The cumulative member count for the entire set of houses was 12110. However, when conducting the interviews, only the relevant members who can relay comprehensive information actively participated as respondents. Therefore the total count of respondents to the survey was 11552 as some members of the families did not answer any section of the survey. With this light, the following steps were executed to improve the technical quality of the dataset while removing any redundancies. The following steps were taken to build up the feature space of the data set of 3020 households.

1. Initially, all the demographic information of the members of the families was gathered. As mentioned earlier, some members of the families did not contribute to the survey. Therefore inserting only the demographic details of such members were treated as redundancy and hence removed from the data set.

2. Then, each individual's response related to one household was merged to create the final data point representing a household. This was executed using the python programming language[25]. Once the responses were merged, the final data set consisted of 3020 data points representing each household.

3. Next, the missing values were identified through manual filtering. These data points were coded accordingly (refer to the variable definition and coding system.xlsx file from the repository).

4. Finally, the dimension of the data set was 3020*283. The 283 columns were impact-related variables covered using the questionnaire containing information related to the impact sections mentioned in Background and Summary.

As it then creates fully populated fields for each data point across impact areas cross field studies can be performed. Also, it is more pragmatic because a data point would relate to a unit that uniquely describes a family's experiences and behavioral changes across different pandemic waves. The demographic distribution of the respondents who participated actively in the survey is illustrated in Figure 5 in comparison with the demographic distribution of the country. As the main intuition of the survey was to capture the impact of the pandemic and different measures taken by the regulatory bodies, the survey was conducted focusing on the regions affected by the COVID-19 pandemic. Therefore, Figure 5 emphasizes one of the key features of the survey conducted, where there is more emphasis on people who would presumably be more affected.

Figure 5 shows the percentage distribution of age groups, education level, gender, and ethnic identity of the 11552 individuals. Accordingly, the sample comprises children, youth, adults, and senior citizens. Also, it has a more female representation and an adequate balance between the binary gender groups.(Note: the demographic information included in the data set is related to the main respondent of a family unit for your reference).

Technical Validation

Removal of indirect identifiers

Privacy and anonymity are important and critical considerations when publishing a database.[26,27], hence were taken care of by the following steps.

1. Excluding the information related to the geographical location of each household's DS and GN divisions which will help to trace the locations of the households.

2. Merging the discrete age values into a few age groups in order to disallow the potential for re-identification.

3. Improving the ambiguity of data regarding marital status and ethnicity by adding random noise to original data.

4. Re-categorizing the status and the sectors of employment to a few groups.

All these steps were conducted under the supervision of a specialist related to the field.

Data Reliability through McDonald's Omega Coefficient

Coefficient omega was proposed 40 years back[28] as a reliability measure of homogeneous items from a measurement instrument. In contrast to a measure like Cronbach's alfa[29] this is a reliability estimate that provides a more accurate estimate of the internal consistency of a multi-item scale[30]. Furthermore, the Omega coefficient can account for multidimensionality, making it a more accurate measure of internal consistency for complex test structures.

In literature McDonald's Omega coefficient has been used to check the reliability of socioeconomic data, including data related to the impact of the COVID-19 pandemic on the economy, education, human mobility[31], psychological behavior[32] and access to health resources. As the overall mobility of a human being depends on multiple indicators such as migration status, length of residence and travel behavior this measure can be used to evaluate how consistent the indicators are in measuring the underlying construct of mobility[31].

With this light, the McDonald's Omega coefficient is used to evaluate the internal consistency and the reliability of each impact section of the presented data. The results tabulated in Table 4 reflect the technical quality of the data to the user. From Table 4, it can be seen that the omega value of all impact sections is higher than 0.5 (/1), depicting strong reliability. The impact sections covering Education and household income and economy have an Omega coefficient exceeding 0.9. An extended analysis of the reliability of the data related to the income and economy impact section was conducted by extracting the data related to subsection changes in food consumption which resulted in a coefficient value of 0.936. In addition, the data related to mobility, cultural behaviour, psychological state, and health services and information are within the acceptable range of reliability. Here the coefficient is calculated for the raw data without conducting any pre-processing. Therefore, high reliability for the above impact section can be achieved by pre-processing the data according to the task of interest.

