Open Data Resources for Fighting COVID-19

Teodoro Alamo[*], Daniel G. Reina[†], Martina Mammarella[‡], Alberto Abella[§]

April 15, 2020

Abstract

We provide an insight into the open data resources pertinent to the study of the spread of Covid-19 pandemic and its control. We identify the variables required to analyze fundamental aspects like seasonal behaviour, regional mortality rates, and effectiveness of government measures. Open data resources, along with data-driven methodologies, provide many opportunities to improve the response of the different administrations to the virus. We describe the present limitations and difficulties encountered in most of the open-data resources. To facilitate the access to the main open-data portals and resources, we identify the most relevant institutions, at a world scale, providing Covid-19 information and/or auxiliary variables (demographics, mobility, etc.). We also describe several open resources to access Covid-19 data-sets at a country-wide level (i.e. China, Italy, Spain, France, Germany, U.S., etc.). In an attempt to facilitate the rapid response to the study of the seasonal behaviour of Covid-19, we enumerate the main open resources in terms of weather and climate variables.

CONCO-Team: The authors of this paper belong to the CONtrol COvid-19 Team, which is composed of different researches from universities of Spain, Italy, France, Germany, United Kingdom and Argentina. The main goal of CONCO-Team is to develop data-driven methods for the better understanding and control of the pandemic.

Keywords: Covid-19, Coronavirus, SARS-CoV-2, Open data, Data driven methods, Seasonal behaviour, Government measures

[*]Departamento de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain (e-mail: talamo@us.es)

[†]Departamento de Ingeniería Electrnica, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain (e-mail: dgutierrezreina@us.es)

[‡]Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, Turin, Italy (e-mail: martina.mammarella@ieiit.cnr.it).

[§]FIWARE Foundation. Germany (e-mail: alberto.abella@gmail.com)

Contents

1 Introduction

We provide in this document a survey on the main open-resources for addressing the Covid-19 pandemic from a data science point of view. Since the number of institutions and research teams working nowadays against the virus is growing at a very fast pace, it is impossible to provide an exhaustive list of all the meaningful open-data providers. At a global world scope, it is easier to identify the relevant open-data resources. However, the enumeration of the regional institutions providing local information is so extensive that we address it specifically only for some countries (like China, Italy, Spain, and the U.S, among others). We focus on the variables that have an effect on the evolution and control of the disease at a global and regional scale. That is, we do not cover in this document the data specifically related to medical treatments, vaccines, etc. We do provide open resources for the number of hospitalized cases, intensive care units (ICU) cases, number of tests,etc. These variables are very relevant to monitor the evolution of the pandemic and also to evaluate the actions taken by the decision-makers.

With this document, we try to make accessible many significant open-data resources on Covid-19 for the scientific community. In many situations, identifying adequate sources is difficult, especially for non-expert data scientists. For example, the GitHub repository contains many meaningful data-sets of global and regional scope, but it might be challenging to discover them without adequate guidance. Besides, the reliability of the data source provider can be a concern. Therefore, this paper is aimed at providing a big picture of the available data source providers for analyzing Covid-19 propagation. We have tried to find stable and reliable resources so that the utility of this paper endures in time.

The paper is organized as follows. We first analyze in Section 2 the different variables that have a significant effect on the evolution and control of the epidemic (demographics, mobility, weather conditions, government measures, etc.). The opportunities that open data resources on Covid-19 offer to fight the pandemic are highlighted, from a data-driven perspective, in Section 3. Different limitations and inaccuracies of the currently available sources, along with the difficulties encountered when using them in a data-science context are discussed in Section 4. The most relevant open data institutions at a global scale, addressing the Covid-19 pandemic, are enumerated in Section 5. More functionally, in Section 6, we identify open source communities that facilitate access to the required data. In Section 7 we identify open-data sets related to specific Covid-19 variables at a global and regional scale. The open access to auxiliary variables of interest to model specific aspects of the pandemic, like seasonal behaviour or local mortality rate is described in Section 8. The open e-learning on Covid-19 is illustrated by means of some examples in Section 9. We finish the paper with a section of conclusions.

2 Covid-19

Coronavirus disease 2019 (Covid-19), technically known as SARS-CoV-2, is an infectious disease that was first identified in December 2019 in Wuhan, the capital of China's Hubei province. The World Health Organisation (WHO) declared the 201920 coronavirus outbreak a Public Health Emergency of International Concern on 30 January 2020 and a pandemic on 11 March 2020.

The virus is mainly spread during close contact and by small droplets produced when those infected cough, sneeze or talk. These small droplets may also be produced during breathing; however, they rapidly fall to the ground or surfaces and are not generally spread through the air over large distances. People may also become infected by touching a contaminated surface and then their face. Although animals can be infected with Covid-19, there is not any evidence to suggest the transmission of the virus from them to human beings.

The virus can survive on surfaces for up to 72 hours such as plastic and stainless steel [42]. It is most contagious during the first three days after onset of symptoms, although spread may be possible before symptoms appear and in later stages of the disease [4]. The time from exposure to onset of symptoms is typically around five days but may range from two to 14 days.

Recommended measures to prevent infection include frequent hand washing, social distancing (maintaining physical distance from others, especially from those with symptoms), covering coughs and sneezes with a tissue or inner elbow and keeping unwashed hands away from the face. The use of masks is recommended for those who suspect they have the virus and their caregivers. On 29 January, the WHO provided some guidelines on the use of masks [21]. Nevertheless, recommendations for mask use by the general public vary. In addition, the use of plastic gloves has also been recommended. Currently, there is no vaccine or specific antiviral treatment for Covid-19. Management involves treatment of symptoms, supportive care, isolation and experimental measures.

2.1 Covid-19 cases in the world

To monitor the spread of Covid-19, the different regional institutions are measuring the number of confirmed cases, deaths, recovered, hospitalized cases, intensive unit care (IUC) cases, etc. Because of the incubation period, all these variables are related with the number of infected cases in a delayed way. The main objective is to calculate the basic reproductive number R_0 of the well-known Susceptible-Infectious-Recovered (SIR) model. Several works have calculated R_0 for some outbreaks of specific locations. The estimated values are ranging from 2 to 3 [27]. However, only limited data have been used in the majority of works. Achieving an accurate model of the reproduction of the virus is a challenging task, which involves many variables. Unfortunately, the open data sets available nowadays are locally collected, imprecise with different criteria (lack of standardization on data collection), inconsistent data models and incomplete.

One of the limitations of the data sets is that only cases confirmed by a laboratory test are included in them. The standard method of diagnosis is by real-time reverse transcription polymerase chain reaction (rRT-PCR) from a nasopharyngeal swab. The infection can also be diagnosed from a combination of symptoms, risk factors and a chest CT scan showing features of pneumonia. Thus, the infected cases without a positive laboratory test are not considered confirmed cases, on a general basis, in the time-series data available in the different open-source repositories. The same problem can be encountered when analyzing death cases. In many situations, only the ones that were previously confirmed by a laboratory test are included in the data sets.

There are relevant variables that are not properly measured. For example, the fraction of infected non-asymptomatic cases in a given population can be only estimated by means of massive tests, which are rarely being done because of lack of resources. Some exceptions

are the massive tests done in small towns (for example in the north of Italy) in which the tests indicated that the fraction of asymptomatic cases in the population could be significant (comparable or even larger than the symptomatic cases). Therefore, asymptomatic cases play an important role in virus transmission [4]. Furthermore, important inaccuracies have been reported on the use of fast tests. It is an important issue since their massive usage can improve the counting of real cases.

The above limitations on the available data-sets have to be taken into consideration in any data-driven method to model or forecast the future spread of the pandemic.

2.2 Covid-19 mortality

Being able to predict the number of patients that will develop life-threatening symptoms is important since the disease frequently requires hospitalisation (and even Intensive Care Unit admission) and challenges the healthcare system capacity [19]. One of the most important ways to measure the burden of Covid-19 is mortality. The probability of dying when getting infected depends on different factors [47], [34], [25]:

- Demographics [33]: Age, gender, prevalence of diabetes, high blood pressure, obesity, etc.
- Health System [25]: Availability of artificial respiration equipment, intensive care units, specialized medical surveillance and treatment, etc.

