Open data on Covid-19 in the Spanish autonomous communities: reutilization in spatial epidemiology studies

José-Antonio Salvador-Oliván; Severino Escolano-Utrilla

Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86882

Recommended citation:

Salvador-Oliván, José-Antonio; **Escolano-Utrilla, Severino** (2022). "Open data on Covid-19 in the Spanish autonomous communities: reutilization in spatial epidemiology studies". *Profesional de la información*, v. 31, n. 4, e310410.

https://doi.org/10.3145/epi.2022.jul.10

Manuscript received on June 23th 2022
Accepted on July 8th 2022

José-Antonio Salvador-Oliván ✉
https://orcid.org/0000-0001-8568-3098

Universidad de Zaragoza
Facultad de Medicina
Departamento de Ciencias de la Documentación e Historia de la Ciencia
Pedro Cerbuna, 12. 50009 Zaragoza, Spain
jaso@unizar.es

Severino Escolano-Utrilla
https://orcid.org/0000-0002-3489-0692

Universidad de Zaragoza
Facultad de Filosofía y Letras
Departamento de Geografía y Ordenación del Territorio
San Juan Bosco, 7. 50009 Zaragoza, Spain
severino@unizar.es

Abstract

The Covid-19 pandemic has highlighted the need for governments and health administrations at all levels to have an open data registry that facilitates decision-making in the planning and management of health resources and provides information to citizens on the evolution of the epidemic. The concept of "open data" includes the possibility of reutilization by third parties. Space and time are basic dimensions used to structure and interpret the data of the variables that refer to the health status of the people themselves. Hence, the main objective of this study is to evaluate whether the autonomous communities' data files regarding Covid-19 are reusable to analyze the evolution of the disease in basic spatial and temporal analysis units at the regional and national levels. To this end, open data files containing the number of diagnosed cases of Covid-19 distributed in basic health or administrative spatial units and temporal units were selected from the portals of the Spanish autonomous communities. The presence of infection-related, demographic, and temporal variables, as well as the download format and metadata, were mainly evaluated. Whether the structure of the files was homogeneous and adequate for the application of spatial analysis techniques was also analyzed. The results reveal a lack of standardization in the collection of data in both spatial and temporal units and an absence of, or ambiguity in, the meaning of the variables owing to a lack of metadata. An inadequate structure was also found in the files of seven autonomous communities, which would require subsequent processing of the data to enable their reuse and the application of analysis and spatial modeling techniques, both when carrying out global analyses and when comparing patterns of evolution between different regions.

Keywords

Open data; Covid-19; Pandemics; Coronavirus; Reutilization of open data; Health information; Health data portals; Autonomous communities; Spain; Spatial epidemiology; Geolocation.

Funding

This work has been financed by the I+D+i Project granted by the Spanish *Ministry of Science and Innovation*, titled "The spatial–temporal spread of Covid-19 (Sars-CoV-2) in Spain and its multiscale relationship with patterns of daily mobility and sociodemographic vulnerability" (Ref. PID2020-115904RB-I00)

Conflicts of interest

The authors declare that there are no conflicts of interest.

1. Introduction

The term "open data" first appeared in 1995 in a paper from an American scientific agency concerning the dissemination of geophysical and environmental data, in which the authors promoted complete and open exchange of scientific information between different countries. Years later, in 2007, a group of open software activists established the concept of open data, defining public data as a common good founded on the principles of openness, participation, and collaboration (**Chignard**, 2013).

The *Open Government Working Group* recommended eight principles related to the access to and use of public data. These principles state that public data are open when the following conditions are met: (1) complete, (2) primary, (3) timely, (4) accessible, (5) machine-processable, (6) nondiscriminatory, (7) nonproprietary, and (8) unlicensed (**Dawes**, 2010).

It is important to distinguish data availability from data usability, although the goal of an open data policy should be to achieve both. Since most open data policies are enforced by legislative bodies without specific guidelines regarding how data should be published in a manner aiding public use, administrators often focus more on data availability than usability, creating a gap between data quality and system quality that makes it impossible for users to utilize the data's full potential (**Park**; **Gil-García**, 2022).

Given this concept of open data, an open data portal should provide data that can be reutilized by professionals, other data users, and the general public. Having an update mechanism, using a data management system, and providing API availability, aspects that more than half of the portals in Spain in 2021 lack, are some of the minimum characteristics that portals must meet for their data to be usable by professionals (**Abella**; **Ortiz-de-Urbina-Criado**; **De-Pablos-Heredero**, 2022).

Different models have been developed to assess the degree of open data reutilization. Meloda 5 is a metric based on various dimensions related to legal license, access to data, technical standards, standardization, content related to geolocation, frequency of updates, dissemination, and reputation (**Abella**; **Ortiz-de-Urbina-Criado**; **De-Pablos-Heredero**, 2019). Another, simpler model was developed by Tim Berners-Lee, in which the quality of data is measured by assigning five stars according to their availability (levels 1, 2, and 3) and reusability (levels 4 and 5) (**Berners-Lee**, 2009). Other requirements that the data must meet for reutilization are: integrity (completeness), accessibility and visibility, ease of use and understanding, updating, validity and usefulness, quality, and granularity (primary data) (**Lourenço**, 2015).

The expansion and effects of the Covid-19 pandemic, and the initial ignorance regarding many aspects of the infection, showed the value of having adequate and publicly available data on a suitable spatial and temporal scale to be able to utilize them advantageously for planning and modeling purposes (**Gardner** *et al*., 2021).

The *World Health Organization* (*WHO*), the *European Centre for Disease Prevention and Control* (*ECDC*), and *Johns Hopkins University* rapidly developed datasets to analyze pandemic trends and assess the risks of spread and contagion in various countries and regions around the world (**Ashofteh**; **Bravo**, 2020). At the international level, a large amount of open-format data has been provided to be utilized or reutilized by researchers in order to predict the spread of the disease and identify potentially vulnerable populations or areas (**Ball**, 2020). Many resources and open data repositories related to Covid-19 have been created by governments and institutions (**Hu** *et al*., 2020; **Pecoraro**; **Luzi**, 2021). **Alamo** *et al*. (2020) present the main datasets available on the Internet related to Covid-19 and list the most relevant institutions that submit up-to-date information, accompanied by a description of their characteristics.

In Spain, it is the autonomous communities (ACs) that manage health information in their territories and have the capacity and responsibility to publish data related to the pandemic via their specific open data portals on coronavirus (**Díez-Garrido**; **Melero-Lázaro**, 2022). CovidDATA-19 collects data on confirmed daily cases, deaths, ICU cases, and AC discharge from the Datadist data source (**Ferrer-Sapena** *et al*., 2020). The *Government of Spain*, through the *Carlos III Health Institute*, also publishes open data on the number of cases by diagnostic technique and province of residence, as well as the number of hospitalizations, ICU admissions, and deaths by sex, age, and province (*ISCIII*, 2022).