Usage Notes

To the best of the authors' knowledge, this dataset is the first of its kind in Sri Lanka as well as in the world which has focused on capturing the different socioeconomic impacts, changes to socioeconomic status, human behavior, and mobility patterns during the COVID-19 pandemic throughout different phases of the pandemic for diverse groups of people. Also, this is a dataset that compares people's behavior, and the impact on different aspects of their lives during three identified phases and compares it to the pre-pandemic period. This enables the study of the evolution of behaviors and impacts throughout the prolonged duration of the pandemic. The survey also covers the impacts on a wide variety of fields: education; income level and employment status; mobility patterns and behavior; food consumption patterns, cultural, health, and psychological state. As this dataset contains information on such diverse fields it enables the identification of correlation patterns and linkages between different aspects of people's lives during the pandemic. The accuracy of this dataset can be assured as this is collected through a door-to-door field survey under close supervision by a group of well-trained enumerators. Furthermore, the dataset could provide insight into how the pandemic has played a role in the current economic crisis in Sri Lanka.

All the coding for the variables/features is present in the excel workbook. The demographic information of the main respondent of a family unit can be found in the data set for reference. In addition, the household economic condition and income-related data can be found from the variables named 3B1_1PP to 3B1_1TW. For further information, users can contact the corresponding author. Finally, for more details about the project and other findings, you can reach the project website through (https://ai4covid.lk/)

Potential Uses and Research Paths

The dataset is focused on the impact of the COVID-19 pandemic and the interventions, on different households in the high-risk areas of Sri Lanka. The impact is assessed by concentrating on socioeconomic and behavioral impact areas. As mentioned in the Data Usage section the demographics of the dataset is shaped by the condition of selecting high-risk areas of the country. As the data set is more focused on analyzing the impact of the pandemic, the underlying factors and how deeply the population affected the condition to focus on the high risked regions is vital. Therefore, users should be well aware of the fact that the demographic distribution of the survey data is conditioned on the highly affected regions and it will not reflect the actual demographic distribution of the country.

The users should be well aware of the limitation of the use of this dataset. The authors discourage the use of this dataset to conduct quantitative comparisons among different demographic groups, genders, ethnic groups, and geographical regions. For instance, it is not encouraged to use the data to regulate spread, transmission, or relative comparison-based analysis. The authors address the users not to employ the data for provincial or district-level comparative analysis. It is encouraged to use the data to study the impact of the pandemic at a national level considering that the minimum level of resolution offered for data extraction is a household. Furthermore, as the minimum resolution offered is a household the individual-level information cannot be extracted from the data. Therefore the individual-level analysis is discouraged. In addition, the survey attempts to extract the sense of feelings/perceptions during the identified waves of the pandemic. Therefore, the provided data relies on the self reported perception of the 'impact' of the COVID-19 pandemic which is a common limitation as well as a feature of all qualitative research.

The user of the data should be aware of the fact that the data can be basically used to identify the demographics of the population affected by the COVID-19 pandemic in the country. This allows the users to identify their peers who were affected by the pandemic in contrast to the actual population of the country. As this represents a cross-section of the nation's affected

community, it can relay information on how certain demographics were more deeply affected in some areas of impacts as opposed to others. Furthermore, conditioned on the pandemic has affected certain communities, it can be used to identify the impact zones, how deep was the impact, and the underlying factors. In addition, it allows the intra and intercomparison among different groups, in terms of impact sections. Further, this analysis can be extended to examine the impact on different fields in scrutiny-level factors.

The data presented in this paper was collected in the late stage of 2021, months prior to the economic crisis of the country boiled over at the beginning of 2022. As the data consists of abundant information on the socioeconomic impact of the pandemic on households, it may have captured the dynamics leading up to the economic crisis at a granular level. Therefore, it might have certain indicators as to how the pandemic had stimulating and/or amplifying effects on the economic crisis that proceeded immediately after the data collection. This allows the users to incorporate the dataset in analysis related to the pandemic impact on the economy of a country or on the economic crisis specifically in Sri Lanka.

The organization of the dataset will allow the use of data analytic-based approaches to analyze different impact sections. These techniques can be incorporated by the users to effectively cluster the data based on different factors and scenarios. This will allow the users to identify the dependencies between diverse impact sections. The conclusions taken by such analysis mentioned in this section will eventually help the policymakers to effectively handle scenarios like this in the future. They can be used to draft procedures that would effectively contain disease propagation while having a minimal negative impact on the lifestyle of the population.

