COVID-19: Open-data resources for monitoring, modeling, and forecasting the epidemic

(Article begins on next page)

29 June 2022

Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic

Teodoro Alamo [1,*,†] 🆔, **Daniel G. Reina** [2,†] 🆔, **Martina Mammarella** [3,†] 🆔 **and Alberto Abella** [4,†] 🆔

[1]   Departamento de Ingeniería de Sistemas y Automática, Escuela Superior de Ingenieros, Universidad de Sevilla, 41092 Sevilla, Spain

[2]   Departamento de Ingeniería Electrónica, Escuela Superior de Ingenieros, Universidad de Sevilla, 41092 Sevilla, Spain; dgutierrezreina@us.es

[3]   Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, 10129 Turin, Italy; martina.mammarella@ieiit.cnr.it

[4]   FIWARE Foundation, 10587 Berlin, Germany; alberto.abella@gmail.com

*   Correspondence: talamo@us.es

†   These authors contributed equally to this work.

Abstract: We provide an insight into the open-data resources pertinent to the study of the spread of the Covid-19 pandemic and its control. We identify the variables required to analyze fundamental aspects like seasonal behavior, 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, on a global scale, providing Covid-19 information and/or auxiliary variables (demographics, mobility, etc.). We also describe several open resources to access Covid-19 datasets at a country-wide level (i.e., China, Italy, Spain, France, Germany, US, etc.). To facilitate the rapid response to the study of the seasonal behavior of Covid-19, we enumerate the main open resources in terms of weather and climate variables. We also assess the reusability of some representative open-data sources.

Keywords: Covid-19; coronavirus; SARS-CoV-2; open data; data-driven methods; machine learning; seasonal behavior; government measures

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 presently 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. On a global scale, we identify the most relevant sources. 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 US, among others). We focus on the variables that have possible effects on the evolution and control of the disease at a global and regional scale [1], i.e., we do not cover in this document the data specifically related to medical treatments, vaccines, etc. [2]. 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 to evaluate the actions taken by the decision-makers [1].

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, GitHub repository contains many meaningful datasets of global and regional scope, but it might be challenging to discover them without adequate guidance. Moreover, 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 and control. 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 on 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 datasets 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 behavior or local mortality rate, is described in Section 8. In Section 9, we discuss the reusability of the available datasets. Finally, a concluding Section 10 is included.

2. Covid-19

Coronavirus disease 2019 (Covid-19), technically known as SARS-CoV-2, is an infectious disease that was first identified on 31 December 2019 in Wuhan, the capital of China's Hubei province. The World Health Organization (WHO) declared the 2019–20 coronavirus outbreak a Public Health Emergency of International Concern on 30 January 2020 and a pandemic on 11 March.

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. The virus is most contagious during the first 4–6 days after onset of symptoms [3], although spread is possible in asymptomatic conditions [4] and in later stages of the disease [3]. The time from exposure to onset of symptoms (incubation period) is typically around 5 days but may range from 2 to 14 days [5]. Recommended measures to control the pandemic include social distancing, mobility constraints, pro-active testing and isolation of detected cases [6].

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 care unit (ICU) cases, etc. Because of the incubation period [5], all these variables are related with the number of infected cases in a *delayed* way. One of the main objectives of those institutes is to estimate the basic reproductive number R_0, which serves to characterize the spread of the virus [7]. Several works have calculated R_0 for some outbreaks of specific locations. The estimated values are ranging from 2 to 3 [8]. However, only limited data have been used in most works. On the other hand, achieving an accurate model of the virus reproduction is a challenging task, which involves many variables and validation steps. Unfortunately, the open datasets available presently are locally collected, imprecise with different criteria (lack of standardization on data collection), inconsistent with data models, and incomplete.

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

Moreover, there are relevant variables that are not accurately measured. For example, the fraction of infected non-asymptomatic cases in a given population can be only estimated by means of massive tests or by effective contact-tracing methods. The massive tests carried out in small towns, for example in the north of Italy, 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 use can improve the detection of real cases.

The above limitations on the available datasets must 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 hospitalization and even ICU in the worst case, challenging the healthcare system capacity [9]. One of the most important ways to measure the burden of Covid-19 is through mortality. The probability of dying when getting infected depends on different factors [10–12]:

- Demographics [13]: age, gender, prevalence of diabetes, high blood pressure, obesity, and other risk factors [14].
- Health System [12]: availability of artificial respiration equipment, ICU, specialized medical surveillance and treatments, etc.

On the one hand, several studies have reported a higher level of mortality for older people [13], even more aggravated in men. Thus, protection strategies should be focused on more vulnerable age and gender groups.

Moreover, the capacity of each regional health system to cope with the pandemic is time-varying. Most of the countries, which 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 US) [15]. The main objective in the control strategies, e.g., contention and mitigation 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 Behavior of Covid-19

Many respiratory viruses have a seasonality because lower temperature and lower humidity help facilitate the transmission of the virus [16,17]. There is no clear evidence that Covid-19 is going to behave seasonally, reducing its transmission in summer. Indeed, during the summer season in the Southern hemisphere, e.g., in some regions of South America and Australia, significant Covid-19 outbreaks have been already reported. In [18], the authors show that on March 2020, the areas with significant community transmission of Covid-19 had distribution roughly along the 30–50º N' corridor, at consistently similar weather patterns consisting of average temperatures of 5–11 ºC and low specific (3–6 g/kg), absolute humidity (4–7 g/m^3). In [19], 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 model. Results indicate that increments of one-degree Celsius in temperature and one per cent in relative humidity lower R_0 by 0.0225 and 0.0158, respectively. The authors developed a web application (http://covid19-report.com/#/r-value), where R_0 values for major worldwide cities can be obtained from temperature and humidity.

2.4. Current Actions to Control Covid-19 Pandemic

For the control community, the different confinement, pro-active testing and isolation strategies that can be implemented by a government clearly constitute control inputs to the system [20]. Many of these strategies to slow or stop the spread of Covid-19 are being implemented worldwide, with different intensities. However, these are not the unique actions that a government can undertake 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 [21] 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 two-fold. On 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 to mitigate the effects of the pandemic on health system, economy and society.

It is not simple to determine the effect of the possible anti-measures to be undertaken by the regional governments for several reasons: (i) various 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 several factors, like demographics and weather conditions of the specific region under consideration; (iii) the available data are, 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 this 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. In the following, current methods for contention and mitigation of the spread of the virus are described.

2.4.1. Social Distancing

Following the emergence of this novel coronavirus SARS-CoV-2 and its spread outside China, many 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 wide-scale social distancing including local and national lockdowns. 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 world's student population [22]. Another important aspect has been tackled by the New York Times: how income affects people's abilities to stay home and practice social distancing [23]. Wealthier people not only have more job security and benefits but also may be better able to avoid becoming sick. In [24], authors use a semi-mechanistic Bayesian hierarchical model to attempt inferring the impact of these interventions across 11 European countries. They assume that changes in the reproductive number, i.e., a measure of transmission, are an immediate response to these interventions being implemented rather than broader gradual changes in behavior. 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 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 trends in cases and deaths are closely monitored.

2.4.2. Reducing Mobility

Mobility of people is crucial to understand the spread of the virus. Higher mobility implies higher number of contacts among people [25]. Furthermore, national and international mobility explains the rapid spatial propagation of the virus worldwide. The authors in [26] 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 one week (one serial interval).

Sensor technology can be a crucial tool to obtain mobility measures [27]. Presently, everyone has a mobile phone equipped with several sensors, including GPS, that can 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 must be taken into consideration for data anonymization. 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 [28].

2.4.3. Testing

The distinction between diagnosed and non-diagnosed is important because non-diagnosed individuals are more likely to spread the infection than diagnosed ones. Indeed, the latter are typically isolated, and this can explain misperceptions of the case fatality rate and of the seriousness of the epidemic phenomenon [9]. The main problem for developing massive tests and serology studies is the scarcity of resources, especially in some countries. Accurate testing requires specific labs to analyze RT-PCR tests. On the other hand, the market of rapid tests is under development [29,30]. Some countries are carrying out serology-based testing. Serology tests are blood-based tests that can be used to identify whether people have been exposed to a particular pathogen by looking at their immune response. In this case, the objective is to have a big picture of the state of population with respect to Covid-19. For instance, to check if herd immunity has been reached in some locations [31].

2.4.4. Tracing Contacts

Tracing the contacts of infected people is crucial to isolate potential infected individuals [3]. Once a person is confirmed as an infected one, tracing people contacted with in the last few days can help to reduce the propagation of virus. However, tracing contacts is a challenging task. Manual registers can require an amount of resources unaffordable for most countries. Therefore, technology should play an important role [3,32], in particular mobile devices [33] and wireless technologies, such as WiFi and Bluetooth.

