Analysis of COVID-19 distant learning measures: an approach based on Bayesan Network

Nicolas Lazzari - nicolas.lazzari2@studio.unibo.it

June 8, 2022

Abstract

During the first outbreak of the COVID-19 virus Unesco and Unicef started to collect data on distant learning measures taken by countries around the world. In the following report we will make use of Bayesan Network to analyse that data. This will be done by taking into consideration the technological development of each country, in order to provide estimated on the number of students affected by each measure, and an approximation of the incidence of the taken measures.

Introduction

During the whole COVID-19 emergency Unesco and Unicef have collected information on implemented distant learning measure all over the world. The data is publicly available [1] and is composed of 210 different countries. Each country is represented by its extended name and its ISO 3166 code, EC group, SDG group and WB income group. The analysed data has been collected from 16/02/2020 to 31/12/2021 and includes three main indicators on daily implemented distant learning measures: *internet*, *radio* and *television*. Additionally a school closure status is included, namely if schools were on an academic break, fully open, closed due to COVID-19 or partially open (i.e. open only for a subset of the overall students).

We will compute the probability of a certain measure being implemented and relate it to the probability of being reached by the corresponding media. By using this information we'll be able to estimate the probability that a students was forced at home and the probability of not being able to benefit from the implemented measures. The former will be used to estimate the probability that a student, forced into distant learning, had experienced mental health conditions such as anxiety and depressive symptoms(2).

Data integration

While Unesco and Unicef data provide the backbone of the whole analysis, supplementary data is required in order to compute the required information. The provided data contains some demographic information for each country, such as the number of students, the number of enrolments for each education level and the number of teachers. However this type of data is inconsistent: school age population and the different levels of education enrolments takes into account different portions of the students (tertiary education students are taken into account in enrolments but not into school age population) making it difficult to use data consistently. For this very reason demographic data from The World Bank [2] will be used. We won't take into consideration the number of out-of-school students since those information are usually estimated instead of provided by the country authorities. We will only use the number of students enrolled in primary, lower secondary and upper secondary education, along with the different proportions of males and females.

Data on radio, television and internet availability will be taken from different sources. Unfortunately most of the data publicly available is old and seems outdated, we will try to use the latest available information for each media. The number of households with a television for each country, hence approximately the number of students that would be able to be reached by that kind of media, are taken from globaldatalab.org [3]. Unfortunately there are 62 missing countries. Those includes small countries that can be safely ignored without skewing too much the data (such as Andorra) along with big countries (Australia, Arab Emirates) which are instead interesting to analyse (figure 1).

[1] `https://covid19.uis.unesco.org/data`

[2] `https://databank.worldbank.org/source/education-statistics`

[3] `https://globaldatalab.org/areadata/view/tv/?levels=1&interpolation=0&extrapolation=0&nearest_real=0`

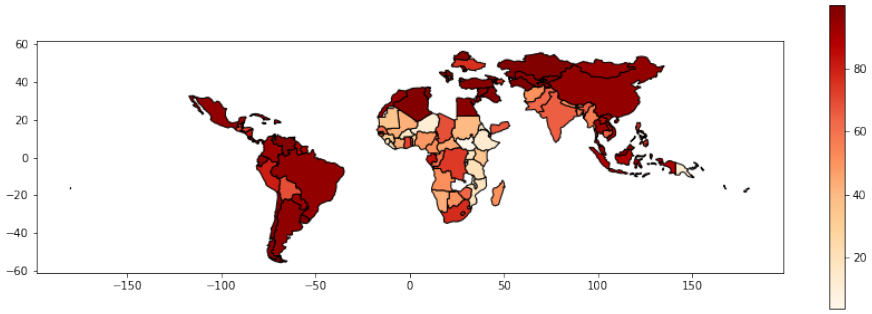

Figure 1: % of households with a TV according to (**?**)

We will try to estimate missing data by taking the mean or the median value from countries that belong to the same groups. On table 1 the Mean Squared Error obtained by using different combinations of groupings is shown. Each grouping have been tested by randomly sampling 100 countries. The best estimation technique is obtained by using the mean of countries in both the same *SDG* and *UN* group. Those result that yields a Na are groupings for which some countries don't have any other country in the same combination of groups. Even using this kind of method, some countries ends up with unrealistic values (i.e. 39% for Australia). This happens because other countries in the same group are highly dissimilar. We will manually fix those instances by using additional data from nationmaster.com [4]. On figure 2 we can see how the world looks like after approximating the missing values.