Code availability

The coding of the measures is available in (HARVARD Dataverse). as the name "variable definition and coding system".

References

1. MF, K. *et al. Emerg. infectious diseases* **27**, 9, https://doi.org/10.3201/eid2703.202589 (2021).
2. Ritchie, H. *et al.* Coronavirus pandemic (covid-19). *Our World Data* (2020). Https://ourworldindata.org/coronavirus.
3. Ilangarathna, G. A. *et al.* A comprehensive overview of education during three covid-19 pandemic periods: Impact on engineering students in sri lanka. *Educ. Sci.* **12**, 197 (2022).
4. Zheng, F., Q.and Jones & Leavitt, S. e. a. Hit-covid, a global database tracking public health interventions to covid-19. *Sci. Data* **7**, 286 (2020).
5. Desvars-Larrive, A. *et al.* A structured open dataset of government interventions in response to covid-19. *Sci. data* **7**, 1–9 (2020).
6. Singh Chauhan, R. *et al.* A database of travel-related behaviors and attitudes before, during, and after covid-19 in the united states. *arXiv e-prints* arXiv–2103 (2021).
7. Brouard, S. *et al.* 'citizens' attitudes under covid19', a cross-country panel survey of public opinion in 11 advanced democracies. *Sci. data* **9**, 1–10 (2022).
8. Shiraef, M. A. *et al.* Covid border accountability project, a hand-coded global database of border closures introduced during 2020. *Sci. data* **8**, 253 (2021).
9. Porcher, S. Response2covid19, a dataset of governments' responses to covid-19 all around the world. *Sci. Data* **7**, 423 (2020).
10. Cunningham, T. J., Fields, E. C. & Kensinger, E. A. Boston college daily sleep and well-being survey data during early phase of the covid-19 pandemic. *Sci. Data* **8**, 1–6 (2021).
11. Elena, M., Baldassarre, D., Johanna, G. G. M., Elena, R. & Maria, R. Public perceptions of multiple risks during the covid-19 pandemic in italy and sweden. *Sci. Data* **7**, 434 (2020).
12. Yamada, Y., Ćepulić, D. & Coll-Martín, T. e. a. Covidistress global survey dataset on psychological and behavioural consequences of the covid-19 outbreak. *Sci. Data* **8** (2021).
13. T., S. N. Y. & et al., S. N. A real-time survey on the psychological impact of mild lockdown for covid-19 in the japanese population. *Sci. Data* **7**, 372 (2020).
14. Sari, G. K. B. e. a., E.and Kağan. Dataset on social and psychological effects of covid-19 pandemic in turkey. *Sci. Data* **9**, 441 (2020).
15. Bailon, C. B. O. e. a., C.and Goicoechea. Covidaffect, real-time monitoring of mood variations following the covid-19 outbreak in spain. *Sci. Data* **7**, 365 (2020).

16. Pepe, E. *et al.* Covid-19 outbreak response, a dataset to assess mobility changes in italy following national lockdown. *Sci. data* **7**, 1–7 (2020).

17. Kang, S. L. Y. e. a., Y.and Gao. Multiscale dynamic human mobility flow dataset in the u.s. during the covid-19 epidemic. *Sci. Data* **7**, 390 (2020).

18. Grashuis, J., Skevas, T. & Segovia, M. S. Grocery shopping preferences during the covid-19 pandemic. *Sustainability* **12**, 5369 (2020).

19. Jakob, C. E., Kohlmayer, F., Meurers, T., Vehreschild, J. J. & Prasser, F. Design and evaluation of a data anonymization pipeline to promote open science on covid-19. *Sci. data* **7**, 1–10 (2020).

20. Trung, T. *et al.* Dataset of vietnamese student's learning habits during covid-19. *Data Brief* **30**, 105682 (2020).

21. Desvars-Larrive, E. H. N. e. a., A.and Dervic. A structured open dataset of government interventions in response to covid-19. *Sci. Data* **7**, 285 (2020).

22. Hillier, A., Cannuscio, C. C., Griffin, L., Thomas, N. & Glanz, K. The value of conducting door-to-door surveys. *Int. J. Soc. Res. Methodol.* **17**, 285–302, 10.1080/13645579.2012.733173 (2014). https://doi.org/10.1080/13645579.2012.733173.