On the one hand, several studies have reported a higher level of mortality for older people [33] and even more aggravated in men. Thus, protection strategies should be focused on more vulnerable age and gender groups.

On the other hand, the capacity of each regional Health System to cope with the pandemic is time-varying. Most of the countries that had already suffered in a severe way the pandemic had their hospitals and physicians overwhelmed by the numbers of critical cases (e.g. Italy, Spain, the U.S.) [10]. The main objective in the control of the disease is to prevent the saturation or overload of the health system because it will be directly translated into a significant increase in mortality.

2.3 Seasonal behaviour of Covid-19

Many respiratory viruses have a seasonality because lower temperature and lower humidity help facilitate the transmission of the virus. There is no clear evidence that Covid-19 is going to behave seasonally, reducing its transmission in summer. In the summer season in the Southern hemisphere, in some regions of South America and Australia, significant Covid-19 outbreaks have been already reported. In [39] the authors show that in March 2020, the areas with significant community transmission of Covid-19 had distribution roughly along the 30-50 N corridor at consistently similar weather patterns consisting of average temperatures of 5-11C, combined with low specific (3-6 g/kg) and absolute humidity (4-7 g/m3). In [43], the authors study the relationship between temperature, humidity and the transmission rate of Covid-19. They used data collected from all the cities in China with more than 100 cases. The authors use a lineal regression framework as a model. Results indicate that one-degree Celsius increase in temperature and one per cent increase in relative humidity lower R_0 by

0.0225 and 0.0158, respectively. The authors developed a web application[1], where R_0 values for major worldwide cities can be obtained from temperature and humidity.

2.4 Actions to control Covid-19 pandemic

For the control community, the different confinement strategies that can be implemented by a government clearly constitute control inputs to the system [3]. Public health and social measures to slow or stop the spread of Covid-19 are being implemented, with different intensities, worldwide.

However, these are not the unique actions that a government can undertake in order to control the pandemic. For example, forcing the population to wear masks or scarves, and plastic gloves might have an inhibitory effect on the spread of the virus [13] and has not a significant impact on the economy (provided masks are produced at large scale).

From a control point of view, the objective is twofold. On the one hand, it is important to assess the effectiveness of the different measures against the spread of the virus. On the other hand, actions should be planned in advance in order to mitigate the effects of the pandemic on the health system, the economy and society.

It is not simple to determine the effect of the possible anti-measures to be undertaken by the regional governments. This is not a simple task for several reasons: (i) several inhibitory actions are generally implemented simultaneously. Therefore, it cannot be evaluated which one has more impact, ii) the efficacy of the anti-measures depends on a number of factors, like demographics and weather conditions of the specific region under consideration, iii) the available data is, in many situations, imprecise and incomplete. The difficulties in predicting the effects of the Covid-19 anti-measures on the regional evolution of disease is one side of the problem. Another one is the inherent time-delay system nature of the dynamics of the disease. The effects of the undertaken measures are observed only weeks later. Another issue is the level of fulfilment of the confinement measures found in each country.

2.4.1 Social distancing

Following the emergence of a novel coronavirus (SARS-CoV-2) and its spread outside of China, Europe is now experiencing large epidemics. In response, many European countries have implemented unprecedented non-pharmaceutical interventions including case isolation, the closure of schools and universities, banning of mass gatherings and/or public events, and most recently, wide-scale social distancing including local and national lockdowns. In [9], authors use a semi-mechanistic Bayesian hierarchical model to attempt to infer the impact of these interventions across 11 European countries. They assume that changes in the reproductive number a measure of transmission - are an immediate response to these interventions being implemented rather than broader gradual changes in behaviour. In particular, this model estimates these changes by calculating backwards from the deaths observed over time to estimate transmission that occurred several weeks prior, allowing for the time lag between infection and death.

One of the key assumptions of the model is that each intervention has the same effect on the reproduction number R_0 across countries and over time. This allows leveraging a greater amount of data across Europe to estimate these effects. It also means that these

[1] `http://covid19-report.com/#/r-value`

results are driven strongly by the data from countries with more advanced epidemics, and earlier interventions, such as Italy and Spain. The main conclusion of this research was that it is critical that the current interventions remain in place and trends in cases and deaths are closely monitored in the coming days and weeks to provide reassurance that transmission of SARS-Cov-2 is slowing.

Many governments around the world closed the educational institutions in an attempt to contain the spread of the Covid-19 pandemic, impacting over 91% of the worlds student population [40].

Another important aspect has been tackled by the New York Times: how income affects peoples abilities to stay home and practice social distancing [41]. Wealthier people not only have more job security and benefits but also may be better able to avoid becoming sick.

2.4.2 Mobility measures

The authors in [44] use the Baidu Mobility Index, measured by the total number of outside travels per day divided by the resident population, to find that reducing the number of outings can effectively decrease the new-onset cases; a 1% decline in the outing number will reduce about 1% of the new-onset-cases growth rate in 7 days (one serial interval). A first quantitative assessment of the impact of the Italian Government on the mobility and the spatial proximity of Italians, through the analysis of a large-scale dataset on de-identified, geo-located smartphone users can be found in [35].

Technology can play an important role to obtain mobility measures [32]. Nowadays, everyone has a mobile phone equipped with a number of sensors, including GPS, that are able to collect data about people mobility. Furthermore, the Internet and mobile phone operators can use their telecommunications towers to gather mobility patterns. Of course, citizen privacy is an issue that has to be taken into consideration for data anonymization.

2.4.3 Massive tests

The distinction between diagnosed and non-diagnosed is important because non-diagnosed individuals are more likely to spread the infection than diagnosed ones, since the latter are typically isolated, and can explain misperceptions of the case fatality rate and of the seriousness of the epidemic phenomenon [19].

3 Data driven ways to fight the pandemic

Currently, the majority of data available on Covid-19 is used for describing the pandemic in terms of reports and visualizations[2]. Although these techniques are useful to highlight the magnitude of the crisis, they are not enough for mitigating the problem. Also, these are insufficient for decision-makers to anticipate the response to the virus propagation and evaluate the effectiveness of the implemented actions. For this reason, it is obvious that more efficient approaches are needed rapidly to i) model and forecast the spread and the consequences of the pandemic and ii) evaluate mitigation approaches that have been carried out. Data-driven models, see for example [19], [20], can be such solution [46],[15]. Many

[2]For example, `https://againstcovid19.com/singapore/dashboard`

data-based techniques can be applied. Nevertheless, regression models, as in [6], are the main tools for forecasting the effect of Covid-19, ranging from classical machine learning approaches like linear regression [36] and Gaussian processes [16] to sophisticated models based on neuronal networks [11]. Variable screening can also help to detect the parameters that impact more on the propagation of the virus. These techniques require sufficient and high-quality data to provide a good estimation. Depending on the model used, the quantity of data can vary notably from hundreds to millions of samples. Moreover, a wide variety of data is necessary for accomplishing a good model of a complex system like the Covid-19 pandemic. Data from different disciplines is required, which hinders the data collection task. We highlight three pillars of data-driven approaches for fighting Covid-19: i) informative variables for developing an accurate model, ii) objectives of the model: characterizing Covid-19 pandemic, epidemic models and forecasting, etc. and iii) its use for efficient decision making.