Group	MSE using mean	MSE using median
UN	506.61	495.30
SDG	451.13	581.27
WB	466.37	528.23
UN ∩ SDG	**419.63**	450.39
UN ∩ WB	Na	Na
SDG ∩ WB	Na	Na
SDG ∩ WB ∩ UN	Na	Na

Table 1: Mean Squared Error obtained estimating TV in households

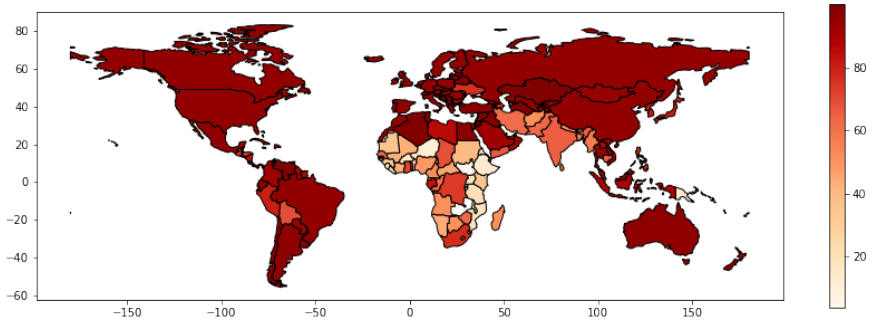

Figure 2: % of households with a TV after estimation

Internet usage is taken from worldbank.orf [5], taking the most recent value for each country. Radio coverage is the most difficult information to obtain, there aren't any free datasets which tries to enumerate or even approximate the portion of population that can access content by radio, however according to

[4] `https://www.nationmaster.com/country-info/stats/Media/Households-with-television`
[5] `https://data.worldbank.org/indicator/IT.NET.USER.ZS?view=chart`

(1) virtually 95% of worldwide population is reached by radio media. We will be using that value for all the different countries.

Bayesan Network

The Bayesan Network structure is not algorithmically inferred by the data but rather manually deducted from it, based on empirical knowledge. On figure 3 the structure is represented visually. We will model 14 variables whose conditional dependency follows from the concept they describe: the variable `Country`, the most interconnected one, is dependant on the variable `Group`, which refers to the *SDG* grouping. `InternetAccess`, `RadioAccess` and `TvAccess` all refer to the population of a country that can benefit from a given mass media, hence they are dependant on `Country`. In a similar way to the media access variables, `Online`, `Radio` and `Tv` refer to the implemented measures by a specific `Country`. `Education` models the probability of being in a particular education level (primary, lower secondary and upper secondary) and is hence dependant on the probability of living in a country (`Country`) along with the probability of either being a male or a female (`Gender`, which is conditionally dependent on `Country`). The probability of being a student and that of spending time at home because of school closures (`Student` and `AtHome`, respectively) both depends on the country. Finally, mental health conditions (`Anxiety` and `Depression`) are both dependant on the probability of spending time at home (`AtHome`) because of restrictive measures and school closures.

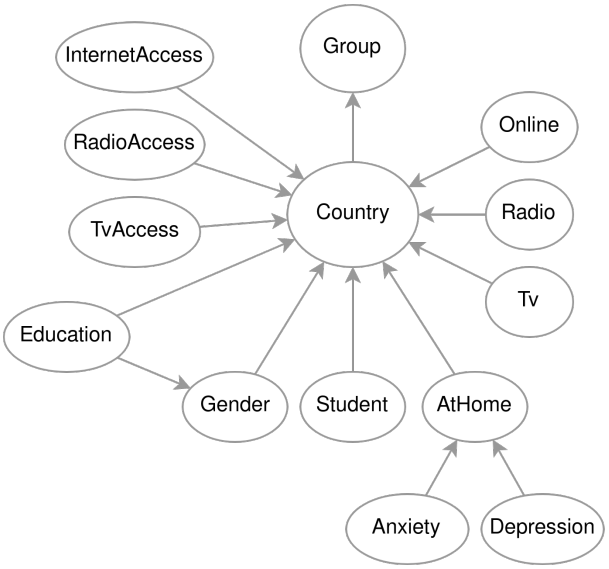

Figure 3: Bayesan network diagram

Inference and analysis

With the previously defined network we can perform inference on the network in order to estimate probabilities that would otherwise be very difficult to analyze. We will perform inference by using the Variable Elimination method available on the python library `pgmpy`.

Implemented measures

If we try to asses the probability of distant learning measure implementation world-wide we obtain the probability distributions of table 2. Online and television distant measures are the most implemented ones, of course being able to teach via radio is an extremely unconventional way of spreading education and avoiding any effort into the implementation of such measure while giving more importance to other measures might be the most profitable approach in terms of educational outcomes.