23. Epidemiology Unit Ministry of Health. Covid-19 epidemiology sri lanka. Available athttps://www.epid.gov.lk/web/images/pdf/corona_monthly_summery/esummery-december.pdf (2021).

24. Ilangarathna, G. A. *et al.* Socioeconomic Impact of COVID-19 during three pandemic waves. 10.7910/DVN/CXMJSM (2023).

25. Van Rossum, G. & Drake Jr, F. L. *Python reference manual* (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

26. Gritzalis, S. Enhancing web privacy and anonymity in the digital era. *Inf. Manag. & Comput. Secur.* (2004).

27. Ciriani, V., Di Vimercati, S. D. C., Foresti, S. & Samarati, P. Theory of privacy and anonymity (2009).

28. Zhang, Z. & Yuan, K.-H. Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educ. Psychol. Meas.* **76**, 387–411, 10.1177/0013164415594658 (2016). PMID: 29795870, https://doi.org/10.1177/0013164415594658.

29. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334, 10.1007/BF02310555 (1951).

30. Hayes, A. & Coutts, J. Use omega rather than cronbach's alpha for estimating reliability. but.... *Commun. Methods Meas.* **14**, 1–24, 10.1080/19312458.2020.1718629 (2020).

31. Dunn, T. J., Baguley, T. & Brunsden, V. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* **105**, 399–412, https://doi.org/10.1111/bjop.12046 (2014). https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjop.12046.

32. Milic, M. *et al.* Validity and reliability of the serbian covid stress scales. *PLOS ONE* **16**, 1–15, 10.1371/journal.pone.0259062 (2021).

Acknowledgements

This research was funded by International Development and Research Center, Canada, grant number 109586-001. The APC was funded by the International Development and Research Center, Canada

Author contributions statement

G.A.I., S.Y., M.P., G.T., R.G., V.H., J.E., P.E.,S.D., Developing the study instrument, G.A.I. Writing the original draft, M.M., G.A.I., Y.R., Data cleaning, G.A.I. Considering privacy and anonymity practices, All authors contributed to the data collection and reviewing of the manuscript.

Competing interests

The authors declare no conflict of interest.