3. Data-Driven Techniques to Fight the Pandemic

Currently, most data available on Covid-19 is used for describing the pandemic in terms of reports and visualizations (For example, https://againstcovid19.com/singapore/dashboard). Although these techniques are useful to highlight the magnitude of the crisis, they are not enough for contending and 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. Classic epidemic models are also useful to obtain mathematical models for epidemics [7]. However, many parameters of these models, such as infected rate and basic reproduction number, require data-driven approaches to estimate them accurately. Also, classic epidemic models, which are normally based on curve fitting techniques, require

data on different phases of the epidemic to obtain the parameters. For these reasons, 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, e.g., [9,34]) can be such solution [35,36]. Many data-based techniques can be applied [37], ranging from classical statistical and machine learning approaches, e.g., linear regression [38,39] and Bayesian inference [40], to sophisticated models based on neural networks [41]. These techniques require sufficient and high-quality data to provide a good estimation. Depending on the methodology used, the quantity of data can vary notably from hundreds to millions of samples. Moreover, a wide variety of data can be necessary for accomplishing an accurate model of a complex and dynamic system like the Covid-19 pandemic. Therefore, data from different disciplines are 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 the 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 (The following list can be improved including other disciplines and variables).

    - *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, serology studies, 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 either names, e.g., countries, cities, etc., or GPS coordinates, i.e., longitude and latitude.
    - *Demographic variables*: population and density of population by location. These variables are required for normalization of the rest of the variables. Other parameters are 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, etc.
    - *Contamination variables*: air pollution, i.e., fine particulate matter $PM_{2.5}$.
    - *International and national mobility and connectivity*: number of international and national flights, number of train connections international and national mobility patterns, traffic patterns, etc.

- **The use of data to estimate the state of the epidemic and develop forecasting models:** By using the aforementioned variables, different models can be developed to estimate the current state of the pandemic and anticipate the response to the propagation of Covid-19. Examples of estimation and forecasting analyses are:

    - Estimation of the infected population.
    - Estimation of economic impact.
    - Forecast of impact in health system through number of infected.
    - Assessing the impact in terms of mortality.
    - Analysis of seasonal behavior.

- **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. Among them, the most relevant are:

  – Assessing the effectiveness of the measures.
  – Planning 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. The main obstacles are addressed in the following sections.

4.1. Variety of Formats

Since there is no a common shared open database on Covid-19, the different sources and variables required to undertake a given analysis are often addressed by 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 [42]. Another issue is related to the wide range of disciplines from which the data sources are coming from. Indeed, these disciplines can be familiar with very different formats and data representation. For instance, some available APIs (Application Programming Interfaces) to get data on Covid-19 provide JavaScript Object Notation (JSON) files. This format is widely used in computer science for web applications. However, for instance, mathematicians and epidemiologists could not 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 [43]. 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

In the WHO global (Covid-19 surveillance document https://www.who.int/publications-detail/global-surveillance-for-human-infection-with-novel-coronavirus-(2019-ncov)), a confirmed case is defined 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 was tested at a laboratory level. Thus, most reports of infection are extremely filtered by the complex and limited testing process. Furthermore, very few datasets 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 (see e.g., [3,44]) not only because it is one key to the estimation of the total number of infected cases, but because it plays a fundamental role in the spread of the virus [3].

4.4. Mortality Rate Is Difficult to Estimate

During the most severe periods of the virus spread in a country, in many situations the number of death cases reported by the administration differs considerably from the real one. This is 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, New York City has reported 5330 more deaths than expected in April 2020 (https://www.nytimes.com/interactive/2020/04/10/upshot/coronavirus-deaths-new-york-city.html), only 3350 of these can be accounted for Covid-19 reasons. These figures suggest that there exists an undercounting on the real number of deaths. Another example is reported for Spain, where the "*Sistema de Monitorización de la Mortalidad diaria (MoMo)*" (https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/MoMo/Paginas/MoMo.aspx) system registers the total number of deaths under any circumstance. The report on 7 April indicates an increase of more than 50% of unexpected deaths in the month before. 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 [45]. Consequently, the comparison of mortality rates between countries makes 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 must take into consideration the fact that the patients who die on any given day were infected much earlier. Thus, the denominator of the mortality rate should be the total number of patients infected at the same time as those who died [45].

Another important parameter to evaluate mortality is to have stratified data, according to gender and age groups. However, such information is not provided by most data sources.

4.5. Not Availability of Individual Case Data

To better understand the disease and to improve models and strategies to fight Covid-19, each case should be tracked with its own timeline, i.e., for each case, relevant information about when symptoms appeared, medical treatments, evolution, degree of isolation, etc., should be available on a country-wide level. Then, 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 possibility is supported by the opinion of many experts and members of the open-source community (See, for example, https://github.com/jgehrcke/covid-19-germany-gae).

An effort of obtaining individual case data can be found in [46]. The authors carried out a survey of 24 questions related to the impact of Covid-19 (Covid19Impact) on citizens in Spain (https://survey123.arcgis.com/share/d29378b51fe8496d8dd77f08ce73973f). The survey was responded to by 146,728 participants over a period of less than two days (i.e., 44 h). The questions were about social contact behavior, financial impact, working situation, and health status. The results of the survey show the negative impact of 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. A similar work has been pursued in UK and the results can be found in [47], where the authors created the Real-World Worry Dataset of 5000 texts (https://github.com/ben-aaron188/covid19worry). The data analysis suggests that people in the UK especially worry about their family and the economic situation.

4.6. Changing and Non-Uniform Criteria

Since the governments are continuously adjusting their response to the virus, it is common to find out abrupt changes in the trend of a time series because a new methodology has been implemented. For example, on 12 February, a sudden spike of 15,152 new Covid-19 cases in China was observed and it

was related to the modified method used for diagnosis, i.e., a combination of SARS-CoV-2 nucleic acid test and clinical Covid-19 features [48].

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 represented by the number of ICU cases. There might be regions reporting the accumulative number of confirmed cases that required ICUs, and others the number of ICUs used by Covid-19 patients. Something similar happens with the number of laboratory tests. They can refer either to the total number of tests carried out or to the number of individuals tested. Indeed, in many situations, the sources do not describe accurately the meaning of the counts.

4.7. Changing Database Structure and Locations

The open-data sources on Covid-19 are constantly improving. To provide more meaningful information, new variables are incorporated into datasets. 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 are 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 how the governments are reporting the data related to Covid-19. Furthermore, there are some concerns about the transparency of countries regarding the data provided.

4.9. Rush in Academia Publications

Many scientific papers are being rapidly published even without peer-review, which is a sub-optimal way to publish science, and more studies are being based on data that is essentially non-peer-reviewed that may have a potential for bias or may contain genuine errors in research methodologies.

5. Open-Data Institutions Providing Worldwide Covid-19 Data

Numerous institutions of different nature, e.g., global institutions, European Union (EU) institutions, universities, newspapers, etc., are providing daily reports on the evolution of the Covid-19 pandemic. In this section, we enumerate those that from our experience, resulted to be the most relevant and reliable ones. In particular, 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. In this section, we describe the nature and characteristics of the information provided, detailing the specifics of the datasets only for the most relevant ones.

5.1. World Health Organization

The primary role of WHO is to direct international health within the United Nations' system and to lead partners in global health responses. In the framework of the Covid-19 pandemic, WHO is providing continuous updates about the current situation all around the world (https://www.who.int/westernpacific/emergencies/covid-19). In [49], 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,. Moreover, it also addresses myth busters related to Covid-19, in order to provide a reliable source of information (see [50]).

5.2. Johns Hopkins University

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/) since the beginning. This university provides a https://coronavirus.jhu.edu/map.html daily update on the global map of the pandemic. The dataset provided by the Johns Hopkins University (JHU) (see Section 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 of University of Oxford that is working on the Covid-19 pandemic and on the policy responses we see around the world. One of their projects related to the study of Covid-19 is focused on tracking what governments around the world are responding to the pandemic and how they compare to others (Further information on the actions developed can be found at: https://www.bsg.ox.ac.uk/news/coronavirus-research-blavatnik-school). 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 Data Portal (EDP) (https://data.europa.eu/euodp/en/data/), which is the official open-data portal of the European Union, gives access to open data published by EU institutions and bodies. EDP acts as single access point to open data and it is published by national open-data portals and institutions in the EU Member States as well as by other non-EU countries. There are numerous datasets on EDP that reference "covid" or "corona". Also, less specific datasets describing former health infections, epidemics or pandemics are also provided (https://www.europeandataportal.eu/en/highlights/covid-19).

To promote research on Covid-19, the European Union has opened a specific data portal, called Covid-19 Data Portal https://www.covid19dataportal.org/. The datasets included in the portal are divided into six categories, such as sequences, expression data, protein, structures, literature and other resources.

In the follows, some of the most relevant European research centers, which have been tackling with the Covid-19 outbreak, are briefly presented.