- **Wish-list of variables:** the list of variables is large since many aspects should be taken into consideration to develop accurate models. The considered variables can be divided into different categories according to their discipline[3].
  - Covid-19 variables: regional time series of the number of confirmed cases, suspicious cases, deaths, recovered, number of tests, hospitalized cases, ICU cases, isolated positive cases, etc. When possible, the data should be divided per gender, age range, etc.
  - Geographic variables: Locations of Covid-19 variables. The locations can be obtained from names (i.e. countries, cities. etc.) and/or GPS coordinates: longitude and latitude.
  - Demographic variables: Population and density of population by location. These variables are required for normalization of the rest of the variables. Also, the age structure of the population, the prevalence of secondary health conditions related to higher Covid-19 mortality, etc.
  - Health system variables: Total number of ICU beds, number of doctors and nurses, personal protective equipment (PPE), respirators, number and types of tests.
  - Government measures: Social distancing, movement restrictions, lockdowns, etc.
  - Weather variables: Temperature, relative humidity, radiation, among others.
- **The use of data in epidemic and forecasting models:** By using the aforementioned variables, different models can be developed to anticipate the response to the propagation of Covid-19. Examples of forecasting analysis:
  - Estimation of the infected population.
  - Forecast of impact in health-system.
  - Assessing the impact in terms of mortality.
  - Analysis of seasonal behaviour.

[3]The following list can be incomplete since other disciplines and variables can also be included.

- **Decision making:** The final objective of the data-driven models is developing useful tools for helping governments and institutions to anticipate the response to the Covid-19 propagation and evaluate their actions.
  - Assessing the effectiveness of the measures.
  - Planing ahead government actions.

4 Limitations and challenges raised by the available data

There exist different issues that can hinder the use of open data to address the challenges raised by the Covid-19 pandemic:

4.1 Variety

The different sources and variables required to undertake a given analysis is often addressed assembling several data sets into a single one. Although the increased quantity of data sources presents new opportunities, working with such a variety of data reinforces the validity challenges [29]. Another issue is that the sources of data are coming from a wide range of disciplines that can be familiar with very different formats and data representation. For instance, some available APIs to get data on Covid-19 provide JSON files. This format is widely used in computer science in web applications. However, for instance, mathematicians or epidemiologists cannot be familiar with such format.

4.2 Time-varying nature

The needs of the outbreak require immediate response, which translates in obtaining the latest information available. This raises some important challenges. For example, government measures are changing rapidly. Often information is outdated by the time it has been identified. The number of countries implementing or amending measures increases daily [2]. The daily availability of the data can be an issue for working with multiple data sources simultaneously.

4.3 Confirmed cases is not a reliable metric

The World Health Organization defines, in its global Covid-19 surveillance document, a confirmed case as a person with laboratory confirmation of Covid-19 infection, irrespective of clinical signs and symptoms. At the outbreak of the pandemic, the access to massive tests was very limited, and often, only a reduced fraction of the hospitalized cases were tested at a laboratory level. Furthermore, very few data sets provide information about the number of suspected cases.

Even under the hypothesis that everyone with minor symptoms is tested, this would only provide an estimate of the symptomatic cases of the disease. The study of the fraction of asymptomatic cases is an active field of research (e.g. [28]) not only because it is key to the estimation of the total number of infected cases, but because it plays a fundamental role in the spread of the virus.

4.4 Mortality rate is difficult to estimate

During the most severe periods of the virus spread in a country, the number of death cases reported by the administration differs in many situations considerably from the real one because only the deaths with previous laboratory confirmation of the disease are included. Thus, the study of national death registers suggests that there are notably and unexpected increases in death rates according to the historical numbers. For instance in New York City, it has been reported 5330 more deaths than expected in the last month[4], only 3350 of these can be accounted for Covid-19 reasons. These figures suggest that there exist undercounting on the real number of deaths. Another example can found in Spain, where the *"Sistema de Monitorizacin de la Mortalidad diaria (MoMo)*[5] system registers the total number of deaths under any circumstances. The last report on April 7th indicates an increase of more than 50% of unexpected deaths in the last month. Such increment is even more significant in men, where it reaches more than 60%.

The mortality rates are much more difficult to estimate since the estimates are often based on the number of deaths relative to the number of confirmed cases of infection, which can be a small fraction of the real ones [5]. Consequently, mortality rates comparisons between countries make compulsory the implementation of correcting factors based on the estimation of Covid-19 infected cases and deaths non registered by the respective administrations.

Also, when considering the increase of mortality due to saturation of the health care system, one has to take into consideration the fact that the patients who die on any given day were infected much earlier, and thus the denominator of the mortality rate should be the total number of patients infected at the same time as those who died [5].

Another important parameter to evaluate mortality is to have stratified data according to age groups. However, such information is not provided by the majority of data sources.

4.5 Not availability of individual case data

In order to better understand the disease, and improve the models and strategies to fight Covid-19, every case should be tracked with its own timeline. That is, for each case, relevant information about when symptoms appeared, medical treatments, evolution, degree of isolation, etc. should be available on a country-wide level. This data should be published anonymously, with a de-identification process to prevent personal identity from being revealed. The data, and the time corresponding to the change of each individual, should be published by an official source in a structured way, at least, with daily frequency. This is the opinion of many experts and members of the open-source community[6].

An effort of obtaining individual case data can be found in [31]. The authors carried out a survey of 24 questions related to the impact of Covid-19 (Covid19Impact) on citizens in Spain[7]. The survey was responded to by 146.728 participants over a period of fewer than two days (44 hours). The questions were about social contact behaviour, financial impact, working situation and health status. The results of the survey show the negative impact of

[4] `https://www.nytimes.com/interactive/2020/04/10/upshot/coronavirus-deaths-new-york-city.html`

[5] `https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/MoMo/Paginas/`

[6] See, for example, `https://github.com/jgehrcke/covid-19-germany-gae`.

[7] `https://survey123.arcgis.com/share/d29378b51fe8496d8dd77f08ce73973f`

Covid-19 on the life of citizens. It is a clear example of how the collaboration of the citizens can be relevant to gather information on the effects of Covid-19.

4.6 Changing and non-uniform criteria

Since the governments are continuously adjusting their response to the virus, it is not unusual to find out abrupt changes in the trend of a time series because a new methodology has been implemented. For example, on 12th February, a sudden spike in new cases in China with Covid19 (15.152 new cases) was due to changed diagnosis method: a combination of SARSCoV2 nucleic acid test and clinical Covid19 features [45].

Another relevant issue is that regions in the same country may provide data under the same label, but with a different meaning. A good example is the number of UIC cases; there might be regions reporting the accumulative number of confirmed cases that required these units, and others the number of UIC units used by Covid-19 patients. Something similar happens with the number of laboratory tests. Sometimes they refer to the total number of tests carried out and sometimes to the number of individuals that were tested (in many situations, the sources do not describe accurately the meaning of the counts).

4.7 Changing data-base structure and locations

The open-data sources on Covid-19 are constantly improving. In order to provide more meaningful information, new variables are incorporated into data sets. This translates into a change in the structure of the data, which requires adjusting the code to download and process the information. When regional data is collected from the official open-data portals of different countries, a surveillance effort is required to keep track of the different modifications. In many situations, the new data-files appear in different locations with different names.

4.8 Government transparency

There are important differences in the ways that the governments are reporting the data related to Covid-19. Furthermore, there are some concerns about the transparency of countries regarding the data provided.

5 Open data institutions providing worldwide Covid-19 data

Some institutions are providing daily reports on the evolution of the pandemic in the world. They are of different nature: global institutions, European Union institutions, universities, newspapers, etc. We enumerate here the most relevant ones, at least from our experience. We highlight the ones that provide updated information on a regular basis in the open-data repository with easy access. Some of the enumerated institutions are making a great effort to provide consolidated data, describing in a rather exhaustive form, the sources and limitations of the provided datasets. We describe in this section the nature and characteristics of the

information provided, without detailing the specifics of the data sets that can be recovered from them (this is done for relevant data sets in subsequent sections).

5.1 World Health Organization (WHO)

The primary role of the World Health Organization (WHO) is to direct international health within the United Nations' system and to lead partners in global health responses. In the framework of the pandemic Covid-19, the WHO is providing continuous updates about the current situation all around the world[8]. In [22], the WHO provides guidelines to follow in the privacy of our house as well as in public, Q&A pages on the most common questions about the virus, how it spreads and how it is affecting people worldwide, and it addresses myth busters related to Covid-19, in order to provide a reliable source of information (see [23]).

