This document contains the list of sources that were used to create the Sub-National Covid-19 Incidence and Determinants Dataset. The original datasets had different levels of granularity and time frequencies. Our aim was to obtain regional-level weekly measures for each variable of interest. Some variables were only measured yearly, so we took the most recent data available, while others were only available at the nation-level, so we assigned national estimates to sub-national regions. Other data were available at more granular levels than our objective: they were measured with greater time frequency or for smaller geographic regions than those of interest. In these cases, data were aggregated using the most suitable measures (averages, medians, maxima, etc.).

Our choice of the geographical scope for our analyses was largely determined by the level of granularity we had in the databases containing information on our outcome variables, which are explained in greater detail below. For most of the European data, our unit of analysis is NUTS2 regions; however, for some countries we only have observations at the NUTS1 level. For non-European countries, data was gathered at the first administrative sub-national level. For India, Mexico and the USA, we had data at the state-level and for Canada we had data by provinces. In Chile, Japan and Peru, data was gathered at the administrative region level, and in Russia the data was available at the federal level.

Data on COVID cases and deaths

To collect the covid cases, different data sources from GitHub repositories were used:

1. The Joint Research Center (JRC) COVID-19 Repository retrieved from: https://github.com/ec-jrc/COVID-19
2. The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved from https://github.com/nytimes/covid-19-data
3. Berry, I., O'Neill, M., Sturrock, S. L., Wright, J. E., Acharya, K., Brankston, G., Harish, V., Kornas, K., Maani, N., Naganathan, T., Obress, L., Rossi, T., Simmons, A. E., Van Camp, M., Xie, X., Tuite, A. R., Greer, A. L., Fisman, D. N., & Soucy, J.-P. R. (2021). A sub-national real-time epidemiological and vaccination database for the COVID-19 pandemic in Canada. *Scientific Data*, 8(1). doi: https://doi.org/10.1038/s41597-021-00955-2. Retrieved from: https://github.com/ccodwg/Covid19Canada
4. Dong, E., Du, Hongru & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), doi: https://doi.org/10.1016/S1473-3099(20)30120-1. Retrieved from: https://github.com/CSSEGISandData/COVID-19.
5. The COVID-19 Mexico Time Series produced by Mario Rovero Zavala and retrieved from: https://github.com/mariorz/covid19-mx-time-series.

Some of the data contained missing information on cumulative cases or deaths, which we imputed either through interpolation or using national-level data from the Oxford Covid-19 Government Response Tracker (explained in greater detail below). The raw data was manipulated to create new indicators, such as weekly indicators of COVID cases and deaths.

Data on excess of deaths

We created a dataset that compares the mortality difference at the regional level for 5 countries and also provides the excess of deaths at the country level for the rest of the world from 3 different repositories:

1. The Economist's Tracker for Covid-19 Excess Deaths retrieved from: https://github.com/TheEconomist/covid-19-excess-deaths-tracker
2. The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved from: https://github.com/nytimes/covid-19-data
3. The Financial Times Coronavirus Excess Mortality Data. Retrieved from: https://github.com/Financial-Times/coronavirus-excess-mortality-data.

Data on airline mobility

Monthly airline mobility data were obtained from Sabre Travel Data, a private company dataset on airport traffic based on data collected directly from the airline industry (https://www.sabre.com/products/travel-data/). The original Sabre dataset included the full origin and destination matrix of passenger flows across airports in our areas of interest in the last quarter of 2019 and on a monthly basis for 2020.

We matched the airports in the Sabre dataset to the specific sub-national areas where they were located. Some airports of major cities or regions lied in neighboring administrative regions, so we matched them to the major region they served. Some areas in our data do not have airports, so they could not be matched to the Sabre dataset. Although these areas do not have airports, they will be likely exposed to air travel from airports in neighboring regions. However, given that we do not address other forms of travel explicitly in our analyses (train, private vehicle, etc.), we do not attempt to model their exposure to air passengers and we assign them null values for all air passenger flow and network metrics.

We computed simple measures of the absolute number of inbound and outbound air passengers for each region in our dataset for the last quarter of 2019 and for each month of 2020. We also computed the share of inbound passengers coming China specifically in the last quarter of 2019.

In addition to air passenger flow metrics, for the last quarter of 2019 and for each month of 2020 we computed network measures to capture the importance of each region in the network of international air travel. We created a directed graph where each node was an airport in the Sabre dataset, and a directed edge between two airports existed when someone in the Sabre database had purchased a ticket for a route connecting the two airports. Some indices were calculated as weighted measures computed by assigning different values to the edges based on either the number of passengers on the route, the kilometers between origin and destination, or the USD cost of the route. We computed the following measures:

1. **Betweenness** is the number of shortest paths on which a node lies. We computed measures based on the three sets of weights.
2. **Degree** is the number of edges originating in a node (out-degree) or pointing to the node (in-degree). We also computed the total degree, which is the sum of the two.
3. **Centrality** is the eigenvector centrality of a node, which is a relative measure of incoming connections scaled by the centrality of the nodes from which those connections originate. A high score means that a node is connected to nodes that also have high scores.
4. **Diversity** is defined as the Shannon entropy of the weights of its incident edges, using the three separate sets of weights.
5. **Hub scores and authority scores** are two symmetric concepts computed using Jon Kleinberg's algorithm that was originally developed for the context of web search. The authority score of a

node is the sum of the hub scores of all the nodes that have edges directly pointing to it. The hubscore of a node is the sum of the authority scores of all nodes to which it has edges pointing to.