Figures & Tables

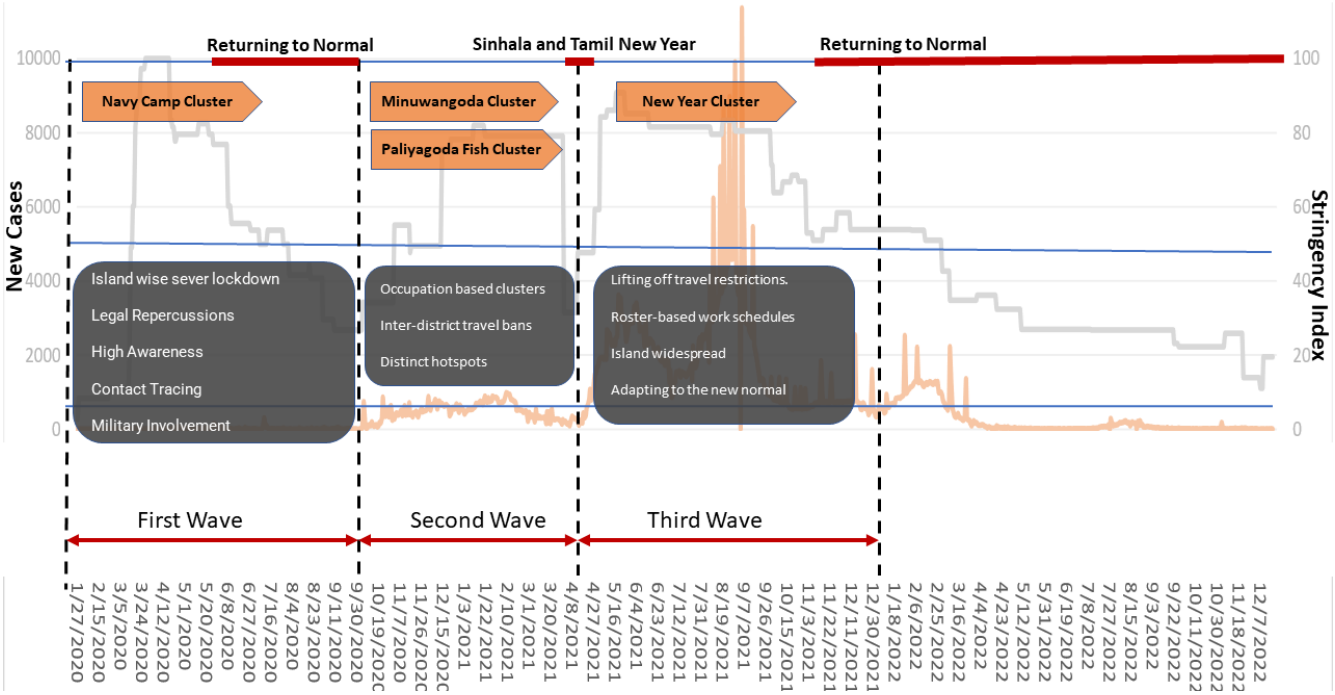

Figure 1. The unique features and labels of the three distinct waves of pandemic

Paper	Location	Sample size	Purpose	Duration	Collection method
[10]	USA	1899	Impact on Mental Health and Sleep Patterns	20/03/2020 - 23/06/2020	Online Survey
[12]	global	173426	Psychological and Behavioral Responses	30/03/2020 - 30/05/2020	Online Survey
[11]	Italy,Sweden	4154	Public Risk Perception	08/2020 - 11/2020 (15 days/month)	Online Survey
[13]	Japan	11333	Factors associated with Psychological Distress	11/05/2020 - 12/05/2020	Online Survey
[15]	Spain	999	Mood Changes	28/03/2020 - 21/06/2020	Online Survey
[14]	Turkey	2817	Intolerance to Distress and Anxiety	13/04/2020 - 25/11/2020	Online Survey
[8]	Global	2817	Policy Changes Affecting International Travel and Immigration	-	Qualtrics Survey
[9]	Global	-	Public Health and Economic Measures	01/01/2020 - 01/10/2020	ACAPS
[21]	Global	-	Government Non-pharmaceutical Interventions (NPIs)	31/12/2019 - 15/07/2020	Written Sources
[4]	Global	200	Public Health Interventions	-	WHO PHSM Database
[17]	USA	-	Mobility Changes and Interactions	01/02/2020 onwards	GPS navigation system
[16]	Italy	170,000	Impact of Interventions	18/01/2020 - 17/04/2020	GPS navigation system