5.2 Johns Hopkins University (JHU)

Johns Hopkins experts in global public health, infectious disease, and emergency preparedness have been at the forefront of the international response to Covid-19 (see `https://coronavirus.jhu.edu/`). This university provides a daily update on the global map of the pandemic. The data set provided by JHU (see sub-subsection 7.1.1) is one of the most frequently used by researchers and journal media.

5.3 University of Oxford

The Blavatnik School of Government is a department in University of Oxford. They are working on to understand the Covid-19 pandemic and the policy responses we see around the world. One of their projects related to the study of Covid-19 is tracking what governments around the world are doing, and how they compare to others[9].

Regarding the comparison of confinement strategies developed by governments, they have created a common index named Stringency Index. This index is based on data obtained by the Oxford Covid-19 Government Response Tracker (OxCGRT), which systematically collects information on several different common policy responses governments have taken.

5.4 European Union

The European Union Open Data Portal[10] gives access to open data published by EU institutions and bodies. The European Data Portal (EDP) acts as single access to point to open data that is published by national open data portals and institutions in the EU Member States and additional countries. There are numerous datasets on EDP that reference "covid" or "corona". Also, less specific data sets describing former health infections, epidemics or pandemics are also provided[11].

[8] `https://www.who.int/westernpacific/emergencies/covid-19`.

[9] Further information on the actions developed can be found at: `https://www.bsg.ox.ac.uk/news/coronavirus-research-blavatnik-school`.

[10] `https://data.europa.eu/euodp/es/data/`

[11] `https://www.europeandataportal.eu/en/highlights/covid-19`.

5.4.1 Joint Research Centre

The Joint Research Centre (JRC) is the European Commission's science and knowledge service[12] which employs scientists to carry out research in order to provide independent scientific advice and support to EU policy.

5.4.2 European Center for Disease Prevention and Control

The European Center for Disease Prevention and Control (ECDC), established in 2004 after the 2003 SARS outbreak and located in Solna, Sweden, is an independent agency of the European Union (EU) whose mission is to strengthen Europe's defences against infectious diseases. ECDC publishes numerous scientific and technical reports covering various issues related to the prevention and control of Transmission (medicine) communicable diseases. Towards the end of every calendar year, ECDC publishes its Annual Epidemiological Report, which analyses surveillance data and infectious disease threats. As well as offering an overview of the public health situation in the EU, the report offers an indication of where further public health action may be required in order to reduce the burden caused by communicable diseases. As other organizations, ECDC is closely monitoring the Covid-19 pandemic, providing risk assessments, public health guidance, advice on response activities to the EU Member States and the EU Commission, and daily-updated data on current outbreak [14].

For EU level surveillance, ECDC requests EU/EEA countries and the UK to report laboratory-confirmed cases of Covid-19 within 24 hours after identification. This is done through the Early Warning and Response System (EWRS).

5.4.3 European Centre for Medium-Range Weather Forecasts

The European Centre for Medium-Range Weather Forecasts (ECMWF) is an independent intergovernmental organization supported by 34 states based in Reading [7]. ECMWF is both a research institute and a 24/7 operational service, producing and disseminating numerical weather predictions to its Member States, Co-operating States and the broader community. ECMWF also archives data and makes this available to authorized users. Some data is also made available under licence, and some are publicly available.

5.5 United Nations (UN)

Good examples of open-data provided by the United Nations are reported in [30]. Moreover, [24] contains the most up-to-date Covid-19 cases and latest trend plot. It covers China, Canada, Australia (at province/state level), and the rest of the world (at country level, represented by either the country centroids or their capitals) and the US at county-level.

5.6 The New York Times

The New York Times is releasing a series of data files with cumulative counts of Covid-19 cases in the United States, at the state and county level, over time. The time series data

[12] `https://ec.europa.eu/knowledge4policy/organization/jrc-joint-research-centre_en`.

is compiled from state and local governments and health departments. Since January 2020, The Times has tracked cases of coronavirus in real time as they were identified after testing. The data have been used to power maps and generate reports about the outbreak. The data begins with the first reported coronavirus case in Washington State on Jan. 21, 2020. The Times publishes regular updates to the data in a GitHub repository[13].

5.7 Our World In Data

Our World in Data (OWID) is an online scientific publication that focuses on large global problems such as poverty, disease, hunger, climate change, war, existential risks, and inequality. Covid-19 data provided by Our World in Data can be found at their open-data portal[14].

5.8 Google

The multinational technology company Google has developed a visual Covid-19 map, where also relevant information can be found worldwide and by country `https://google.com/covid19-map/`. The map is continuously updated, and the data is taken from Wikipedia[15]. They also present statistics about the number of confirmed cases, cases per 1 million of people (normalized data), number of people recovered, and deaths.

Another tool developed by Google that can be used to obtain data about Covid-19 is Google DataSet Search[16]. Numerous data sets can be found by the term Covid-19. The application allows users to filter the data sets by several fields such as last updated, download format, usage rights, topic, and accessibility.

5.9 ACAPS

ACAPS is an independent information provider (`https://www.acaps.org`). It is not affiliated to the UN or any other organization. The ACAPS Analysis Team is mainly dedicated to researching and analyzing global and crisis specific data. ACAPS was established in 2009 as a non-profit, non-governmental project with the aim of providing independent, groundbreaking humanitarian analysis to help humanitarian workers, influencers, fundraisers, and donors make better decisions. They provide regional reports on the pandemic, and additional information like description of the worldwide measures against the spread of the virus available at `https://www.acaps.org/what-we-do/reports` and in [2].

5.10 Organization for Economic Co-operation and Development (OECD)

The Organisation for Economic Co-operation and Development (OECD)[17] is an international organization that, together with governments, policymakers and citizens, has the goal

[13] `https://github.com/nytimes/covid-19-data`.
[14] `https://ourworldindata.org/coronavirus`.
[15] `https://en.wikipedia.org/wiki/Template:2019%E2%80%9320_coronavirus_pandemic_data`
[16] `https://datasetsearch.research.google.com/`.
[17] `https://www.oecd.org`.

of establishing evidence-based international standards and finding solutions to a range of social, economic and environmental challenges. From improving economic performance and creating jobs to fostering strong education and fighting international tax evasion, they provide a forum and knowledge hub for data and analysis, exchange of experiences, best-practice sharing, and advice on public policies and international standard-setting. OECD provides different reports and data about the pandemic[18]: government actions, economic impact, etc.

5.11 MCR: Centre for Global Infectious Disease Analysis

The MRC Centre for Global Infectious Disease Analysis (MRC GIDA) is an international resource and centre of excellence for research and capacity building on the epidemiological analysis and modelling of infectious diseases, and to undertake applied collaborative work with national and international agencies to support policy planning and response operations against infectious disease threats.

The MCR presents reports on Covid-19 under five categories[19]: i) Weekly-forecasts, ii) resources, iii) information, iv) video updates, and v) publications.

Furthermore, in collaboration with several departments of Imperial College London (Imperial College Covid-19 Response Team) and Oxford University, they have developed a model[20] for estimating the number of infections and the impact of non-pharmaceutical interventions on Covid-19 in 11 European countries [16].

5.12 The Institute for Health Metrics and Evaluation (IHME)

The IHME is an independent global health research center at the University of Washington[21]. They have developed a model to determine the extent and timing of deaths and excess demand for hospital services due to Covid-19 in the US [10]. The work use data on confirmed Covid-19 deaths from WHO and local and national governments; data on hospital capacity and utilization for US states; and observed Covid-19 utilization data from different locations. The model for death rate is based on curve-fitting using Wuhan data, considered more stable. Then, from the projected death rates, the model estimates hospital service utilization using an individual-level microsimulation model. The model simulates the deaths, individuals requiring admission in hospital and date of admission, and the age-specific fraction of admissions requiring ICU care, by using the average age pattern. For the simulation of deaths, they used data from Italy, China, South Korea, and the US (they did not have enough data only from the USA at the time of developing the model). The main outputs of the model are: the number of deaths, bed and ICU occupancy, and ventilator use. The main conclusion of the work is that the epidemic will cause an unmanageable load on health systems in the US in the following weeks.