Despite the enormous effort involved in the creation of these portals, the data are used to describe the incidence in report format, visualizations, and downloads only on a province-by-province basis, but are insufficient to apply techniques with which to build spatial and temporal models that allow the prediction of the evolution of the pandemic and the analysis of the effectiveness of containment measures in more precise, homogeneous, and comparable basic units, such as municipalities or Basic Health Zones (BHZs). Therefore, it is necessary to visit the open data portals of the ACs to determine whether it is possible to reutilize Covid-19 data to which spatial analysis techniques can be applied. These techniques allow for the identification of spatial–temporal patterns that can, in turn, model or predict the propagation and incidence in spatial units with municipality or BHZ resolution (**Escolano-Utrilla**; **Salvador-Oliván**, 2022).

Some studies of the open data portals on coronavirus in the ACs have evaluated what data they contain and the download format (**Díez-Garrido**; **Melero-Lázaro**, 2022; **Martín-Fernández** *et al*., 2021), but none of them have assessed whether it is possible to reutilize them in specialized discipline research and, more specifically, in studies that seek to apply spatial analysis techniques.

The development of spatial–temporal models relating to incidence, risk, and propagation requires data that meet certain characteristics in terms of their content and form of expression. The objective of this study is thus to determine whether the portals of the ACs contain reusable Covid-19 data files to analyze the evolution of the pandemic in basic spatial and temporal analysis units, at both the AC and national levels.

2. Spatial epidemiology and health geography

Epidemiological studies aim to explain the distribution of a disease in a population by identifying people at risk, the time of onset of the disease, and their location (**Rothman**; **Greenland**; **Lash**, 2011). Geographical location has become an essential element in understanding the dynamics and evolution of an epidemic, thus spatial epidemiology has emerged as a novel approach to understanding and controlling current epidemics.

Moreover, health geography also addresses the study of spatial distribution and the incidence of diseases, but with a territorial approach; that is, the territory as a physical space and as a social construction situates it at the center of research. Hence, the thematic domains of health geography are broader than those of space epidemiology, since they also include health systems content (the distribution of health equipment, resources, services, etc.) as well as demographic, environmental, and planning content (**Souris**, 2019).

Studies of spatial epidemiology and health geography have contributed significantly to knowledge on how the Covid-19 pandemic has spread and the factors that influence its uneven incidence and social and territorial risk. Among other aspects, the research developed in this field has improved the knowledge of spatiotemporal patterns of geographical propagation and the transmission routes of the pandemic (**Franch-Pardo** *et al*., 2021) and has made it possible to identify local critical points (**Hu** *et al*., 2020) and calculate the number of new infections per case (*R* number), as well as to assess the incidence of multiple socioeconomic, demographic, environmental, and infection risk factors (**Elliott**; **Wartenberg,** 2004; **Kang** *et al.*, 2020; **Xu**; **Kraemer**, 2020; **Kobayashi** *et al*., 2021; **Paez**, 2021).

Although spatial epidemiology and health geography have different areas and objectives of investigation, they share numerous methods, techniques, data, and means of spatial and spatial –temporal analysis, since geographical space lies at the core of both disciplines (**Souris**, 2019). For data on health events –changes in the physiological or health status of individuals in a population– to be usable in these disciplines, they must include thematic as well as location information.

Most studies of spatial epidemiology and health geography apply spatial analysis procedures implemented in geographic information system (GIS) programs or similar systems that are capable of analyzing digital geographic information (**Kirby**; **Delmelle**; **Elbert**, 2017). The most widespread data model in GIS uses points, lines, polygons, pixels, and other digital objects to represent the location and shape of the real phenomena and objects (people, territories, environmental variables, etc.) under study. The location description usually follows some standard of data and metadata and is connected to the thematic information, usually presented in relational tables, by means of a geocode. Spatial information is considered to be part of the public infrastructure, maintained by institutions dedicated specifically to the creation and dissemination of geographical information (*Ministry of Transport, Mobility, and Urban Agenda*, date unknown).

Data regarding health events are taken from individuals whose location can be registered. Individual data allow for the reconstruction of trajectories and behaviors in great detail and in local contexts, but are available only in rare circumstances. Individual data on health events are normally published by aggregating the information into spatial units with the highest possible resolution. In turn, these units can merge into larger ones –often resulting in a loss of spatial resolution– depending on the objectives of each study or the spatial scope of health planning. From a practical point of view, it is advisable that the data refer to specific spatial units, such as BHZs, Health Areas, or others, or administrative units (census sections, districts, or municipalities) represented by their geocode and whose spatial information is updated and accessible. This option also facilitates the use of other (demographic, socioeconomic, and environmental) information referring to these same spatial units.

3. Methods

Data sources

The Spanish autonomous communities' open data portals on coronavirus were used as analysis units. These portals were selected because their objective is to disseminate datasets in reusable formats so that other citizens and researchers can also use them (**García-García**; **Curto-Rodríguez**, 2018).

They were located through a Google search using the terms "open data," "coronavirus," and the name of the autonomous community, and the website of the open data portal of each AC was accessed during the month of May 2022.

These portals contain multiple Covid-19 data files, so only those that met the following inclusion criteria, suitable to address the objective of this study, were selected:

- Files with data on diagnosed cases of Covid-19 in BHZs (districts or health centers) or administrative units (municipalities, counties, or provinces).
- Files with data on diagnosed cases of Covid-19 in temporal units (data collected in days, weeks, or different periods or waves of the pandemic).

The reason for having information with high spatial and temporal resolution is due to its strategic value for the management of health resources and for the adoption of measures to control the pandemic at the local level at which health systems operate.

When some ACs did not have files that met the two inclusion criteria, those that met either of them were selected.

The following exclusion criteria were applied:

- Files with global data for the entire AC, not disaggregated into spatial units.
- Files with final cumulative data exclusively, or only with figures from the last 7 or 14 days.

Variables and data extraction

On the basis of the *Meloda 5* metric for the reutilization of open datasets (**Abella**; **Ortiz-de-Urbina-Criado**; **De-Pablos-Heredero**, 2019) and variables used to obtain epidemiological models (**Alamo** *et al*., 2020), the variables necessary to analyze and spatially represent the spatial–temporal evolution and situation of the pandemic in basic spatial units were selected.

The variables were classified into six categories:

(1) portal identification data;

(2) geographical variables, of crucial importance since the place where an individual lives is considered to be a potential determinant of morbidity and mortality (**Diez-Roux**, 2001);

(3) data on Covid-19 infections;

(4) temporal variables;

(5) demographic variables, also important to determine the density and demographic and socioeconomic structures of the population since they have been recognized as a risk factor in the transmission of the disease and in the incidence rate (**Priyadarsini**; **Suresh**, 2020; **Sy** *et al*., 2021); and

(6) other variables, such as the download format or the existence of metadata that facilitate the interpretation of the data in the portals.