5.4.1. Joint Research Center

The Joint Research Center (JRC) is the European Commission's science and knowledge service (https://ec.europa.eu/knowledge4policy/organisation/jrc-joint-research-centre_en), which employs scientists to carry out research 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 EU agency, whose mission is to strengthen Europe's defenses against infectious diseases. ECDC publishes numerous scientific and technical reports covering various issues related to the prevention and control of infectious diseases. Towards the end of every calendar year, ECDC publishes its Annual Epidemiological Report, which analyzes surveillance data and infectious disease threats. In addition to 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 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 EU Member States and the EU Commission, and daily updated data on current outbreak [51].

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

5.4.3. European Center for Medium-Range Weather Forecasts

The European Center for Medium-Range Weather Forecasts (ECMWF) is an independent intergovernmental organization supported by 34 states based in Reading [52]. ECMWF is both a research institute and a 24/7 operational service, producing and disseminating numerical weather predictions to EU Member States, Co-operating States and the broader community. ECMWF also archives data and makes them available to authorized users. Some data are also made available under license, and some are publicly available.

5.5. United Nations

Good examples of open data provided by the United Nations (UN) are reported in [53]. Moreover [54] contains the most up-to-date Covid-19 cases and latest trend plot. It covers China, Canada, Australia at province/state level whereas the rest of the world, including US, is covered at country level, represented by either the country centroids or their capitals.

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 US, at state and county level, over time. The time-series data are compiled from states, local governments and health departments. Since January 2020, The NY Times has tracked cases of coronavirus in real time as they were identified after testing. Then, these data have been used to power maps and generate reports about the outbreak. The data collection began with the first reported coronavirus case in Washington State, on 21 January 2020. Since then, the NY Times publishes regular updates of data in a (GitHub repository https://github.com/nytimes/covid-19-data).

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 OWID can be found at their open-data portal https://ourworldindata.org/coronavirus.

5.8. Africa Centers for Disease Control and Prevention

Africa Centers for Disease Control and Prevention (CDC) is a specialized technical institution of the African Union established to support public health initiatives of Member States and strengthen the capacity of their public health institutions to detect, prevent, control and respond quickly and effectively to disease threats (https://africacdc.org/). They provide reports on status, mitigation strategies and guidelines on Covid-19 at https://africacdc.org/covid-19/covid-19-resources/.

5.9. 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 exploited are taken from Wikipedia (https://en.wikipedia.

org/wiki/Template:2019%E2%80%9320_coronavirus_pandemic_data). They also present statistics about the number of confirmed cases, cases per million people (normalized data), number of people recovered, and deaths.

Another relevant tool developed by Google, which can be used to obtain data about Covid-19, is the Google DataSet Search (https://datasetsearch.research.google.com/). Numerous data sets can be found looking for the term *Covid-19*. The application allows users to filter the datasets by several fields, such as last updated, download format, usage rights, topic, and accessibility, etc.

5.10. ACAPS

ACAPS, initially known as The Assessment Capacities Project, is an independent information provider helping humanitarian actors respond more effectively to disasters (https://www.acaps.org). ACAPS was established in 2009 as a non-profit, non-governmental project with the aim of providing independent, ground-breaking humanitarian analysis to help humanitarian workers, influencers, fundraisers, and donors make better decisions. It is not affiliated to the UN or any other organization but is a non-profit project of a consortium of two NGOs, i.e., the Norwegian Refugee Council and Save the Children, and it receives support from several international sources, e.g., the Humanitarian Aid and Civil Protection organization. The ACAPS analysis team is mainly dedicated to researching and analyzing global and crisis specific data. 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 [43].

5.11. Organization for Economic Co-Operation and Development

The Organization for Economic Co-operation and Development (OECD) (https://www.oecd.org) is an international organization that, together with governments, policymakers and citizens, has the goal 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, experiences exchange, best-practice sharing, and advice on public policies and international standard-setting. OECD provides different reports and data about government actions and economic impact due to the pandemic, which can be found at http://www.oecd.org/coronavirus/en/.

5.12. Medical Research Council Center for Global Infectious Disease Analysis

The Medical Research Council Center for Global Infectious Disease Analysis (MRC GIDA) of the Imperial College of London is an international resource and center of excellence for research and capacity-building on the epidemiological analysis and modeling 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 MRC presents reports on Covid-19 under five categories (https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-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 developed a model (The updates of the model can be accessed at https://github.com/ImperialCollegeLondon/covid19model) for estimating the number of infections and the impact of non-pharmaceutical interventions on Covid-19 in eleven European countries [40].

5.13. The Institute for Health Metrics and Evaluation

The Institute for Health Metrics and Evaluation (IHME) is an independent global health research center at the University of Washington (http://www.healthdata.org/). 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 [15]. The work uses: (i) data on confirmed Covid-19 deaths from WHO and from local and national governments; (ii) data on hospital capacity and use for US states; and (iii) observed Covid-19 use data from different locations. A web service, where the projections of the model can be determined for each country and for the following four months, is available (https://covid19.healthdata.org/projections). The information provided is: (i) hospital 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.

5.14. New England Complex Systems Institute (NECSI)

It is a research institution in the US Its focus is on advancing the study of complex systems (https://necsi.edu/). They have developed a portal https://www.endcoronavirus.org/ with the following goals: (i) stop the spread of Covid-19, (ii) consult governments, (iii) institutions and individuals, (iv) provide useful data and guidelines, and (v) crush the curve. The portal includes guidelines and reports on governments, communities, medical institutions, companies, families and individuals.

5.15. MIDAS Network

MIDAS is a global network of scientists and practitioners from academia, industry, government, and non-governmental agencies, who develop and use computational, statistical and mathematical models to improve the understanding of infectious disease dynamics as it relates to pathogenesis, transmission, effective control strategies, and forecasting (https://midasnetwork.us/covid-19/). They have created a portal for Covid-19 modeling, which provides an important and reliable catalog of data resources, including datasets, webinars, and funding announcements.

5.16. Covid-19 Data Hub

The Covid-19 Data Hub project has been funded by the Institute for Data Valorization IVADO, Canada (https://ivado.ca/en/). The goal of the project is to provide the research community with a unified data hub by collecting worldwide fine-grained case data merged with demographics, air pollution, and other exogenous variables helpful for a better understanding of Covid-19 (https://covid19datahub.io/). In addition, they provide R package to download Covid-19-related datasets.

5.17. Science.gov

Science.gov, a gateway portal to US government science information with free access to research and development results and scientific and technical information from scientific organizations across 13 federal agencies, uses software that supports federated search in real time, over 70 information sources (e.g., databases) across the leading federal science and technology agencies in the United States. Using a combination of search terms for Covid-19, Science.gov has provided a link https://www.science.gov/coronavirus.html off its homepage that the public can use to quickly access federally funded research on the Covid-19 disease. Upon linking to the coronavirus research results, users can access freely available peer-reviewed literature (journal articles and accepted manuscripts).

5.18. United States National Institute of Standards and Technology

The National Institute of Standards and Technology (NIST) is a physical sciences laboratory and a non-regulatory agency of the United States Department of Commerce. Its mission is to promote innovation and industrial competitiveness. NIST's activities are organized into laboratory programs that also include information technology. For the Covid-19 pandemic, they provide a dedicated open portal where it is possible to search for specific datasets (randr19.nist.gov) related to the virus outbreak. Moreover, in collaboration with Allen Institute for Artificial Intelligence (AI2), the National Library of Medicine (NLM), Oregon Health & Science University (OHSU), and the University of Texas Health Science Center at Houston (UTHealth), NIST has formed the so-called TREC-COVID challenge, which is currently building a set of Information Retrieval (IR) test collections based on the CORD-19 datasets (see Section 6.3 for further details on CORD-19 competition) and the Text Retrieval Conference (TREC) model. Additional information on this challenge can be found at https://ir.nist.gov/covidSubmit/.

5.19. United States National Institutes of Health

The National Institutes of Health, which represents the primary agency of the United States government responsible for biomedical and public health research, is one of the most prominent source of data on the Covid-19 pandemic (https://www.nih.gov/health-information/coronavirus). In particular, the NIH Office of Data-Science Strategy provided a portal dedicated to open-access data and computational resources related to the Covid-19 fight available at https://datascience.nih.gov/covid-19-open-access-resources, seeking to provide the research community with links to open-access data (see e.g., https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/), computational, and supporting resources.

5.20. 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 [55].

5.21. EuroMOMO

It is a European mortality monitoring activity, aiming to detect and measure excess deaths related to seasonal influenza, pandemics and other public health threats (https://www.euromomo.eu/). They report weekly bulletins (https://www.euromomo.eu/bulletins/2020-18/) on excess of mortality of European countries.

5.22. World Bank Open Data

The World Bank Group (WBG) is a family of five international organizations that make leveraged loans to developing countries. The World Bank's activities are mainly focused on developing countries, in fields such as education, health, agriculture, etc. During the Covid-19 pandemic, WBG help developing countries strengthen their pandemic response and health care systems. Furthermore, WBG has highlighted the importance of *data* to support countries in managing the global Covid-19 outbreak, including

in their open-data portal, i.e., the World Bank Open Data (https://data.worldbank.org/), an entire section (https://www.worldbank.org/en/who-we-are/news/coronavirus-covid19?intcid=wbw_xpl_banner_en_ext_Covid19) dedicated to Covid-19 and datasets (http://datatopics.worldbank.org/universal-health-coverage/coronavirus/) with real-time data, statistical indicators, and other types of data that are relevant to the coronavirus pandemic, particularly focused on the economic and social impacts of the pandemic and the World Bank's efforts to address them.