A web service, where the projections of the model can be determined for each country and for the following four months, is available [22]. The information provided is: i) hospital

[18] http://www.oecd.org/coronavirus/en/.

[19] https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/

[20] The updates of the model can be accessed at https://github.com/ImperialCollegeLondon/covid19model

[21] http://www.healthdata.org/

[22] https://covid19.healthdata.org/projections

resources needs, including the number of beds, the number of ICU beds, and ventilators; ii) the number of death per day, and iii) the total number of deaths.

5.13 Open Data Watch

Open Data Watch is a non-profit, non-governmental organization founded by three development data specialists (`https://opendatawatch.com/`). It monitors progress and provides information and assistance to guide the implementation of open data systems. The Open Data Watch team is experienced in the development of data management and statistical capacity-building in developing countries. They have collected data from different sources all around the world related to the Covid-19 pandemic. Indeed, to address the ongoing need for data-driven decision making, Open Data Watch has put together some articles, organized by the stages of the data value chain: availability, openness, dissemination, and use and uptake. These papers are updated as new information becomes available. These references and related links can be found in [12].

6 Open source communities

This section covers repositories of open source communities. Open sources communities are dedicated to joining people with similar interests. These have been widely developed in the software field, where many professionals and practitioners join their efforts to achieve bigger goals on software projects. These communities are playing a very active role in facilitating access to Covid-19 data sets from official open portals all over the world.

6.1 GitHub

GitHub is a company (a subsidiary of Microsoft) for hosting software development using Git. It provides control versions and project management, among other tools. Numerous open software projects are daily posted free of charge. Since Covid-19 outbreak, many projects, including the data used for these projects, have been posted. The majority of the data sets included in this paper can be obtained from GitHub.

Examples:

- Open Covid-19 Dataset: `https://github.com/open-covid-19/data`
- Covid-19 Data Processing Pipelines and Datasets: `https://github.com/covid19-data/covid19-data`
- `https://github.com/pomber/covid19`

6.2 Kaggle

Kaggle is a community for data scientist and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Regarding Covid-19 pandemic, the portal opens a new challenge weekly to work on Covid-19 data[23]. The challenge consists of forecasting confirmed cases and fatalities for the following week. Furthermore, some data analysis posts can be found for each competition[24]. The challenges opened up to the date (10/04/2020) can be found at.

`https://www.kaggle.com/c/covid19-global-forecasting-week-1`
`https://www.kaggle.com/c/covid19-global-forecasting-week-2`
`https://www.kaggle.com/c/covid19-global-forecasting-week-3`
`https://www.kaggle.com/c/covid19-global-forecasting-week-4`

7 Covid-19 Data Sets

This section presents the main available data sets that can be found on the Internet related to Covid-19. The section is divided into two parts. First, we present international data sets that provide global information related to the virus impact of each country, such as number of total/new confirmed cases and number of total/new confirmed death. Second, we include a number of regional data sets, where local information can be found. Although the information can be redundant on several data sets, we believe that it could be interesting to validate the developed models/analysis.

7.1 International data sets

In this section, we briefly introduce the institution that provides the data and includes the link (URL) to access it easily.

7.1.1 Johns Hopkins University Data Set

Johns Hopkins experts in global public health, infectious disease, and emergency preparedness have been at the forefront of the international response to Covid-19[25]. This university provides a daily update on the global map of the pandemic[26].

The Covid-19 data set of Johns Hopkins University can be downloaded in .csv format from the folder `/csse_covid_19_data/csse_covid_19_time_series` at Github repository `https://github.com/CSSEGISandData/COVID-19`. Six .csv files can be downloaded

- `time_series_covid19_confirmed_global.csv`
- `time_series_covid19_deaths_global.csv`
- `time_series_covid19_recovered_global.csv`
- `time_series_covid19_confirmed_US.csv`
- `time_series_covid19_deaths_US.csv`

[23] `https://www.kaggle.com/tags/covid19`.
[24] For example: `https://www.kaggle.com/frlemarchand/covid-19-forecasting-with-an-rnn`
[25] `https://coronavirus.jhu.edu/`.
[26] `https://coronavirus.jhu.edu/map.html`.

- `time_series_covid19_recovered_US.csv`

The global files refer to worldwide Covid-19 data. A reduced number of countries are further divided into regions (e.g. China and Australia). However, most of them, like Spain, or Italy, are not divided into regions. The US data .csv files correspond to the United States. All the data refer to accumulated cases. That is, cases up to the date of the row in which the data is consigned. Furthermore, the geographical coordinates of each region/country are also available.

JHU also provides information on how the data could be smoothed in order to analyze if the "curve has flattened". Flattening the curve involves reducing the number of new Covid-19 cases from one day to the next. This helps prevent healthcare systems from becoming overwhelmed. The data plots that can be found at `https://coronavirus.jhu.edu/data/new-cases` are obtained by means of a 5-day moving average that is obtained for each day by averaging the values of that day, the two days before, and the two next days. This approach helps to avoid major events (such as a change in reporting methods) from skewing the data.

7.1.2 Geographical Distribution of Covid-19 Worldwide (ECDC Data Set)

This dataset is sourced from the European Centre for Disease Prevention and Control (ECDC). The ECDC publishes full time-series data for the number of confirmed Covid-19 cases and deaths daily for countries around the world.

Each day the ECDC collects data from 6am to 10am CET and publishes this data via its Covid-19 dashboard[27]. This data set is then also made publicly available through downloadable files in different formats[28].

This data set can be also downloaded from the open-portal of *Our World in Data*:

- Total confirmed cases:
  `https://covid.ourworldindata.org/data/ecdc/total_cases.csv`
- Total deaths:
  `https://covid.ourworldindata.org/data/ecdc/total_deaths.csv`
- New confirmed cases:
  `https://covid.ourworldindata.org/data/ecdc/new_cases.csv`
- New deaths:
  `https://covid.ourworldindata.org/data/ecdc/new_deaths.csv`
- All four metrics:
  `https://covid.ourworldindata.org/data/ecdc/full_data.csv`
- Population data:
  `https://covid.ourworldindata.org/data/ecdc/locations.csv`

[27] `https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html`
[28] `https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-wo`

7.1.3 Covid-19 Testing (*Our World in Data* data set)

Our World in Data publishes useful information on how the different countries are carrying out laboratory tests to detect Covid-19 cases[29]. The data set on the number of tests carried out in the world is published by *Our World in Data* in the GitHub repository Owid/covid.19-data[30].

7.2 Examples of regional data sets

The majority of the following data sets can be found in GitHub, searching by the term Covid-19.

7.2.1 Argentina

The Argentina Ministry of Health provides daily updates on the Covid-19 spreading, including data on the number of infected people divided by regions[31].

7.2.2 Australia

The data corresponding to the different regions of Australia can be downloaded from the Johns Hopkins University Github repository[32]. The regional Australian Covid-19 data is integrated into the global time series .csv files. The available information is: i) confirmed cases, ii) death cases, iii) recovered cases. See Sub-Subsection 7.1.1 for further details.

7.2.3 China

The data corresponding to the different regions of China can be downloaded from the Johns Hopkins University Github repository[33]. The regional Covid-19 Chinese data is integrated into the global time series .csv files. The available information is: i) confirmed cases, ii) death cases, iii) recovered cases. See Sub-Subsection 7.1.1 for further details.

The National Health Commission of the People's Republic of China daily updates the available information on the situation in China[34].