If we try to compute the probability of multiple measures being implemented, on table 3, we find out that the most likely combination is the one in which online and tv measure are both implemented ($\approx 63.4\%$). The addition of radio measures to the previous ones is also quite likely ($\approx 29\%$). It's more

	Yes	No
Online	0.9508	0.0492
Radio	0.3219	0.6781
Tv	0.9524	0.0476

Table 2: Distant learning implemented measures

likely that none of the distant learning measures have been implemented ($\approx 1.55\%$) rather than only tv ($\approx 0.8\%$) or radio ($\approx 0.6\%$). We can realistically suppose that in those areas in which an internet connection is not available, tv or radio signals might still be available, which means that some students probably couldn't access any type of distant learning, even though they might have been able to benefit for some of that media.

	$P(Tv, Radio, Online)$
¬Tv, ¬Radio, ¬Online	0.0155
¬Tv, ¬Radio, Online	0.0206
¬Tv, Radio, ¬Online	0.0069
¬Tv, Radio, Online	0.0047
Tv, ¬Radio, ¬Online	0.0083
Tv, ¬Radio, Online	0.6338
Tv, Radio, ¬Online	0.0186
Tv, Radio, Online	0.2917

Table 3: Distant learning implemented measures

Measure accessibility

Taking into consideration the probability of a measure being implemented is not a good measure when trying to understand the effectiveness of a country response to the emergency. We also need into account the probability of using an implemented measure. On table 4 we can see the probability distribution of being able to access implemented measures. Most of the students are reached by radio as a media if it is implemented, but from table 2 we already know that we don't have many countries using that kind of measure. Television, on the other hand, would reach the vast majority of the population, which is good sign since most countries did implement education through television. The most worrying result, however, is on internet measures: while the majority of those students that can benefit from online learning measures can indeed access them, 41% of the student population can't make use of that kind of measures.

Implemented Measure	Access to measure	*No* access to measure
Tv	0.7854	0.2156
Radio	0.95	0.05
Internet	0.5899	0.4101

Table 4: Conditional probability distribution on measure access

Online measures accessibility

As one could expect, the fact that more than 4 in 10 students can't access online measures isn't equally spread all over countries. On figure 4 we can see the probability of being able to access online measures if they have been implemented. More than half of the students population on central and southern Asia and on southern Africa can't benefit from online learning, even though it has been implemented by their country's institutions.

If, instead of grouping countries by their geographical position, we group them by their economic indicators (figure 5) we can see that most of the students from low income and low middle income countries can't access online learning.

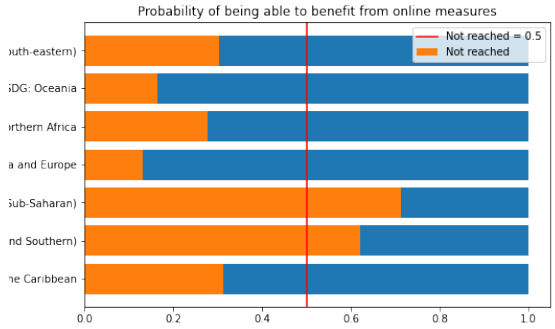

Figure 4: Probability of being reached by SDG groups

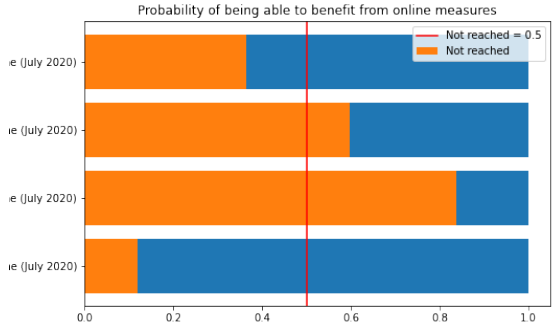

Figure 5: Probability of being reached by WB groups

School closures

Many students, in particular from low income regions, can't access online learning. If we check the probability distribution of being forced to online learning and compute the marginal probability of staying at home we obtain $P(\texttt{AtHome} = \text{yes}) = 0.532$. During the analyzed period the chances of school being closed for students is higher than 50%. On figure 6 we can see that probability all over the world. Most of Europe and Central Africa have a low probability of distant learning, however a lot of densely populated areas, such as India and USA, display an higher probability of being forced into distant learning.