This dataset is of particular relevance to assess the correlation among the health emergency and the extraordinary shock the global economy is facing, trying to reply to the question: *how is the deadly virus impacting global poverty?* (https://blogs.worldbank.org/opendata/impact-covid-19-coronavirus-global-poverty-why-sub-saharan-africa-might-be-region-hardest). Indeed, estimating how much global poverty will increase because of Covid-19 is challenging and comes with a lot of uncertainty. To answer this question, they propose a model based on household survey data provided by PovcalNet (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx) (an online tool provided by the World Bank for estimating global poverty) and extrapolate forward using the growth projections from the recently launched World Economic Outlook. Comparing these Covid-19-impacted forecasts with the forecasts from the previous edition of the World Economic Outlook provides an assessment of the impact of the pandemic on global poverty, assuming that the pandemic does not change inequality within countries.

6. Open-Source Communities

This section covers repositories of open-source communities, which 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 datasets from official open portals all over the world.

6.1. GitHub

GitHub is a subsidiary company 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 the Covid-19 outbreak, many projects and related datasets have been posted. Most of those included in this paper can be obtained from GitHub. Some examples are: (i) Open Covid-19 Dataset https://github.com/open-covid-19/data); (ii) Covid-19 Data Processing Pipelines and Datasets https://github.com/covid19-data/covid19-data; and (iii) JSON time series of coronavirus cases dataset https://github.com/pomber/covid19.

6.2. Harvard Dataverse Repository

Harvard Dataverse is a free data repository, open to all researchers from any discipline, both inside and outside of the Harvard community. Researchers and practitioner can share, archive, cite, access, and explore research data (https://support.dataverse.harvard.edu/). They have opened a link at https://dataverse.harvard.edu/dataverse/2019ncov for works related to Covid-19, where both the papers and the data used for the analysis can be found.

6.3. Kaggle

Kaggle is a community for data scientist and machine learning practitioners. Kaggle allows users to find and publish datasets, to explore and build models in a web-based data-science environment, to work with other data scientists and machine learning engineers, and to enter competitions to solve data-science challenges. Regarding the Covid-19 pandemic, the portal opens a new challenge

weekly to work on Covid-19 data (https://www.kaggle.com/tags/covid19). The challenge consists of forecasting confirmed cases and fatalities for the following week. Furthermore, some data analysis posts can be found for each competition. (For example https://www.kaggle.com/frlemarchand/covid-19-forecasting-with-an-rnn). The challenges opened up to the date (10 April 2020) can be found at the following links: (i) 18 March: https://www.kaggle.com/c/covid19-global-forecasting-week-1; (ii) 25 March: https://www.kaggle.com/c/covid19-global-forecasting-week-2; (iii) 1 April: https://www.kaggle.com/c/covid19-global-forecasting-week-3; and (iv) 8 April: https://www.kaggle.com/c/covid19-global-forecasting-week-4.

Moreover, the Covid-19 Open Research Dataset Challenge (CORD-19) competition has been launched, aimed at developing text and data mining tools that can help the medical community to develop answers to high priority scientific questions https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. The available dataset, based on data sources provided by the Center for Security and Emerging technology of Georgetown University (http://cset.georgetown.edu) is composed by a corpus of more than 44,000 full-text documents, about Covid-19/SARS-CoV-2 and related coronaviruses.

Another relevant competition based on Covid-19 data is the UNCOVER COVID-19 Challenge https://www.kaggle.com/roche-data-science-coalition/uncover. In this case, the objective is modeling solutions to key questions that were developed and evaluated by a global front-line of healthcare providers, hospitals, suppliers, and policy makers. In this case, the challenge is promoted by Hoffmann-La Roche Limited (Roche Canada).

6.4. Zindi

Zindi is the first data-science competition platform in Africa. Zindi hosts an entire data-science ecosystem of scientists, engineers, academics, companies, NGOs, governments and institutions, focused on solving Africa's most pressing problems. Regarding the Covid-19 pandemic, they have opened a competition aimed at building an epidemiological model that predicts the spread of Covid-19 throughout the world. The target variable is the cumulative number of deaths caused by Covid-19 in each country by each date. The challenge can be found at https://zindi.africa/competitions/predict-the-global-spread-of-covid-19/data.

7. Covid-19 Datasets

This section presents the main available datasets that can be found on the Internet related to Covid-19. The section is divided into two parts. First, we present international datasets 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 several 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 Datasets

In this section, we briefly introduce the institutions that provide international datasets, also including the link (URL) to an easier access to them.

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 (https://coronavirus.jhu.edu/). JHU provides a daily update of the global map of the pandemic, which can be found at https://coronavirus.jhu.edu/map.html.

The JHU Covid-19 dataset can be downloaded in .csv format from the GitHub repository https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series. In this folder, five different .csv files can be downloaded: (i) global number of confirmed cases https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv; (ii) global number of deaths https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv; (iii) global number of recovered https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv; (iv) total number of confirmed cases in US https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv; and (v) total number of deaths in US https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_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, whereas, most of them like Spain or Italy, are not. The US data .csv files correspond to the United States. In both cases, all the data refer to accumulated cases, i.e., cases up to the date of the row in which the data is consigned. Furthermore, the geographical coordinates of each region/country are also provided.

The data plots, which can be recovered at https://coronavirus.jhu.edu/data/new-cases, are obtained by means of a 5-days moving window, averaging the values of that day, the two days before, and the two days after. 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

The Geographical Distribution of Covid-19 Worldwide Dataset is sourced from the ECDC, which publishes full time-series data for the number of confirmed Covid-19 cases and deaths daily for various countries around the world. On daily basis, the ECDC collects data from 6 a.m. to 10 a.m. CET and publishes this data via its Covid-19 dashboard (https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html). Then, this dataset is also made publicly available through downloadable files in different formats (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide). This dataset can be also downloaded from the open portal of *Our World in Data*. In particular, it is possible to recover the following datasets: (i) total confirmed cases https://covid.ourworldindata.org/data/ecdc/total_cases.csv; (ii) total deaths https://covid.ourworldindata.org/data/ecdc/total_deaths.csv; (iii) new confirmed cases https://covid.ourworldindata.org/data/ecdc/new_cases.csv; (iv) new deaths https://covid.ourworldindata.org/data/ecdc/new_deaths.csv; (v) all four metrics https://covid.ourworldindata.org/data/ecdc/full_data.csv; and (vi) population data https://covid.ourworldindata.org/data/ecdc/locations.csv.

7.1.3. Covid-19 Data Hub

The Covid-19 Data Hub project makes all the data available at https://github.com/covid19datahub/COVID19. The dataset includes a large range of variables such as Covid-19 variables (confirmed cases, death, etc.), population, density, ICU, number of tests, ventilators, testing policy and contact tracing, among others.

7.1.4. MIDAS Network Dataset

The MIDAS network publishes an open dataset with several data resources to study the Covid-19 pandemic https://github.com/midas-network/COVID-19). The resources are divided into different

sections, such as data catalog, parameter estimates, software tools, and documents. In particular, a collection of .csv files can be found in the catalog section about the situation of each country.

7.1.5. Covid-19 Testing (*Our World in Data* Dataset)

Our World in Data publishes useful information on how the different countries are carrying out laboratory tests to detect Covid-19 cases (https://ourworldindata.org/covid-testing). The dataset on the number of tests carried out globally is published by *Our World in Data* in the following GitHub repository Owid/covid.19-data (https://github.com/owid/covid-19-data/tree/master/public/data/testing).

7.2. Examples of Regional Datasets

Most of the following datasets can be found in GitHub, searching by the term *Covid-19*. Table 1 contains the links to access the regional respositories.

Table 1. Some examples of regional Covid-19 data resources.