Table 1. Literature Review

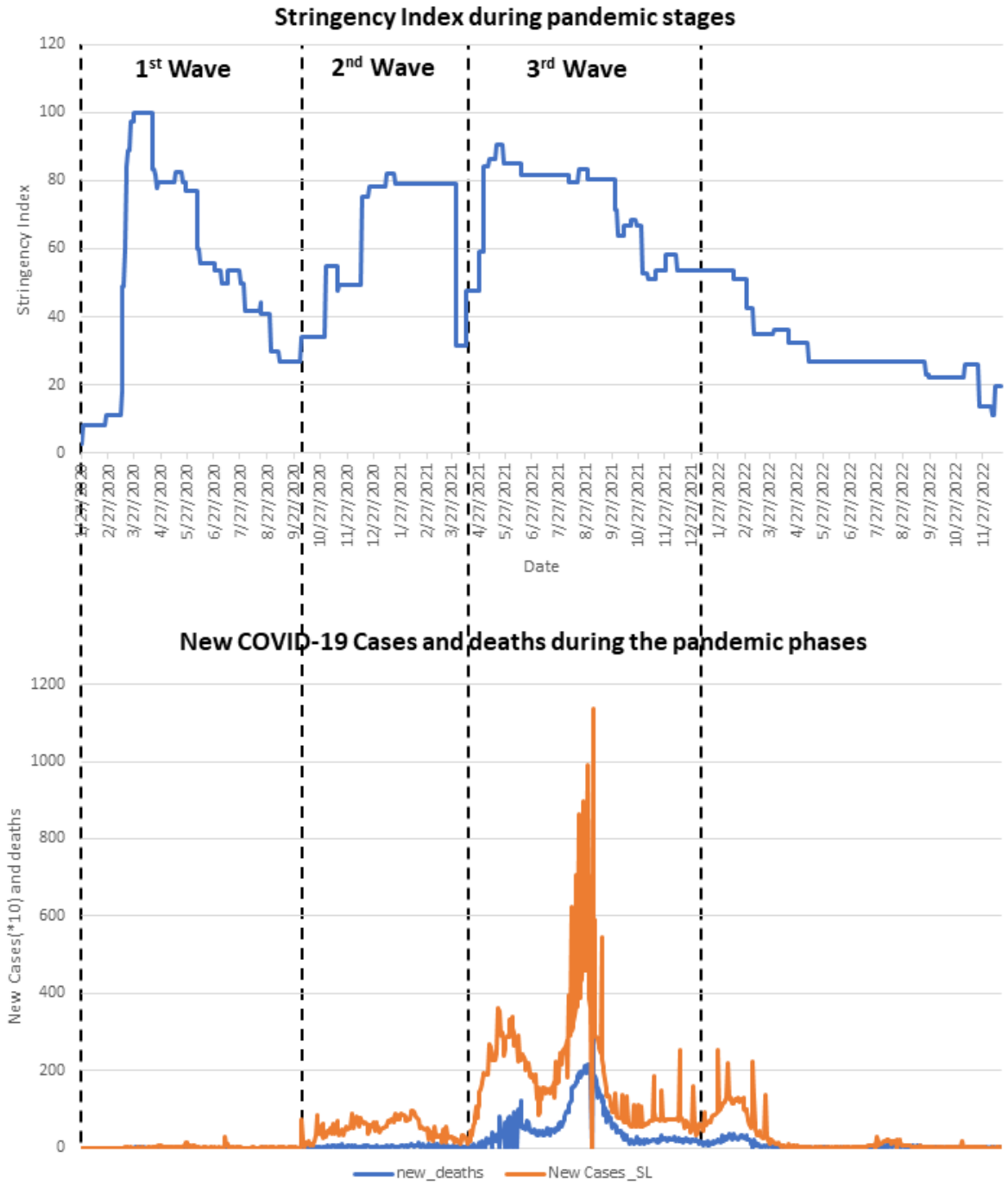

Figure 2. Fluctuation of the stringency index in Sri Lanka during the respective waves,the time intervals of the three phases and the recorded daily cases and deaths.