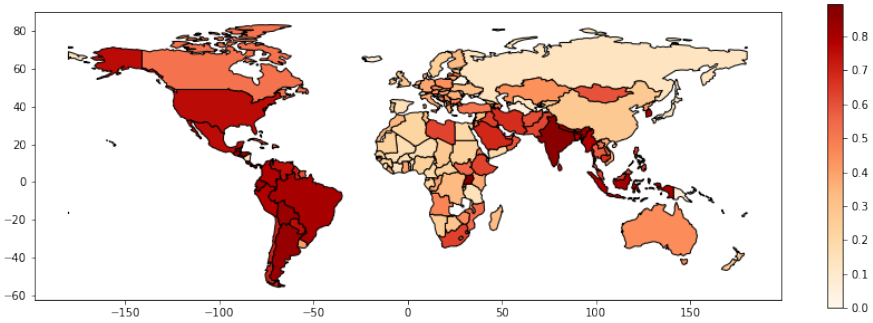

Figure 6: Probability of being at home in each country

If we perform the same analysis as we did before, grouping by geographical position or income level (respectively on figure 7 and 8), we can see that the probability of being forced at home are similar all over the world, with the exception of Latin America and Central and Southern Asia. Quite unexpectedly

countries from different income groups shows a similar probability of distant learning. That is because the period of analysis ranges from the initial phase of the pandemic to the end of 2021. At the beginning of the new school year, around September and October 2020, many countries brought students back into school without any limitation. Indeed, if we look at the probability of a school being closed due to COVID-19 (on figure 9), we see that school closure changes over time. Probability quickly ramps up to local maxima, $P(AtHome = yes) \approx 0.9$, during the first pandemic round of March 2020, and oscillates between $P(AtHome = yes) = 0.3$ and $P(AtHome = yes) = 0.5$ after June 2020. This gives us a clear insight on the fact that, even if some areas were more affected than other, globally the vast majority of the students experienced some type of school closures.

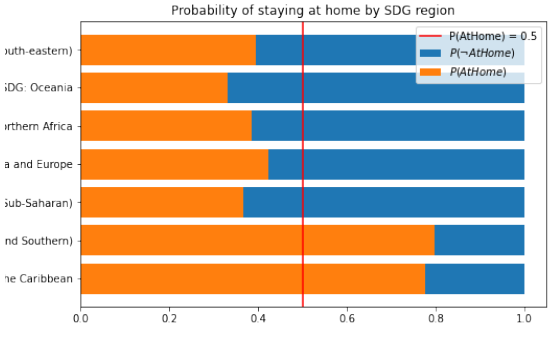

Figure 7: Probability of being at home by SDG group

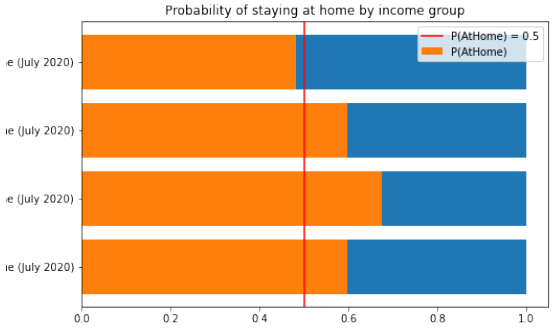

Figure 8: Probability of being at home by WB group

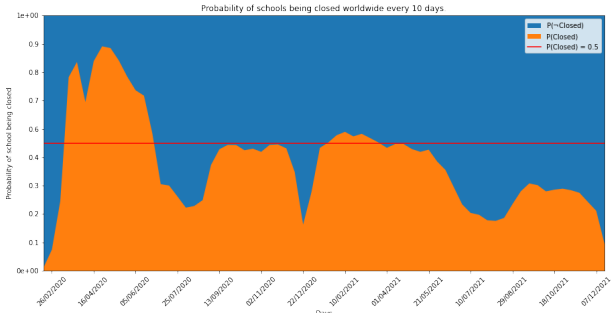

Figure 9: Probability of school being closed over the analyzed period.

Moreover, since we have different distributions of students between different grades all over the world, we also expect a specific subset of the population on being affected more. Indeed on table 5 we can see how the probability of being at home is conditioned by the education grade. The most affected grade is primary school. This is mostly because less developed countries have a lower attendance rate in higher educational grades than primary school, as we can see on figure 10.