Of the files downloaded from the ACs that met the inclusion criteria, the presence or absence of the variables were evaluated and the data were collected in an Excel file. The structure of the open data files was also evaluated to determine whether they were appropriate for the application of spatial–statistical procedures. Virtually all geographic information systems use a “list” data structure for the analysis of spatiotemporal data, which has at least three columns:

- one that registers the codes of the spatial units (which are repeated for each time measurement);
- another that records the measurements of the time variable (expressed in appropriate units and format);
- and a third that contains the values of the variable of interest (the phenomenon studied).

Data in table form, with spatial units in rows and temporal units in columns, can be used for synchronous analysis and can also be transformed into “list” format by means of a transposition operation. For this reason, the selected files were classified into three types:

- adequate structure if each row contained an identifiable territorial unit (municipality, health district, or BHZ) and its number of cases according to a temporal distribution (days or weeks);
- inadequate structure, but that can be reutilized with data processing; and
- inadequate structure owing to a lack of essential data, i.e., where it is not possible to reutilize the data.

The complete list of selected variables and the *Excel* file are available on the *Open Science Framework* (*OSF*) at *https://osf.io/r6byj*

4. Results

4.1. General characteristics

All the ACs, except Extremadura and the autonomous Cities of Ceuta and Melilla, offered open data files relating to the coronavirus pandemic. Of the 16 ACs, only 9 state that the license type for the use of data files is Creative Commons whenever authorship is recognized (Attribution or By), while only one, Castile–La Mancha, specifies that it must be shared as the original license.

> Open data are those that are not only shared but can be reused by third parties to carry out studies in their scientific disciplines

Half of the ACs with open data present metadata that describe the meaning and content of the variables in the files, i.e., Andalusia, the Balearic Islands, the Canary Islands, Castile and Leon, Catalonia, the Valencian Community, Navarre, and the Basque Country.

All the ACs publish data in machine-readable, nonproprietary format, so there are no barriers to reutilization of the available data. The data are in csv format in all of them, while only the Balearic Islands and Catalonia offer their data in rdf format as well.

Table 1. General characteristics of the open data files of the autonomous communities

Autonomous community	Legal license	Metadata	Data format								
			rdf	csv	json	PC-axis	ods	xml	rss	tsv	*Excel*
Andalusia		✓		✓	✓	✓	✓				✓
Aragon				✓							✓
Asturias	CC BY			✓							
Balearic Islands		✓	✓	✓	✓			✓	✓	✓	
Canary Islands	CC BY	✓		✓							
Cantabria	CC BY			✓							✓
Castile–La Mancha				✓							
Castile and Leon	CC BY-SA	✓		✓	✓			✓			✓
Catalonia	CC BY	✓	✓	✓				✓	✓	✓	
Valencian Community		✓		✓	✓			✓		✓	
Galicia	CC BY			✓							
La Rioja				✓	✓			✓			✓
Madrid	CC BY			✓	✓						
Murcia	CC BY			✓							
Navarre		✓		✓	✓			✓		✓	✓
Basque Country		✓		✓	✓		*	✓			✓

4.2. Geographical and temporal data of the files

Table 2 summarizes the main characteristics of the file of each AC that presents a higher level of spatial and temporal granularity, declaration of cases, and completeness of the data (period of time covered).

Table 2. Characteristics of the Covid-19 data file with greater granularity and completeness in each autonomous community

Autonomous community	Spatial unit	Diagnosed cases	Temporal resolution	Period
Andalusia	Health districts	Diagnostic tests	Weeks	Complete
Aragon	BHZ	Total	Days	Complete
Asturias	Health areas	Total	Weeks	Until 12/2021
Balearic Islands	BHZ	Total	Days	Complete
Canary Islands	Municipalities	Total	Days	01/2021–03/2022
Cantabria	Municipalities	Total	No	Acute phase
	No	Total	Days	Acute phase
Castile–La Mancha	Provinces	Total	No	Acute phase
	No	Diagnostic tests	Days	Acute phase
	Municipalities	Diagnostic tests (> 60)	Two weeks	03/2022–
Castile and Leon	BHZ	Totals, diagnostic tests, and PCR	Days	Complete
Catalonia	Basic Health Area	Totals, diagnostic tests	Days	Complete
Valencian Community	Health departments	Diagnostic tests	Days	Complete
Galicia	Health areas	Total	Days	Complete
La Rioja	BHZ	Total	No	Complete
Madrid	BHZ	Total	Days/weeks	Complete
Murcia	BHZ	Total	Days	Acute phase
Navarre	BHZ	Total	Days	Acute phase
Basque Country	BHZ	Total	Days	13/05/2020–present

Andalusia provides the number of cases diagnosed by diagnostic tests for active infection (PCR and antigen tests) by health districts, with their names and *National Statistics Institute* (*INE*) codes and by weeks, from March 2020 to March 2022. As of March 2022, data for people 60 years of age or older are collected in another file (which includes data since the beginning of the pandemic). Therefore, the two files cover the entire period, and also contain the description of the included variables.

For Aragon, a data file with the total number of cases distributed daily and by Basic Health Zones from the beginning of the pandemic to the present can be downloaded. As of mid-May 2020, confirmed cases were diagnosed using PCR and antigen testing, which resulted in an increase from days prior, in which cases were only diagnosed using PCR tests. The drawback is that a file must be downloaded for each day until March 2022, after which the files are downloaded by weeks, collecting the daily data for that week. There are no metadata.

Asturias has a file with the total number of cases (although the means of diagnosis is not indicated) distributed weekly by health areas from the beginning of the pandemic until December 2021.

The Balearic Islands has a data file with the total number of cases diagnosed by PCR and antigen tests distributed by BHZ (with their geocodes) and days, according to age and sex, from the beginning of the pandemic to the present. Metadata are available.

The Canary Islands has a file with the number of cases and deaths, according to age and sex, distributed by municipalities (both residence and assigned) and days, from January 2021 to March 2022. Consequently, it does not cover the entire period of the acute phase of the pandemic. Each record corresponds to a case, and in each the sex and age group are indicated.

Cantabria does not have a single file with data disaggregated into spatial and temporal units; rather, it maintains two files: one which collects the total number of cumulative cases by municipalities, and another which has the total number of cases per day but for the whole community. In both cases, the time period runs until March 2022.