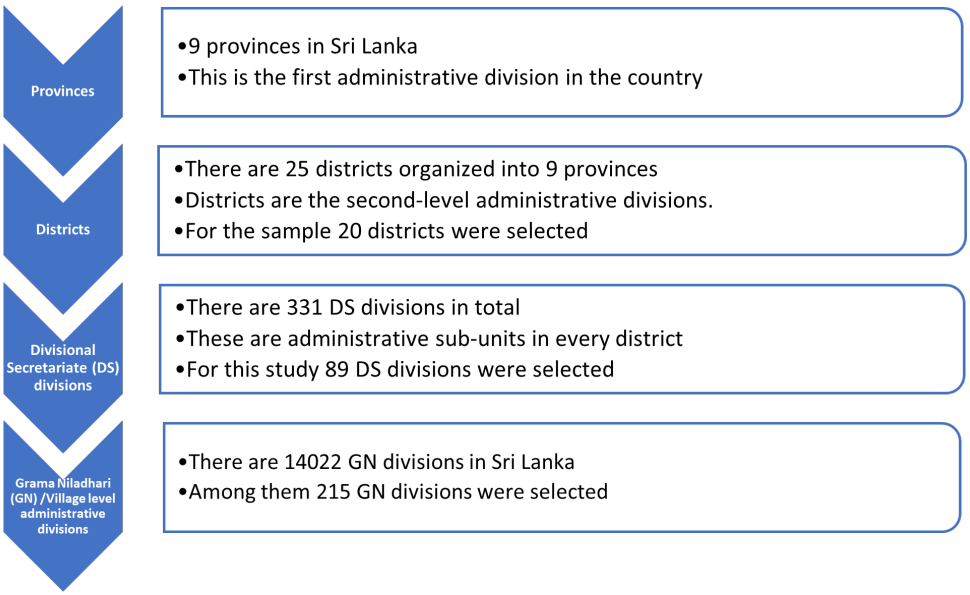

Figure 3. The administrative levels in Sri Lanka

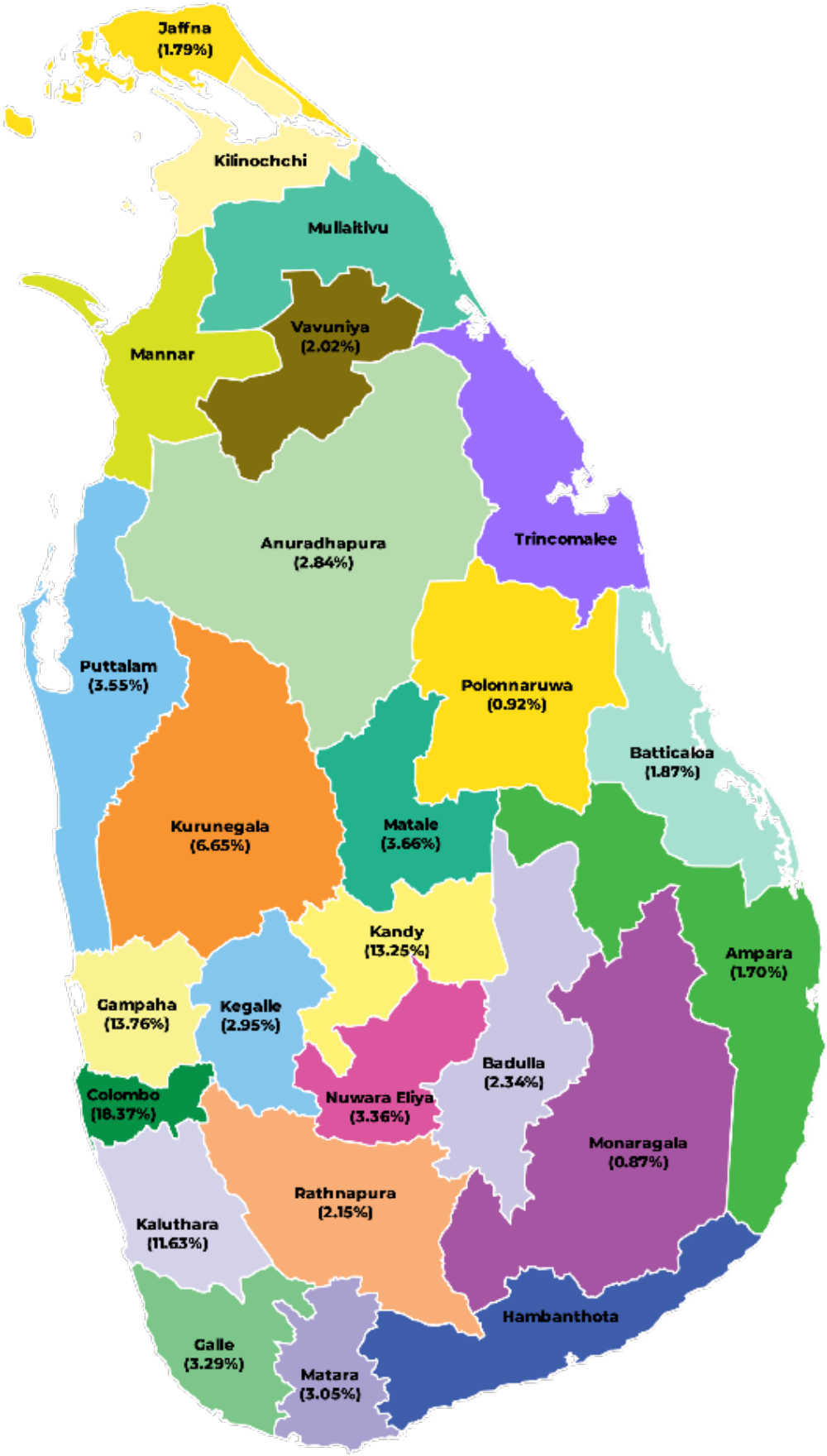

Figure 4. Percentage of the respondents in different districts in Sri Lanka

Figure 5. Percentage distribution of the sample demographics of respondents in comparison with the actual population of the country.