Data on socio-economic, demographic and public health factors

Data on socio-economics, demographics, public health and co-morbidities were taken from:

1. OECD regional statistics: OECD (2020), Regional Statistics, https://data.oecd.org
2. Sorci, G., Faivre, B., & Morand, S. (2020). Explaining among-country variation in COVID-19 case fatality rate. *Scientific Reports*, 10(1), 1-11.
3. Diabetes and global ageing among 65-99-year-old adults: *Findings from the International Diabetes Federation Diabetes Atlas*, 9th edition
4. World Health Organization (2020) Global Health Observatory (GHO) Data.
5. NCD Risk Factor Collaboration (2017). "Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults". *Lancet*, 390:2627-2642.
6. NCD Risk Factor Collaboration (2017). "Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants". *Lancet*, 389:37-55.
7. NCD Risk Factor Collaboration (2017). "Repositioning of the global epicentre of non-optimal cholesterol". *Nature*. 582, pages73–77.

The OECD dataset was measured at the sub-national level, while all other resources were gathered at the national level.

Data on Government policy responses

Data on policy responses to the COVID outbreak were taken from the Oxford COVID-19 Government Response Tracker:

Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)." *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01079-8.

The dataset contains daily information on 19 indicators of government responses. The data is gathered daily and at the country level, except for data in the USA and the UK where information on sub-national territories is available. Each policy indicator is coded in two variables: one is an ordinal measure of its strictness and the other is an indicator of whether it was applied at the state level or at a regional level.

At the time of download, the dataset included indicators on three main areas of interest: containment and closure policies, health system policies and economic support policies. We focused on the first two and considered the following indicators: school closing, workplace closing, cancelling of public events, restrictions on gatherings, closing of public transport, stay at home requirements, restrictions on internal movement and international travel controls, public information campaigns, testing policy, contact tracing, emergency investment in healthcare, investment in vaccines, facial coverings and vaccination policy.

The Oxford Covid-19 Government Response Tracker also summarizes the various indicators into aggregate indices to provide general pictures of government activity. In our dataset, we included the Stringency index, which is a summary of the eight containment and closure policies along with the health system index of public information campaigns. The index is calculated as follows: for each component indicator, the daily

recorded value is divided by the maximum possible value that the indicator can take, then it is scaled by a factor of 0.5 if the policy was applied at the regional, rather than national, level. Then, a simple average of the nine transformed component indicators is computed. Additionally, we computed a revised version of the Stringency Index that did not include restrictions on internal movement.

We aggregated all the daily policy indicators and composite indices into weekly measures by taking their maximum weekly value. In other words, we considered the maximum recorded level of strictness for each policy as the weekly value. Whenever sub-national data was available, we matched at the sub-national level. For cases where only national data was recorded, we assigned national values to sub-national regions.

The Oxford Covid-19 Government Response Tracker also includes information on reported COVID cases and deaths measured at the national level and taken from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. We used this information to impute some missing values for our outcome variables.

Data on local mobility

Local mobility data were gathered from the Google COVID Mobility Reports, and retrieved from: https://www.google.com/covid19/mobility.

This dataset includes daily trends in mobility across different categories of places by region. Daily data is provided as a percentage change compared to a fixed reference point for each day of the week in each region. The reference represents the "normal" amount of mobility on a given day of the week and is calculated as the median mobility for each day of the week in the five weeks between January the 3rd and February 6th, 2020. The reference values of mobility never change in order to aid comparisons across weeks. We aggregated data into weekly estimates by taking the average, maximum and median percent change for each indicator in each week.

Data on Facebook Social Connectedness

The data were retrieved from:

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259-80. DOI: 10.1257/jep.32.3.259.

Originally, this data was a pairwise data frame that shows the SCI between any two regions in the world that may share a connection between any two users. From this data, we computed the measures of betweenness, centrality and diversity, as defined for the network measures of the air passenger flows. In addition to this, we computed the median, mean, minimum and maximum values of the Social Connectedness Index (SCI) for each area, as defined in Bailey et al. (2018).

Data on temperature and air quality

Data on temperatures were taken from Iowa Environmental Mesonet (IEM) compiled at the University of Iowa, and available at: https://mesonet.agron.iastate.edu/

The air quality data were obtained from the Air Quality Open Data Platform Worldwide COVID-19 dataset, available at: http://aqicn.org/data-platform/covid19/.

For regions that could not be matched, because for example they lacked weather stations, we matched information from the closest available region. Daily data were aggregated by computing the weekly maxima, minima and averages for each region.

Data on interpersonal distance

These data were gathered through graphical questions administered in surveys and published in the paper:

Sorowska et al. (2017), "Preferred interpersonal distances: A global comparison". *Journal of Cross Cultural Psychology*, 48(8). DOI: https://doi.org/10.1177/0022022117698039

Data on intergenerational cohabitation

These data were gathered in the European Union Labour Force Survey (EU-LFS) and were retrieved from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Age_of_young_people_leaving_their_parental_household&oldid=494351#Geographical_differences