	Source	GitHub Repositories
Argentina	Ministry of Health https://www.argentina.gob.ar/coronavirus/informe-diario	Covid19arData https://github.com/SistemasMapache/Covid19arData
Australia	Australian Health Department https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers	covid-19-au https://github.com/covid-19-au/covid-19-au.github.io
China	China National Health Commission http://en.nhc.gov.cn/DailyBriefing.html	JHU https://github.com/CSSEGISandData/COVID-19/, Midas-China https://github.com/midas-network/COVID-19/tree/master/data/cases/china
France	Public France Health System https://www.santepubliquefrance.fr	opencovid19-fr https://github.com/opencovid19-fr/data/blob/master/README.en.md, FRANCE-COVID-19 https://github.com/cedricguadalupe/FRANCE-COVID-19
Germany	Robert Koch Institute https://www.rki.de/EN/Home/homepage_node.html	covid-19-germany-gae https://github.com/jgehrcke/covid-19-germany-gae
Iceland	Government of Iceland https://www.covid.is/data	gaui-covid19 https://github.com/gaui/covid19
Italy	Italian Civil Protection Department http://www.protezionecivile.gov.it/attivita-rischi/rischio-sanitario/emergenze/coronavirus	pcm-dpc https://github.com/pcm-dpc/COVID-19
Paraguay	Ministry of Public Health and Soc. Welfare https://www.mspbs.gov.py/reporte-covid19.html	covidpy-rest https://github.com/torresmateo/covidpy-rest/blob/master/data/covidpy.csv
South Africa	National Inst. Communicable Diseases https://www.nicd.ac.za/	covid19za https://github.com/dsfsi/covid19za
South Korea	Centers for Disease Control and Prevention https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030	COVID19-Korea https://github.com/parksw3/COVID19-Korea
Spain	Ministry of Health https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov-China/home.htm	datadista-Covid-19 https://github.com/datadista/datasets/tree/master/COVID%2019
United Kingdom	Public Health England https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases	covid-19-uk-data https://github.com/tomwhite/covid-19-uk-data
United States	Centers for Disease Control and Prevention https://www.cdc.gov/	JHU https://github.com/CSSEGISandData/COVID-19/, Nytimes https://github.com/nytimes/covid-19-data

7.2.1. Africa

For the African continent, reliable datasets can be found at https://github.com/dsfsi/covid19africa as GitHub repository [56].

7.2.2. Argentina

The Argentina Ministry of Health provides daily updates on the Covid-19 spread, including data on the number of infected people divided by regions https://www.argentina.gob.ar/coronavirus/informe-diario). The data can be downloaded from the GitHub repository Covid19arData https://github.com/SistemasMapache/Covid19arData.

7.2.3. Australia

The Health Department Health Department (https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers) of the Australian Government publishes the Covid-19 data. The data corresponding to the different regions of Australia can be downloaded from the JHU GitHub repository (https://github.com/CSSEGISandData/COVID-19/). The regional Australian Covid-19 data are integrated into the global time-series .csv files, which include information on confirmed cases, number of deaths, and number of recovered. See Section 7.1.1 for further details. An additional GitHub repository is available at https://github.com/covid-19-au/covid-19-au.github.io.

7.2.4. China

The data corresponding to the different regions of China can be downloaded from the JHU GitHub repository https://github.com/CSSEGISandData/COVID-19/. The regional Covid-19 Chinese data is integrated into the global time-series .csv files. Moreover, the National Health Commission of the People's Republic of China updates daily the available information on the situation in China http://en.nhc.gov.cn/DailyBriefing.html). Relevant information about the pandemic in China can also be found at the MIDAS GitHub repository: https://github.com/midas-network/COVID-19/tree/master/data/cases/china.

7.2.5. France

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

- Covid-19 Hospital Data https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/: hospitalized cases, ICU cases, deaths per department, region, gender and age range [57].
- Covid-19 Emergency Room Admissions https://www.data.gouv.fr/en/datasets/donnees-des-urgences-hospitalieres-et-de-sos-medecins-relatives-a-lepidemie-de-covid-19/: hospitalized cases, ICU cases, deaths per department, region, gender and age range [58].
- Covid-19 Laboratory Tests https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-tests-de-depistage-de-covid-19-realises-en-laboratoire-de-ville/: number of positive and negative laboratory tests per department, gender, and age group [59].

Moreover, French Covid-19 datasets can be found in these two GitHub repositories: (i) Covid-19 epidemic French national data https://github.com/opencovid19-fr/data/blob/master/README.en.md; and (ii) Projet d'historisation du nombre de cas par région du Covid-19 https://github.com/cedricguadalupe/FRANCE-COVID-19. Finally, the national mortality register can be accessed at https://www.insee.fr/fr/information/4470857#tableau-figure1_radio1 to compare the number of deaths with previous years.

7.2.6. Germany

The main official open-data provider in Germany is the Robert Koch Institute (https://www.rki.de/EN/Home/homepage_node.html), a public health institute in Germany. It provides, by means of a catalog of infectious diseases (https://www.rki.de/DE/Content/Infekt/infekt_node.html), pertinent information on each disease listed in the catalog, e.g., SARS. In particular for Covid-19, data on risk assessments, spread of the epidemic, epidemiological studies, etc., can be found at https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/nCoV.html. Moreover, it provides also daily reports of Covid-19 outbreak in Germany [60]. Additional data on Covid-19 case numbers in Germany, divided by state over time, can be found at the GitHub repository https://github.com/jgehrcke/covid-19-germany-gae. The national mortality register can be found at Destatis https://www.destatis.de/EN/Themes/Society-Environment/Population/Deaths-Life-Expectancy/_node.html.

7.2.7. Iceland

During this pandemic, it has been reported (https://ourworldindata.org/covid-testing) that Iceland is one the best countries in terms of data on testing population. The Iceland government publishes all information at the following link: https://www.covid.is/data. A GitHub repository can be found at https://github.com/gaui/covid19.

7.2.8. Italy

The Italian Civil Protection Department http://www.protezionecivile.gov.it/attivita-rischi/rischio-sanitario/emergenze/coronavirus, i.e., the national body in Italy that deals with the prediction, prevention and management of emergency events, daily updates a GitHub repository organized by regions and provinces, where the Covid-19 time-series 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 (Available at https://github.com/pcm-dpc/COVID-19/raw/master/dati-regioni/dpc-covid19-ita-regioni.csv) provides the number of confirmed cases, deaths, recovered, hospitalized, confined at home and ICU cases, in addition to the number of daily tests. Furthermore, GEDI *Gruppo Editoriale*, a relevant Italian media conglomerate, provides a portal where those data are arranged in several interactive graphs, which include also the impact on the local mobility (https://lab.gedidigital.it/gedi-visual/2020/coronavirus-in-italia/). The national mortality register http://dati.istat.it/Index.aspx?QueryId=19670 of Italy can be consulted to evaluate the magnitude of the epidemic with respect the number of deaths in previous years.

7.2.9. Paraguay

The official portal for data reports on Covid-19 for Paraguay can be found at https://www.mspbs.gov.py/reporte-covid19.html. The data are provided by the Public Health system and 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 also found at https://github.com/torresmateo/covidpy-rest/blob/master/data/covidpy.csv as GitHub repository.

7.2.10. South Africa

The information on Covid-19 spreading in South Africa can be found at https://github.com/dsfsi/covid19za [61,62], 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. These data have been used in [62] to determine what data should be included in a public repository amid the Covid-19 outbreak and how this data should be disseminated within a public dashboard.

7.2.11. South Korea

Korea Centers for Disease Control and Prevention (KCDC) provides data sets on Covid-19 cases regularly at https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030. A specific GitHub repository is available at https://github.com/parksw3/COVID19-Korea.

7.2.12. Spain

The regional Covid-19 Spanish data are collected by the Spanish government and they are available at the national open-data portal https://datos.gob.es/. Different health datasets can be searched at its open-data catalog (https://datos.gob.es/es/catalogo?theme_id=salud). The specific search *Covid* provides datasets related to the global Spanish data classified into regions, e.g., *Evolución de enfermedad por el coronavirus (Covid-19)*, or specific of a particular Spanish region, e.g., *Evolución del coronavirus (Covid-19) en Euskadi*. In the GitHub repository https://github.com/datadista/datasets/tree/master/COVID%2019, the Covid-19 time series by regions (CCAA) can be downloaded. Also, auxiliary information, like number of available ICUs per region before the pandemic outbreak, age distribution of confirmed cases, etc., can be found there. Furthermore, similar data can be also found at https://www.epdata.es/ searching by the term *Covid-19*. It is important to highlight that each of the different regions might report case numbers with different criteria. The national mortality register can be accessed at MoMo https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/MoMo/Paginas/MoMo.aspx.

7.2.13. United Kingdom

The UK 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 UK. The PHE took on the role of the Health Protection Agency, the National Treatment Agency for Substance Misuse and several other health bodies. The official open-data resource provided by the UK 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: (i) *total number of confirmed cases and deaths in the UK*; (ii) *deaths by country*: England, Scotland, Wales and North Ireland; (iii) *deaths by NHS regions*: London, South East, South West, East of England, Midlands, North East and Yorkshire, North West; and (iv) *deaths by UTLA authorities*: daily cases at each of more than 149 different UTLAs. A description of how the confirmed and deaths cases are counted is also available at https://www.gov.uk/guidance/coronavirus-covid-19-information-for-the-public#number-of-cases-and-deaths. The .csv files corresponding to the number of confirmed cases and deaths can also be downloaded from the official public health system (https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases). Additional datasets reporting the UK Covid-19 cases can be found at https://github.com/tomwhite/covid-19-uk-data as GitHub Repository. The national mortality register can be found at Office for National Statistics https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales.