For Castile–La Mancha, there is also no single file disaggregated into spatial and temporal units; once again there are two files: one which includes the total number of final cumulative cases and deaths by province, and another which includes the number of cases confirmed by diagnostic tests distributed daily for the entire AC. Both cover the acute phase of the pandemic. Since March 2022, another data file was created with the number of cases in the last two weeks in people 60 years of age or older, distributed by municipalities.

Castile and Leon provides a file with the total number of cases and those diagnosed by diagnostic and PCR tests, distributed by days and health centers (with their GPS codes and coordinates), as well as the total number of tests performed, extending from the beginning of the pandemic to the present.

For Catalonia, there are four files with data disaggregated into spatial and temporal units. The most useful one is that which contains the total number of cases and those diagnosed by diagnostic tests (PCR and antigen tests) distributed by days in Basic Health Areas (BHAs) throughout the pandemic period to the present. Metadata are available.

The Valencian Community has a file with cases confirmed by diagnostic tests according to age and sex, distributed daily by Health Departments throughout the pandemic to date.

In Galicia, the file contains all the cases diagnosed by diagnostic tests and self-tests distributed by Health Areas, and covers the days during the entire pandemic period to the present.

La Rioja has several data files that are not downloadable, and others that lead to a 404 error page or contain inaccurate data (for example, in a number of cases “<10” appears instead of the exact value). The most useful data file is the one that collects the cumulative number of positive cases by BHZ broken down by PCR, antigen, and CLIA (ELISA/antibody test) tests, with the total number of tests performed. The beginning of the period is not specified, but it extends to the present.

> The Spanish autonomic open data portals on Covid-19 lack standardization in the data collected and a homogeneous structure, which makes it difficult to reuse

For Madrid, there are two files. One of them contains all the diagnosed cases distributed by BHZ and days until 1 July 2020; from July 2, the data are distributed by weeks until March 2022. From March to the present, cases are only recorded in people 60 years of age or older, distributed by BHZ and days.

In Murcia there are two files, one until 30 November 2020, and another from 1 December 2020 to March 2022, with the total number of cases distributed by BHZ and days.

Navarre has a file with the total number of cumulative cases distributed by BHZ and days from the beginning of the pandemic to March 2022. To determine the number of new cases per day and per BHZ, it is necessary to process the data.

In the Basque Country, there is a file that collects the total number of cases distributed by BHZ with their codes and by days, from 15 May 2020 to the present. In this file, the days are listed in the columns, unlike in the rest of the ACs, where they are listed in the rows. It contains the number of cases diagnosed by PCR, antigen tests, and antibody tests, as well as the number of tests performed, and the number of deaths, hospitalizations, and ICU cases, by age group and sex.

The highest level of territorial disaggregation is observed in the nine ACs that provide data at the BHZ scale, including Catalonia, although its denomination is different (Basic Health Area, or BHA). In Andalusia, Asturias, the Valencian Community, and Galicia, the data are aggregated in lower-resolution units, with designations specific to each community. Three ACs (Canary Islands, Cantabria, and Castile–La Mancha) do not provide data by areas of health service management, but rather administrative units such as municipalities or provinces (Fig. 1).

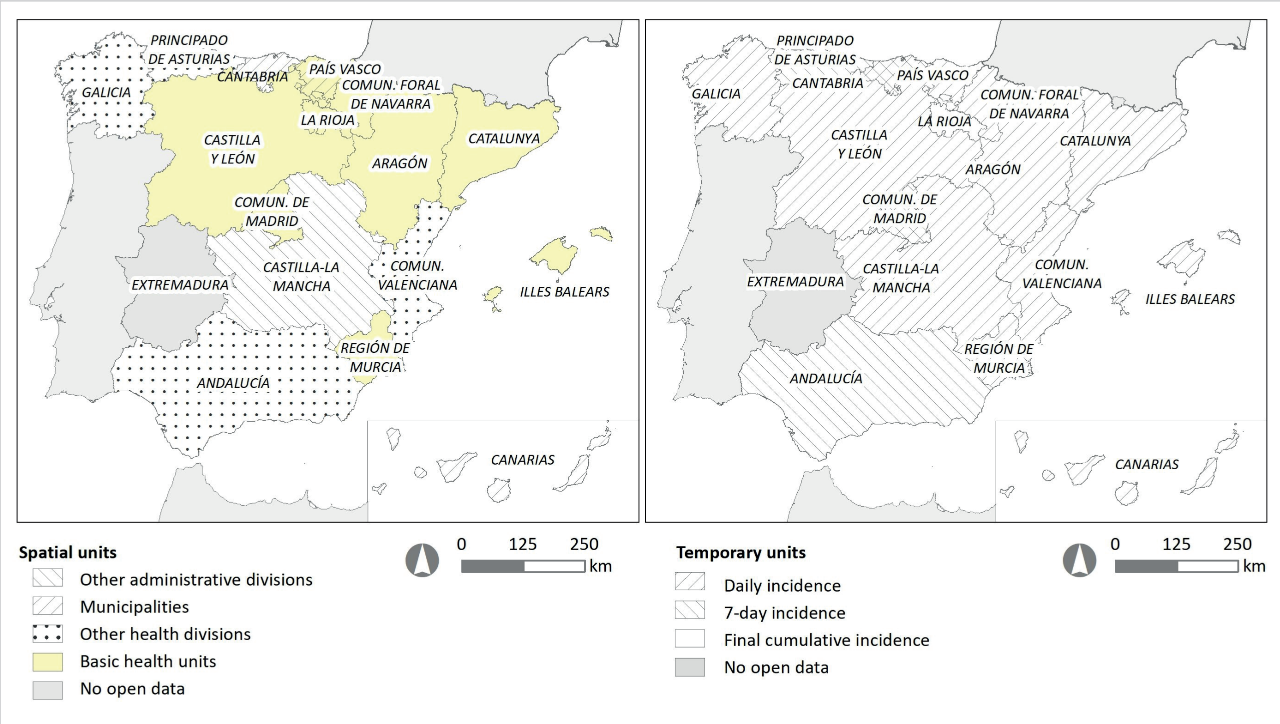

Figure 1. Representation of spatial (left) and temporal (right) units of the open data records of Covid-19 by autonomous communities

Regarding the temporal distribution, the number of cases is recorded by days in all the ACs except in Andalusia and Asturias, where it is recorded by weeks. In Castile–La Mancha, daily data are not registered by spatial units and, when they appear by weeks, the data are the total cumulative cases from the last two weeks only in those 60 years or older. In Cantabria, the data are registered in the same way; the daily data refer to the entire AC, and when the data are by municipality, they are final cumulative data. In La Rioja, the data are also not distributed in temporal units and record the final cumulative cases by BHZ (Fig. 1).

4.3. Structure of the files

Table 3 presents the structure of the files of the ACs, as well as their assessment and data processing requirements with respect to their use in applications of spatial analysis techniques and whether subsequent data processing is necessary.