7.2.4 Italy

The so-called *Dipartimento della Protezione Civile* (Civil Protection Department), i.e. the national body in Italy that deals with the prediction, prevention and management of emergency events, daily updates a GitHub repository, where the Covid-19 time series by regions, and provinces, can be downloaded (`https://github.com/pcm-dpc/COVID-19`). The csv file corresponding to the daily data of each of the 20 Italian regions[35] provides the number of

[29] `https://ourworldindata.org/covid-testing`.
[30] `https://github.com/owid/covid-19-data/tree/master/public/data/testing`.
[31] `https://www.argentina.gob.ar/coronavirus/informe-diario`.
[32] `https://github.com/CSSEGISandData/COVID-19/`.
[33] `https://github.com/CSSEGISandData/COVID-19/`.
[34] `http://www.nhc.gov.cn/xcs/xxgzbd/gzbd_index.shtm`
[35] Available at `https://github.com/pcm-dpc/COVID-19/raw/master/dati-regioni/dpc-covid19-ita-regioni.csv`.

confirmed cases, deaths, recovered, hospitalized, confined at home and ICU cases. Moreover, the daily number of tests is also available.

In addition, GEDI *Gruppo Editoriale*, a relevant Italian media conglomerate, provides a portal where those data are arranged in several interactive graphs, including also the impact on the local mobility[36].

7.2.5 France

The data corresponding to France is provided by the different regions and published by the Public France Health System at the official open data portal `https://www.data.gouv.fr/`. Among the different data sets available under the search of the term Covid, three are highlighted by the portal (organized into .csv files):

- Covid-19 Hospital Data: Hospitalized cases, IUC cases, deaths per department, region, gender and age range [17].
- Covid-19 Emergency Room Admissions: Hospitalized cases, IUC cases, deaths per department, region, gender and age range [37].
- Covid-19 Laboratory Tests: Number of positive and negative laboratory tests per department, gender, and age group [18].

French Covid-19 data sets in GitHub are:

- Covid-19 epidemic French national data[37].
- *Projet d'historisation du nombre de cas par rgion du Covid-19*[38].

7.2.6 Germany

The main official open data provider in Germany is the Robert Koch Institute[39], a public health institute in Germany. It provides, by means of a catalogue of infectious diseases[40], pertinent information on each one (e.g. SARS). In particular, risk assessments, the spread of the epidemic, epidemiological studies, etc. for the Covid-19 can be found[41]. Moreover, it provides daily situation reports of Covid-19 in Germany [26].

Covid-19 case numbers in Germany by state, over time, can be found at the GitHub repository `https://github.com/jgehrcke/covid-19-germany-gae`.

7.2.7 Paraguay

The official portal for data reports on Covid-19 is `https://www.mspbs.gov.py/reporte-covid19.html`. The data is provided by the Public Health system. The reports are stratified by age and gender, including data about the number of cases, number of deaths, and recovered people. Furthermore, data on Covid-19 spreading in Paraguay can be found at `https://github.com/torresmateo/covidpy` as GitHub repository.

[36] `https://lab.gedidigital.it/gedi-visual/2020/coronavirus-in-italia/`.
[37] `https://github.com/opencovid19-fr/data/blob/master/README.en.md`
[38] `https://github.com/cedricguadalupe/FRANCE-COVID-19`.
[39] `https://www.rki.de/EN/Home/homepage_node.html`.
[40] `https://www.rki.de/D.E./Content/InfAZ/InfAZ_marginal_node.html`.
[41] `https://www.rki.de/D.E./Content/InfAZ/S/SARS/SARS.html?nn=2386228`

7.2.8 South Africa

The information on Covid-19 spreading in South Africa can be found at https://github.com/dsfsi/covid19za as GitHub repository. The repository, named Covid-19 Data for South Africa is maintained and hosted by Data Science for Social Impact research group, led by Dr Vukosi Marivate, at the University of Pretoria.

7.2.9 Spain

The regional Covid-19 Spanish data is collected by the Spanish government. It is available at the national open data portal (https://datos.gob.es/). Different health data sets can be searched at its open data catalogue[42]. The specific search "Covid" provides data sets related to the global Spanish data classified into regions (e.g. *Evolucin de enfermedad por el coronavirus (Covid-19)*) or specific of a particular Spanish region (e.g. *Evolucin del coronavirus (Covid-19) en Euskadi*). In the GitHub repository https://github.com/datadista/datasets/tree/master/C the Covid-19 time series by regions (CCAA) can be downloaded. Also, auxiliary information like available IUC posts per region before the outbreak of the epidemic, age distribution of confirmed cases, etc. can be found there. Furthermore, similar data can be found at https://www.epdata.es/ searching by the term Covid-19. It is important to notice that each of the different regions might report case numbers with different criteria.

7.2.10 United Kingdom

The U.K. government is collecting data and making them officially available by the Public Health England (PHE), i.e. the executive agency of the Department of Health and Social Care in the U.K. The PHE took on the role of the Health Protection Agency, the National Treatment Agency for Substance Misuse and a number of other health bodies. The official open-data resource provided by the U.K. government can be found at https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases. This dashboard is showing reported cases by Upper Tier Local Authority in England (UTLA). An excel file with relevant information can be downloaded from the dashboard. The information is organized at different levels:

- U.K.: Total number of confirmed cases and deaths in the U.K.
- Deaths by country: England, Scotland, Wales and North Ireland.
- Deaths by NHS regions: London, South East, South West, East of England, Midlands, North East and Yorkshire, North West.
- Deaths by UTLA authorities: Daily cases at each of more than 149 different Upper Tier Local Authorities.

A description of how the confirmed and deaths cases are counted is also available at https://www.gov.uk/guidan The .csv files corresponding to the number of confirmed cases and deaths can also be downloaded from the official public health system[43].

[42] https://datos.gob.es/es/catalogo?theme_id=salud.

[43] https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases

Additional data set reporting the UK Covid-19 cases can also be found at https://github.com/tomwhite/co as GitHub Repository.

7.2.11 United States

The data corresponding to the United States can be obtained from the *2019 Novel Coronavirus Covid-19 (2019-nCoV)* data set from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). This data set is available in GitHub repository https://github.com/CSSEGISandData/COVID-19/, which is updated daily by JHU-CSSS [24].

Another relevant source for the U.S. is the Centers for Disease Control and Prevention[44]. This entity publishes different data on the Covid-19 cases by state and auxiliary information as the number of tests carried out. The CDC also publishes weekly surveillance reports[45].

Moreover, the COVID Tracking Project collects and publishes[46] the testing data available for U.S. states and territories, divided by states. Similar information can also be obtained from https://coronavirus.1point3acres.com/en, including Canada.

Last, the New York Times is releasing a series of data files with cumulative counts of Covid-19 cases in the United States, at the state and county level, over time. These data can be found at https://github.com/nytimes/covid-19-data as a Github repository.

8 Data sets of relevant variables for Covid-19 analysis

In this section, we include data sets relevant for the study and development of models of Covid-19, such as demography, government measures, weather and climate data. These are variables that have been demonstrated to impact on the virus propagation.

8.1 Demographics data sets

- Population:
  - Population of countries at 2018: The data set Covid-19 data, available at the EU Open Data Portal provides the population information per country at 2018.
  - European countries: Data set "Population on 1 January" from the EU open data portal.
  - List of countries by their population 2020: Available at Kaggle[47]. The data set contains not only population values, but also other features of each country.
- Population density:
  - European Environment Agency: The data set "Population density dis-aggregated with Corine land cover 2000" provides a GeoTiFF format file[48].

[44] https://www.cdc.gov/.
[45] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/.
[46] https://covidtracking.com/data.
[47] https://www.kaggle.com/tanuprabhu/population-by-country-2020.
[48] https://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2.

- Age Structure: *Our World in Data* provides a report on the present situation of age structure in the planet and by countries [38]. The corresponding data set is available at `https://ourworldindata.org/age-structure`.

8.2 Data sets on government measures

ACAPS (see subsection 5.9) publishes reports and data-sets on government measures on Covid-19 at `https://www.acaps.org/projects/covid19`. Updated reports can be downloaded from [2]. Moreover, the ACAPS #COVID19 Government Measures Dataset [1] puts together the measures implemented by governments worldwide in response to the Covid-19 pandemic.

The researched information available falls into five categories: i) social distancing; ii) movement restrictions; iii) Public Health measures; iv) social and economic measures; and v) lockdowns. Each category is broken down into several types of measures.