Education grade	$P(AtHome \mid Education\ grade)$
Primary	0.5264
Lower secondary	0.2570
Upper secondary	0.2167

Table 5: Probability of being at home conditioned by the educational grade

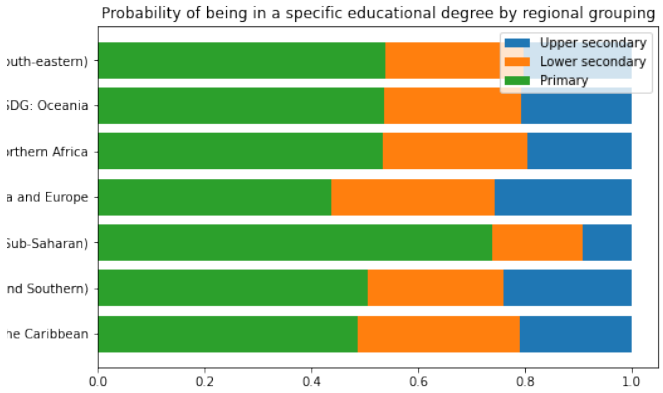

Figure 10: Probability of being in a particular education grade by SGD group

Mental health conditions

Lastly we will analyse the impact that distant learning had on students. Many students reported higher level of anxiety, depression and sleep disorders during and after the lockdown restrictions, when compared to the situation before COVID-19 outbreak(2). On table 6 the probabilities health condition problems when students needs to undergo distant learning are compared to the probabilities before COVID-19. It's clear how distant learning affected mental health conditions of many students during the COVID-19 outbreak. If we take into account the data we showed before, it's even more visible how it mostly affected low-income areas where mental healthcare is not always available. In addition, as we've seen before, students in primary schools are among the most affected ones by school closures. When joined with table 6 data, it raises a big question: how much are those students going to be affected in the following years? On figure 11 the probability of mental health condition is analysed throughout the whole period in which data is available. Depression and anxiety are not independent one from another, when experiencing one it's probable to experience also the other one, and it's difficult for experts to recognize those symptoms, let alone when asked to self-assess it, as done in (2). For this reason the shown plot should be considered as an extreme approximation of reality. It still gives a good measure, however, of the state of things. During the first pandemic round nearly half of the worldwide students might have experienced some kind of mental illness, was it mild depression or crippling anxiety.

Education grade	$P(HealthCondition \mid \neg AtHome)$	$P(HealthCondition \mid AtHome)$
Anxiety	0.2403	0.4286
Depression	0.297	0.513

Table 6: Probability of incurring into health condition when forced into distant learning

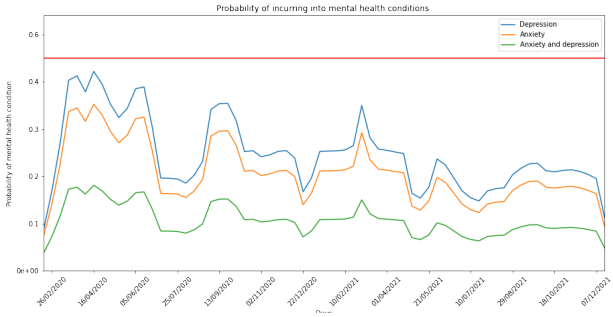

Figure 11: Probability of mental health conditions during the analysed period

(Unsolicited) Personal thoughts

While economical and social outcomes of COVID-19 have been widely analyzed, mental health issues on students might have been severely underestimated. Every student life have been completely twisted during the lockdown, and when the situation became more controllable, they've been asked to get back to their previous routine without taking into account how challenging that would've been. We surely don't have enough data to approximate the actual probabilities of mental health conditions. But it's clear how the data at our disposable is enough to warn about the problems that might end up under-analyzed. UNESCO and UNICEF have been constantly worried about the burden that students had to take when getting back to their normal lives, yet not much have been done lately.
The use of Bayesan Networks, and of Artificial Intelligence and technology in general, won't help fixing those problems or propose any easy solution. This simple analysis is yet to be considered comprehensive or fully representative of reality. The results obtained are enough to raise awareness on such a crucial and delicate subject. With the amount of data that gets stored daily, in particular the one collected during the lockdown, much more deep and insightful information can be retrieved. The data on how students communicated between each other, used online platforms, looked at different content and even performed online shopping before and after the lockdown is not only valuable in building better recomender and profiling tools, but it's crucial in understanding social and psychological issues that might have been considered superficial.

References

[1] Statistics on Radio | United Nations Educational, Scientific and Cultural Organization, Apr. 2013. [Online; accessed 4. May 2022].

[2] Lyubetsky, N., Bendersky, N., Verina, T., Demyanova, L., and Arkhipova, D. IMPACT of distance learning on student mental health in the COVID-19 pandemic. *E3S Web Conf. 273* (2021), 10036.