7.2.14. United States

The data corresponding to the United States can be obtained from the *2019 Novel Coronavirus Covid-19 (2019-nCoV)* dataset from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). This dataset is available as GitHub repository at https://github.com/CSSEGISandData/COVID-19/, which is daily updated by JHU-CSSS itself [54]. Another relevant source for the US is the Centers

for Disease Control and Prevention (CDC) (https://www.cdc.gov/). This entity publishes different data on Covid-19 cases by state and auxiliary information as the number of tests carried out. The CDC also publishes weekly surveillance reports, which can be found at https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/. Moreover, the Covid Tracking Project (https://covidtracking.com/data) collects and publishes the testing data available for the US states and territories, divided by states. Similar information can 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 US, at state and county level, over time. These data can be found at https://github.com/nytimes/covid-19-data as GitHub repository.

8. Data Sets of Relevant Variables for Covid-19 Analysis

In this section, we include datasets relevant for the study and development of models of Covid-19, such as demography, government measures, weather, and climate data. These are variables that are under research to evaluate their influence on the virus propagation.

8.1. Demographics Datasets

Demographics datasets are of significant importance for Covid-19 analysis. In this section, they have been arranged in three main groups: (i) population; (ii) population density; and (iii) age structure. We highlight the following datasets on population:

- *European Countries population*: the dataset Eurostat *Population on 1 January* dataset from the EU open-data portal, available at https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_pjan&lang=en, provides the population information per country at 2019.
- *Global population*: the Population reference Bureau has published at https://www.prb.org/worldpopdata/ 2019 World Population Data Sheet.
- *List of countries by their population 2020*: the global population at 2020 per country can be retrieved on Kaggle (https://www.kaggle.com/tanuprabhu/population-by-country-2020). The dataset contains not only population values but also other features for each country.

Moreover, there are also some portals that provide demographic information, e.g., population pyramid as at https://www.populationpyramid.net/.

For the population density, the European Environment Agency provided the *Population density disaggregated with Corine land cover 2000* dataset as a GeoTiFF format file, which can be found at https://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2.

Last, about age structure, *Our World in Data* provides a report on the present situation on the planet, divided by countries, [63]. The corresponding dataset is made available at https://ourworldindata.org/age-structure.

8.2. Datasets on Government Measures

For datasets related to the government measures, ACAPS publishes reports and datasets on government measures on Covid-19 at https://www.acaps.org/projects/covid19 (see Section 5.10 for further deatils). In particular, updated reports can be downloaded from [43]. Moreover, the ACAPS #COVID19 Government Measures Dataset [43] 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.

Another source for government measures to fight the Covid-19 outbreak is made available by the OxCGRT group at Oxford University, which provides an online API to access the country stringency data available at https://covidtracker.bsg.ox.ac.uk/about-api whereas the full dataset can be found at https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker.

The EpidemicForecasting.org website provides a dataset on mitigation measures carried out by countries, which can be found at http://epidemicforecasting.org/about. In addition to that, they provide a simulator, i.e., the GLEAMviz simulator, which allows exploration of realistic epidemic spreading scenarios at the global scale (http://www.gleamviz.org/simulator/).

An application named CHIME, i.e., Covid-19 Hospital Impact Model for Epidemics, has been developed by the Penn Medicine academic medical center from the University of Pennsylvania. This app 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/.

Last, the Open Government Partnership (OPG) organization has created a list of open government approaches to fight Covid-19 available at https://www.opengovpartnership.org/collecting-open-government-approaches-to-covid-19/. In particular, these approaches are organized by country and regions, and a brief description and related URL are also provided for each one.

8.3. Weather Datasets and Applications

In this section, we focus on datasets related to weather, which are provided by several organizations all around the world, as described in the follows.

The first group of organizations providing weather datasets are the following EU providers:

- *European Center for Medium-Range Weather Forecasts (ECMWF)*: is a research institute and an operational service, producing global numerical weather predictions and other data. It operates two services from the EU's Copernicus Earth observation program, the Copernicus Atmosphere Monitoring Service (CAMS) and the Copernicus Climate Change Service (C3S). Two main services are provided by the ECMWF. The first one is the 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 (CDS) (https://cds.climate.copernicus.eu/), which provides information about the past, present and future climate in terms of Essential Climate Variables (ECVs) and derived climate indicators. The second ECMWF service is the Copernicus Climate Change Service (C3S*), which has worked with environmental software experts B-Open (https://www.bopen.eu/) to develop an application that allows health authorities and epidemiology centers to explore whether temperature and humidity affect the spread of the coronavirus. This application is freely accessible from the C3S Climate Data Store [64].

- *European Commission's Joint Research Center (JRC)*: different open-data projects at JRC can be of interest for the scientific community fighting Covid-19. We highlight here the most relevant one, represented by the Photovoltaic Geographical Information System (PVGIS). The focus of PVGIS is the 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 (https://ec.europa.eu/jrc/en/pvgis) allows access to meteorological data pertinent to the study of the seasonal behavior of the pandemic. Three tools are available: (i) Photovoltaic Performance; (ii) Solar Radiation; and (iii) Typical Meteorological Year (TMY tool).

The second group of organizations providing weather datasets are US providers: (i) the National Oceanic and Atmospheric Administration (NOAA); and (ii) the National Aeronautics and Space Administration (NASA). 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 through its open climate data portal (Climate Data Online (CDO): https://www.ncdc.noaa.gov/cdo-web/) 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.

On the other hand, 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 (https://power.larc.nasa.gov/) 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 at https://power.larc.nasa.gov/data-access-viewer/, which is a responsive web mapping application providing data sub-setting, charting, and visualization tools in an easy-to-use interface.

Last, there are many online APIs that provide weather data (See, for example, the list presented in https://datarade.ai/data-categories/weather-data/overview). Some of them can be used free of charge for a limited number of requests. As an example, see *World Weather online* (https://www.worldweatheronline.com/developer/api/historical-weather-api.aspx).

8.4. Mobility Data Sets

This section includes datasets related to mobility of people.

- **Mobility reports:** Google has developed Covid-19 Community Mobility Reports, in which each 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. A similar tool has been developed by Apple and it can be found at https://www.apple.com/covid19/mobility. The reports can be obtained filtering by country. One important difference with respect to the Google app is that the raw data can be retrieved in the form of .csv files. Last, the GeoDS Lab (Department of Geography at University of Wisconsin-Madison) has developed a web application to identify mobility pattern changes in the US [65]. The application can be accessed at https://geods.geography.wisc.edu/covid19/physical-distancing/.
- **Airport connectivity:** FLIRT is a tool that allows the obtaining of data about commercial flights. It shows direct flights from a selected location, and can simulate passengers taking multi-leg itineraries. The data can be downloaded in different formats (.csv, JSON, etc.) at https://flirt.eha.io/.
- **Contact tracing:** Another important source of data related to mobility for modeling the pandemic is human behavior inferred from wireless technologies, such as cell communications, WiFi and Bluetooth, among others. On this line, CRAWDAD is the Community Resource for Archiving Wireless Data At Dartmouth, a wireless network data resource for the research community. This repository contains wireless trace data from many contributing locations, and staff to develop better tools for collecting, anonymizing, and analyzing the data. The repository can be accessed at http://crawdad.org/index.html and it allows filtering of the data, for instance, Human Behavior Modeling and Opportunistic Connectivity, among other fields.

9. Reusability of Open-Data Sources

To maximize the value of the data sources about Covid-19, it is necessary that data sources are not only available but also have a set of characteristics that make them reusable. Due to the global affection of the pandemic, data sources are most of the cases coming from public institutions. These open government data should follow the eight principles of open data as reported in [66].

MELODA 5 [67] is a metric to assess the reusability of open datasets. This metric considers 8 dimensions that affects the reusability of a dataset, which are listed hereafter:

1. *Legal license*: assesses the legal rights given to the reusers of the dataset.
2. *Technical format*: assesses the digital storage format in which the data is stored and released.
3. *Access*: assesses the possibilities offered to reusers to interact with the dataset to retrieve the necessary set of data.
4. *Standardization*: assesses how popular and agreed are the fields composing the dataset and its description.
5. *Geolocalization*: assesses the geographical content of the released data.
6. *Updating frequency*: assesses the frequency of updating of the dataset.
7. *Dissemination*: assesses the efforts and resources done by the publishing entity to makes popular the released datasets.
8. *Prestige*: assesses the reputation of the publishing entity for the reusers of their data. (For Covid-19, this dimension cannot be set due to the novelty of the phenomenon).

According to these dimensions, the assessment of the main data sources mentioned in previous sections is reported in the following list: (In this list the prestige dimension has not been removed and therefore there are 6 points of difference with next table).