The OxCGRT of Oxford University provides an online API to access the country stringency data at `https://covidtracker.bsg.ox.ac.uk/about-api`. In addition, the full data set can be found at `https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-governmen`

The academic medical center PennMedicine from the University of Pennsylvania has developed an application named CHIME (COVID-19 Hospital Impact Model for Epidemics). It is designed to assist hospitals and public health officials to understand hospital capacity needs as they relate to the COVID-19 pandemic. The application is based on a data model available at `https://code-for-philly.gitbook.io/chime/`.

The Open Government Partnership (OPG) organization has created a list of open government approaches to fight Covid-19 `https://www.opengovpartnership.org/collecting-open-government-ap` The approaches are separated by country and regions. Also, a brieft description and URL of each one can found.

8.3 Weather Data Sets and Applications

8.3.1 European Union Providers

- **ECMWF:** European Centre for Medium-Range Weather Forecasts. It is a research institute and an operational service, producing global numerical weather predictions and other data. It operates two services from the EUs Copernicus Earth observation programme, the Copernicus Atmosphere Monitoring Service (CAMS) and the Copernicus Climate Change Service (C3S). Services provided by the ECMWF:
  - European Climate Data Store: The European Commission has entrusted ECMWF with the implementation of the Copernicus Climate Change Service (C3S). The mission of C3S is to provide authoritative, quality-assured information to support adaptation and mitigation policies in a changing climate. At the heart of the C3S infrastructure is the Climate Data Store[49] (CDS), which provides information about the past, present and future climate in terms of Essential Climate Variables (ECVs) and derived climate indicators.

[49] `https://cds.climate.copernicus.eu/`.

  - The Copernicus Climate Change Service (C3S*) has worked with environmental software experts B-Open[50] to develop an application that allows health authorities and epidemiology centres to explore whether temperature and humidity affect the spread of the coronavirus. This application is freely accessible from the C3S Climate Data Store [8].

- European Commissions Joint Research Centre (JRC): Different open-data projects at JRC can be of interest for the scientific community fighting Covid-19. We highlight here the following one:

  - Photovoltaic Geographical Information System (PVGIS): The focus of PVGIS is research in solar resource assessment, photovoltaic (PV) performance studies, and the dissemination of knowledge and data about solar radiation and PV performance. The PVGIS web application[51], allows to access to meteorological data pertinent to the study of the seasonal behaviour of the pandemic. Three tools are available: i) Photovoltaic Performance, ii) Solar Radiation, iii) Typical Meteorological Year (TMY tool).

8.3.2 NOAA

The National Oceanic and Atmospheric Administration (NOAA) is an American scientific agency within the United States Department of Commerce that focuses on the conditions of the oceans, major waterways, and the atmosphere. It provides thought its open climate data portal[52] provides free access to global historical weather and climate data in addition to station history information. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, etc.

8.3.3 NASA

NASA's goal in Earth science is to observe, understand, and model the Earth system to discover how it is changing. From an open-data perspective, NASA's project Prediction of Worldwide Energy Resource (POWER) can be very useful to recollect time series and monthly means of the most relevant weather and climate variables for a given location.

POWER project[53] was initiated to improve upon the current renewable energy data set and to create new data sets from new satellite systems. The POWER project targets three user communities: (1) Renewable Energy, (2) Sustainable Buildings, and (3) Agroclimatology. The access to the information can be done through the Data Access Viewer[54], which is a responsive web mapping application providing data subsetting, charting, and visualization tools in an easy-to-use interface.

[50] `https://www.bopen.eu/`.
[51] `https://ec.europa.eu/jrc/en/pvgis`
[52] Climate Data Online (CDO): `https://www.ncdc.noaa.gov/cdo-web/`.
[53] `https://power.larc.nasa.gov/`
[54] `https://power.larc.nasa.gov/data-access-viewer/`

8.3.4 Weather Online APIs

There are many online APIs that provide weather data[55]. Some of them can be used free of charge for a limited number of requests. As an example, see *Worl Weather online*[56].

8.4 Mobility data sets

Google has developed Covid-19 Community Mobility Reports. Each Community Mobility Report is broken down by location and displays the change in visits to places like grocery stores and parks. The reports can be obtained by location at `https://www.google.com/covid19/mobility/`. As a result, a PDF document can be downloaded containing figures and trends.

9 Open e-learning on Covid-19

Just to illustrate that there are also many open e-learning resources on Covid-19, we provide here some examples:

- European Laboratory for Learning and Intelligent Systems (ELLIS). Online workshop[57]: Ellis against Covid-19: Machine Learning in Covid-19 Research.
- IEEE Control System Society Italian Chapter: Online workshop on Modeling and Control of the Covid-19 outbreak[58] entitled "How dynamical models can help control the epidemic outbreak".
- Oxford Mathematics Public Lecture: How do mathematicians model infectiouss disease outbreaks?[59] by Robin Thomson. Mathematics Institute. Oxford University.
- Simulating an epidemic: Animation[60] on the effect of different social distancing and isolating measures (on a non epidemiological model) by Grant Sanderson.
- Outbreak simulator: A simplified model of a disease process. The goal is to learn how epidemics unfold in general[61].
- Coronavirus: Why You Must Act Now Politicians, Community Leaders and Business Leaders: What Should You Do and When?[62].

10 Conclusions

In this paper, we provide a review of relevant open-data sources for better understanding the worldwide spread of the Covid-19. We enumerate the variables required to obtain consistent

[55] See, for example, the list presented in `https://datarade.ai/data-categories/weather-data/overview`.
[56] `https://www.worldweatheronline.com/developer/api/historical-weather-api.aspx`.
[57] `https://www.youtube.com/watch?v=0jg_NNwF7k4`.
[58] `http://www.ieeecss.it/events.html`.
[59] `https://livestream.com/oxuni/Thompson`.
[60] `https://youtu.be/gxAaO2rsdIs`.
[61] `https://meltingasphalt.com/interactive/outbreak/`
[62] `https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca`.

epidemiological and forecasting models. We focus not only on the specific Covid-19 time-series but also on a set of auxiliary variables related to the study of its potential seasonal behaviour, the effect of age structure and prevalence of secondary health conditions in the mortality, the effectiveness of government actions, etc.

We analyze the present situation of the available Covid-19 open-data. Unfortunately, it is far from ideal because of a good number of issues like data inconsistency, changing criteria, a large diversity of sources, non-comparable metrics between countries, delays, etc. Despite the difficulties, the availability of open data resources on Covid-19 and related variables provides many opportunities to different communities. In particular, epidemiologists, data-driven researches, health care specialists, machine learning community, data scientists, etc. With the goal of facilitating these communities the access to the required open-sources, we identify the principal open-data entities pertinent to the study of Covid-19. We enumerate different open data sets (and their corresponding repositories) related to Covid-19 cases at a worldwide scale, but also at a more regional level. We provide specific information about the data resources for a selection of countries that have been selected because of the intensity with which the pandemic has impacted them, or for their relevance in the seasonal study of Covid-19 (south-hemisphere). Finally, we provide other open resources that facilitate the incorporation of demographics, weather and climate variables, etc.

10.1 Updates and Contributors

To keep pace with the pandemic, this paper will be updated regularly in arXiv[63] during the Covid-19 pandemic. We thank feedback and collaboration from the community to enrich and update the paper. Contributors can contact the CONCO-Team via e-mail: conco.team@gmail.com.

References

[1] ACAPS. COVID19 government measures dataset. `https://www.acaps.org/covid19-government-measures-dataset`, 2020.

[2] ACAPS. Report on COVID19 government measures updates. `https://www.acaps.org/special-report/covid-19-government-measures-update`, 2020.

[3] Roy M. Anderson, Hans Heesterbeek, Don Klinkenberg, and T. Déirdre Hollingsworth. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*, 395(10228):931–934, 2020.

[4] Yan Bai, Lingsheng Yao, Tao Wei, Fei Tian, Dong-Yan Jin, Lijuan Chen, and Meiyun Wang. Presumed asymptomatic carrier transmission of COVID-19. *Research Letter: https://jamanetwork.com/journals/jama/article-abstract/2762028*, 2020.