Table 3. Possibility of reutilizing the autonomous communities' data files to apply spatial analysis techniques

Autonomous community	Structure	Requires data processing
Andalusia	×	Redundant information
Aragon	×	One file per day or week has to be downloaded. The variable "day" must be included in the global file
Asturias	×	The basic health areas have to be moved to rows
Balearic Islands	✓	
Canary Islands	×	Each row corresponds to a single case. Necessary to add the number of cases in a day
Cantabria	××	Data are not recorded in spatial units
Castile–La Mancha	××	Missing data
Castile and Leon	✓	
Catalonia	✓	
Valencian Community	×	Departments of Health need to be moved to rows (data transposition)
Galicia	✓	
La Rioja	××	Missing data
Madrid	✓	
Murcia	✓	
Navarre	×	Submits the number of cumulative cases for each day and BHZ. Deduction necessary to find out the number of daily cases and add them to aggregate them in weeks or other periods
Basque Country	×	Days need to be moved to rows (they are in columns)

(✓ Adequate structure;; × Requires data processing; ×× Cannot be reutilized)

Only six ACs have a file with the appropriate structure, where each row corresponds to the number of cases per BHZ and per day or other period. In five ACs, data processing is necessary to adapt the structure. Of these, the file from Andalusia presents the least difficulties, since the process simply consists of eliminating rows with redundant information (the Health Districts along with the number of cases appear, as do the provinces, which contain the total number of cases from the districts that belong to each province). For Aragon, it is necessary to download as many files as there are days, and after March 2022, as many as there are weeks; in each file, there is a sheet with several tables, so the data of the number of cases and BHZ have to be copied and pasted into a single file, and the day to which the file belongs should be included as a variable.

In the Canary Islands file, each row corresponds to a single case per municipality along with the date of diagnosis, so it is necessary to add up the number of cases on the same day and in the same municipality. However, in Navarre, the number of cumulative cases for each day and BHZ is provided, thus making it necessary to obtain the number of daily cases in each BHZ per day by subtracting the value of cumulative cases up to the current day from the previous day.

In the Asturias file, there are eight columns that correspond to the number of cases in each of the eight Health Areas; it is necessary to convert the eight columns to a single column with the names of all the Health Areas, repeating for each day. In the Valencian Community, something similar happens: there are 24 columns, each compiling the number of cases per Department of Health, meaning that all the columns must be converted into a single one by compiling the names of the departments. Finally, in the Basque Country, the opposite situation occurs: each row is a separate BHZ and the columns are the days, so in this case, it is also necessary to move each day to a single column with as many rows as there are days for representation and temporal analysis with geographic information systems.

The data in the files from Cantabria, Castile–La Mancha, and La Rioja could not be used since variables are missing, related either to the territory or to the temporal distribution of the cases..

5. Discussion

The results of this work reveal the lack of standardization and heterogeneity of the data models provided by the different ACs. This mainly affects the case diagnosis procedure (PCR, antigen tests, or antibody tests), temporal registration of number of cases (daily, weekly, by wave, or cumulative), the spatial units (BHZs, Health Districts, Health Areas, municipalities, counties, or provinces), and the period of time covered (from the beginning of the pandemic to the present, the acute phase of the pandemic, and from March 2022 with the new Surveillance and Control Strategy), in addition to other data that complement the epidemiological analysis of the pandemic, such as the number of deaths or ICU admissions, the number of tests performed, total population or by age group and sex, etc.

In general, the problems found in our study in the open data portals are consistent with those obtained by **Gardner** *et al*. (2021) in the collection of Covid-19 data:

- ambiguity of the parameter definitions;
- different rates of updating variables between the portals; and
- deficiencies in the system's structure, which is probably of a hierarchical nature, obtaining the data at the local level (BHZs or municipalities) and adding them to Health Areas or provinces.

In addition, other drawbacks that hinder the reutilization of data and that are consistent with those indicated by **Alamo** *et al*. (2020) were identified: a lack of standardization in data collection, incomplete data (including only cases confirmed by a laboratory test, which also concerns the number of deaths), and a lack of precision in the measurement of key variables (for example, not collecting asymptomatic cases, which play an important role in the transmission of the virus).

Although there are a variety of formats for downloading data, in all the ACs, the files can be downloaded in a format suitable for further processing in a database or spreadsheet application, such as Excel or other similar programs (xls, xlxs, csv).

The problems posed by the lack of data normalization between ACs and the poor registration in each of them are discussed below, focusing on the core variables in spatial epidemiology: cases with Covid infection, geographical location, and temporal distribution.

5.1. Covid variables

One of the problems identified is the absence of, and/or ambiguity in, the definition of these variables, caused by a lack of metadata or an insufficient description of their meaning. The key thematic variable for the analysis of any epidemic is the number of infected cases. The *World Health Organization* defines a confirmed case as any person in whom infection has been confirmed by laboratory testing (*World Health Organization*, 2022). This fact conditions the actual estimate of infected cases, because not all people who were symptomatic and infected could access tests, especially during the early months of the pandemic, and those who were asymptomatic or mildly symptomatic were not tested.

In most of the AC files, the total number of cases appears without indicating how they were diagnosed, whether they were confirmed through diagnostic tests (PCR, antigen tests, or serological tests), and how many diagnostic tests were employed (one, several, or all), thus making any analysis of the evolution of the pandemic over time (by applying a new methodology in case accounting) and comparisons between ACs more complex.

As of 28 March 2022, following the acute phase of the pandemic, the new Covid-19 Surveillance and Control Strategy came into force, which represents a significant change in registration method, since only cases confirmed by diagnostic testing in 60 years or older, immunosuppressed, and pregnant populations are included (*Ministry of Health*, 2022). However, there are ACs in which all cases continue to be registered from this date, while in others, only those cases in people 60 years or older are still registered.

There is also a lack of uniformity regarding the number of laboratory tests, which may refer to the total number of tests performed or the number of people examined, without accurately describing the meaning of the counts (**Alamo** *et al*., 2020).

To carry out any study, it is necessary to know what a variable measures and how it has been measured. Data files should provide well-structured, organized, and detailed metadata that are of high quality and ensure the correct and appropriate interpretation of the data (**Barcellos**; **Bernardini**; **Viterbo**, 2022), so that the data can be understood and used by other researchers (**Kubler** *et al*., 2018; **Wu**; **Ma**; **Zhang**, 2021).

In this sense, FAIR data principles (findability, accessibility, interoperability, and reusability) define the characteristics that data, tools, vocabularies, and infrastructures must include to facilitate reutilization by third parties, ensuring rapid and appropriate access to complete and organized data (**Wilkinson** *et al*., 2016). Also noteworthy is the existence of the Metadata Quality Assessment (MQA) tool, developed by the data.europea.eu consortium to assess the quality of metadata of public open data and how to improve them (*Publications Office of the European Union*, 2020).