- *2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by JHU CSSE*: 31
- *Our world in data. Coronavirus Source Data*: 34
- *Argentina*: 30
- *Australia*: 34
- *China*: 31
- *Italy*: 38
- *France*: 34
- *Germany*: 37
- *Paraguay*: 25
- *South Africa*: 40
- *Spain*: 37
- *United Kingdom*: 35
- *United States*: 32

Although the maximum score for MELODA 5 is 61 points [67], in Table 2 the *Prestige* dimension of the publishing institution regarding Covid-19 is not included due to the novelty of the situation. To obtain a fair comparison, this criterion has been removed from the analysis, thus leaving only 7 dimensions for the data sources. Accordingly, a maximum of 55 points can be achieved. From this table, it is clear that none of the sources score results higher than 35 points, a value that can be considered good but far from optimum. We shall highlight that some sources are releasing their data with a license that restrict commercial use (they are not open data). (See definition of open data at http://opendefinition.org). Hence, a score of 1 has been set for them on *License* dimension. Another remarkable point is the general lack of

an API to access individual data in the data sources. This forces the reusers to update the full dataset daily. For this reason, most of the data sources score 1 in *Access* dimension. It is also remarkable the general lack of geolocalization contents for most of the data sources. A mere indication of the region/area is the most common geographic content. Consequently, 3 is the more frequent score for *Geolocalization* dimension. Regarding technical format, .csv is the most popular, together with some sources using JSON file formats. This last format provides additional key identification for each value. Although many sources include a definition of the field, no Standardization effort is detected for sharing the same information between sources. In fact, there is a myriad of different field names and contents. Hence, the *Dissemination* dimension score has been considered the maximum for those sources that have a website to disseminate the data sources.

Table 2. Total score corresponding to the first 7 reusability dimensions of MELODA 5 for different open institutional data sources. (AR: Argentina; AU: Australia; CN: China; DE: Germany; FR: France; GB: United Kingdom; IT: Italy; JHU: Johns Hopkins University; OWID: Our World In Data; PY: Paraguay; SP: Spain; US: United States; ZA: South Africa).

	AR	AU	CN	DE	FR	GB	IT	JHU	OWID	PY	SP	US	ZA
License	6	3	3	6	6	6	6	3	6	1	6	6	6
Technical Format	1	3	3	6	3	6	6	3	3	1	6	3	6
Access	1	1	1	3	1	1	1	1	1	1	1	1	6
Standardization	1	3	3	1	3	1	1	3	3	1	3	1	1
Geolocalization	3	3	3	3	3	3	6	3	3	3	3	3	3
Updating Frequency	6	6	6	6	6	6	6	6	6	6	6	6	6
Dissemination	6	6	6	6	6	6	6	6	6	6	6	6	6
TOTAL	24	25	25	31	28	29	32	25	28	19	31	26	34

10. Conclusions

In this paper, we provide a review of relevant open-data sources for better understanding the worldwide spread of Covid-19. We enumerate the variables required to obtain consistent epidemiological and forecasting models. In particular, 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 behavior, 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 researchers, 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. Furthermore, we enumerate different open datasets, and their corresponding repositories, related to Covid-19 cases on a global scale, but also at a regional/local level. In addition, 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, e.g., south-hemisphere countries. Finally, we provide other open resources that facilitate the incorporation of demographics, weather and climate variables, etc.

References

1.	Lakshmi Priyadarsini, S.; Suresh, M. Factors influencing the epidemiological characteristics of pandemic COVID-19: A TISM approach. *Int. J. Healthc. Manag.* **2020**, 1–10. [CrossRef]

2.	Le, T.T.; Andreadakis, Z.; Kumar, A.; Román, R.G.; Tollefsen, S.; Saville, M.; Mayhew, S. The COVID-19 vaccine development landscape. *Nat. Rev. Drug Discov.* **2020**, *10*. [CrossRef]

3.	Ferretti, L.; Wymant, C.; Kendall, M.; Zhao, L.; Nurtay, A.; Abeler-Dörner, L.; Parker, M.; Bonsall, D.; Fraser, C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **2020**, eabb6936. [CrossRef] [PubMed]

4.	Bai, Y.; Yao, L.; Wei, T.; Tian, F.; Jin, D.Y.; Chen, L.; Wang, M. Presumed Asymptomatic Carrier Transmission Of COVID-19. *Res. Lett.* **2020**, *323*, 1406–1407. Available online: https://jamanetwork.com/journals/jama/article-abstract/2762028 (accessed on 10 April 2020).

5.	Lauer, S.A.; Grantz, K.H.; Bi, Q.; Jones, F.K.; Zheng, Q.; Meredith, H.R.; Azman, A.S.; Reich, N.G.; Lessler, J. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **2020**. [CrossRef]

6.	Hellewell, J.; Abbott, S.; Gimma, A.; Bosse, N.I.; Jarvis, C.I.; Russell, T.W.; Munday, J.D.; Kucharski, A.J.; Edmunds, W.J.; Sun, F.; et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **2020**, *8*, e488–e496. [CrossRef]

7.	Martcheva, M. *An Introduction to Mathematical Epidemiology*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2015.

8.	Liu, Y.; Gayle, A.A.; Wilder-Smith, A.; Rocklöv, J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J. Travel Med.* **2020**, *27*, taaa021. [CrossRef]

9.	Giordano, G.; Blanchini, F.; Bruno, R.; Colaneri, P.; Di Filippo, A.; Di Matteo, A.; Colaneri, M. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **2020**, 1–6. [CrossRef]

10.	Zhou, F.; Yu, T.; Du, R.; Fan, G.; Liu, Y.; Liu, Z.; Xiang, J.; Wang, Y.; Song, B.; Gu, X.; et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **2020**, *395*, 1054–1062. [CrossRef]

11.	Peeri, N.C.; Shrestha, N.; Rahman, S.; Zaki, R.; Tan, Z.; Bibi, S.; Baghbanzadeh, M.; Aghamohammadi, N.; Zhang, W.; Haque, U. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: What lessons have we learned? *Int. J. Epidemiol.* **2020**. [CrossRef]

12. Ji, Y.; Ma, Z.; Peppelenbosch, M.P.; Pan, Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob. Health* **2020**, *8*, e480. [CrossRef]

13. Onder, G.; Rezza, G.; Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **2020**. [CrossRef] [PubMed]

14. Leung, C. Risk factors for predicting mortality in elderly patients with COVID-19: A review of clinical data in China. *Mech. Ageing Dev.* **2020**, 111255. [CrossRef] [PubMed]

15. COVID, I.; Murray, C.J. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv* **2020**. [CrossRef]

16. Lowen, A.C.; Mubareka, S.; Steel, J.; Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog.* **2007**, *3*, e151. [CrossRef]

17. Chan, K.; Peiris, J.; Lam, S.; Poon, L.; Yuen, K.; Seto, W. The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* **2011**, *2011*, 734690. [CrossRef]

18. Sajadi, M.M.; Habibzadeh, P.; Vintzileos, A.; Shokouhi, S.; Miralles-Wilhelm, F.; Amoroso, A. Temperature and Latitude Analysis to Predict Potential Spread and Seasonality for Covid-19. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3550308 (accessed on 18 April 2020).

19. Wang, J.; Tang, K.; Feng, K.; Lv, W. High Temperature and High Humidity Reduce the Transmission of COVID-19. 2020. Available online: https://ssrn.com/abstract=3551767 (accessed on 16 April 2020).

20. Anderson, R.M.; Heesterbeek, H.; Klinkenberg, D.; Hollingsworth, T.D. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **2020**, *395*, 931–934. [CrossRef]

21. Eikenberry, S.E.; Mancuso, M.; Iboi, E.; Phan, T.; Eikenberry, K.; Kuang, Y.; Kostelich, E.; Gumel, A.B. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *arXiv* **2020**, arXiv:2004.03251.

22. UNESCO. Covid-19 Educational Disruption and Response. 2020. Available online: https://en.unesco.org/covid19/educationresponse (accessed on 20 April 2020).

23. Valentino-DeVries, J.; Lu, D.; Dance, G.J. Location Data Says It All: Staying at Home During Coronavirus Is A Luxury. *The New York Times*, 3 April 2020.

24. College London, I. Report 13— Estimating the Number of Infections and the Impact of Non-Pharmaceutical Interventions on COVID-19 in 11 European Countries. 2020. Available online: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-europe-npi-impact/ (accessed on 10 April 2020).

25. Zhang, J.; Litvinova, M.; Liang, Y.; Wang, Y.; Wang, W.; Zhao, S.; Wu, Q.; Merler, S.; Viboud, C.; Vespignani, A.; et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **2020**. [CrossRef]

26. Wang, J.; Tang, K.; Feng, K.; Lv, W. When is the COVID-19 Pandemic Over? Evidence from the Stay-at-Home Policy Execution in 106 Chinese Cities. 2020. Available online: https://ssrn.com/abstract=3561491 (accessed on 15 April 2020).