[5] David Baud, Xiaolong Qi, Karin Nielsen-Saines, Didier Musso, Léo Pomar, and Guillaume Favre. Real estimates of mortality following COVID-19 infection. *The Lancet infectious diseases*, 2020.

[63] https://arxiv.org/

[6] Giuseppe C. Calafiore, Carlo Novara, and Corrado Possieri. A modified SIR model for the COVID-19 contagion in Italy, 2020.

[7] European Centre for Medium-Range Weather Forecasts. Ecmwf Forecasts. https://www.ecmwf.int/en/forecasts, 2020.

[8] Copernicus Climate Change Service. C3S helps health experts explore how temperature and humidity affect virus spread. https://climate.copernicus.eu/c3s-helps-health-experts-explore-how-temperature-and-humid 2020.

[9] Imperial College London. Report 13 - Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-eur 2020.

[10] IHME COVID, Christopher J.L. Murray, et al. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*, 2020.

[11] Raj Dandekar and George Barbastathis. Neural network aided quarantine control model estimation of covid spread in wuhan, china. *arXiv*, pages arXiv–2003, 2020.

[12] Open Data Watch. What is being said: Data in the time of COVID-19. https://opendatawatch.com/what-is-being-said/data-in-the-time-of-covid-19/, 2020.

[13] Steffen E. Eikenberry, Marina Mancuso, Enahoro Iboi, Tin Phan, Keenan Eikenberry, Yang Kuang, Eric Kostelich, and Abba B. Gumel. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *arXiv preprint arXiv:2004.03251*, 2020.

[14] European Centre for Disease Prevention and Control. Situation dashboard: latest available data. https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html, 2020.

[15] Yaqing Fang, Yiting Nie, and Marshare Penny. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. *Journal of Medical Virology*, 2020.

[16] Seth Flaxman, Swapnil Mishra, Axel Gandy, et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *Imperial College COVID-19 Response Team*, 30, 2020.

[17] Santé Publique France. Donnes hospitalires relatives l'pidmie de COVID-19. https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covi 2020.

[18] Santé Publique France. Donnes relatives aux tests de dpistage de COVID-19 raliss en laboratoire de ville. https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-tests-de-depistage-de-covid-1 2020.

[19] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, Marta Colaneri, et al. A SIDARTHE model of COVID-19 epidemic in Italy. *arXiv preprint arXiv:2003.09861*, 2020.

[20] Antonio Gómez Expósito, José Antonio Rosendo Macías, and Miguel Ángel González Cagigal. Modelado y análisis de la evolución de una epidemia vírica mediante filtros de kalman: el caso del COVID-19 en España. 2020.

[21] World Health Organization. Advice on the use of masks in the community, during home care and in healthcare settings in the context of the novel coronavirus (2019-nCoV) outbreak: interim guidance, 29 January 2020. Technical report, World Health Organization, 2020.

[22] World Health Organization. Coronavirus disease 2019 - Situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reportst, 2020.

[23] World Health Organization. Myth busters. https://www.who.int/emergencies/diseases/novel-corona 2020.

[24] John Hopkins Univeristy Center for Systems Science and Engineering. Coronavirus COVID-19 global cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b 2020.

[25] Yunpeng Ji, Zhongren Ma, Maikel P. Peppelenbosch, and Qiuwei Pan. Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health*, 8(4):e480, 2020.

[26] Robert Koch Institut. Coronavirus disease 2019 (COVID-19) daily situation report of the Robert Koch Institute. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Archiv.ht 2020.

[27] Ying Liu, Albert A. Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 2020.

[28] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance*, 25(10), 2020.

[29] Stephen J. Mooney, Daniel J Westreich, and Abdulrahman M. El-Sayed. Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)*, 26(3):390, 2015.

[30] United Nations. Publish existing data following open data guidelines. https://covid-19-response.unstatshub.org/open-data/publish-existing-data-as-open-data/, 2020.

[31] Nuria Oliver, Xavier Barber, Kirsten Roomp, and Kristof Roomp. The covid19 impact survey: Assessing the pulse of the COVID-19 pandemic in Spain via 24 questions. *arXiv preprint arXiv:2004.01014*, 2020.

[32] Nuria Oliver, Emmanuel Letouzé, Harald Sterly, Sébastien Delataille, Marco De Nadai, Bruno Lepri, Renaud Lambiotte, Richard Benjamins, Ciro Cattuto, Vittoria Colizza, Nicolas de Cordes, Samuel P. Fraiberger, Till Koebe, Sune Lehmann, Juan Murillo, Alex Pentland, Phuong N. Pham, Frdric Pivetta, Albert A. Salah, Jari Saramki, Samuel V. Scarpino, Michele Tizzoni, Stefaan Verhulst, and Patrick Vinck. Mobile phone data and COVID-19: Missing an opportunity? *arXiv preprint arXiv:2003.12347*, 2020.

[33] Graziano Onder, Giovanni Rezza, and Silvio Brusaferro. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA*, 2020.

[34] Noah C. Peeri, Nistha Shrestha, Siddikur Rahman, Rafdzah Zaki, Zhengqi Tan, Saana Bibi, Mahdi Baghbanzadeh, Nasrin Aghamohammadi, Wenyi Zhang, and Ubydul Haque. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *International Journal of Epidemiology*, 2020.

[35] Emanuele Pepe, Paolo Bajardi, Laetitia Gauvin, Filippo Privitera, Brennan Lake, Ciro Cattuto, and Michele Tizzoni. COVID-19 outbreak response: a first assessment of mobility changes in Italy following national lockdown. *medRxiv*, 2020.

[36] Gaetano Perone. An ARIMA model to forecast the spread of COVID-2019 epidemic in Italy. *arXiv preprint arXiv:2004.00382*, 2020.

[37] Santé Publique France. Dones des urgences hospitalires et de sos mdecins relatives l'pidmie de COVID-19. https://www.data.gouv.fr/en/datasets/donnees-des-urgences-hospitalieres-et-de-sos-medeci
2020.

[38] Hannah Ritchie and Max Roser. Age structure. *Our World in Data*, 2020. https://ourworldindata.org/age-structure.

[39] Mohammad M. Sajadi, Parham Habibzadeh, Augustin Vintzileos, Shervin Shokouhi, Fernando Miralles-Wilhelm, and Anthony Amoroso. Temperature and latitude analysis to predict potential spread and seasonality for Covid-19. *Available at SSRN 3550308*, 2020.

[40] UNESCO. Covid-19 educational disruption and response. https://en.unesco.org/covid19/educationresponse, 2020.

[41] Jennifer Valentino-DeVries, Denise Lu, and Gabriel J.X. Dance. Location data says it all: Staying at home during coronavirus is a luxury. *The New York Times*, 2020.

[42] Neeltje van Doremalen, Trenton Bushmaker, Dylan H. Morris, Myndi G. Holbrook, Amandine Gamble, Brandi N. Williamson, Azaibi Tamin, Jennifer L. Harcourt, Natalie J. Thornburg, Susan I. Gerber, et al. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine*, 2020.

[43] Jingyuan Wang, Ke Tang, Kai Feng, and Weifeng. High temperature and high humidity reduce the transmission of COVID-19. *Available at SSRN:https://ssrn.com/abstract=3551767*, 2020.

[44] Jingyuan Wang, Ke Tang, Kai Feng, and Weifeng. When is the COVID-19 pandemic over? Evidence from the stay-at-home policy execution in 106 Chinese cities. *Available at SSRN: https://ssrn.com/abstract=3561491*, 2020.

[45] Yishan Wang, Hanyujie Kang, Xuefeng Liu, and Zhaohui Tong. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. *Journal of Medical Virology*, 2020.

[46] Sheng Zhang, MengYuan Diao, Wenbo Yu, Lei Pei, Zhaofen Lin, and Dechang Chen. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Diseases*, 93:201–204, 2020.

[47] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 2020.