5.2. Temporal distribution

There is a lack of homogeneity in the time period recorded in the series of each AC. Although most records extend from the beginning of the pandemic to the present, some cover only the acute phase (from February–March 2020 to March 2022), only until December 2021, from January 2021 to March 2022, or from May 2020. This disparity makes it impossible to carry out a global analysis and compare, from a synchronous point of view, the behavior of the pandemic between ACs.

The temporal resolution of cases is also variable among the ACs; in some, the availability of data is daily, while in others it is weekly or even biweekly. This problem has mainly been observed when data are obtained from multiple sources (**Alamo** *et al*., 2020).

It is important that the dates of notification and registration of cases be a true reflection of reality. The delay in reporting cases and the lack of regular updating of data, at times taking several days or a weekend, may affect the accuracy and reliability of the data on the recorded dates and changes in patterns or trends not attributable to the observed phenomenon. There is also a lack of information on whether the date used for the registration of cases in the file has the same meaning in all ACs or is different, referring to the date of onset of symptoms, the date of performance of the diagnostic test, or the date of diagnosis minus three days (**Martín-Fernández** *et al*., 2021).

5.3. Spatial resolution

Another feature of the open data on Covid-19 is the lack of homogeneity of spatial units between the different ACs. One of the factors identified that have limited the use of spatial techniques to detect patterns of spread during the pandemic has been the lack of availability of data expressed in high-resolution spatial units and the bias of Covid-19 data, along with the scarcity of demographic data on a large scale or with high spatial resolution (**Fatima** *et al*., 2021). The results of our study reveal a lack of normalization in the availability of data in spatial units at different scales.

> The application of spatio-temporal analysis techniques to open data on Covid-19 is very useful to obtain geographic patterns and models that explain the dynamics, transmission and incidence of the disease

Article 62 of *General Health Law 14/1986* from April 25 (*España*, 1986) states that

> "to achieve maximum operation and effectiveness in the operation of services at the primary level, the Health Areas will be divided into Basic Health Zones" ["*para conseguir la máxima operatividad y eficacia en el funcionamiento de los servicios a nivel primario, las Áreas de Salud se dividirán en Zonas Básicas de Salud*"].

To carry out spatial and temporal studies regarding the spread, incidence, risk, and other aspects of interest related to the pandemic, it is essential to have information expressed in equivalent spatial units, such as BHZs or other administrative types of greater resolution, which should be comparable between ACs, cover the entire national territory, and be as stable as possible. This is also a necessary condition for the efficient management of health resources and for the adoption of measures to control the pandemic.

Recording in higher-resolution spatial units, such as census sections or municipalities and even BHZs, makes it possible to aggregate them, manually or automatically, into other smaller and larger units (health sectors, or others) for analysis on another scale. In contrast, operating in the opposite direction (for example, from provinces to health sectors or from BHZs to municipalities) is not direct and requires adopting certain assumptions about the distribution of the phenomenon studied and applying complex analytical techniques.

For most ACs, comprehensive data are available on the number of cases disaggregated by BHZ and day. Castile and Leon is the community with the best results in terms of transparency and information on the epidemiological situation of Covid-19; in that community, BHZs with GPS coordinates are noted. The worst community is noted to be Castile–La Mancha (**Díez-Garrido**; **Melero-Lázaro**, 2022). This is consistent with the analysis of our results, as the latter contains the final cumulative number of cases per province or the total number of cases per day for the entire autonomous community, which is not useful for subsequent data reutilization. In addition, it contains errors regarding the dates used: the 559 rows correspond to all the months of the year 2022, while the file was downloaded in May.

It is difficult to work with open data directly from official sources of the ACs, since some do not publish data, multiple methods of data collection are used, and they are not structured in a uniform way, which requires them to be reprocessed (**Ferrer-Sapena** *et al*., 2020). The open data files available come from various sources containing different variables (level of aggregation and meaning) that are organized and structured differently (in rows or columns). These deficiencies make it difficult to use them globally and for synchronous and diachronic analyses throughout the national territory. In addition, the existence of multiple files for each AC, some of which include repeated data, requires more time and resources (people and economic) to create and maintain them.

5.4. Reutilization at the national level and proposed data file structure

Currently, to carry out studies on the incidence and spread of Covid-19 throughout the national territory, it is necessary to group the data of the different ACs into a single file. The following ACs would not be included in this file: Extremadura and Ceuta and Melilla (as they do not have open data on Covid-19); Cantabria, Castile–La Mancha, and La Rioja (owing to lack of spatial or temporal data); and the Canary Islands (as it only has data by municipalities). In the rest of the ACs, the files selected in this study would have to undergo a process of restructuring and calculation of the data already described in the "Results" section, as well as normalization of the spatial variables, which would consist of grouping the Health Areas by Health Districts, and of the temporal variables, grouping the days into weeks, since this is the level of disaggregation that all the files can share.

To carry out spatial epidemiology studies at the country level, all ACs should have open data in reusable files with a homogeneous structure and the presence of metadata that accurately describe the meaning of the variables. The essential variables of the file must be related to: (a) infection: the number of diagnosed cases distributed according to different diagnostic tests (PCR, antigen tests, antibody tests), as well as the number of tests performed to confirm the diagnosis, and other variables such as hospitalizations or ICU cases and deaths; (b) spatial location: higher-resolution units such as BHZs, with their names and coordinates or geocodes; (c) temporal variables: days (full date); and (d) demographic: age and sex groups.

In this way, each row of the file would correspond to the number of cases diagnosed (distributed by test performed and by age group and sex) in a BHZ per day. The file would also exhibit desirable characteristics related to the integrity or completeness of the data (in terms of both the variables collected and the period of time covered), accuracy (diagnostic tests used for the declaration of confirmed cases), granularity (data provided on a larger scale of spatial disaggregation, by BHZ, and temporal, by days), and timeliness (data should be updated daily, without delays, so that it does not negatively affect the development of predictive models).

6. Conclusions

This article has addressed one of the fundamental principles of open data: the possibility of reutilization by third parties for studies of spatial epidemiology. From the analysis and evaluation of the open data related to Covid-19 available and accessible in the portals of the ACs of Spain, various problems have been identified that make it impossible to reutilize them for the development of spatial–temporal models of the coronavirus infection. Of these, the most important are inconsistency of data, ignorance and disparity of criteria in the meaning of some variables, basic units of nonhomogeneous analysis, lack of updating, different time series, and change of criterion in the diagnosis of confirmed cases. These problems make it impossible to compare the evolution of Covid-19 among the different ACs.