27. Oliver, N.; Lepri, B.; Sterly, H.; Lambiotte, R.; Delataille, S.; De Nadai, M.; Letouzé, E.; Salah, A.A.; Benjamins, R.; Cattuto, C.; et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci. Adv.* **2020**. [CrossRef]

28. Pepe, E.; Bajardi, P.; Gauvin, L.; Privitera, F.; Lake, B.; Cattuto, C.; Tizzoni, M. COVID-19 outbreak response: A first assessment of mobility changes in Italy following national lockdown. *medRxiv* **2020**. [CrossRef]

29. Kashir, J.; Yaqinuddin, A. Loop mediated isothermal amplification (LAMP) assays as a rapid diagnostic for COVID-19. *Med. Hypotheses* **2020**, *141*, 109786. [CrossRef] [PubMed]

30. Nguyen, T.; Duong Bang, D.; Wolff, A. 2019 novel coronavirus disease (COVID-19): Paving the road for rapid detection and point-of-care diagnostics. *Micromachines* **2020**, *11*, 306. [CrossRef]

31. Fine, P.; Eames, K.; Heymann, D.L. "Herd immunity": A rough guide. *Clin. Infect. Dis.* **2011**, *52*, 911–916. [CrossRef] [PubMed]

32. Rao, A.S.S.; Vazquez, J.A. Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine. *Infect. Control. Hosp. Epidemiol.* **2020**, 1–18. [CrossRef]

33. Oliver, N.; Letouzé, E.; Sterly, H.; Delataille, S.; De Nadai, M.; Lepri, B.; Lambiotte, R.; Benjamins, R.; Cattuto, C.; Colizza, V.; et al. Mobile phone data and COVID-19: Missing an opportunity? *arXiv* **2020**, arXiv:2003.12347.

34. Gómez Expósito, A.; Rosendo Macías, J.A.; González Cagigal, M.Á. *Modelado y Análisis de la Evolución de una Epidemia Vírica Mediante Filtros de Kalman: El Caso del COVID-19 en España*; Technical report; Universidad de Sevilla: Sevilla, Spain, 2020.

35. Zhang, S.; Diao, M.; Yu, W.; Pei, L.; Lin, Z.; Chen, D. 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. *Int. J. Infect. Dis.* **2020**, *93*, 201–204. [CrossRef]

36. Fang, Y.; Nie, Y.; Penny, M. Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis. *J. Med. Virol.* **2020**, *92*, 645–659. [CrossRef]

37. Mahalle, P.; Kalamkar, A.B.; Dey, N.; Chaki, J.; Shinde, G.R. Forecasting Models for Coronavirus (COVID-19): A Survey of the State-of-the-Art. *TechRxiv* **2020**. Available online: https://www.techrxiv.org/articles/Forecasting_Models_for_Coronavirus_COVID-19_A_Survey_of_the_State-of-the-Art/12101547/1 (accessed on 25 April 2020).

38. Perone, G. An ARIMA model to forecast the spread of COVID-2019 epidemic in Italy. *arXiv* **2020**, arXiv:2004.00382.

39. Calafiore, G.C.; Novara, C.; Possieri, C. A modified SIR model for the COVID-19 contagion in Italy, 2020. *arXiv* **2020**, arXiv:2003.14391.

40. Flaxman, S.; Mishra, S.; Gandy, A.; Unwin, H.J.; Coupl, H.; Mellan, T.A.; Zhu, H.; Berah, T.; Eaton, J.W.; Guzman, P.N.; et al. Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *arXiv* **2020**, arXiv:2004.11342.

41. Dandekar, R.; Barbastathis, G. Neural Network aided quarantine control model estimation of COVID spread in Wuhan, China. *arXiv* **2020**, arXiv:2003.09403.

42. Mooney, S.J.; Westreich, D.J.; El-Sayed, A.M. Epidemiology in the era of big data. *Epidemiology* **2015**, *26*, 390. [CrossRef]

43. ACAPS. Report on COVID19 Government Measures Updates. 2020. Available online: https://www.acaps.org/special-report/covid-19-government-measures-update (accessed on 12 April 2020).

44. Mizumoto, K.; Kagaya, K.; Zarebski, A.; Chowell, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* **2020**, *25*, 2000180. [CrossRef]

45. Baud, D.; Qi, X.; Nielsen-Saines, K.; Musso, D.; Pomar, L.; Favre, G. Real estimates of mortality following COVID-19 infection. *Lancet Infect. Dis.* **2020**. [CrossRef]

46. Oliver, N.; Barber, X.; Roomp, K.; Roomp, K. The covid19 impact survey: Assessing the pulse of the COVID-19 pandemic in Spain via 24 questions. *arXiv* **2020**, arXiv:2004.01014.

47. Kleinberg, B.; van der Vegt, I.; Mozes, M. Measuring emotions in the COVID-19 real world worry dataset. *arXiv* **2020**, arXiv:2004.04225.

48. Wang, Y.; Kang, H.; Liu, X.; Tong, Z. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. *J. Med. Virol.* **2020**, *92*, 538–539. [CrossRef]

49. World Health Organization. Coronavirus Disease 2019—Situation Reports. 2020. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed on 25 April 2020).

50. World Health Organization. Myth Busters. 2020. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters (accessed on 10 April 2020).

51. European Centre for Disease Prevention; Control. Situation Dashboard: Latest Available Data. 2020. Available online: https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html (accessed on 20 April 2020).

52. Centre for Medium-Range Weather Forecasts, E. ECMWF Forecasts. 2020. Available online: https://www.ecmwf.int/en/forecasts (accessed on 22 April 2020).

53. Nations, U. Publish Existing Data Following Open Data Guidelines. 2020. Available online: https://covid-19-response.unstatshub.org/open-data/publish-existing-data-as-open-data/ (accessed on 12 April 2020).

54. Johns Hopkins 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). 2020. Available online: https://coronavirus.jhu.edu/map.html (accessed on 25th April 2020).

55. Open Data Watch. What Is Being Said: Data in the Time of COVID-19. 2020. Available online: https://opendatawatch.com/what-is-being-said-data-in-the-time-of-covid-19/ (accessed on 15 April 2020).

56. Marivate, V.; Nsoesie, E.; Bekele, E.; Africa Open COVID-19 Data Working Group. Coronavirus COVID-19 (2019-nCoV) Data Repository for Africa. *Zenodo* **2020**. [CrossRef]

57. France, S.P. Données Hospitalières Relatives à L'épidémie de COVID-19. 2020. Available online: https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/ (accessed on 25 April 2020).

58. Santé Publique France. Doneés des Urgences HospitalièRes et de SOS MéDecins Relatives à L'épidémie de COVID-19. 2020. Available online: https://www.data.gouv.fr/en/datasets/donnees-des-urgences-hospitalieres-et-de-sos-medecins-relatives-a-lepidemie-de-covid-19/ (accessed on 25 April 2020).

59. France, S.P. Données Relatives aux Tests de Dépistage de COVID-19 Réalisés en Laboratoire de Ville. 2020. Available online: https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-tests-de-depistage-de-covid-19-realises-en-laboratoire-de-ville/ (accessed on 25 April 2020).

60. Robert Koch Institut. Coronavirus Disease 2019 (COVID-19) Daily Situation Report of the Robert Koch Institute. 2020. Available online: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Archiv.html (accessed on 27 April 2020).

61. Marivate, V.; de Waal, A.; Combrink, H.; Lebogo, O.; Moodley, S.; Mtsweni, N.; Rikhotso, V.; Welsh, J.; Mkhondwane, S. Coronavirus disease (COVID-19) case data-South Africa. *Zenodo* **2020**. [CrossRef]

62. Marivate, V.; Combrink, H.M. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. *Data Sci. J.* **2020**, *19*, 19. [CrossRef]

63. Ritchie, H.; Roser, M. Age structure. *Our World in Data*. 2020. Available online: https://ourworldindata.org/age-structure (accessed on 15 April 2020).

64. Climate Change Service, C. C3S Helps Health Experts Explore How Temperature and Humidity Affect Virus Spread. 2020. Available online: https://climate.copernicus.eu/c3s-helps-health-experts-explore-how-temperature-and-humidity-affect-virus-spread?q=coronavirus-and-climate-c3s-helps-health-experts-explore-how-temperature-and-humidity-affect-virus&utm_campaign=COVID19&utm_medium=posts&utm_source=social_media&fbclid=IwAR2h6Z3mUf1L7AvHhyOJ6F2PKIB4gmYIJyZuCS8sVaWgMIxlmjbQe_jkYaU (accessed on 17 April 2020).

65. Gao, S.; Rao, J.; Kang, Y.; Liang, Y.; Kruse, J. Mapping County-Level Mobility Pattern Changes in the United States in Response to COVID-19. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3570145 (accessed on 18 April 2020).

66. Tauberer, J.; Lessig, L. The 8 Principles of Open Government Data. 2007. Available online: http://www.opengovdata.org/home/8principles (accessed on 24 April 2020).

67. Abella, A.; Ortiz-de Urbina-Criado, M.; De-Pablos-Heredero, C. MEloda 5: A Metric to Assess Open Data Reusability. 2019. Available online: http://www.elprofesionaldelainformacion.com/contenidos/2019/nov/abella-ortiz-pablos.pdf (accessed on 18 April 2020).