Section	Subsection	Description
Demographic Information	Gender	• The basic demographic information of the main respondent is recorded
	Age	
	Marital	
	Education Level	
	Ethnicity	
	Employement Status	
	Employement Sector	
Impact on Education	Changes in the education pattern	• Information regarding the frequency of attendance of respondents to online and in-class education. • Amount of educational work carried out in comparison with pre-pandemic. • New learnings • Preference between online and in-person classes • The devices used
	Access to resources	• Device ownership and capability to purchase devices. • Interruptions due to connectivity issues, power failures, and device malfunctions. • The technical know-how • Package limitations.
	Feelings about self-directed learning at home	• Feelings on, – Lacking social aspects of in person education. – The Learning Environment. – Availability of human resources and support.
Impact on access to health services and information	Accessing to health services	• Ability to attend health clinics and access doctors with reasoning. • Availability of medicine. • Receival of health guidelines
	Information regarding the pandemic	• The feeling about the reliability of the received pandemic related information. • The sources of information.
Impact on Income and Economic Status	Impact on livelihood and household income	• Changes in employment routine. • Changes to income. • Additional Measures taken for compensation.
	Financial supports during the identified periods	• Receival of financial assistance and other benefits from the government and other organizations.

	Change in food consumption during identified pandemic phases compared to the pre-pandemic period	• Availability of food in terms of quantity. • Reasons for increased or decreased consumption.
Impact on Mobility	Frequency of accessing food sources	• An assessment of the modes/ways used to fulfill the food needs is explored. • The respondents were questioned on how often they visited nearby shops, the main supermarket in the city, a mini supermarket nearby, home delivery, their own garden, government/NGOs, online apps, from neighbors, fairs, and others to collect supplies needed.
	Frequency and methods of accessing financial services	• Study the localities to access financial services during the pandemic, including Bank located in town/city, Bank located nearby, Local financial Institutes, ATMs in the city, ATMs nearby, Online banking, Post office, and other.
	Frequency of access to the leisure activities	• How often people participated in outdoor activities Going on trips, Recreational places, Social Gatherings, and other places during a pandemic.
	Frequency of using modes of transportation	• An exploration to identify the dominant modes of transportation including public transport, shared vehicle, Use of hired three-wheeler, Use own motor vehicles, Use of bicycles, Walking, Other during COVID-19.
	Frequency of residing at following places to attend for work	• Study to analyze the way/mode that people used to attend work. (Work from home, come to work from home, come to work from rented out place/ boarding place, come to work from a common hostel, Other).
	Frequency of using the following mode of transportation for work	• Study the likelihood of using different transportation methods to report their work stations including the Use of public or Shared vehicles (School vans, staff vehicles, private vehicles provided by the institute, Use hired three-wheeler, Use own motor vehicles, Use of bicycles, Walking, Other).
	Frequency of experiencing the levels of interaction in work	• How often level of people interacted in their workplace Work alone at home, Work with few known groups of people, Work with regular few known people and limited unknown people, Work with irregular few unknown people, Work with irregular large groups, Other.
Psychological Impact	Effect of lockdown on one's emotions, feelings, and various aspects of life	• How respondents feel about the pandemic and lockdown periods.

	Effect of lockdown on relationships	• How the pandemic affects relationships among family members, neighbors, and office peers
Impact on Cultural behavior	Impact on cultural engagements and behaviors	• How the pandemic affects cultural beliefs and engagements.

Table 2. The main impact sections and the sub-sections of the questionnaire

Administrative Divisions	Criteria to select the divisions	Number of selections at each administrative level
Provinces	All the provinces in Sri Lanka	9
Districts	Considering the severity of the disease spread and containment strategies enforced during the pandemic phases (https://www.dgi.gov.lk/)	20
Divisional Secretariate (DS) divisions	Highly populated DSs from the selected districts (http://www.statistics.gov.lk)	89
Grama Niladhari (GN) divisions	Using the dependency-Ratio (which illustrates the portion of the population over 60 years of age and under 14 years) in the risk map, Department of Census and Statistics, Sri Lanka (,http://www.statistics.gov.lk/ref/Riskmaps) High risk GNs were selected from the DSs. Further, GNs with the following sectors were prioritized; Fisheries, agriculture, estate sector,and industrial areas	202
Households	With the help of respective village level administrative officer using the list of registered voters.	15 from each GN division

Table 3. Criteria followed to select different administrative levels and households

Impact Section	Omega Coefficient
Education	0.928
Health Services and Information	0.644
Income and Economic Status	0.936
Mobility	0.572
Psychology	0.644
Cultural	0.528

Table 4. McDonald Omega Coefficient for Each Impact Section