Having open data via portals is an important asset that is useful for multiple purposes, but the data must meet a series of characteristics that allow for their efficient use and thereby generate added value in research and create useful knowledge in decision-making. It is unnecessary to have a large quantity of files for each AC; the more there are, the more time and effort is needed to update them. It is sufficient to have one file for each AC that has the same structure and that collects all the important variables, which will facilitate more efficient maintenance and result in greater reliability of the data. The ACs register and publish the original data; if these are not reliable and current, their use by researchers or as secondary sources of information by other bodies, such as the Spanish *Ministry of Health*, the *European Commission*, or the *WHO*, will be completely useless.

> Not so many data files on Covid-19 are necessary in each autonomous community of Spain; one that records the main data at a higher level of temporal and spatial resolution is sufficient

7. References

Abella, Alberto; **Ortiz-de-Urbina-Criado, Marta**; **De-Pablos-Heredero, Carmen** (2019). "Meloda 5: A metric to assess open data reusability". *Profesional de la información*, v. 28, n. 6, e280620.
https://doi.org/10.3145/epi.2019.nov.20

Abella, Alberto; **Ortiz-de-Urbina-Criado, Marta**; **De-Pablos-Heredero, Carmen** (2022). "Criteria for the identification of ineffective open data portals: pretender open data portals". *Profesional de la información*, v. 31, n. 1, e310111.
https://doi.org/10.3145/epi.2022.ene.11

Alamo, Teodoro; **Reina, Daniel G.**; **Mammarella, Martina**; **Abella, Alberto** (2020). "Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic". *Electronics*, v. 9, n. 5, pp. 827.
https://doi.org/10.3390/electronics9050827

Ashofteh, Afshin; **Bravo, Jorge M.** (2020). "A study on the quality of novel coronavirus (Covid-19) official datasets". *Statistical journal of the IAOS*, v. 36, n. 2, pp. 291-301.
https://doi.org/10.3233/SJI-200674

Ball, William P. (2020). "Open data, open science and transparency in the time of Covid 19". *Radical statistics newsletter*, n. 127, pp. 4-9.

Barcellos, Raissa; **Bernardini, Flavia**; **Viterbo, José** (2022). "Towards defining data interpretability in open data portals: Challenges and research opportunities". *Information systems*, v. 106, 101961.
https://doi.org/10.1016/j.is.2021.101961

Berners-Lee, Tim (2009). *Linked data*.
https://www.w3.org/DesignIssues/LinkedData.html

Chignard, Simon (2013). "A brief history of open data". *Paris innovation review.*
http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data

Dawes, Sharon S. (2010). "Stewardship and usefulness: Policy principles for information-based transparency". *Government information quarterly*, v. 27, n. 4, pp. 377-383.
https://doi.org/10.1016/j.giq.2010.07.001

Díez-Garrido, María; **Melero-Lázaro, Mónica** (2022). "La transparencia como medicina. La comunicación de la pandemia a través de los portales de datos abiertos de las autonomías". *Estudios sobre el mensaje periodístico*, v. 28, n. 1, pp. 65-80.
https://doi.org/10.5209/esmp.77172

Diez-Roux, Ana V. (2001). "Investigating neighborhood and area effects on health". *American journal of public health*, v. 91, n. 11, pp. 1783-1789.
https://doi.org/10.2105/AJPH.91.11.1783

Elliott, Paul; **Wartenberg, Daniel** (2004). "Spatial epidemiology: Current approaches and future challenges". *Environmental health perspectives*, v. 112, n. 9, pp. 998-1006.
https://doi.org/10.1289/ehp.6735

Escolano-Utrilla, Severino; **Salvador-Oliván, José-Antonio** (2022). "Global regularity and local variability of the space-temporal patterns of Covid 19 in Aragón (Spain)". *Boletín de la Asociación de Geógrafos Españoles*, n. 93.
https://doi.org/10.21138/bage.3276

España (1986). "Ley 4/1986, de 25 de abril, General de Sanidad". *BOE*, n. 102, 29 abril.
https://www.boe.es/eli/es/l/1986/04/25/14/con

Fatima, Munazza; **O'Keefe, Kara J.**; **Wei, Wenjia**; **Arshad, Sana**; **Gruebner, Oliver** (2021). "Geospatial analysis of Covid-19: A scoping review". *International journal of environmental research and public health*, v. 18, n. 5, 2336.
https://doi.org/10.3390/ijerph18052336

Ferrer-Sapena, Antonia; **Calabuig, José-Manuel**; **Peset, Fernanda**; **Sánchez-del-Toro, Isabel** (2020). "Trabajar con datos abiertos en tiempos de alarma: el caso de uso de covidDATA-19". *Profesional de la información*, v. 29, n. 4, e290421.
https://doi.org/10.3145/epi.2020.jul.21

Franch-Pardo, Ivan; **Desjardins, Michael R.**; **Barea-Navarro, Isabel**; **Cerdà, Artemi** (2021). "A review of GIS methodologies to analyze the dynamics of Covid-19 in the second half of 2020". *Transactions in GIS*, v. 25, n. 5, pp. 2191-2239.
https://doi.org/10.1111/tgis.12792

García-García, Jesús; **Curto-Rodríguez, Ricardo** (2018). "Divulgación de información pública de las comunidades autónomas españolas (2013-2017): portal de datos abiertos, portal de transparencia y web institucional". *Profesional de la información*, v. 27, n. 5, pp. 1051-1060.
https://doi.org/10.3145/epi.2018.sep.09

Gardner, Lauren; **Ratcliff, Jeremy**; **Dong, Ensheng**; **Katz, Aaron** (2021). "A need for open public data standards and sharing in light of Covid-19". *The lancet infectious diseases*, v. 21, n. 4, e80.
https://doi.org/10.1016/S1473-3099(20)30635-6

Hu, Tao; **Guan, Weihe-Wendy**; **Zhu, Xinyan**; **Shao, Yuanzheng**; **Liu, Lingbo**; **Du, Jing**; **Liu, Hongqiang**; **Zhou, Huan**; **Wang, Jialei**; **She, Bing**; **Zhang, Luyao**; **Li, Zhibin**; **Wang, Peixiao**; **Tang, Yicheng**; **Hou, Ruizhi; Li, Yun**; **Sha, Dexuan**; **Yang, Yifan**; **Lewis, Ben**; **Kakkar, Devika**; **Bao, Shuming** (2020). "Building an open resources repository for Covid-19 research". *Data and information management*, v. 4, n. 3, pp. 130-147.
https://doi.org/10.2478/dim-2020-0012

ISCIII (2022). *Covid-19 en España*. ISCIII Centro Nacional de Epidemiología.
https://cnecovid.isciii.es/

Kang, Dayun; **Choi, Hyunho**; **Kim, Jong-Hun**; **Choi, Jungsoon** (2020). "Spatial epidemic dynamics of the Covid-19 outbreak in China". *International journal of infectious diseases*, v. 94, pp. 96-102.
https://doi.org/10.1016/j.ijid.2020.03.076

Kirby, Russell S.; **Delmelle, Eric**; **Erberth, Jan M.** (2017). "Advances in spatial epidemiology and geographic information systems". *Annals of epidemiology*, v. 27, n. 1.
https://doi.org/10.1016/j.annepidem.2016.12.001

Kobayashi, Shinji; **Falcón, Luis**; **Fraser, Hamish**; **Braa, Jorn**; **Amarakoon, Pamod**; **Marcelo, Alvin**; **Paton, Chris** (2021). "Using open source, open data, and civic technology to address the Covid-19 pandemic and infodemic". *Yearbook of medical informatics*, v. 30, n. 1, pp. 38-43.
https://doi.org/10.1055/s-0041-1726488

Kubler, Sylvain; **Robert, Jérémy**; **Neumaier, Sebastian**; **Umbrich, Jürgen**; **Le-Traon, Yves** (2018). "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process". *Government information quarterly*, v. 35, n. 1, pp. 13-29.
https://doi.org/10.1016/j.giq.2017.11.003

Lourenço, Rui-Pedro (2015). "An analysis of open government portals: A perspective of transparency for accountability". *Government information quarterly*, v. 32, n. 3, pp. 323-332.
https://doi.org/10.1016/j.giq.2015.05.006

Martín-Fernández, Almudena; **Alfaro-Latorre, Mercedes**; **Crespo-Sánchez-Eznarriaga, Belén**; **Fidalgo-García, Luz** (2021). "Disponibilidad de datos abiertos relacionados con la Covid-19 en los portales web de las comunidades autónomas y del Ministerio de Sanidad". *Revista española de salud pública*, v. 95, e202112190.
https://dialnet.unirioja.es/servlet/articulo?codigo=8254679

Ministry of Health (2022). *Estrategia de vigilancia y control frente a Covid-19 tras la fase aguda de la pandemia*.
https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Nueva_estrategia_vigilancia_y_control.pdf

Ministry of Transport, Mobility, and Urban Agenda. Infraestructura de Datos Espaciales de España (IDEE).
https://plataforma.idee.es/web/idee/inicio

Paez, Antonio (2021). "Reproducibility of research during Covid-19: Examining the case of population density and the basic reproductive rate from the perspective of spatial analysis". *Geographical analysis*, first online.
https://doi.org/10.1111/gean.12307

Park, Sora; **Gil-García, J. Ramón** (2022). "Open data innovation: Visualizations and process redesign as a way to bridge the transparency-accountability gap". *Government information quarterly*, v. 39, n. 1, 101456.
https://doi.org/10.1016/j.giq.2020.101456

Pecoraro, Fabrizio; **Luzi, Daniela** (2021). "Open data resources on Covid-19 in six European countries: Issues and opportunities". *International journal of environmental research and public health*, v. 18, n. 19, 10496.
https://doi.org/10.3390/ijerph181910496

Priyadarsini, S. Lakshmi; **Suresh, M.** (2020). "Factors influencing the epidemiological characteristics of pandemic Covid-19: A TISM approach". *International journal of healthcare management*, v. 13, n. 2, pp. 89-98.
https://doi.org/10.1080/20479700.2020.1755804

Publications Office of the European Union (2020). *Metadata quality assessment methodology. How data.europa.eu measures the quality of all stored metadata.*
https://data.europa.eu/mqa/methodology?locale=en

Rothman, Kenneth J.; **Greenland, Sander**; **Lash, Timothy L.** (2011). *Modern epidemiology. Third edition*. Philadelphia: Lippincott Williams & Wilkins. ISBN: 978 1 45115920 2

Souris, Marc (2019). *Epidémiologie et géographie: Principes, methods et outils de l'analyse spatiale.* London: ISTE. ISBN: 978 1 784065 73 7

Sy, Karla-Therese L.; **White, Laura F.**; **Nichols, Brooke E.** (2021). "Population density and basic reproductive number of Covid-19 across United States counties". *PLoS one*, v. 16, n. 4, pp. 1-11.
https://doi.org/10.1371/journal.pone.0249271

WHO (2022). *Public health surveillance for Covid-19: interim guidance*. World Health Organization.
https://www.who.int/publications/i/item/who-2019-nCoV-surveillanceguidance-2022.1

Wilkinson, Mark D.; **Dumontier, Michel**; **Aalbersberg, Ijsbrand-Jan**; **Appleton, Gabrielle**; **Axton, Myles**; **Baak, Arie**; **Blomberg, Niklas**; **Boiten, Jan-Willem**; **Da-Silva-Santos, Luiz-Bonino**; **Bourne, Philip E.**; **Bouwman, Jildau**; **Brookes, Anthony**; **Clark, Tim**; **Crosas, Mercè**; **Dillo, Ingrid**; **Dumon, Olivier**; **Edmunds, Scott**; **Evelo, Chris T.**; **Finkers, Richard**; **González-Beltrán, Alejandra**; **Gray, Alasdair J. G.**; **Groth, Paul**; **Goble, Carole**; **Grethe, Jeffrey S.**; **Heringa, Jaap**; **Hoen, Peter A. C.'t**; **Hooft, Rob**; **Kuhn, Tobias**; **Kok, Ruben**; **Kok, Joost**; **Lusher, Scott J.**; **Martone, Maryann E.**; **Mons, Albert**; **Packer, Abel L.**; **Persson, Bengt**; **Rocca-Serra, Philippe**; **Roos, Marco**; **Van-Schaik, Rene**; **Sansone, Susanna-Assunta**; **Schultes, Erik**; **Sengstag, Thierry**; **Slater, Ted**; **Strawn, George**; **Swertz, Morris A.**; **Thompson, Mark**; **Van-der-Lei, Johan**; **Van-Mulligen, Erik**; **Velterop, Jan**; **Waagmeester, Andra**; **Wittenburg, Peter**; **Wolstencroft, Katherine**; **Zhao, Jun**; **Mons, Barend** (2016). "The FAIR guiding principles for scientific data management and stewardship". *Scientific data*, v. 3, n. 1, 160018.
https://doi.org/10.1038/sdata.2016.18

Wu, Dan; **Ma, Le**; **Zhang, Hui** (2021). "Evaluation of open health data portals for Covid-19 from the perspective of the user experience". *The electronic library*, v. 39, n. 2, pp. 296-317.
https://doi.org/10.1108/EL-01-2021-0011

Xu, Bo; **Kraemer, Moritz U. G.** (2020). "Open access epidemiological data from the Covid-19 outbreak". *The lancet infectious diseases*, v. 20, n. 5, pp. 534.
https://doi.org/10.1016/S1473-3099(20)30119-5